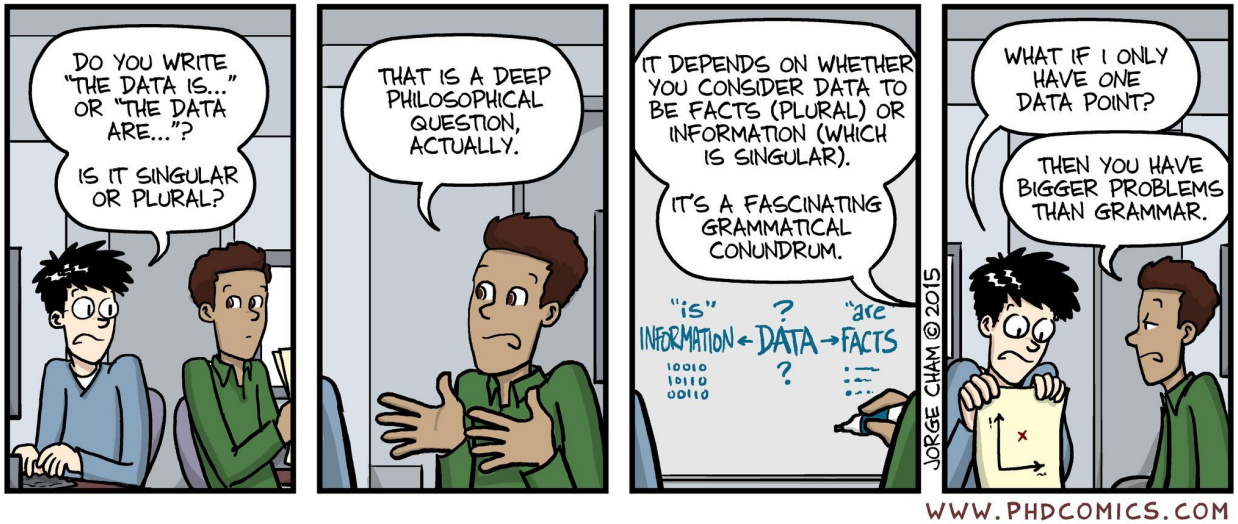
DPV
Data Types

Suzanne Little
suzanne.little@dcu.ie

Today

- What is data?
- Data types



(https://twitter.com/PHDcomics/status/1296758564594126849)

Data is collected information (a working definition)

Data is or Data are ?
https://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular

Discussion:
How many potential data sources can you think of?
Vevox

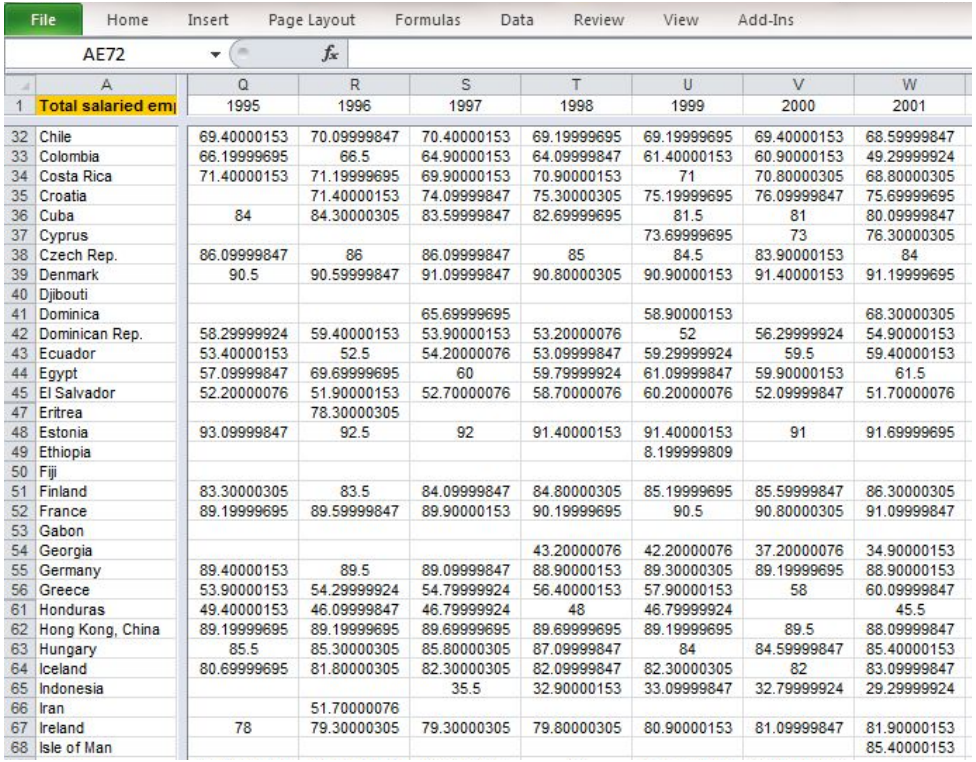
Where does data come from?

Type your answer here...

Submit

20 characters remaining

Data



	A	Q	R	S	T	U	V	W
1	Total salaried em	1995	1996	1997	1998	1999	2000	2001
32	Chile	69.40000153	70.09999847	70.40000153	69.19999695	69.19999695	69.40000153	68.59999847
33	Colombia	66.19999695	66.5	64.90000153	64.09999847	61.40000153	60.90000153	49.29999924
34	Costa Rica	71.40000153	71.19999695	69.90000153	70.90000153	71	70.80000305	68.80000305
35	Croatia		71.40000153	74.09999847	75.30000305	75.19999695	76.09999847	75.69999695
36	Cuba	84	84.30000305	83.59999847	82.69999695	81.5	81	80.09999847
37	Cyprus					73.69999695	73	76.30000305
38	Czech Rep.	86.09999847	86	86.09999847	85	84.5	83.90000153	84
39	Denmark	90.5	90.59999847	91.09999847	90.80000305	90.90000153	91.40000153	91.19999695
40	Djibouti							
41	Dominica			65.69999695		58.90000153		68.30000305
42	Dominican Rep.	58.29999924	59.40000153	53.90000153	53.20000076	52	56.29999924	54.90000153
43	Ecuador	53.40000153	52.5	54.20000076	53.09999847	59.29999924	59.5	59.40000153
44	Egypt	57.09999847	69.69999695	60	59.79999924	61.09999847	59.90000153	61.5
45	El Salvador	52.20000076	51.90000153	52.70000076	58.70000076	60.20000076	52.09999847	51.70000076
47	Eritrea		78.30000305					
48	Estonia	93.09999847	92.5	92	91.40000153	91.40000153	91	91.69999695
49	Ethiopia					8.199999809		
50	Fiji							
51	Finland	83.30000305	83.5	84.09999847	84.80000305	85.19999695	85.59999847	86.30000305
52	France	89.19999695	89.59999847	89.90000153	90.19999695	90.5	90.80000305	91.09999847
53	Gabon							
54	Georgia				43.20000076	42.20000076	37.20000076	34.90000153
55	Germany	89.40000153	89.5	89.09999847	88.90000153	89.30000305	89.19999695	88.90000153
56	Greece	53.90000153	54.29999924	54.79999924	56.40000153	57.90000153	58	60.09999847
61	Honduras	49.40000153	46.09999847	46.79999924	48	46.79999924		45.5
62	Hong Kong, China	89.19999695	89.19999695	89.69999695	89.69999695	89.19999695	89.5	88.09999847
63	Hungary	85.5	85.30000305	85.80000305	87.09999847	84	84.59999847	85.40000153
64	Iceland	80.69999695	81.80000305	82.30000305	82.09999847	82.30000305	82	83.09999847
65	Indonesia			35.5	32.90000153	33.09999847	32.79999924	29.29999924
66	Iran		51.70000076					
67	Ireland	78	79.30000305	79.30000305	79.80000305	80.90000153	81.09999847	81.90000153
68	Isle of Man							85.40000153

log files
social media content
photographs
microblogs
surveys
news
cctv video
movies
television
sales records
clicks
adwords
statistics
audio recordings
playlists
search terms
sensors
pedometer/activity monitor
spectrographs
microscopy
genomes
numbers

Data

Example: a *person* (**object** or **entity** or **instance** or **record** or **row**) has **attributes** (or **features** or **descriptors** or **variables** or **columns**)

- Name
- Passport number
- Birth place
- Eye colour
- Shoe size

Data types

- Structured vs Unstructured
- Quantitative vs Qualitative
- Discrete vs Continuous
- Four levels of data

Structured vs Unstructured

tables, organised, observations,

Row is instance, Column is attribute

Examples:

- company records
- scientific observation

Easier for Machine Learning to work with (kinda)

1	Total salaried em	1995	1996	1997
32	Chile	69.40000153	70.09999847	70.40000153
33	Colombia	66.19999695	66.5	64.90000153
34	Costa Rica	71.40000153	71.19999695	69.90000153
35	Croatia		71.40000153	74.09999847
36	Cuba	84	84.30000305	83.59999847

No hierarchy or arrangement

Raw signals that need processing

Examples:

- tweets & social media posts
- server logs
- media (images, video, etc)

More challenging to work with. How to turn into “Structured”?

Special types of data to watch for

- Temporal (or Time Series)
- Geographic (or Spatial)
- Documents, Images, Video, Audio, 3D
- “Raw” data - unstructured and (sometimes) incidental

Qualitative vs Quantitative

Quality, Label, Trait

Categorical

Limited mathematical functions

Examples:

- Country of origin
- Gender
- Favourite Colour

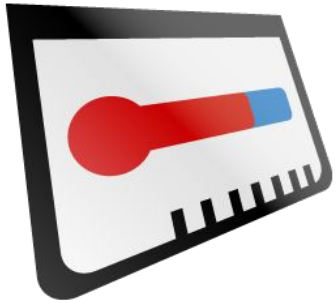
Quantity, Measurement

Numerical

“All the maths!” (well most)

Examples:

- Shoe size
- Temperature
- Bank balance



Quantitative

Discrete vs Continuous

only certain values are valid

ie: there are gaps

usually from counting

Examples:

- Number of times attended
- Number of crimes reported

theoretically any value is possible

depends on measuring device ability

usually from measurements

Examples:

- Cholesterol level
- Time required to complete task

Data types

Structured vs Unstructured

Quantitative vs Qualitative

Discrete vs Continuous

Four levels of data measurement

1. Nominal
2. Ordinal
3. Interval
4. Ratio

NOIR (Stanley Stevens)

Categorical

Nominal (name, label, category)

Gender, Department, Language

Not described by numbers
No maths except equality & set membership
mode but not mean or median

Ordinal (labels plus order)

Temperature (very hot, hot, warm, mild)
Medals (Gold, Silver, Bronze), Scale (Likert - 1 to 10), colour
Can be arranged in an order but not added or subtracted, median but not mean

Qualitative

Measurement

Interval (numbers with distance/space)

We can now talk about “difference” (+/-)

Shoe size (a size 6 foot is not 2 x size 3!)
Temperature (0’C)

Does a zero value mean something?

Ratio (also numbers but with zero)

Can now multiply & divide

Age, Amount of rainfall, Book sales,
Temperature (in Kelvin)

Can you have negative values?

Quantitative

Data

Qualitative or Quantitative? Continuous or Discrete?
Nominal, Ordinal, Interval or Ratio? Any special types?

- Name
- Passport number
- Birth place
- Eye colour
- Shoe size

→ Kahoot

Why do we care?

Type of data determines:

- What statistics are possible/meaningful
- How data can be processed and/or stored
- Which machine learning model can be used
- Which visualisation method to use

References

Reference Sheet for Data Types

“Types of Data”: Chapter 2 of “Principles of Data Science”, Sinan Ozdemir (2016), https://dcu.primo.exlibrisgroup.com/permalink/353DCU_INST/jrp0g3/alma991005580550807206 (DCU Library)

[Statistical perspective] Datacamp Introduction to Statistics Ch1 Data Types