DMV
Data Types

Suzanne Little

Data types

- Recall: Data is collected information (a working definition)
- Structured vs Unstructured
- Quantitative vs Qualitative
- Discrete vs Continuous
- Four levels of data
- Some special data types to watch for

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

Structured vs Unstructured

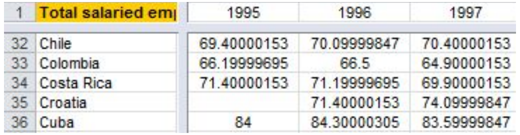
tables, organised, observations,

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

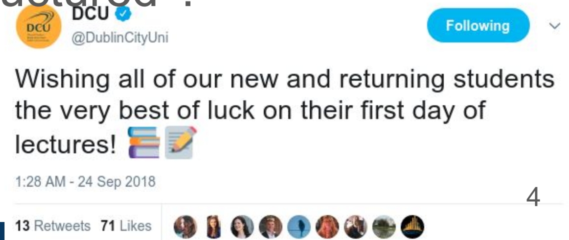
No hierarchy or arrangement

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"?



DCU
@DublinCityUni

Wishing all of our new and returning students the very best of luck on their first day of lectures! 📚📝

1:28 AM - 24 Sep 2018

13 Retweets 71 Likes

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

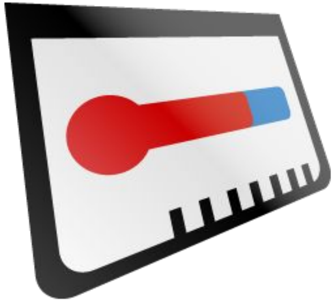
Quantity, Measurement

Numerical

“All the maths!” (well most)

Examples:

- Shoe size
- Temperature
- Bank balance



Quantitative

Discrete vs Continuous

Discrete

only certain values are valid

ie: there are gaps

usually from counting

Examples:

- Number of times attended
- Number of crimes reported

Continuous

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 by order but not added or subtracted, median but not mean

Qualitative

Measurement

Interval (numbers with proportionate spaces)

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

Income, Shoe size,
Temperature (^{o}C, ^{o}F)

“defined interval between values but lacks a zero point”

Ratio (also numbers but with zero)

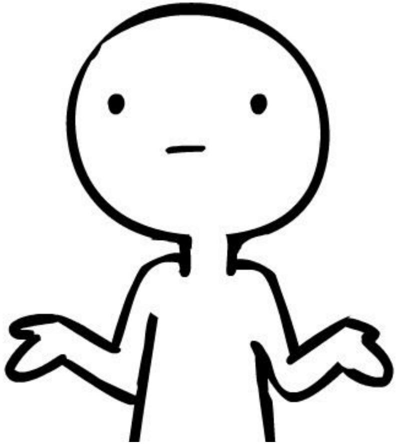
Can now multiply & divide

Age, Amount of rainfall, Book sales,
Temperature (in Kelvin), [normally counting]

Zero has meaning - no negatives

Quantitative

Why do we care?

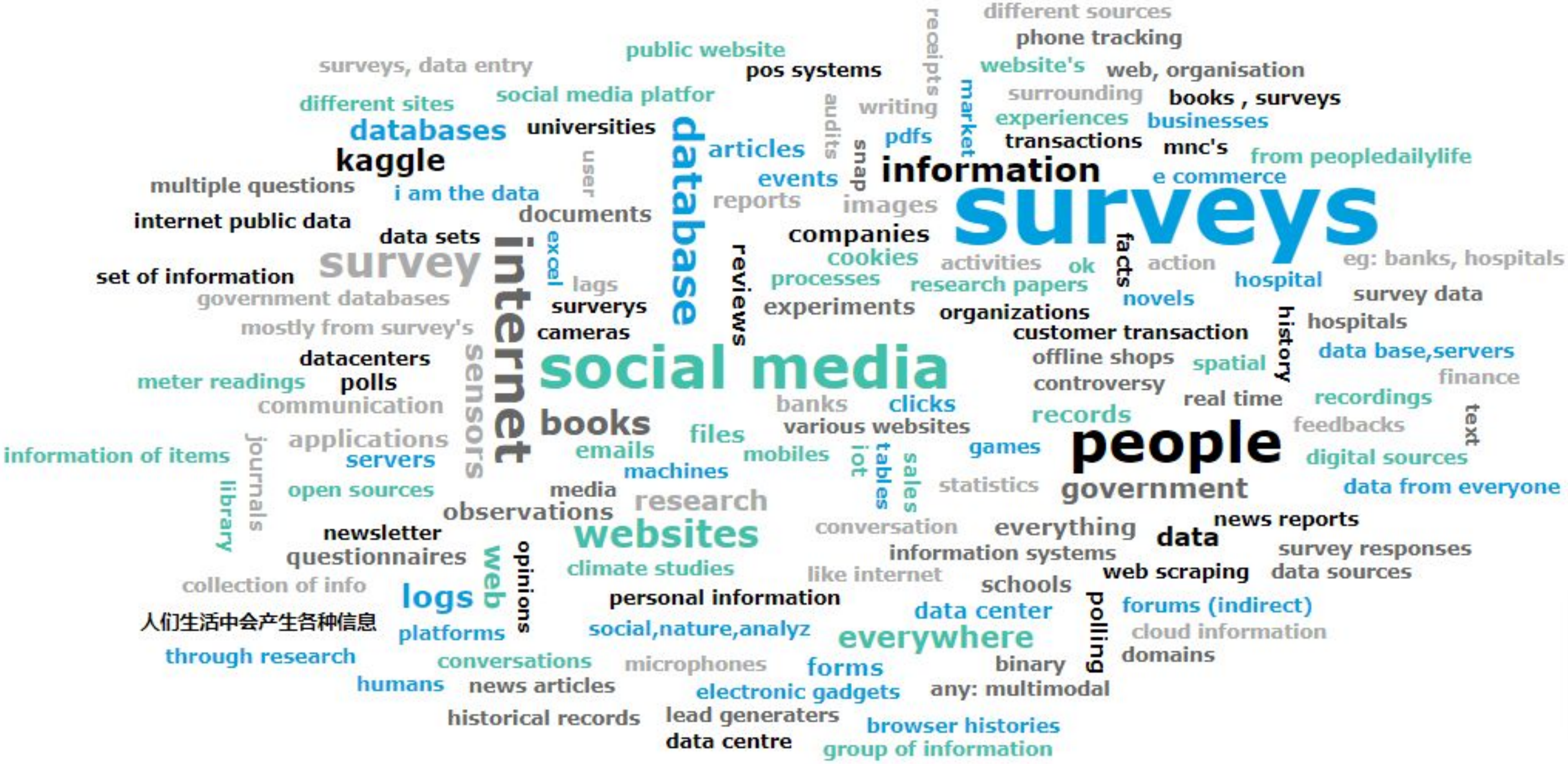


Type of data determines:

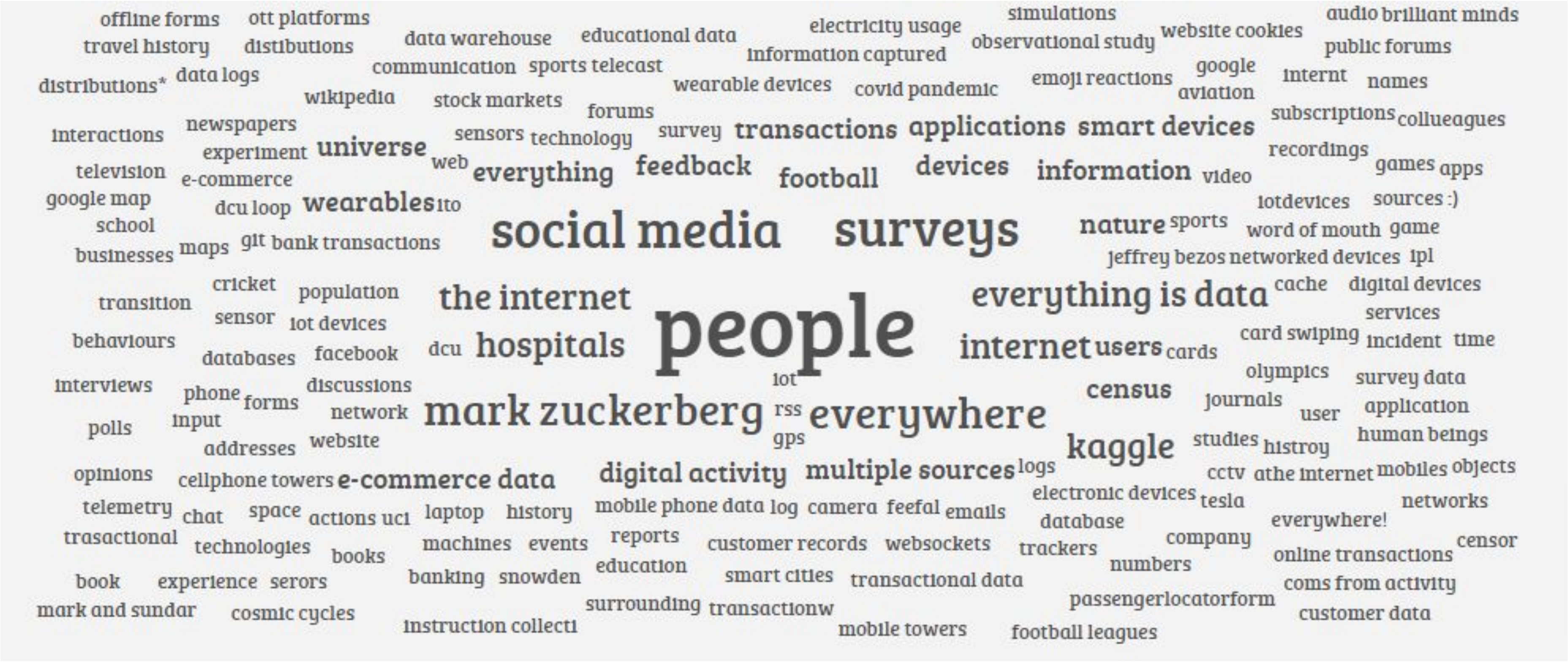
- What statistics are possible/meaningful
- How data can be processed and/or stored
- Which machine learning model can be used
- Which visualisation method to use

let’s try identifying types …
vevox.app

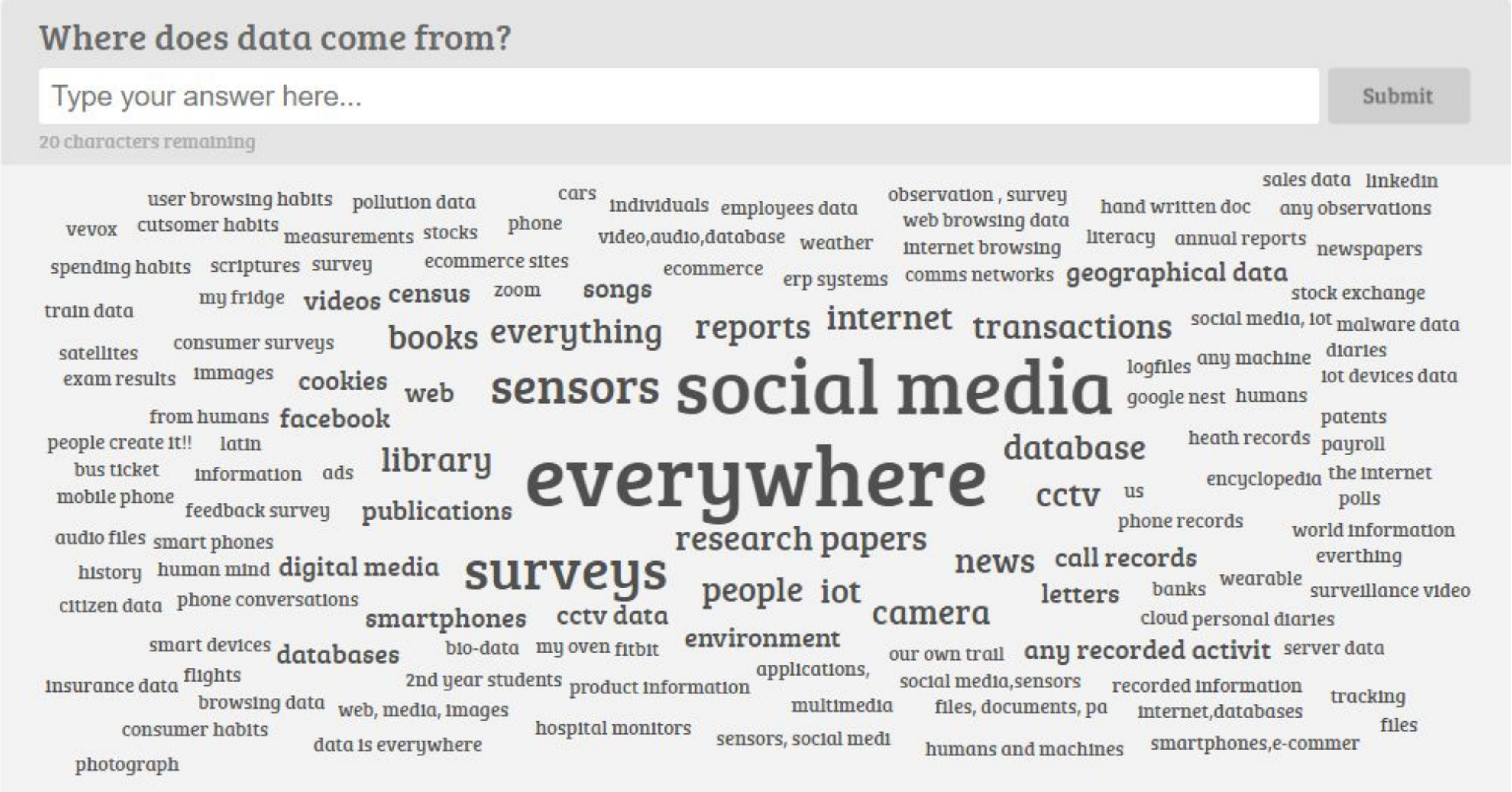
Data Sources


different sources
phone tracking
receipts
public website
surveys, data entry
pos systems
website's
web, organisation
social media platfor
different sites
audits
writing
market
surrounding
books , surveys
experiences
businesses
databases
universities
database
articles
snap
pdfs
transactions
mnc's
from peopledailylife
kaggle
user
information
e commerce
events
multiple questions
i am the data
reports
images
surveys
internet public data
documents
companies
data sets
excel
reviews
cookies
facts
activities
ok
action
eg: banks, hospitals
survey
internet
set of information
lags
processes
research papers
hospital
novels
survey data
government databases
surverys
experiments
organizations
history
hospitals
mostly from survey's
cameras
customer transaction
data base,servers
datacenters
sensors
social media
offline shops
spatial
finance
meter readings
polls
controversy
real time
recordings
communication
banks
clicks
text
books
various websites
records
feedbacks
applications
files
people
information of items
journals
servers
emails
mobiles
iot
tables
sales
games
digital sources
machines
open sources
media
statistics
government
data from everyone
library
research
observations
news reports
newsletter
websites
conversation
everything
data
questionnaires
web
opinions
information systems
survey responses
climate studies
like internet
web scraping
data sources
collection of info
schools
logs
personal information
polling
forums (indirect)
人们生活中会产生各种信息
data center
platforms
social,nature,analyz
everywhere
cloud information
through research
conversations
microphones
forms
binary
domains
humans
news articles
electronic gadgets
any: multimodal
historical records
lead generaters
browser histories
data centre
group of information

2022/2023 Class answers



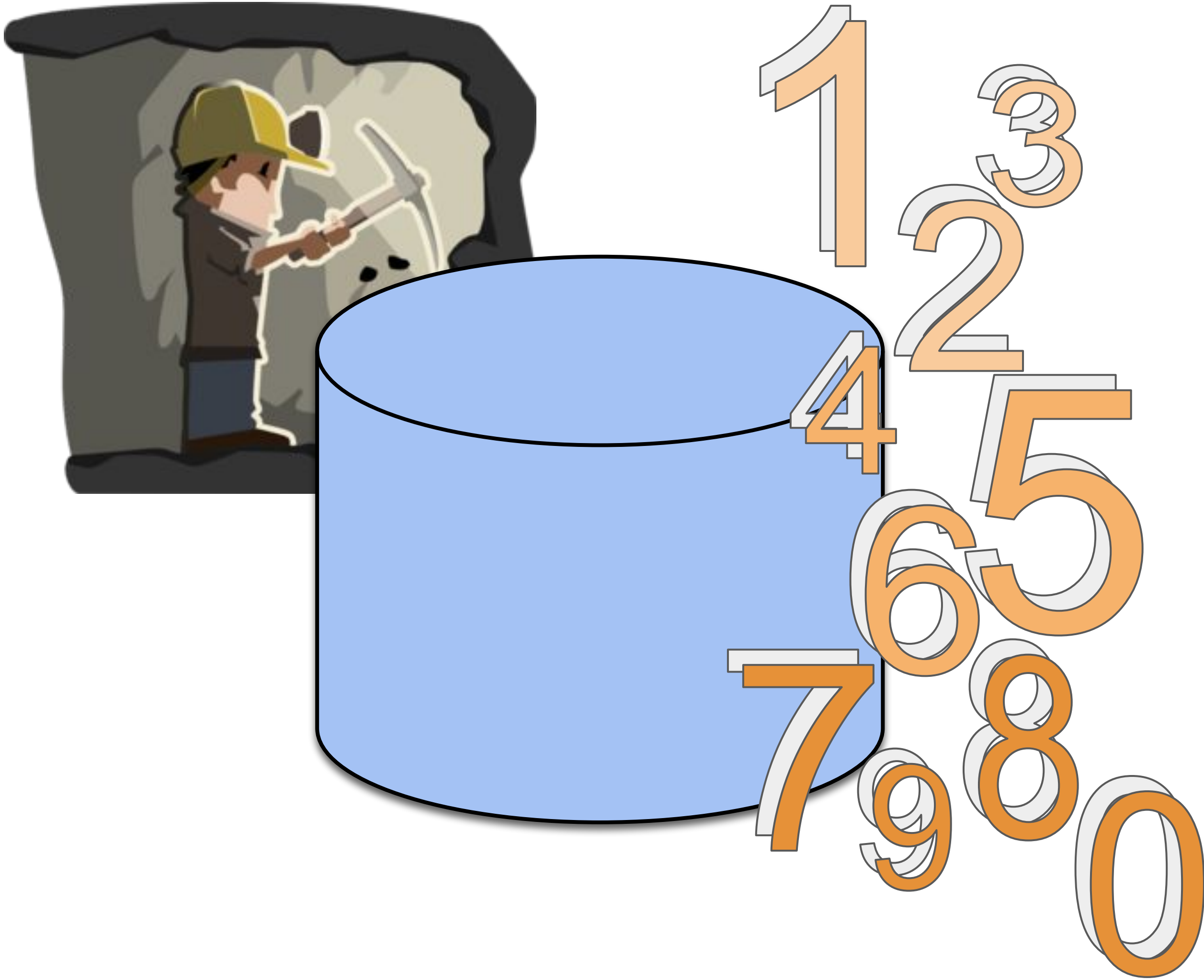
2020/2021 Class answers



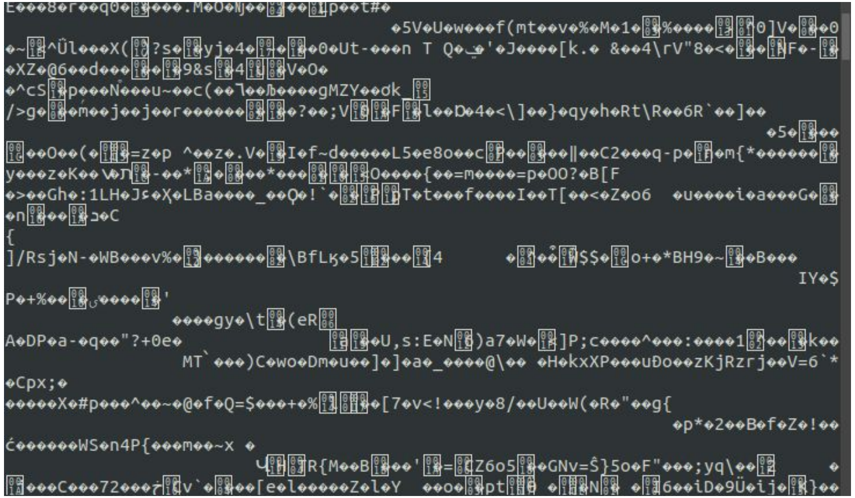
Where does data come from?

Data sources

- Files
- Databases
- “The Internet”
- Open Data



Data sources: Files



- Text or Binary
- Open or Proprietary
- Tabulated data - CSV, TSV, DB
- Other text data
    - JSON -- JavaScript Object Notation
    - XHTML -- web pages
    - KML (https://developers.google.com/kml/documentation/kml_tut?csw=1)
    - many other XML-based! (YAML …)
    - Specialist data formats (GDP, ASX etc.) -- may be proprietary
- List of file formats … (https://en.wikipedia.org/wiki/List_of_file_formats)

Data sources: Databases

- Traditional relational db: Oracle, MySQL, Postgres, etc.
  - Tables (“relations”) of rows and columns
  - Unique key per row
  - Links between rows (“foreign key”)
  - Optimise structures (the database schema)
  - Stored procedures (queries) to speed up responses
  - Most commonly use SQL - Structured Query Language
    - `SELECT CustomerName,City FROM Customers;`
    - `SELECT CustomerName,Age FROM Customers WHERE City=’Dublin’;`
- In memory databases: SAP Hana (http://hana.sap.com/abouthana.html)
- NoSQL, document, column, graph, etc.

Data sources: the Internet

- Crawlers or spiders
  - Scraping data from semi-structured sources
    - Parse HTML
    - Match Patterns to extract data
    - Identify links (repeat)
- URL
  - Files and databases on the web
  - Many libraries and apps will accept either a local path or url
- How many file formats?

Open data

Public data, shared and freely available

Why?

Why not?

Examples of open data

Some examples of projects that use open data are:

- Plantwise - Lose less, feed more (https://www.plantwise.org/)
- Humanitarian OpenStreetMap (https://www.hotosm.org/)
- OpenGLAM (https://openglam.org/)

Or on a less elevated topic … the Great British Public Toilet Map: open geospatial data!

Where to find open data?

- https://data.gov.ie/
- http://www.dublindashboard.ie/
- https://www.google.com/publicdata/directory
- https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/

Also lots of datasets for learning data science:

- https://www.kaggle.com/datasets
- https://github.com/datasets

Exercise

Data.gov is the portal for the US Government's Open Data. Browse the portal and find a dataset to answer the following questions.

1. What format is the dataset available in?
2. How many features (attributes or columns) does the data have?
3. Are the features mostly categorical (qualitative) or numerical (quantitative)?
4. What is a question that this dataset could help you answer? (you don't need to provide the answer!)

What’s on Loop or will be on github.com?

Material from last week plus Formal Data Management Lifecycles document

Slides & Notes on Data Types

Notes on Files

→ Exercise: Data Formats - Files

Notes on Open Data

→ Exercise: using open data

Linked Data (RDF & SPARQL) → Includes Exercise: using SPARQL on DBPedia