Big data to tidy infographics

http://training.theodi.org/BigTidy

David Tarrant · @davetaz

Open Data Science

Day 1: Unlocking data from the web

Day 2: Data management and statistics

Day 3: Big data and data visualisation

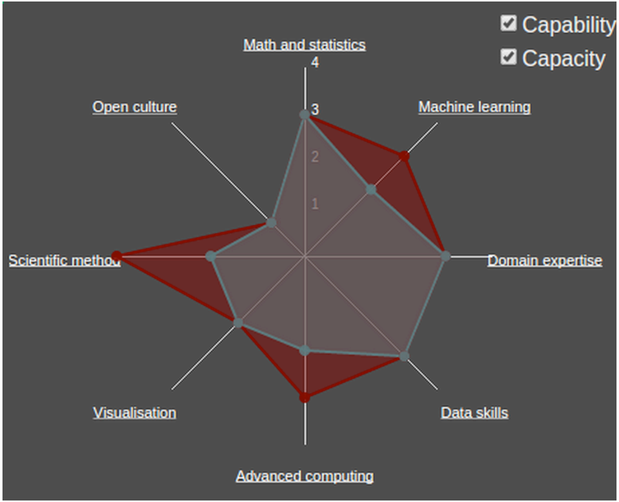
BY SA

Course aim

Equip you with the knowledge and tools to help you upskill as modern data scientists.

cc BY SA

Data science



Today

Big data

Cloud computing

Data visualisation

The *future* bit of data science

cc BY SA

Session 1

Big data, tiny answers

Session 2

Infographics and interaction

Session 1

Big data, tiny answers

cc BY SA

Outcomes

Define and identify big data

Design a strategy for dealing with big data

Apply a number of big data tools

cc BY SA

Exercise

What is Big Data?

BY SA

Big Data

Dataset that are too large and complex to manipulate with standard methods or tools.

cc BY SA

Excel

Workbook WAS limited to 65,536 rows (2^{16} aka 16-Bit)

64-Bit operating system addressing limit is 2^{64}

18,446,744,073,709,551,615

q q t b m t h

What is big data?

Volume

Velocity

Variety

Veracity

CC BY SA

What is big data?

Volume

We create around 4 zettabytes of data day.

That’s 1 sextillion bytes per day (128-Bit OS required)

Velocity

Variety

Veracity

cc BY SA

What is big data?

Volume

Velocity

The data is created quicker than we can process it.

Variety

Veracity

cc BY SA

What is big data?

Volume

Velocity

The data is continuously changing in structure, format and detail.

Variety

Veracity

cc BY SA

What is big data?

Volume

Velocity

Variety

Veracity

The data quality is highly variable and affected by changing perception of truth and fact.

cc BY SA

Big Data

Taken collectively. All digital data is big data. Looking at a facet might reveal that you are looking at a dataset that only conforms to one or two of the **V**s.

Can you name a dataset that shows the characteristics of all 4 **V**s?

cc BY SA

A few more V’s

Value and **V**iability

More data does not mean better results.

In fact often entirely the opposite is true.

Sample selection is critical to all good statistic studies.

Not being able to control selection may lead to an incorrect conclusion.

Conclusion

The majority of datasets are large.

Lots of rows with lots of joins that can be processed. If you know how to exploit computing power available.

Big Data processing: UK Trade Data

cc BY SA

UK Trade Data

Exports

Non-EU

150,000 to 200,000

per month

Imports

Non-EU

190,000 to 220,000

per month

Dispatches

EU

210,000 to 250,000

per month (+estimates)

Arrivals

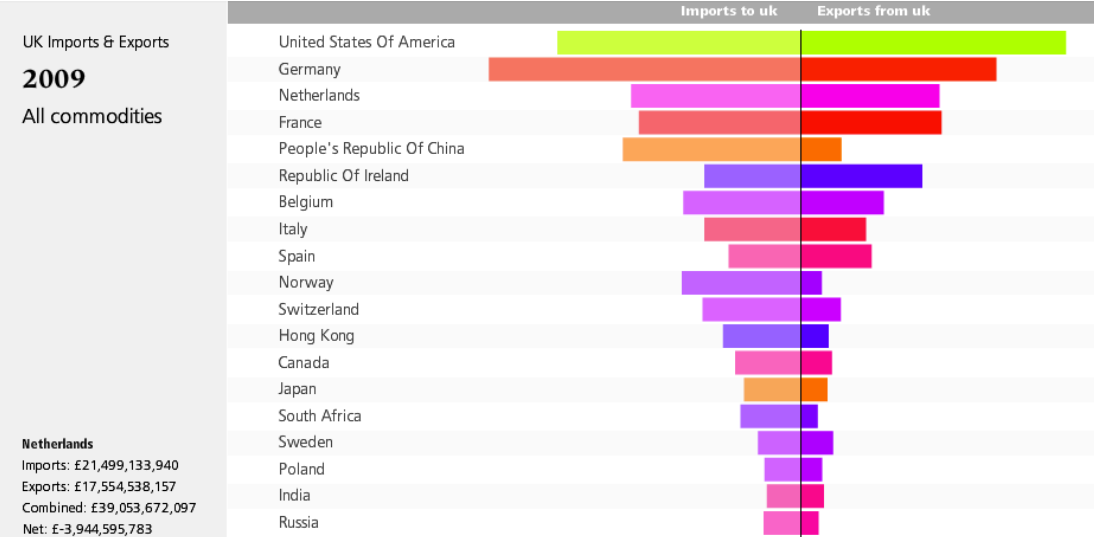
EU

125,000 to 135,000

per month (+estimates)

BY SA

Distilled information



UK Imports & Exports

2009

All commodities

Netherlands

Imports: £21,499,133,940

Exports: £17,554,538,157

Combined: £39,053,672,097

Net: £-3,944,595,783

Exercise

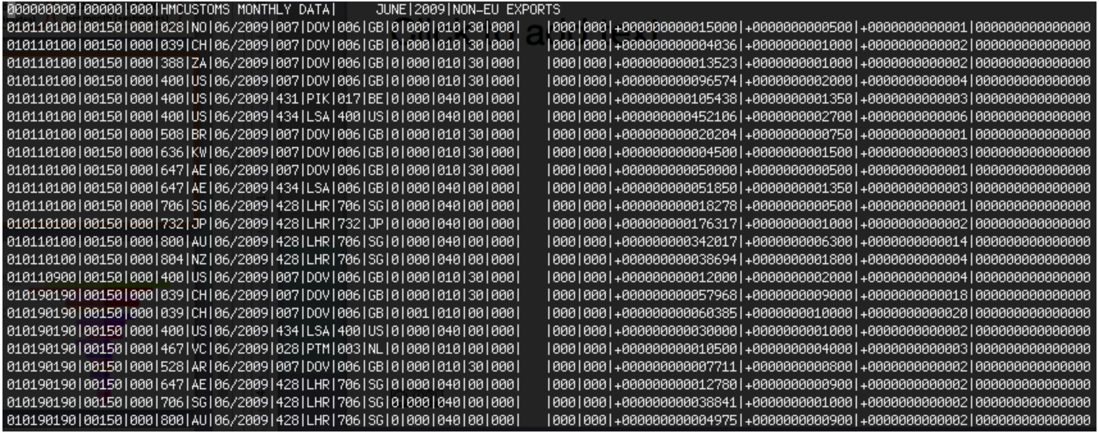
Q: How have imports and exports on Jet Engines changed over the years?

Design a processing pipeline that can answer this question from the data.

bit.ly/uk_trade

CC BY SA

Stage 1: What the format????



```
000000000|00000|000|HMCUSTOMS MONTHLY DATA|     JUNE|2009|NON-EU EXPORTS
010110100|00150|000|028|NO|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000015000|+00000000000500|+00000000000001|0000000000000000
010110100|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000004036|+00000000001000|+00000000000002|0000000000000000
010110100|00150|000|388|ZA|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000013523|+00000000001000|+00000000000002|0000000000000000
010110100|00150|000|400|US|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000096574|+00000000002000|+00000000000004|0000000000000000
010110100|00150|000|400|US|06/2009|431|PIK|017|BE|0|000|040|00|000|    |000|000|+0000000000105438|+00000000001350|+00000000000003|0000000000000000
010110100|00150|000|400|US|06/2009|434|LSA|400|US|0|000|040|00|000|    |000|000|+0000000000452106|+00000000002700|+00000000000006|0000000000000000
010110100|00150|000|508|BR|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000020204|+00000000000750|+00000000000001|0000000000000000
010110100|00150|000|636|KW|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000004500|+00000000001500|+00000000000003|0000000000000000
010110100|00150|000|647|AE|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000050000|+00000000000500|+00000000000001|0000000000000000
010110100|00150|000|647|AE|06/2009|434|LSA|006|GB|0|000|040|00|000|    |000|000|+0000000000051850|+00000000001350|+00000000000003|0000000000000000
010110100|00150|000|706|SG|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000018278|+00000000000500|+00000000000001|0000000000000000
010110100|00150|000|732|JP|06/2009|428|LHR|732|JP|0|000|040|00|000|    |000|000|+0000000000176317|+00000000001000|+00000000000002|0000000000000000
010110100|00150|000|800|AU|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000342017|+00000000006300|+00000000000014|0000000000000000
010110100|00150|000|804|NZ|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000038694|+00000000001800|+00000000000004|0000000000000000
010110900|00150|000|400|US|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000012000|+00000000002000|+00000000000004|0000000000000000
010190190|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000057968|+00000000009000|+00000000000018|0000000000000000
010190190|00150|000|039|CH|06/2009|007|DOV|006|GB|0|001|010|00|000|    |000|000|+0000000000060385|+00000000010000|+00000000000020|0000000000000000
010190190|00150|000|400|US|06/2009|434|LSA|400|US|0|000|040|00|000|    |000|000|+0000000000030000|+00000000001000|+00000000000002|0000000000000000
010190190|00150|000|467|VC|06/2009|028|PTM|003|NL|0|000|010|00|000|    |000|000|+0000000000010500|+00000000004000|+00000000000003|0000000000000000
010190190|00150|000|528|AR|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+0000000000007711|+00000000000800|+00000000000002|0000000000000000
010190190|00150|000|647|AE|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000012780|+00000000000900|+00000000000002|0000000000000000
010190190|00150|000|706|SG|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000038841|+00000000001000|+00000000000002|0000000000000000
010190190|00150|000|800|AU|06/2009|428|LHR|706|SG|0|000|040|00|000|    |000|000|+0000000000004975|+00000000000900|+00000000000002|0000000000000000
```

Stage 2: RTFM

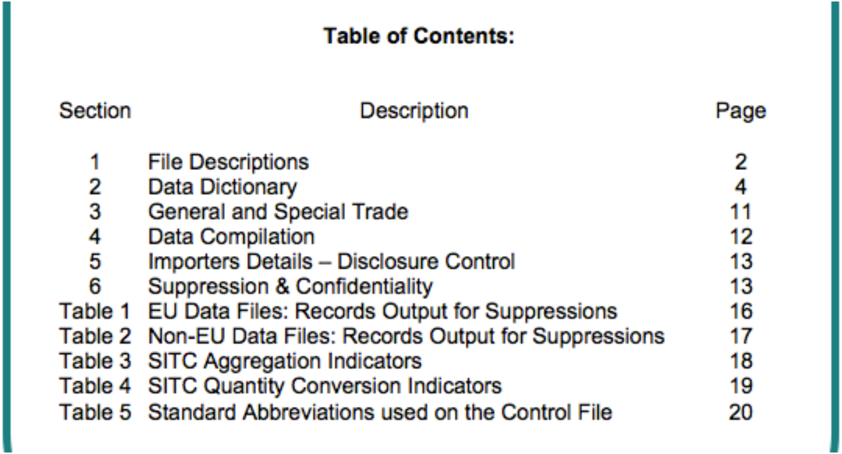
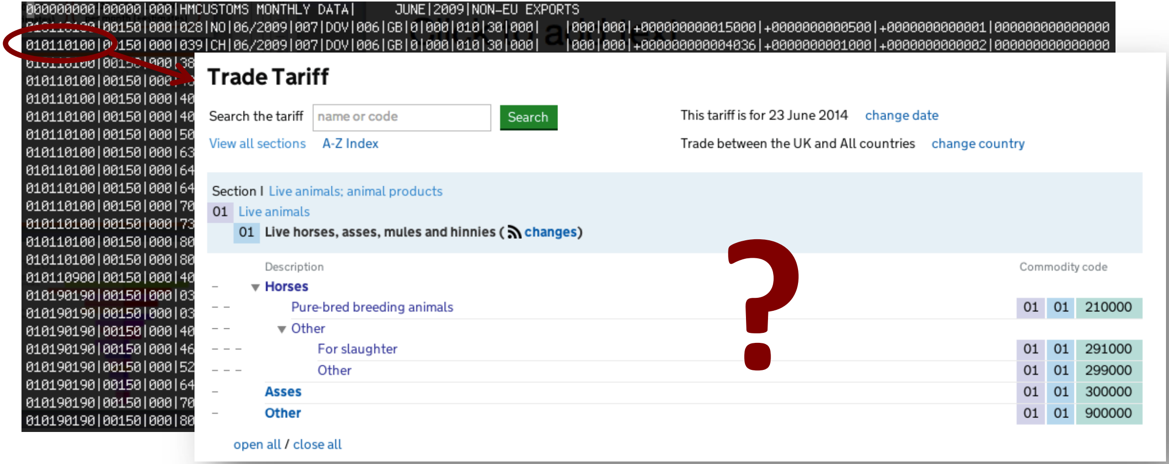


Table of Contents:

Section	Description	Page
1	File Descriptions	2
2	Data Dictionary	4
3	General and Special Trade	11
4	Data Compilation	12
5	Importers Details – Disclosure Control	13
6	Suppression & Confidentiality	13
Table 1	EU Data Files: Records Output for Suppressions	16
Table 2	Non-EU Data Files: Records Output for Suppressions	17
Table 3	SITC Aggregation Indicators	18
Table 4	SITC Quantity Conversion Indicators	19
Table 5	Standard Abbreviations used on the Control File	20

Stage 3: Decode

010110100



```
000000000|00000|000|HMCUSTOMS MONTHLY DATA|     JUNE|2009|NON-EU EXPORTS
010110100|00150|000|028|NO|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+000000000015000|+0000000000500|+0000000000001|000000000000000
010110100|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+000000000004036|+0000000001000|+0000000000002|000000000000000
010110100|00150|000|38
010110100|00150|000
010110100|00150|000|40
010110100|00150|000|40
010110100|00150|000|50
010110100|00150|000|63
010110100|00150|000|64
010110100|00150|000|64
010110100|00150|000|70
010110100|00150|000|73
010110100|00150|000|80
010110100|00150|000|80
010110900|00150|000|40
010190190|00150|000|03
010190190|00150|000|03
010190190|00150|000|40
010190190|00150|000|46
010190190|00150|000|52
010190190|00150|000|64
010190190|00150|000|70
010190190|00150|000|80
```

Trade Tariff

Search the tariff [name or code] Search

This tariff is for 23 June 2014 change date

View all sections A-Z Index

Trade between the UK and All countries change country

Section I Live animals; animal products

01 Live animals

01 **Live horses, asses, mules and hinnies (changes)**

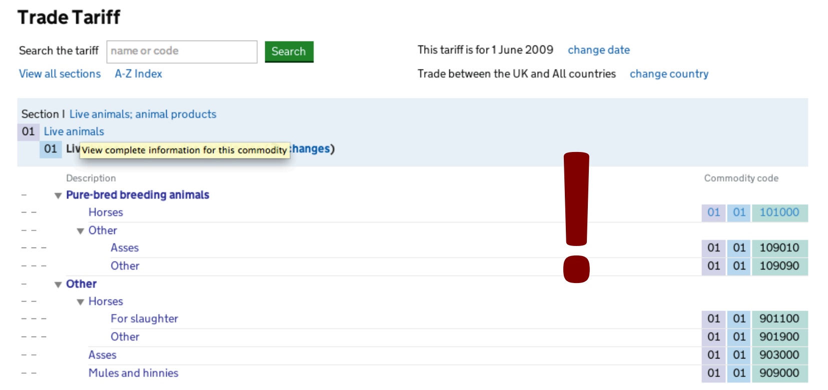
	Description	Commodity code
–	▼ **Horses**	
– –	Pure-bred breeding animals	01 01 210000
– –	▼ Other	
– – –	For slaughter	01 01 291000
– – –	Other	01 01 299000
–	**Asses**	01 01 300000
–	**Other**	01 01 900000

open all / close all

Stage 3b: API?

010110100

https://www.gov.uk/trade-tariff/headings/0101?country=&day=1&month=6&year=2009



Trade Tariff

Search the tariff [name or code] Search

This tariff is for 1 June 2009 change date

View all sections A-Z Index

Trade between the UK and All countries change country

Section I Live animals; animal products

01 Live animals

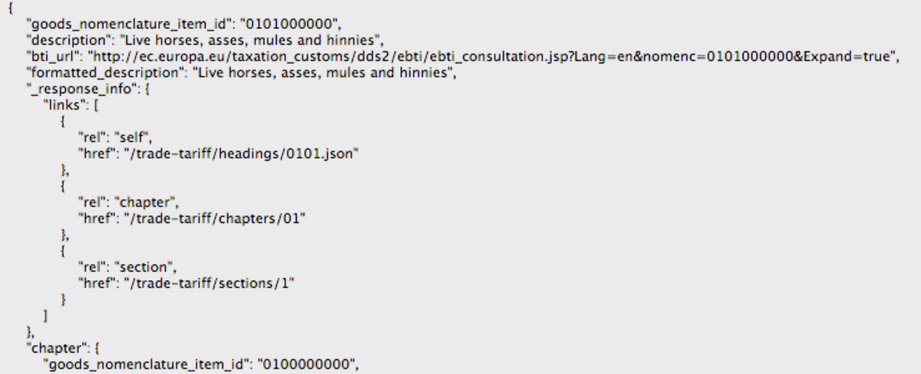
01 Liv View complete information for this commodity :hanges)

	Description	Commodity code
–	▼ Pure-bred breeding animals	
– –	Horses	01 01 101000
– –	▼ Other	
– – –	Asses	01 01 109010
– – –	Other	01 01 109090
–	▼ Other	
– –	▼ Horses	
– – –	For slaughter	01 01 901100
– – –	Other	01 01 901900
– –	Asses	01 01 903000
– –	Mules and hinnies	01 01 909000

The codes for the same things have changed. Meaning that we have to compare the text! Ahhh.

Stage 4: API for data?

https://www.gov.uk/trade-tariff/headings/0101**.json**?country=&day=1&month=6&year=2009



```
{
    "goods_nomenclature_item_id": "0101000000",
    "description": "Live horses, asses, mules and hinnies",
    "bti_url": "http://ec.europa.eu/taxation_customs/dds2/ebti/ebti_consultation.jsp?Lang=en&nomenc=0101000000&Expand=true",
    "formatted_description": "Live horses, asses, mules and hinnies",
    "_response_info": {
        "links": [
            {
                "rel": "self",
                "href": "/trade-tariff/headings/0101.json"
            },
            {
                "rel": "chapter",
                "href": "/trade-tariff/chapters/01"
            },
            {
                "rel": "section",
                "href": "/trade-tariff/sections/1"
            }
        ]
    },
    "chapter": {
        "goods_nomenclature_item_id": "0100000000",
```

Stage 5: Predict scale

(12 * 4) files per year

12 Comcode tables

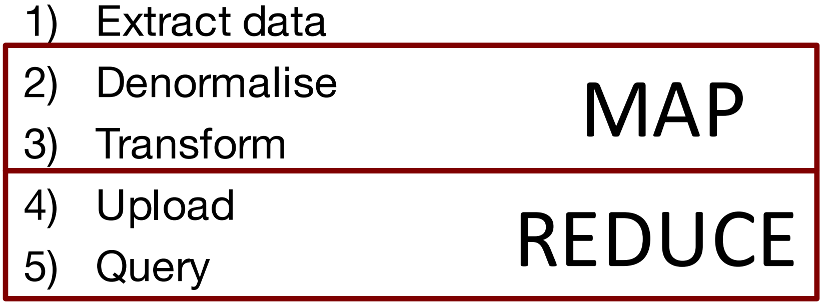
12 Portcode tables

To answer one query you may have to join 48 tables to 24 others to answer it.

This is not how map reduce and big data work.

cc BY SA

A large open data project



1) Extract data

2) Denormalise

3) Transform

4) Upload

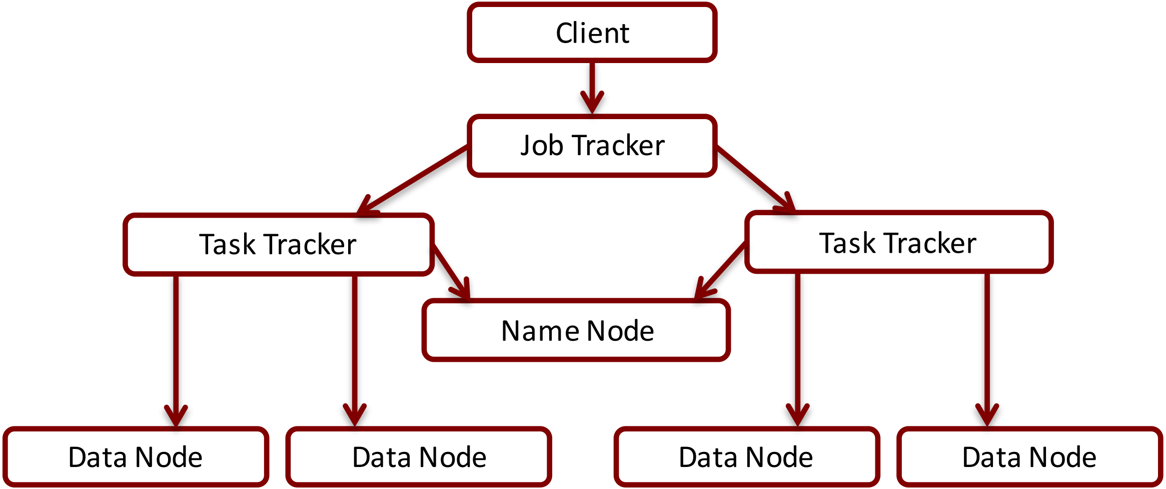
5) Query

MAP

REDUCE

Pivot in the cloud?

Cloud computing



Process pipeline

1) Translate to CSV (exports_makecsv)
2) Filter out supressed data (exports_process_supression)
3) Get ComCode data for that month (get_comcodes)
4) De-Normalise CSV with ComCodes and translate dates to timestamps (expand_csv)
5) Import into Big Query

cc BY SA

DEMO & EXERCISE

Data in Socrata: bit.ly/uk_trade_socrata

Quesitons

Is the UKTrade data big data?

What are the biggest problems with the data?

How would you change your data to use cloud compute platforms?

Session 2

Infographics and interaction

cc BY SA

Outcomes

Describe the key aspects of infographics

Analyse a number of infographics for effectiveness

Create your own interactive online infographic

cc BY SA

Session 1

Big data, tiny answers

Session 2

Infographics and interaction

Outcomes

Define and identify big data

Design a strategy for dealing with big data

Apply a number of big data tools

cc BY SA

Outcomes

Describe the key aspects of infographics

Analyse a number of infographics for effectiveness

Create your own interactive online infographic

cc BY SA

GoScience

Thanks to everyone, we do hope this course has been helpful.

Help us improve by filling in our survey.

bit.ly/odifeedback

cc BY SA

Thank-you