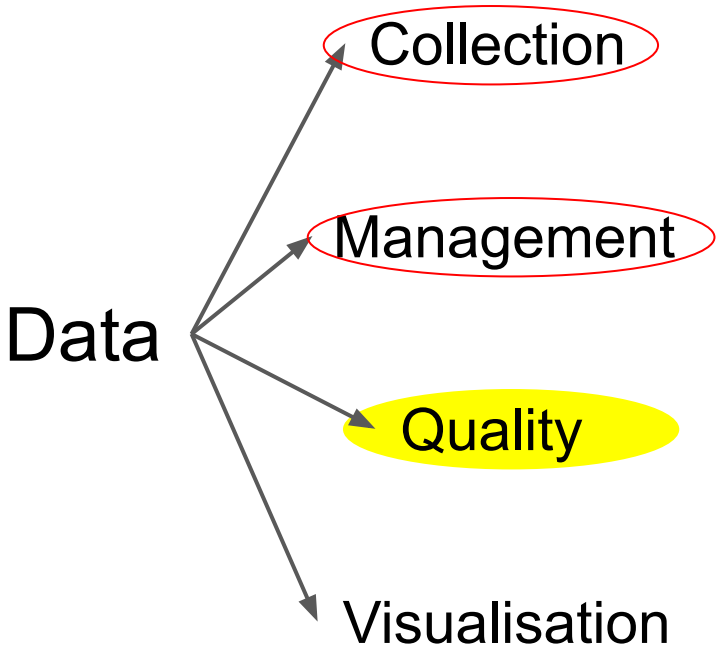
05 Data Quality & Cleaning

suzanne.little@dcu.ie

Recap



Outcomes

1. Analyse the requirements of applications handling large datasets.
2. Demonstrate an ability to efficiently structure a large dataset.
3. Implement data quality measures.
4. Identify and implement appropriate data visualization techniques.

DMV: where are we?

- Introduction: A Data Analytics Pipeline
  - Formal Data Management Lifecycles
- Data Collection: what is data? where does it come from?
  - data from files (text or binary, open or proprietary)
  - data types (SvU, QvQ, DvC, NOIR)
- Big data: 4 Vs
- Open Data
- Metadata - what is it? what is it used for?

Today: Data Quality & Cleaning → Lab Exercises: Spreadsheets, OpenRefine

Don’t forget to register your assignment pair!

Reminder: Housekeeping

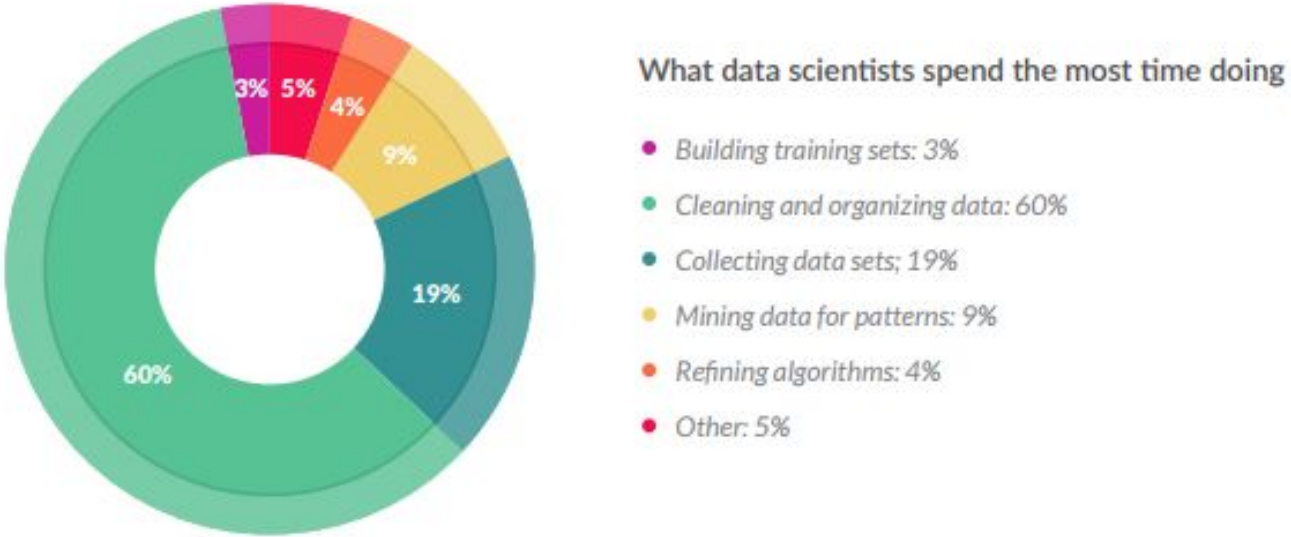
Be polite and respectful to your fellow students and to me (have your phone on silent, stay quiet when I'm talking)

Keep the back 2 rows and the aisle seats free until after 4:15pm

If an alarm sounds follow the exit signs outside to the assembly point

No food or drink allowed in the lecture theatre

I'll generally give you a 10 min break part way through



http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

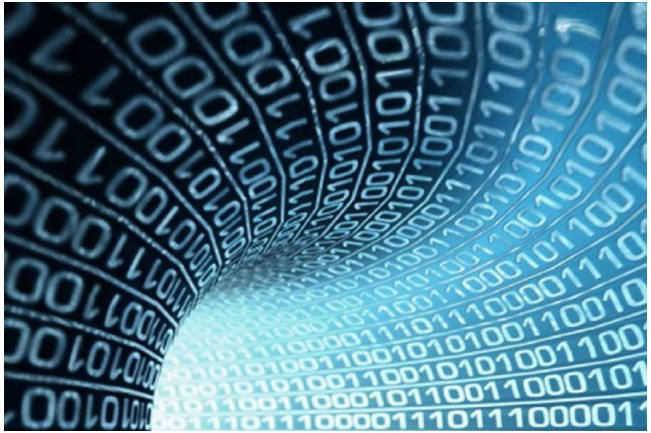
Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning

Objectives:

1. Managing data projects
2. Cleaning data

Data Quality and Cleaning



Computers are absolute (1 or 0)

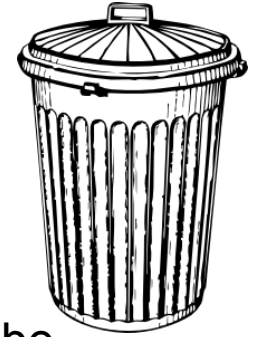
People are complicated

Data generated by people is complicated and often messy

How do we **assess data quality**?

How can we **clean data** for storage and processing?

Data Quality



High quality data is free from both errors and artefacts.

Error: data that is missing or lost due to the capture process and cannot be recovered.

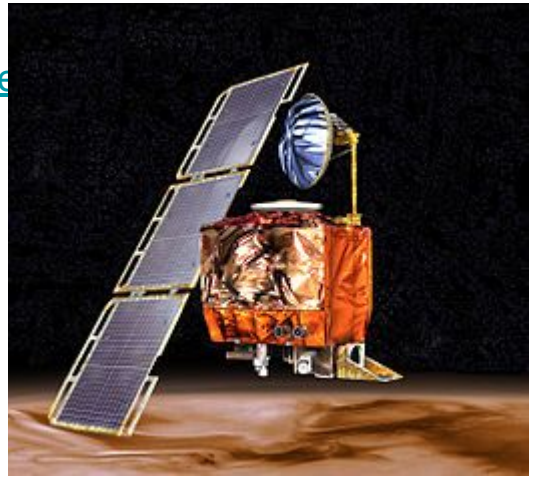
Artefact: something that has been introduced into the dataset during the gathering, processing, integration or cleaning activities.

Poor quality data may be due to **individual** (one off) or **collective** (systemic) issues.

What happens when the data quality is poor?

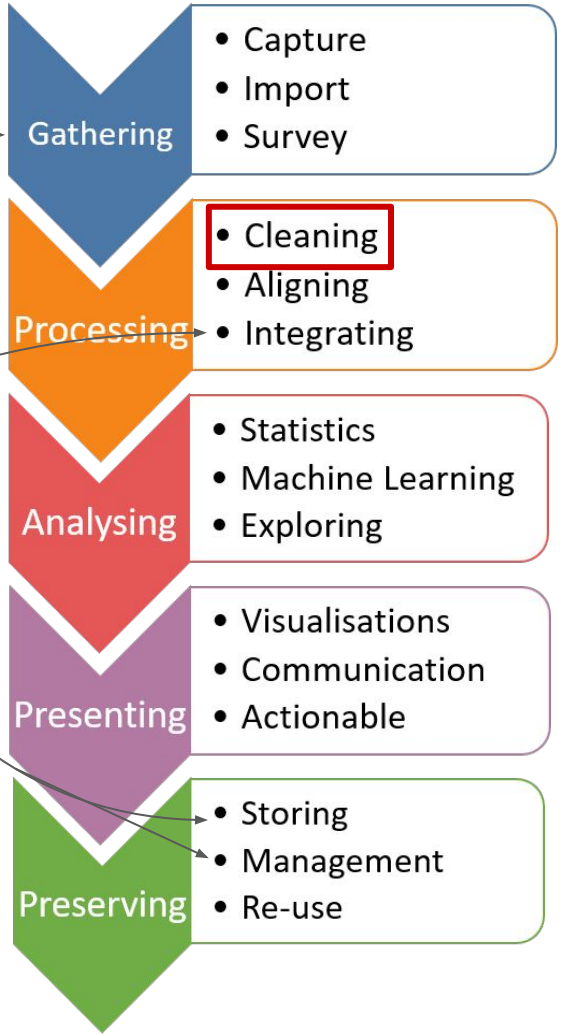
Mars orbiter: In 1999, the $125 million Mars climate orbiter was destroyed when incorrect units used by a contractor and NASA engineers (metric vs imperial) caused it to have the incorrect trajectory.

https://www.vice.com/en_us/article/qkvzb5/the-time-nasa-lost-a-mars-orbiter-be
mixup

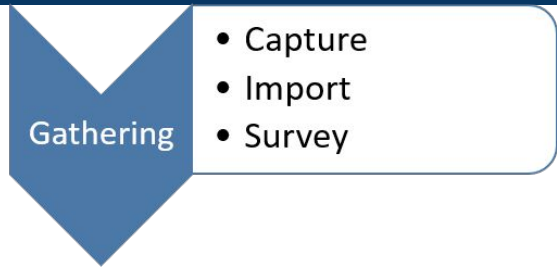


Where do problems occur?

- Data Gathering
- Data Storage
- Data Integration
- Data Analysis



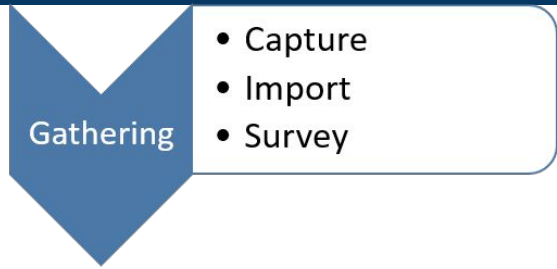
Data Gathering



- Manual entry errors or artefacts caused by typos, lazy people etc.
  - Eg. DUC or DCU
  - Eg. duplicate entry, duplicate entry, duplicate entry
- Poor survey or interface design
  - Eg. What colour is your toothbrush? The only options are red, green or blue! What if it's multicoloured?
  - Eg. A drop down for entering your age only goes back to 1923. No one alive older than 100?
- No standards for format or controlled vocabulary for fields
  - Eg. DCU or Dublin City University?
  - Eg. centimeters or inches?

Think back: the class questionnaires and survey in week 1 …

Data Gathering: solutions



Preemptive:

➔ build in integrity checks and entry constraints (process architecture)
➔ Process management - reward accurate human data entry, data sharing, data stewards, redundancy

Retrospective:

➔ Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
➔ Diagnostic focus (automated detection of glitches)

Data Delivery

Destroying or mutilating information by inappropriate pre-processing

- Inappropriate aggregation
- Nulls converted to default values

Loss of data:

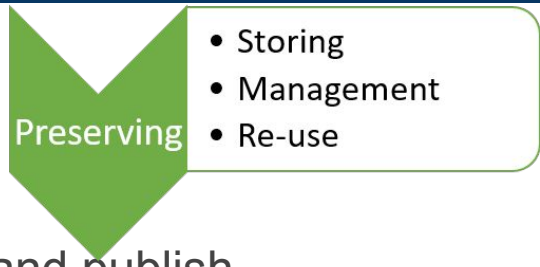
- Buffer overflows
- Transmission problems (network overload)
- No checks

→ “Understanding packet loss for sound monitoring in a smart stadium IoT testbed”

Data Delivery: solutions

- ➔ Build reliable transmission protocols
  - ◆ Use a relay server
- ➔ Verification
  - ◆ Checksums, verification parser
  - ◆ Do the uploaded files fit an expected pattern?
- ➔ Relationships
  - ◆ Are there dependencies between data streams and processing steps
- ➔ Interface agreements
  - ◆ Data quality commitment from the data stream supplier.

Data Storage



- Format conversion errors
  - string or float?
  - 1918 or 2018?
  - rounding or approximation (e.g., database field limited to integers)
- No metadata recorded
  - What does the field mean?

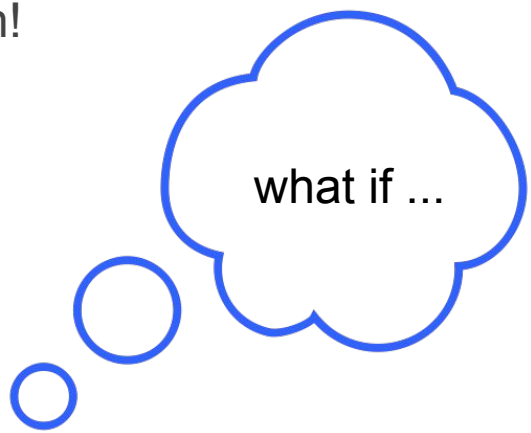
Also possible to have technical issues

- Transmission errors (network dropout)
- Disk failure or corruption

➔ Document and publish

➔ Data exploration and retrospective checking

➔ Assume that the worst might happen!



Data Integration

Combine data sets (acquisitions, across departments, organisations)

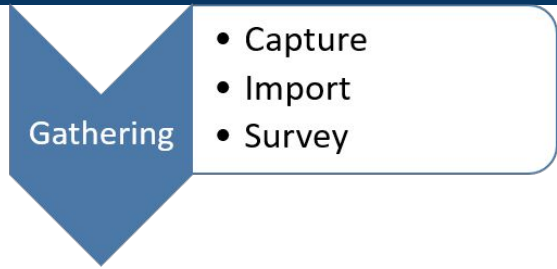
Common source of problems

- Heterogeneous data : no common key, different field formats, approximate matching
- Different definitions : What is a customer, an account, a family, …
- Time synchronization : Does the data relate to the same time periods? Are the time windows compatible?
- Legacy data : IMS, spreadsheets, ad-hoc structures, binary data
- Sociological factors : Reluctance to share – loss of power

Data Integration: solutions

- ➔ Data browsing and exploration
  - ◆ Many hidden problems and meanings : must extract metadata.
  - ◆ View before and after results : did the integration go the way you thought?
- ➔ Commercial Tools
  - ◆ Significant body of research in data integration
  - ◆ Many tools for address matching, schema mapping are available
- ➔ https://www.gartner.com/reviews/market/data-integration-tools

Data Retrieval



Exported data sets are often a *view* of the actual data. Problems occur because:

- Source data not properly understood
- Need for derived data not understood
- Just plain mistakes
  - Inner join vs. outer join
  - Understanding NULL values
- Computational constraints
  - E.g., too expensive to give a full history, we'll supply a snapshot.

Data Mining and Analysis

What are you doing with all this data anyway?

Problems in the analysis

- Scale and performance
- Confidence bounds? 0.95, 0.99?
- Attachment to models
- Insufficient domain expertise
- Casual empiricism (use of an arbitrary number to support a pre-conception)
    - 85% of all statistics are made up on the spot!

> Conan Doyle: "I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts." (Sherlock Holmes)

Data Analysis: solutions

- Data exploration
  - Determine which mod
  - the analysis part of the feedback loop

Engage your brain!
(smell test or commonsense)

Data exploration

◆ Determine which models and techniques are appropriate, find data bugs, develop domain expertise

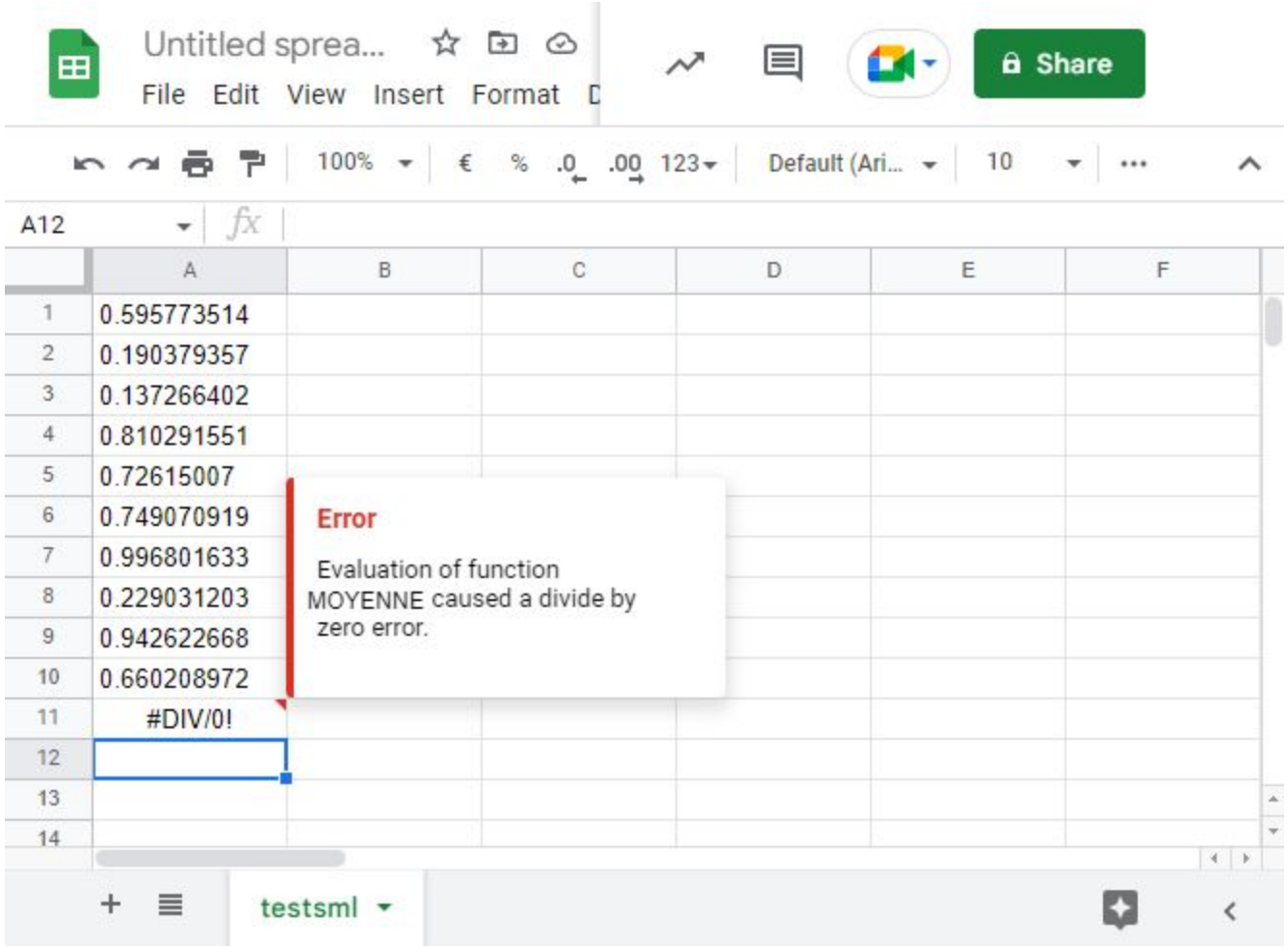
? Continuous analysis

◆ Are the results stable? How do they change?

? Accountability

◆ Make the analysis part of the feedback loop

Questions & Discussion


Untitled sprea...
File Edit View Insert Format
Share
100% € % .0 .00 123 Default (Ari... 10
A12
A B C D E F
1 0.595773514
2 0.190379357
3 0.137266402
4 0.810291551
5 0.72615007
6 0.749070919
7 0.996801633
8 0.229031203
9 0.942622668
10 0.660208972
11 #DIV/0!
12
13
14
Error
Evaluation of function MOYENNE caused a divide by zero error.
testsml

Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning

Conventional measures of data quality

Accuracy : The data was recorded correctly.

Completeness : All relevant data was recorded.

Uniqueness : Entities are recorded once.

Timeliness : The data is recent or kept up to date.

Date published vs Data captured ...

Consistency : The data agrees with itself (internal).

Credibility : The data comes from a recognised (or official) source.

Problems with conventional measures

Unmeasurable: Accuracy and completeness are extremely difficult, perhaps impossible to measure.

Context independent: No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.

Incomplete: What about interpretability, accessibility, metadata, analysis, etc.

Vague: The conventional definitions provide no guidance towards practical improvements of the data.

So what do you do to measure data quality?

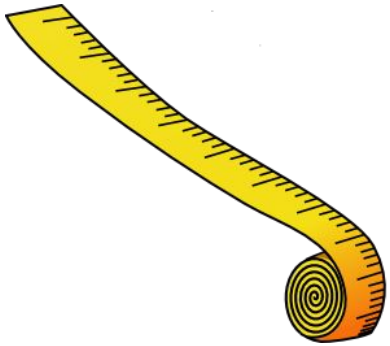
Some options ...

- Inventory (expensive)
- Using a proxy measure such as tracking customer complaints
- Applying formal measures of accessibility
- Using test cases with known results and checking for glitches or errors in analysis
- Successfully completing an end-to-end process (e.g., data ingestion, processing, indexing, querying and summarisation)

Data quality constraints

- Many data quality problems can be captured by **static constraints** based on the schema.
  - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to **problems in workflow**, and can be captured by **dynamic constraints**
  - E.g., orders above $200 are processed by Biller 2
- The constraints follow an 80-20 rule
  - A few constraints capture most cases, thousands of constraints to capture the last few cases.
- Adherence to constraints are **measurable**.

Data quality metrics



- Measure to improve → incentivise
  - Indicates what is wrong and how to improve
  - Realize that DQ is a messy problem, no set of numbers will be perfect
- Types of metrics
  - Static vs. dynamic constraints
  - Operational vs. diagnostic
- A very large number metrics are possible → choose the most important ones
- Warning: Metrics can give incentives for bad behavior → throw away data that doesn’t join.

Methods for data cleaning

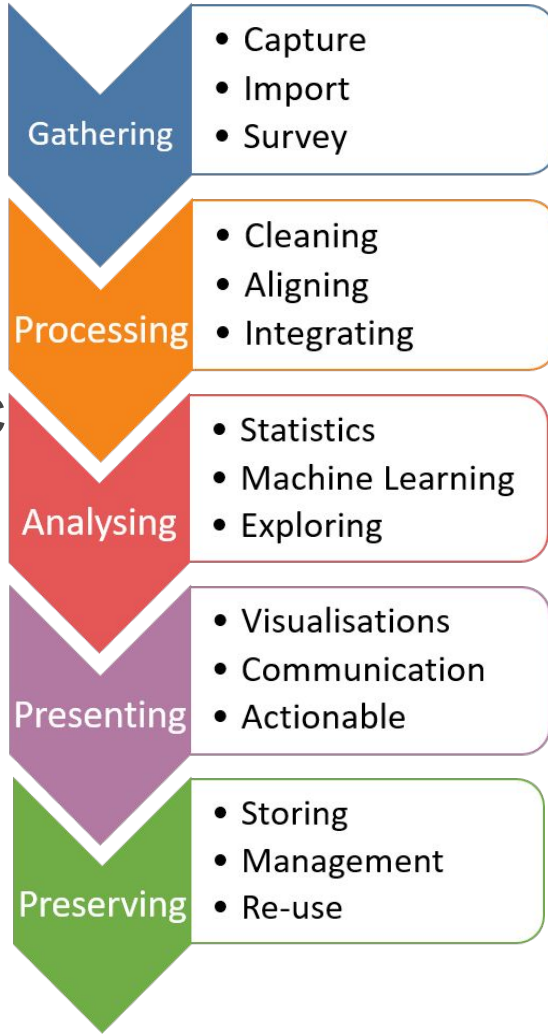
1. Implement process mandates (fix the human problem)
   a. Including schema or rule restrictions to enforce format
2. Custom tools written in a General Purpose language (hack a script)
   a. Good for one-off quick fixes (cleaning that only happens once)
3. Off the shelf tools such as Spreadsheets, OpenRefine (formerly Google Refine)
   a. Form of exploratory data analysis and cleansing

Using something like Jupyter or R-studio is a mix of 2 & 3 as you can interactively explore and refine but also document your quick fix for future use.

Application

- A: House price census missing Collins Ave
- B: Forex data for EUR-GBP missing May 2016
- C: Temperature values entered in ºF rather than ºC
- D: Entries overwritten when merging company records

1. Which phase or task is the likely cause?
2. Is it an error or an artefact?
3. What solutions could you use?
4. What might be possible consequences?



More examples

- Sound monitoring packets lost when network overloaded
- See also http://okfnlabs.org/bad-data/ for a few examples
- "Excel: Why using Microsoft's tool caused Covid-19 results to be lost"
- Mis-calculation examples (not always "bad" data), https://www.bbc.com/news/magazine-27509559

Practical suggestions for initial data quality checks

- Missing values, records or variables - are empty cells no value (0) or no measurement (null)? How should they be handled?
- Erroneous values - typos or values that are clearly out of place (gender value in age column
- Inconsistencies - capitalisation, units of measurement
- Duplicate records
- Out of date - e.g., age will have changed
- Leading or trailing spaces! Windows or Linux end of line characters
- Format of dates - DD/MM/YYYY, MM/DD/YYYY, ?? Excel based or Unix based
- “Sanity checks” - look for extreme values or outliers, count how many records

Tools for Data Cleaning

Many options

Spreadsheets (Google, Excel, etc)

Purpose built tools (RapidMiner, TableauPrep, OpenRefine)

General Purpose Language (Python, R, etc)

Some things to consider:

- How big is your data?
- How frequently are you cleaning? Once off or regularly?
- How will you document your work?

OpenRefine

- Until late 2012, Google Refine, now no longer maintained by Google
- Offline (desktop) tool that runs in the browser
- OpenRefine can:
  - Import/export a range of data types
  - Explore data (use graphs to check distributions and outliers)
  - Clean and transform data
    - Fix errors in fields
    - Use heuristics to group data and spot errors
  - Call services to enhance your data (e.g. Geocoding)
    - What OpenRefine calls *reconciling*
  - Handle large (-ish) datasets
- http://openrefine.org/

Tools for data cleaning -- LLMs?

What about using something like ChatGPT to clean data? Sounds attractive!

Pros: no code experience required, faster results

Cons: Can you trust it? Can you replicate it consistently?

Try the exercise (link on loop) to compare two LLMs

https://www.packtpub.com/article-hub/data-cleaning-made-easy-with-chatgpt
https://robertorocha.info/using-chatgpt-to-clean-data-an-experiment/

If we don't know if the dates are in US or EU format, which of these dates are ambiguous?

Ambiguous dates are those where the day and month can be confused depending on whether the format is US (MM/DD/YY) or EU (DD/MM/YY). For example, dates like "07/11/1982" could either mean July 11 (US) or November 7 (EU).

Here are the ambiguous dates from your list:

1. **27/06/91** – Could be June 27 (EU) or July 6 (US)
2. **06-18-83** – Could be June 18 (US) or 18th June (EU) (but more likely US due to order)
3. **03/10/82** – Could be October 3 (US) or March 10 (EU)
4. **10/10/83** – Could be October 10 (US or EU)
5. **12-05-98** – Could be May 12 (US) or December 5 (EU)
6. **10/10/1985** – Could be October 10 (US or EU)
7. **07/11/1982** – Could be November 7 (EU) or July 11 (US)

These dates are ambiguous because the day and month values are both 12 or less, making it hard to determine which is the day and which is the month without additional context.

Summary

Poor quality data has big consequences

If you can, control for quality in gathering and integration before you do analysis!

Build a “tool kit” of programs, methods and questions to satisfy yourself that your data is clean before you analyse or visualise.

Develop your sense of when, where and how problems might happen (How can you do this? Read the references, think about the causes)

Next steps

Document linked from loop has detailed summary and other links

Suggest doing the European Data Portal Course on Data Cleaning, linked from loop

Four sets of exercises to perform data cleaning -- labs this week & next week:

1. Google Sheets (Oct 10th)
2. OpenRefine (Oct 10th)
3. Python/Pandas in notebooks (Oct 17th)
4. LLMs (Oct 17th)

Next weeks lab: an exam style question to review data cleaning