

Data Organization

Mingjun Zhong

Department of Computing Science

University of Aberdeen

Content

- Data Parsing
- Data Dictionary (metadata)
- Data Integration
- Data Transformation

Data organization

The first task that data scientists face.

- **Data parsing:** Identify the structure of the raw data, so that the data can be read in
- **Data dictionary:** metadata are identified and defined from the data, by exploring the data
- **Data integration:** multiple datasets are integrated into a single data table from different sources
- **Data transformation:** data is transformed from the raw original format into the desired data structure

Data parsing

- The process of identifying the structure of the raw data source.
- The data can be properly read
 - What tools are needed to read in the data?
- Raw data sources can have many different formats: Structured data, Semi-structured data, and Unstructured data.
- **Structured data:** flat files (e.g., CSV, Excel, TXT files, etc.), databases (e.g., SQL, data warehouses, etc.)
- **Semi-structured data:** XML, JSON.
- **Unstructured data:** Text files (e.g., documents like CSV and TXT files, log files, etc), social media, web scraping, real-time streams (streaming data or sensor data), images

Data parsing

Mango; 365,14; 1692,64
Apple; 2568,62; 1183,78
Lemon; 51,65; 685,67
Orange; 1760,75; 128,14
Maple; 880,86; 323,43

- Some file formats can be read using **open-source programs**
- Some can only be read using **proprietary software**
- **Possible barriers even in structured data files**
- A well-structured file format can present a wide variety of encodings:
 - CSV format allows a wide range of delimiters, quotes and escape characters
 - They have to be determined before the file can be loaded
- Multiple tables may appear in a single file; for example, multiple tabs in an excel file

Data parsing

For example, in Artificial Intelligence:

- You have a lot of **images**, **text files**, **log files** on different machines, or **data records** in databases
- Sometimes, your data, like images, is just URLs (e.g., ImageNet)
 - You download them to local machines or just use the data on-the-fly?
- You have crawled large text files and stored them on your machine, you might use **Spark** to process them on a cluster and use **Pandas DataFrame** to organise the data
- Your log files may be stored in a data lake or data warehouse.
- How do you parse them and use them: read into a particular format, organising them for analysis?

Data parsing: CSV

Mango; 365,14; 1692,64
Apple; 2568,62; 1183,78
Lemon; 51,65; 685,67
Orange; 1760,75; 128,14
Maple; 880,86; 323,43

- A common data file format – can be structured or unstructured
- CSV files may need to select the proper parameters to read the data correctly.
- In the example, an ambiguous CSV file in real-world scenario
- Different parameters lead to different output:
 - Choosing semicolon, comma, or space as the field separator?
 - They lead to different structured tables

Data parsing: CSV

Mango; 365,14; 1692,64
Apple; 2568,62; 1183,78
Lemon; 51,65; 685,67
Orange; 1760,75; 128,14
Maple; 880,86; 323,43

- `pandas.read_csv('example1.csv', sep=',', names=['c1', 'c2', 'c3'])`
- `pandas.read_csv('example1.csv', sep=';', names=['c1', 'c2', 'c3'])`

0	Mango; 365	14; 1692	64
1	Apple; 2568	62; 1183	78
2	Lemon; 51	65; 685	67
3	Orange; 1760	75; 128	14
4	Maple; 880	86; 323	43

0	Mango	365,14	1692,64
1	Apple	2568,62	1183,78
2	Lemon	51,65	685,67
3	Orange	1760,75	128,14
4	Maple	880,86	323,43

Data parsing: CSV

name	address	date joined
john smith	1132 Anywhere Lane Hoboken NJ, 07030	Jan 4
erica meyers	1234 Smith Lane Hoboken NJ, 07030	March 2

- This CSV file contains three fields: name, address, and date joined.
- The data is delimited by commas
- The problem is that the `address` field contains commas
- What you can do:
 - Use a different delimiter to separate the field data items
 - Wrap the data in quotes

Data parsing: CSV

```
name, address, date joined  
john smith, 1132 Anywhere Lane Hoboken NJ, 07030, Jan 4  
erica meyers, 1234 Smith Lane Hoboken NJ, 07030, March 2
```

1. Use a different delimiter to separate the field data items

```
name; address; date joined  
john smith; 1132 Anywhere Lane Hoboken NJ, 07030; Jan 4  
erica Meyers; 1234 Smith Lane Hoboken NJ, 07030; March 2
```

2. Wrap the data in quotes

```
name, address, date joined  
john smith, "1132 Anywhere Lane Hoboken NJ, 07030", Jan 4  
erica meyers, "1234 Smith Lane Hoboken NJ, 07030", March 2
```

3. Write a Python program to properly quote the data fields

Data parsing: XML

- In Python, there are three XML parsing models (theoretical):
 - Document Object Model (DOM)
 - Simple API for XML (SAX)
 - Streaming API for XML (StAX)

Document Object Model (DOM)

- The most straightforward model to use
- DOM parser is convenient to use
- Maybe time consuming and need large memory
- Suitable when parsing relatively small documents or need parsing infrequently

Simple API for XML (SAX)

- Lower memory than DOM
- Can deal with arbitrarily large files
- Great for single-pass processing, such as conversion to other formats
- Convenient for parsing incoming XML data in real time
- Inconvenient for handling deeply nested elements
- Cheap in terms of space and time, but more difficult to use than DOM

Streaming API for XML (StAX)

- Built on top of SAX
 - Gives more control over the parsing process
 - More convenient state management
-
- In Python, these models are built in Python packages:
 - `xml.dom`
 - `xml.sax`
 - Choose which one to use depending on the data

Data parsing: JSON

- JSON stands for JavaScript Object Notation
- JSON format is **text-based**: you can create JSON files using the code editor of your choice

Hello_world.json

```
{  
  "greeting": "Hello, world!"  
}
```

Data parsing: JSON

- JSON structure is a dictionary that contains a string as a key and a value
- Key-value pair in a JSON object is separated by a colon (:)
- We can use the package *json* in Python to load JSON files.

```
{
  "name": "Frieda",
  "isDog": true,
  "hobbies": ["eating", "sleeping", "barking"],
  "age": 8,
  "address": {
    "work": null,
    "home": ["Berlin", "Germany"]
  },
  "friends": [
    {
      "name": "Philipp",
      "hobbies": ["eating", "sleeping", "reading"]
    },
    {
      "name": "Mitch",
      "hobbies": ["running", "snacking"]
    }
  ]
}
```


Data dictionary (e.g., metadata)

- Refers to understanding the **contents** of data
- Then translate the contents into **metadata**
- Dataset described by data dictionary or metadata: information about the meaning and type of each attribute in the table
- Data collectors or domain experts provide metadata:
 - Text documents describing the profile of the data
 - Extra headers in the data describing data features
 - Files like CSV describing features
- Data scientists need to understand the data by looking up metadata
- However, metadata/dictionary may be missing (out-of-date)

Data dictionary: table understanding

- Firstly, to understand the table information; need to explore the tables
- How many tables are contained in the data?
- How many features totally in the data?
- How many features in each table?
- How many instances/records in each table?
- Are there any metadata in the data?
- Is the data table relational table?
- Is the data time series or tabular data?

1

Broadband in Contract Price Rises Survey

2

ONLINE Fieldwork: 9th to 11th January 2023

3

4

5

Do you know whether your provider can increase your monthly payment during your minimum contract period?

6

Base: All currently in a fixed term contract and know who their provider is

7

8

9

Total (T)

10

Unweighted base

1304

11

Weighted base

1273

12

Yes, my provider can increase my monthly payment during my minimum contract period

459

36%

13

14

15

No, my provider cannot increase my monthly payment during my minimum contract period

424

33%

16

17

I'm not sure whether my provider can increase my monthly payment during my minimum contract period

390

31%

18

19

20

21

22

23

Do you know whether your provider can increase your monthly payment?

24

Base: All currently not in a fixed term contract and know who their provider is

<

>

Mobile

Broadband

+

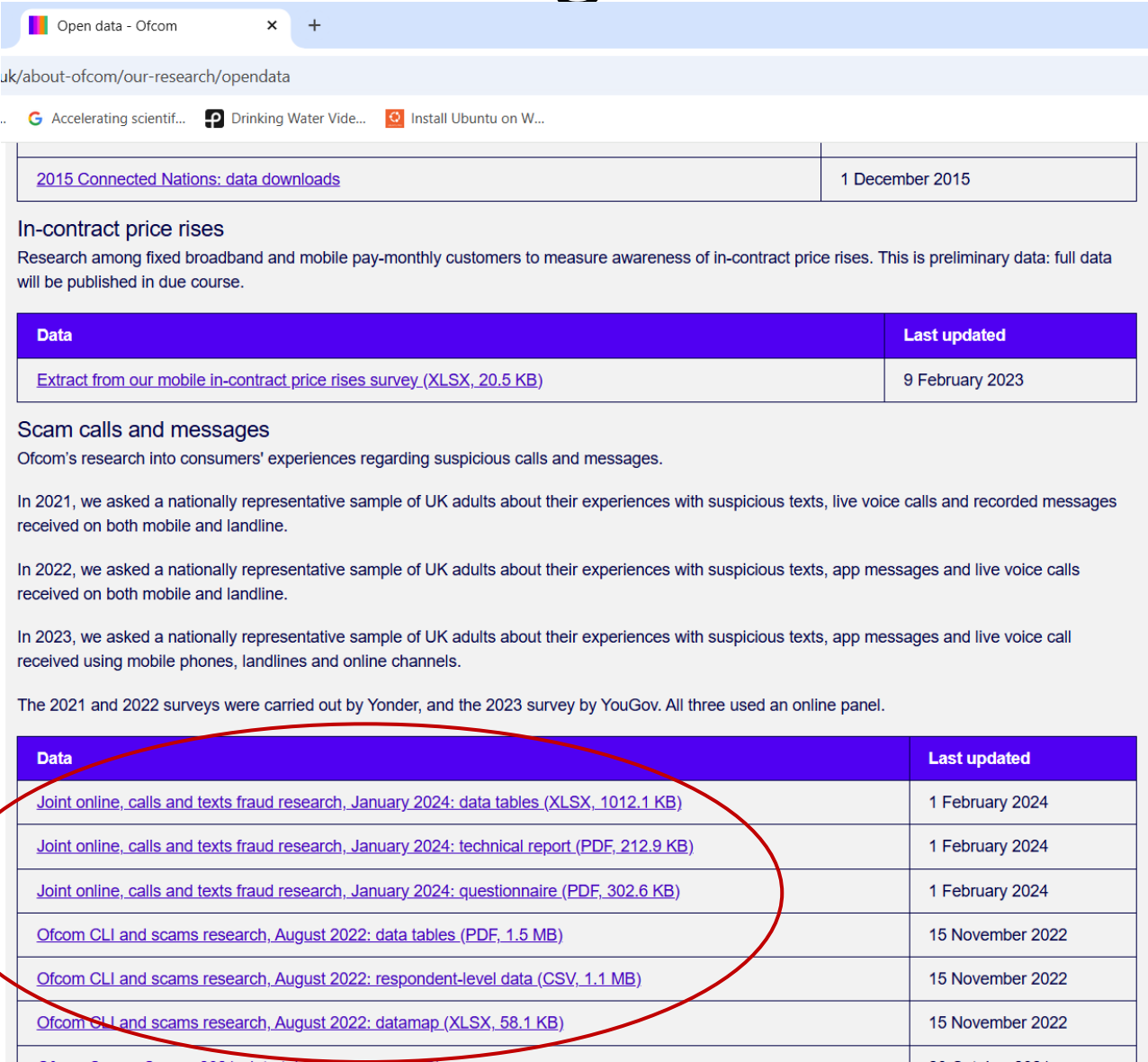
In-contract price rises

Research among fixed broadband and mobile pay-monthly customers to measure awareness of in-contract price rises. This is preliminary data: full data will be published in due course.

Data	Last updated
Extract from our mobile in-contract price rises survey (XLSX, 20.5 KB)	9 February 2023

Data dictionary: table understanding

- Some of these questions may not be answered
- Domain expert may be needed:
 - what are the file names meaning?
 - File names may not be meaningful
 - Headers might be missing
 - What are the feature names meaning?
 - What are the records meaning?



Open data - Ofcom

uk/about-ofcom/our-research/opendata

Accelerating scientif... Drinking Water Vide... Install Ubuntu on W...

2015 Connected Nations: data downloads	1 December 2015
--	-----------------

In-contract price rises
Research among fixed broadband and mobile pay-monthly customers to measure awareness of in-contract price rises. This is preliminary data: full data will be published in due course.

Data	Last updated
Extract from our mobile in-contract price rises survey (XLSX, 20.5 KB)	9 February 2023

Scam calls and messages
Ofcom's research into consumers' experiences regarding suspicious calls and messages.

In 2021, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, live voice calls and recorded messages received on both mobile and landline.

In 2022, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, app messages and live voice calls received on both mobile and landline.

In 2023, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, app messages and live voice call received using mobile phones, landlines and online channels.

The 2021 and 2022 surveys were carried out by Yonder, and the 2023 survey by YouGov. All three used an online panel.

Data	Last updated
Joint online, calls and texts fraud research, January 2024: data tables (XLSX, 1012.1 KB)	1 February 2024
Joint online, calls and texts fraud research, January 2024: technical report (PDF, 212.9 KB)	1 February 2024
Joint online, calls and texts fraud research, January 2024: questionnaire (PDF, 302.6 KB)	1 February 2024
Ofcom CLI and scams research, August 2022: data tables (PDF, 1.5 MB)	15 November 2022
Ofcom CLI and scams research, August 2022: respondent-level data (CSV, 1.1 MB)	15 November 2022
Ofcom CLI and scams research, August 2022: datamap (XLSX, 58.1 KB)	15 November 2022

Data dictionary: table understanding

- Example:

Example: <https://www.ofcom.org.uk/about-ofcom/our-research/opendata>

- Different topics:

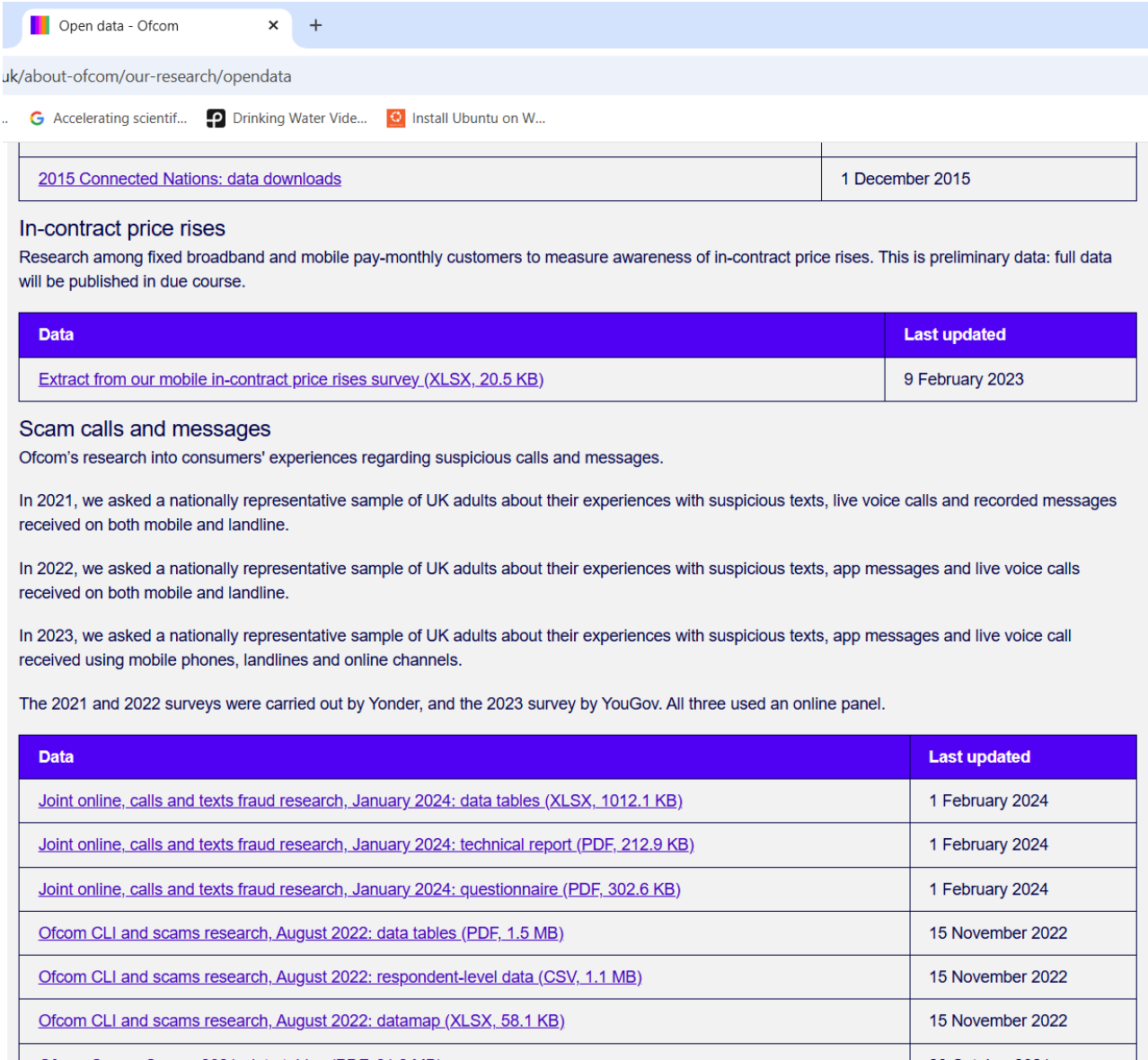
- In-contract price rises

- ✓ One xlsx data file

- Scam calls and messages

- ✓ Multiple data including pdf, xlsx files

- All files with meaningful names



The screenshot shows a web browser window with the URL <https://www.ofcom.org.uk/about-ofcom/our-research/opendata>. The page displays a table of data downloads. The table has two columns: 'Data' and 'Last updated'. The first row shows '2015 Connected Nations: data downloads' with a last updated date of '1 December 2015'. Below this, there is a section titled 'In-contract price rises' with a description: 'Research among fixed broadband and mobile pay-monthly customers to measure awareness of in-contract price rises. This is preliminary data: full data will be published in due course.' This section contains a table with two columns: 'Data' and 'Last updated'. The first row shows 'Extract from our mobile in-contract price rises survey (XLSX, 20.5 KB)' with a last updated date of '9 February 2023'. Below this, there is a section titled 'Scam calls and messages' with a description: 'Ofcom's research into consumers' experiences regarding suspicious calls and messages.' This section contains three paragraphs of text describing the research in 2021, 2022, and 2023. At the bottom, there is a table with two columns: 'Data' and 'Last updated'. The first row shows 'Joint online, calls and texts fraud research, January 2024: data tables (XLSX, 1012.1 KB)' with a last updated date of '1 February 2024'. The second row shows 'Joint online, calls and texts fraud research, January 2024: technical report (PDF, 212.9 KB)' with a last updated date of '1 February 2024'. The third row shows 'Joint online, calls and texts fraud research, January 2024: questionnaire (PDF, 302.6 KB)' with a last updated date of '1 February 2024'. The fourth row shows 'Ofcom CLI and scams research, August 2022: data tables (PDF, 1.5 MB)' with a last updated date of '15 November 2022'. The fifth row shows 'Ofcom CLI and scams research, August 2022: respondent-level data (CSV, 1.1 MB)' with a last updated date of '15 November 2022'. The sixth row shows 'Ofcom CLI and scams research, August 2022: datamap (XLSX, 58.1 KB)' with a last updated date of '15 November 2022'.

2015 Connected Nations: data downloads	1 December 2015
--	-----------------

In-contract price rises

Research among fixed broadband and mobile pay-monthly customers to measure awareness of in-contract price rises. This is preliminary data: full data will be published in due course.

Data	Last updated
Extract from our mobile in-contract price rises survey (XLSX, 20.5 KB)	9 February 2023

Scam calls and messages

Ofcom's research into consumers' experiences regarding suspicious calls and messages.

In 2021, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, live voice calls and recorded messages received on both mobile and landline.

In 2022, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, app messages and live voice calls received on both mobile and landline.

In 2023, we asked a nationally representative sample of UK adults about their experiences with suspicious texts, app messages and live voice call received using mobile phones, landlines and online channels.

The 2021 and 2022 surveys were carried out by Yonder, and the 2023 survey by YouGov. All three used an online panel.

Data	Last updated
Joint online, calls and texts fraud research, January 2024: data tables (XLSX, 1012.1 KB)	1 February 2024
Joint online, calls and texts fraud research, January 2024: technical report (PDF, 212.9 KB)	1 February 2024
Joint online, calls and texts fraud research, January 2024: questionnaire (PDF, 302.6 KB)	1 February 2024
Ofcom CLI and scams research, August 2022: data tables (PDF, 1.5 MB)	15 November 2022
Ofcom CLI and scams research, August 2022: respondent-level data (CSV, 1.1 MB)	15 November 2022
Ofcom CLI and scams research, August 2022: datamap (XLSX, 58.1 KB)	15 November 2022

Data dictionary: feature understanding

- Feature information normally provided by feature names in the table, or additional file explaining the features
- If feature information is not provided, domain expert would help to provide the information, or the information will be inferred from the data
- Feature understanding is a crucial step.
- Data scientists need to understand features before analysing the data

1	<u>Broadband in Contract Price Rises Survey</u>		
2	ONLINE Fieldwork: 9th to 11th January 2023		
3			
4			
5	Do you know whether your provider can increase your monthly payment during your minimum contract period?		
6	Base: All currently in a fixed term contract and know who their provider is		
7			
8		Total (T)	
9	Unweighted base	1304	
10	Weighted base	1273	
11	Yes, my provider can increase my monthly payment during my minimum contract period	459	
12		36%	
13			
14	No, my provider cannot increase my monthly payment during my minimum contract period	424	
15		33%	
16	I'm not sure whether my provider can increase my monthly payment during my minimum contract period	390	
17		31%	
18			
19			
20			
21			
22			
23	Do you know whether your provider can increase your monthly payment?		
24	Base: All currently not in a fixed term contract and know who their provider is		

<

>

Mobile

Broadband

+

Data dictionary: feature understanding

- Example: home broadband speeds data (May 2020)
- What are the meanings of **features** in the table?
- May need metadata or domain experts to explain

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	unit_id	ISP	Technology	PACKAGE (download by upload)	MarketClass	Distance from exchange - DSL only	package_for_weighting	Region	Country	Geography	Download - 24 (ave)	Download - 24 min(ave)	Download	Dow
2	940006	BT	ADSL1		8 A	3698.29	BT 8 ADSL1	Scotland	Scotland	Rural	1.748510753	1.434534194	1.858125	1.7
3	33806	BT	ADSL1		8 B	3102.02	BT 8 ADSL1	East	England	Rural	3.558533778	3.0890032	3.694687	3.3
4	943154	BT	ADSL1		8 B	3034.22	BT 8 ADSL1	East Midlands	England	Rural	3.557283673	2.925570065	3.765915	3.4
5	942412	BT	ADSL1		8 B	3934.59	BT 8 ADSL1	North West	England	Rural	1.313005372	1.268160516	1.342373	1.3
6	947276	BT	ADSL1		8 B	3404.29	BT 8 ADSL1	Northern Ireland	Northern Ireland	Rural	0.972343345	0.888351226	1.014567	0.9
7	947260	BT	ADSL1		8 B	5068.77	BT 8 ADSL1	Scotland	Scotland	Rural	1.66331283	0.828585032	2.027452	1.6
8	940006	BT	ADSL1		8 B	5045.00	BT 8 ADSL1	Scotland	Scotland	Rural	1.701200200	1.414740044	1.810005	1.6

Data dictionary: value understanding

- Data scientists need to understand the values in each feature
- Categorical, numerical, or text data features?
- How many unique elements in each feature?
- Are there any outliers or anomalies in the feature values?
- Do values outside the expected range of values of a given feature exist? (anomalies)

Data dictionary: value understanding

- Example: home broadband speeds data (May 2020)
- What are the meanings of **values** in the table?
- May need metadata or domain experts to explain

The screenshot shows a Microsoft Excel spreadsheet titled "home-broadband-speeds-data-may-2020-weighte...". The ribbon includes File, Home, Insert, Page Layout, Formulas, Data, Review, View, Automate, Help, and Acrobat. The Home tab is active, showing options for Clipboard, Font, Alignment, Number, Styles, Cells, Editing, Add-ins, and Adobe Acrobat. The formula bar shows "unit_id". The data table is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M
	unit_id	ISP	Technology	PACKAGE (download by upload)	MarketClass	Distance from exchange - DSL only	package_for_weighting	Region	Country	Geography	Download - 24 (ave)	Download - 24 min(ave)	Download
1	940006	BT	ADSL1		8 A	3698.29	BT 8 ADSL1	Scotland	Scotland	Rural	1.748510753	1.434534194	1.858125
2	33806	BT	ADSL1		8 B	3102.02	BT 8 ADSL1	East	England	Rural	3.558533778	3.0890032	3.694687
3	943154	BT	ADSL1		8 B	3034.22	BT 8 ADSL1	East Midlands	England	Rural	3.557283673	2.925570065	3.765915
4	942412	BT	ADSL1		8 B	3934.59	BT 8 ADSL1	North West	England	Rural	1.313005372	1.268160516	1.342373
5	947276	BT	ADSL1		8 B	3404.29	BT 8 ADSL1	Northern Ireland	Northern Ireland	Rural	0.972343345	0.888351226	1.014567
6	947260	BT	ADSL1		8 B	5068.77	BT 8 ADSL1	Scotland	Scotland	Rural	1.66331283	0.828585032	2.027452

Data integration

- Refers to combining related information from multiple sources
- Aggregate all the information they need into a single data structure: e.g., patient information aggregated from different hospitals
- Data is often received in instalments, e.g., daily, monthly, etc.
- Different information in different tables:
 - One table contains demographic information about patients
 - The other table contains medical tests of each patient
- New features may be added into the newest instalments
 - Blood tests may be added into the medical test table
- Joining (or record linkage) several tables together, i.e., adding new features to existing table
- Unioning several tables, i.e., adding new rows to an existing table

	A	B
1	Spotlight Projects - Climate Change Survey	
2		
3	Page 1	Table 1 Climate Change
4	Page 2	Table 2 Climate Change
5	Page 3	Table 3 Climate Change
6	Page 4	Table 4 Eco Groups 1 (from Q.10 - climate)
7	Page 5	Table 5 Eco Groups 1 (from Q.10 - climate)
8	Page 6	Table 6 Eco Groups 1 (from Q.10 - climate)
9	Page 7	Table 7 Eco Groups 2 (from Q.11)
10	Page 8	Table 8 Eco Groups 2 (from Q.11)
11	Page 9	Table 9 Eco Groups 2 (from Q.11)
12	Page 10	Table 10 Q.1 Which of the following devices do you h
13	Page 11	Table 11 Q.1 Which of the following devices do you h
14	Page 12	Table 12 Q.1 Which of the following devices do you h
15	Page 13	Table 13 Q.1A Actively Used
16	Page 14	Table 14 Q.1A Actively Used
17	Page 15	Table 15 Q.1A Actively Used
18	Page 16	Table 16 Q.1A Actively Used
19	Page 17	Table 17 Q.1A Actively Used
20	Page 18	Table 18 Q.1A Actively Used
21	Page 19	Table 19 Q.1A Actively Used
22	Page 20	Table 20 Q.1A Actively Used
23	Page 21	Table 21 Q.1A Actively Used
24	Page 22	Table 22 Q.1A Actively Used
25	Page 23	Table 23 Q.1A Actively Used
26	Page 24	Table 24 Q.1A Actively Used
27	Page 25	Table 25 Q.1A Actively Used
28	Page 26	Table 26 Q.1A Actively Used
29	Page 27	Table 27 Q.1A Actively Used
30	Page 28	Table 28 Q.1A Actively Used
<div><div>< ></div><div>INDEX</div><div>p1</div><div>+</div></div>		

Record linkage and table joining

Dataset A										
rec_id	given_name	surname	street_number	address_1	address_2	suburb	postcode	state	date_of_birth	soc_sec_id
rec-0-org	rachael	dent	1	knox street	lakewood estate	byford	4129	vic	19280722	1683994
rec-1-org	isabella	everett	25	pike place	rowethorpe	marsden	2152	nsw	19110816	6653129
rec-10-org	lachlan	reid	5	carrington road	legacy vlge	yagoona	2464	nsw	19500531	3232033
rec-100-org	hayden	stapley	38	tindale street	villa 2	cromer heights	4125	vic	NaN	4620080
rec-1000-org	victoria	zbierski	70	wybalena grove	inverneath	paralowie	5065	nsw	19720503	1267612

Dataset B										
rec_id	given_name	surname	street_number	address_1	address_2	suburb	postcode	state	date_of_birth	soc_sec_id
rec-0-dup-0	rachael	dent	4	knox street	lakewood estate	byford	4129	vic	19280722	1683994
rec-1-dup-0	isabella	everett	25	pike mlace	rowethorpe	marsden	2152	nsw	19110816	6653129
rec-10-dup-0	lachlnn	reid	5	carrington road	legacy vlge	yagoona	2446	nsw	19500531	3232033
rec-100-dup-0	hayden	stapley	NaN	tindale street	villa 2	cromer heights	4125	vic	NaN	4620080
rec-1000-dup-0	victoria	zbierski	70	wybalena grove	inverbeath	paralowie	5065	nsw	19720503	1267612

- Refers to identifying records across multiple tables that corresponding to the same entity, and integrating the information into a single extended table:
 - For example, integrating demographic information and medical tests of a patient
- Primary key (if exists) can be used to joining multiple tables (e.g., patient id in this example)
- However, there are difficulties:
 - the primary keys can change across tables
 - There may not be primary keys

Python Record Linkage Tools: <https://recordlinkage.readthedocs.io/en/latest/index.html>

Table unioning

- Refers to aggregating together row-wise different tables that contain different entities with the same information into one extended table
- Obstacles:
 1. The structure of different tables can differ, so not possible to concatenate tables
 2. Header (feature) names changing across tables
 3. Features added or deleted
 4. Features located in different positions across multiple tables
- We have to match the features across tables

Heterogeneous integration

- Refers to aggregating together different structured sources (relational tables, time series) and different physical locations (websites, repositories) into a single extended structure format
- Data repository can be massive, multiple sources across different locations
- Time consuming and task complex for aggregating all these data
- Query based systems can be used, if a subset of the data is required for analysis
- Data parsing can be implemented by using ETL (Extract, Transform and Load) pipelines; however, data scientists usually need interactions with pipelines

Data transformation

- Data transformation can have different meanings at different scenarios
- Here, we refer to **table transformation** and **information extraction**
- Data models requires that data must have a particular form before being analysed
 - Data table must be a $N \times D$ table, where D is the number of features, and N is the number of records
 - The data must be transformed into this structure
- Often data are not structured, e.g., text data – we must extract information from the data

Table transformation

- This process involves any manipulation on the data that changes the overall “shape” of the data
- Some data rows and/or features are removed from the data table
- The Tundra dataset: remove the rows with the year of observation not recorded.

(<https://github.com/TundraTraitTeam/TraitHub/tree/master>)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		Individual	AccSpecies	OriginalN	Latitude	Longitude	Elevation	SiteName	SubsiteName	Treatment	DayOfYear	Year	DataContr	Comments	ValueKind	Trait	Value	Units	Genus	
4014	37816	37816	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	4.915	mm2/mg	Asterolasia	
4015	37817	37817	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	5.148	mm2/mg	Asterolasia	
4016	37818	37818	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	5.118	mm2/mg	Asterolasia	
4017	37819	37819	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	4.644	mm2/mg	Asterolasia	
4018	37820	37820	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	5.93	mm2/mg	Asterolasia	
4019	37821	37821	Asterolasi	Asterolasi	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	5.359	mm2/mg	Asterolasia	
4020	37822	37822	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	9.973	mm2/mg	Austrolopyrum	
4021	37823	37823	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	9.542	mm2/mg	Austrolopyrum	
4022	37824	37824	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	10.107	mm2/mg	Austrolopyrum	
4023	37825	37825	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	9.418	mm2/mg	Austrolopyrum	
4024	37826	37826	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	10	mm2/mg	Austrolopyrum	
4025	37827	37827	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	11.149	mm2/mg	Austrolopyrum	
4026	37828	37828	Austrolop	Austrolop	-36.9017	147.2894	NA	Victoria, A	Bogong High Pl	none	NA	NA	Megan Go	mean of 20	Individual	Leaf area	12.01	mm2/mg	Austrolopyrum	

Information extraction

- Refers to extracting knowledge from semi-structured and/or unstructured data
- This will form a structured data table
- For example, in NLP, Named Entity Recognition (NER) is used to represent names of people and places
- Data Science tasks

Reference

Key concept materials were taken from the following paper:

- Nazabal, Alfredo, Christopher KI Williams, Giovanni Colavizza, Camila Rangel Smith, and Angus Williams. "Data engineering for data analytics: A classification of the issues, and case studies." *arXiv preprint arXiv:2004.12929* (2020).