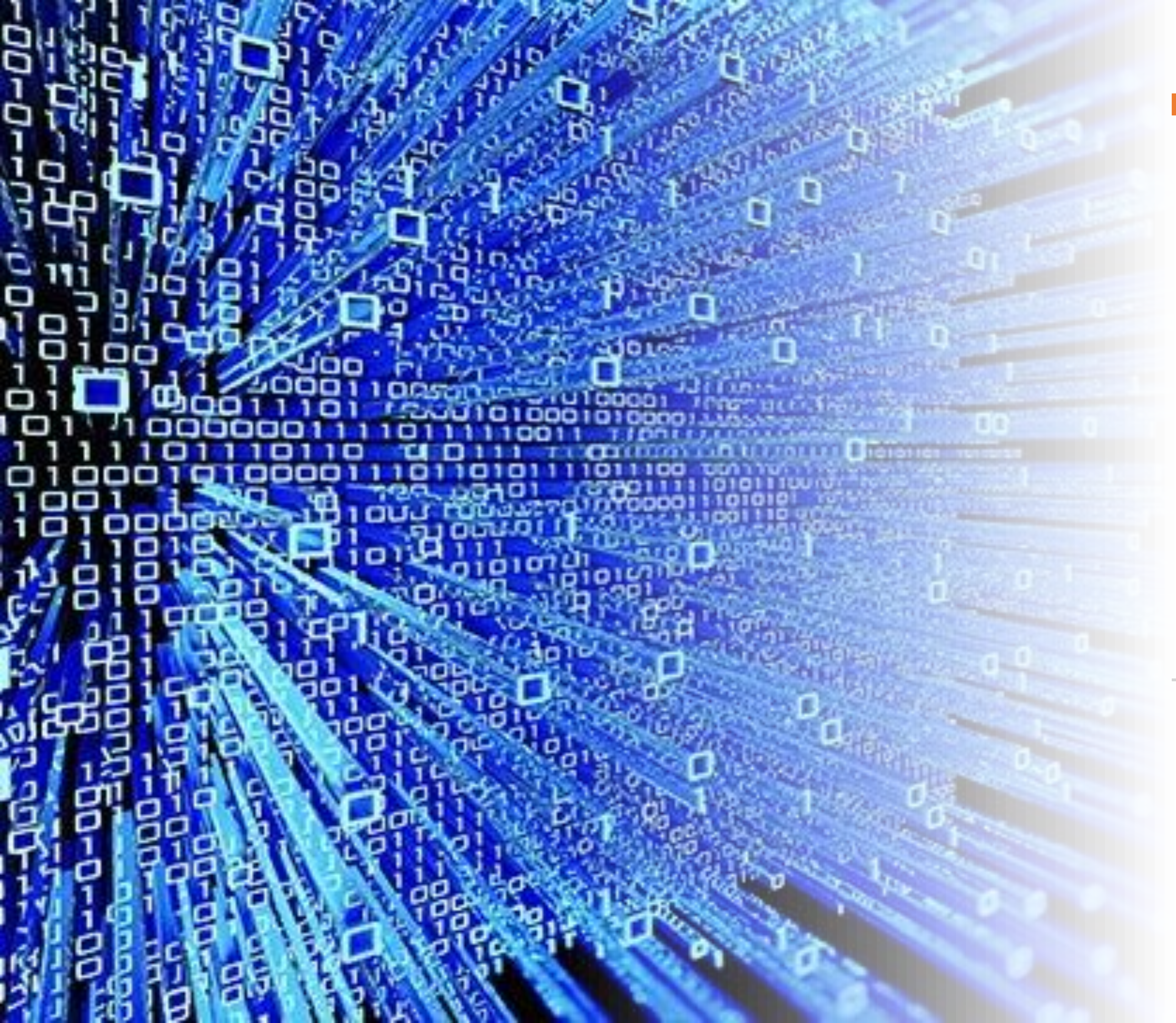
WQD7001



Finding & Getting Data

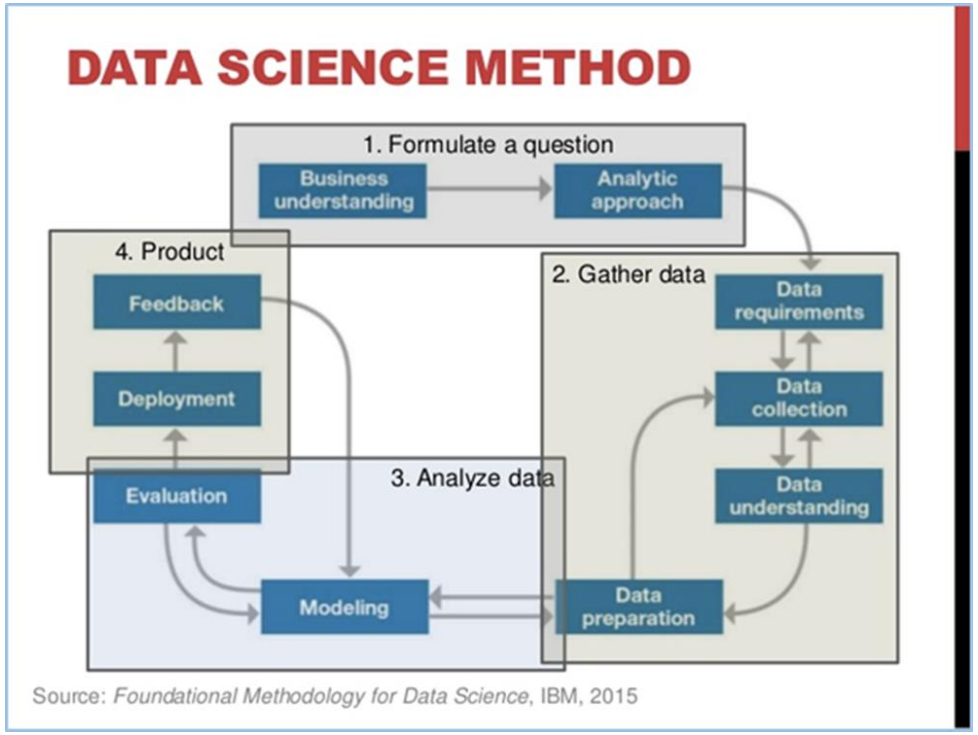
By Dr. Salimah M

Learning Objectives:

1. To map roles and tools to data scientist activities.
2. To explain data literacies.
3. To discuss finding data.
4. To discuss getting data.

The Data

- The **most important** thing in data science is the **<u>question</u>**.
- The **second most important** thing is the **<u>data</u>**.
- Often the data will limit or enable the questions.
- But having data cannot save you if you do not have a question.

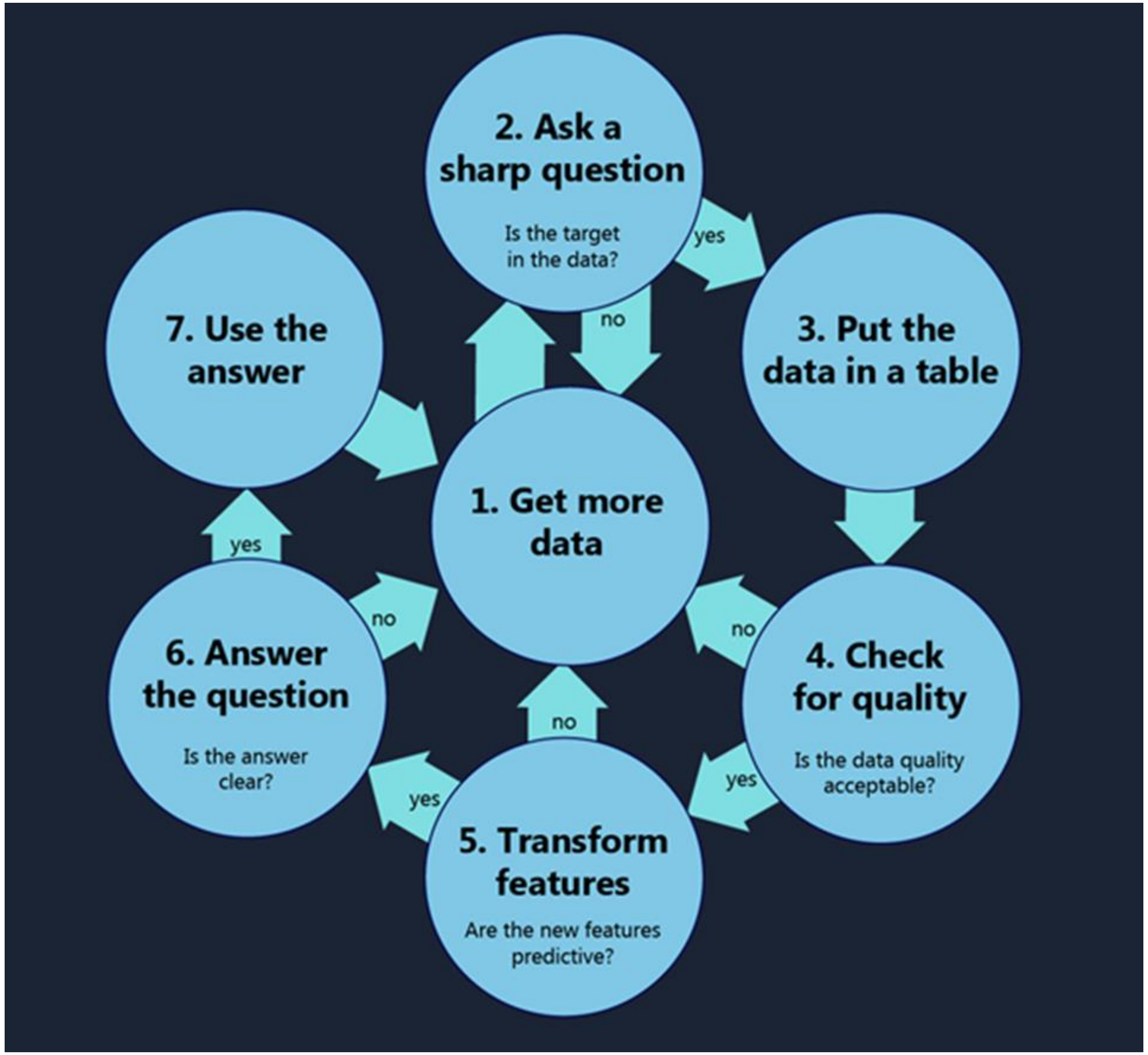


Source: *Foundational Methodology for Data Science*, IBM, 2015

Data

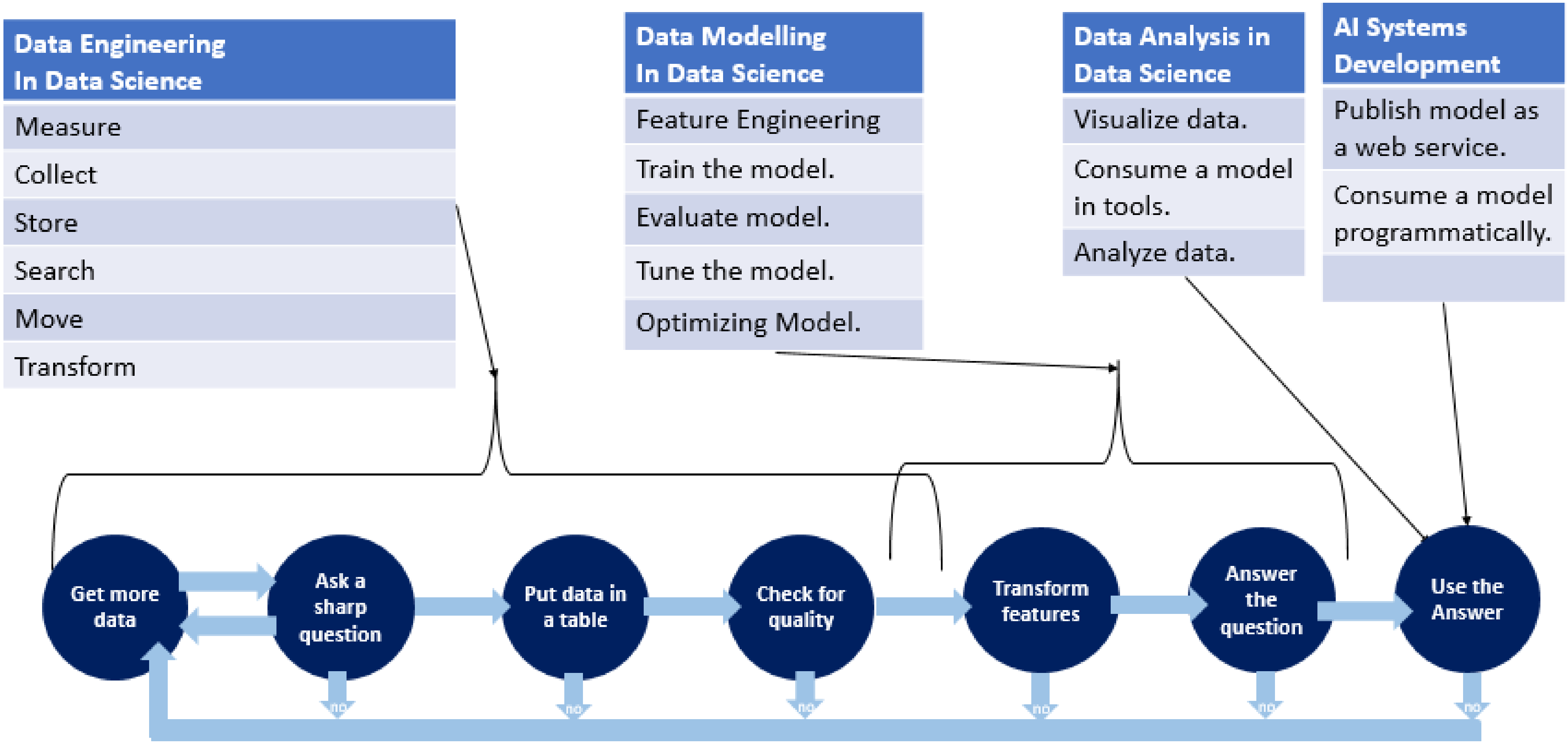
- Data are values of qualitative or quantitative variables, belonging to a set of items.

Data Science Process



https://blogs.technet.microsoft.com/machinelearning/2016/03/28/how-to-do-data-science/

Mapping Data Engineering, Modelling and Analysis to Data Scientist Activities



Data Engineering In Data Science
Measure
Collect
Store
Search
Move
Transform

Data Modelling In Data Science
Feature Engineering
Train the model.
Evaluate model.
Tune the model.
Optimizing Model.

Data Analysis in Data Science
Visualize data.
Consume a model in tools.
Analyze data.

AI Systems Development
Publish model as a web service.
Consume a model programmatically.

Get more data

Ask a sharp question

Put data in a table

Check for quality

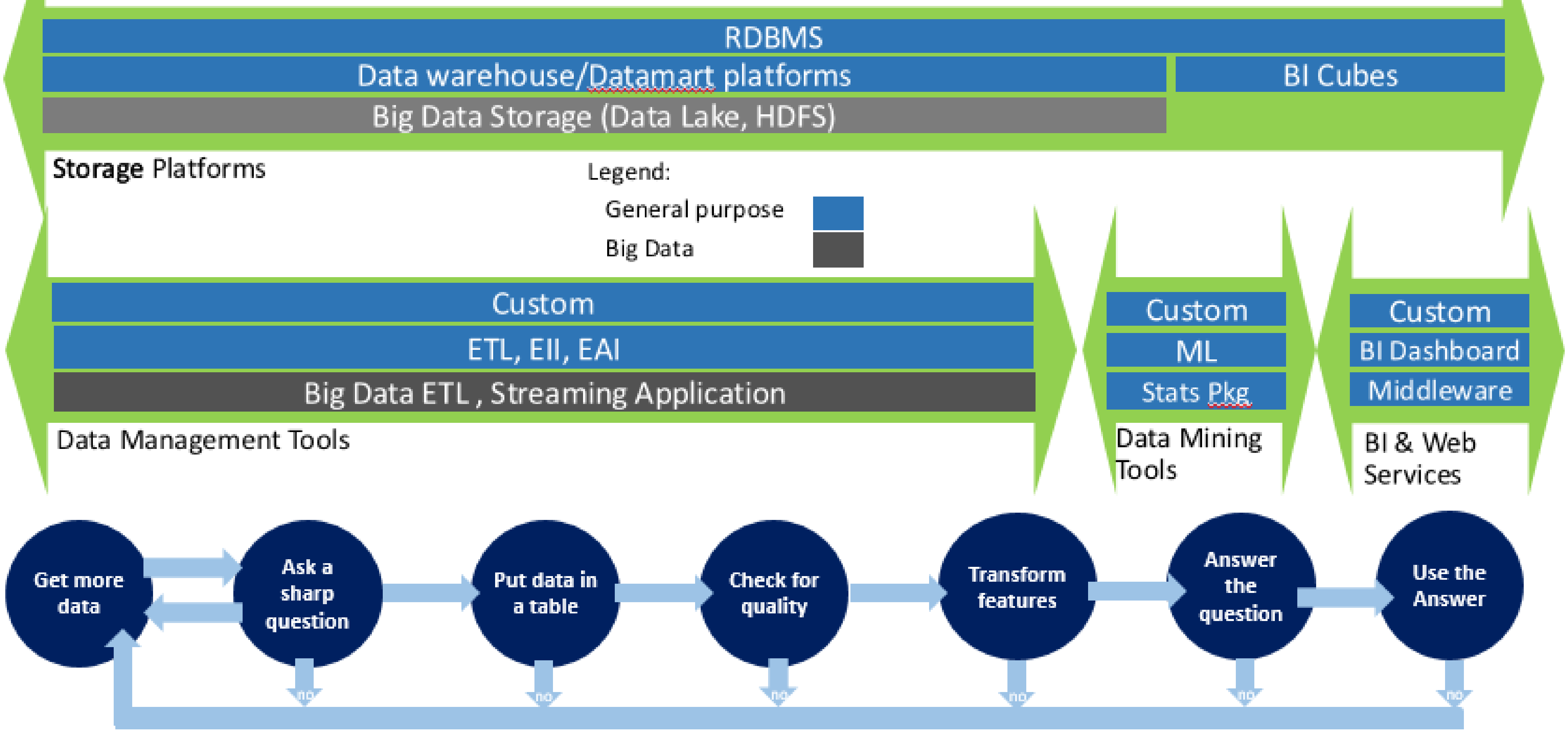
Transform features

Answer the question

Use the Answer

* Credit to: Dr. Dzahar Mansor, National Technology Officer, Microsoft Malaysia

Mapping Generic Tools to Data Scientist Activities



RDBMS

Data warehouse/Datamart platforms

BI Cubes

Big Data Storage (Data Lake, HDFS)

Storage Platforms

Legend:

General purpose 🟦

Big Data ⬛

Custom

ETL, EII, EAI

Big Data ETL , Streaming Application

Data Management Tools

Custom

ML

Stats Pkg

Data Mining Tools

Custom

BI Dashboard

Middleware

BI & Web Services

Get more data

Ask a sharp question

Put data in a table

Check for quality

Transform features

Answer the question

Use the Answer

no

no

no

no

no

no

* Credit to: Dr. Dzahar Mansor, National Technology Officer, Microsoft Malaysia

Data Scientist Fundamental Skills

- Data = NUMBER, LABEL.
- Transformation may be required:
  - Some NUMBER are LABEL.
  - Some LABEL can be NUMBER.

- Sharp question must be answered with a NUMBER or LABEL..
- Must define a TARGET NUMBER or LABEL.
- Example: What will be the stock price next week.
  - Sharp question.
  - Target is stock price.

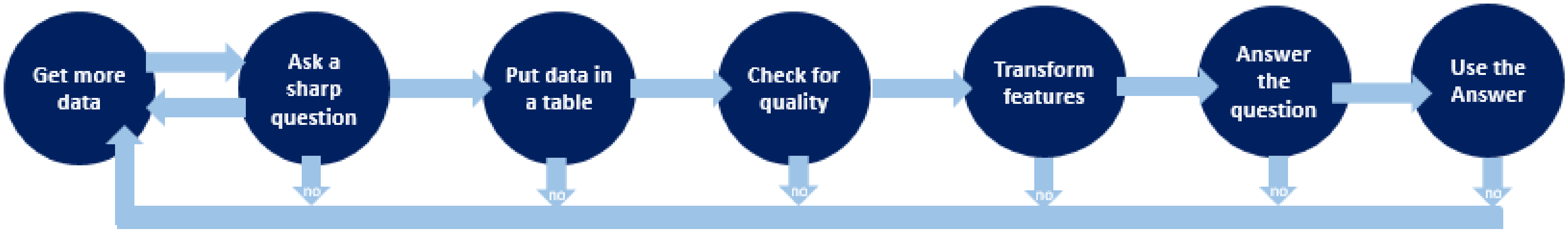
- Build one table.
- One TARGET per row.
- In building the table from various sources, may need to:
  - Aggregate.
  - Distribute.
  - Compute.
  - Measure.
  - Estimate.
  - Leave Blanks.

- No short cuts, need to "walk through the table column by column".
- Some typical cleansing techniques::
  - Unify (labels (NO, No, no to N)., Unify the meanings into standard labels (mutants, villains into BAD).
  - Convert string representation of number to a number (5' 6" to 66 (inches)).
- Handle values that are still missing.

- Feature Engineering – messaging data into a form suitable prediction.
- Some typical techniques:
  - Multiply columns (default).
  - Column Difference .
  - Data-specific e.g Images (SIFT), Text (TF-IDF)
  - Domain specific e.g Econometric, agricultural, sociological,. Etc.
  - Deep learning e.g. Images, text, audio

- 5 question Data Scientists ask:
  - 1. How much / how many?
  - 2. Which category?
  - 3. Which groups?
  - 4. Is it weird?
  - 5. Which action?

- Same ways to use the answer:
  - Make a web service.
  - Make a decision.
  - Set a price.
  - Publish code.
  - Write a report on the results.
  - Visualize (e.g. build a dashboard etc).



* Credit to: Dr. Dzahar Mansor, National Technology Officer, Microsoft Malaysia

Big or Small Data?
You need the RIGHT data

It’s What You Do with It

“The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data …”

John Tukey (Wikiquote)

John Tukey (1915-2000)
American Statistician

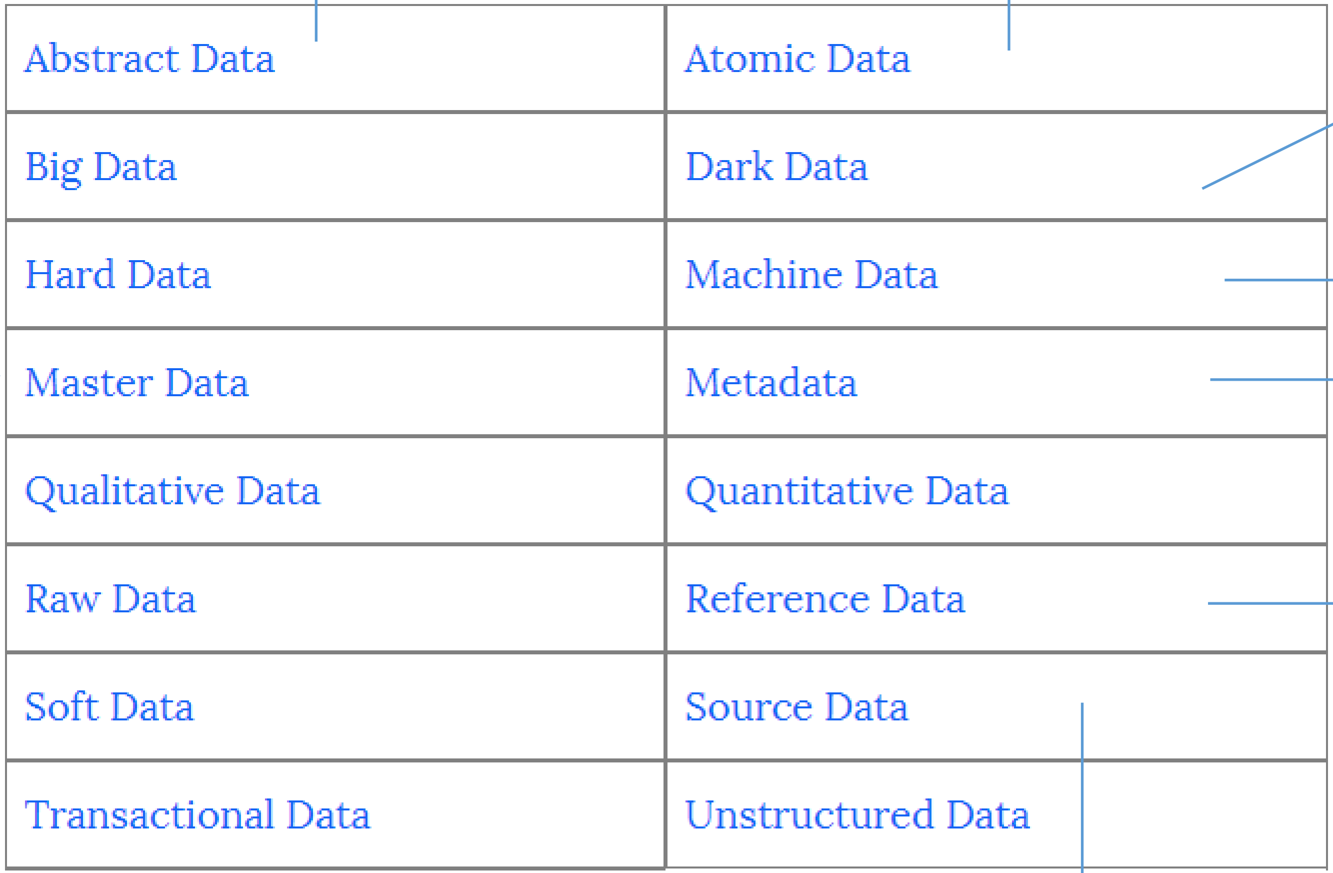
UBER success story

- Uber’s success isn’t a function of the big data it collects.
- That big data has enabled the company to enter new markets and fulfill new jobs in the lives of its customers.
- Uber’s success results from something very different: the **small**, **right** data it needed to do something very simple — **dispatch cars**.

Data Literacy – Data Types

https://simplicable.com/new/data

All About Data



Abstract Data	Atomic Data
Big Data	Dark Data
Hard Data	Machine Data
Master Data	Metadata
Qualitative Data	Quantitative Data
Raw Data	Reference Data
Soft Data	Source Data
Transactional Data	Unstructured Data

Data based on an intuitive concept

Data that can't be deconstructed to parts or sources.

Eg Base salary, unit sales, post codes

Data that an organization has collected but hasn't used.

Data that is generated by machines without human involvement.

Data that describes other data.

Data that is used to structure and constrain other data, typically stable information with a known set of values that rarely change.

- A list of valid states for a country
- Constants such as PI
- List of valid months and days of the week

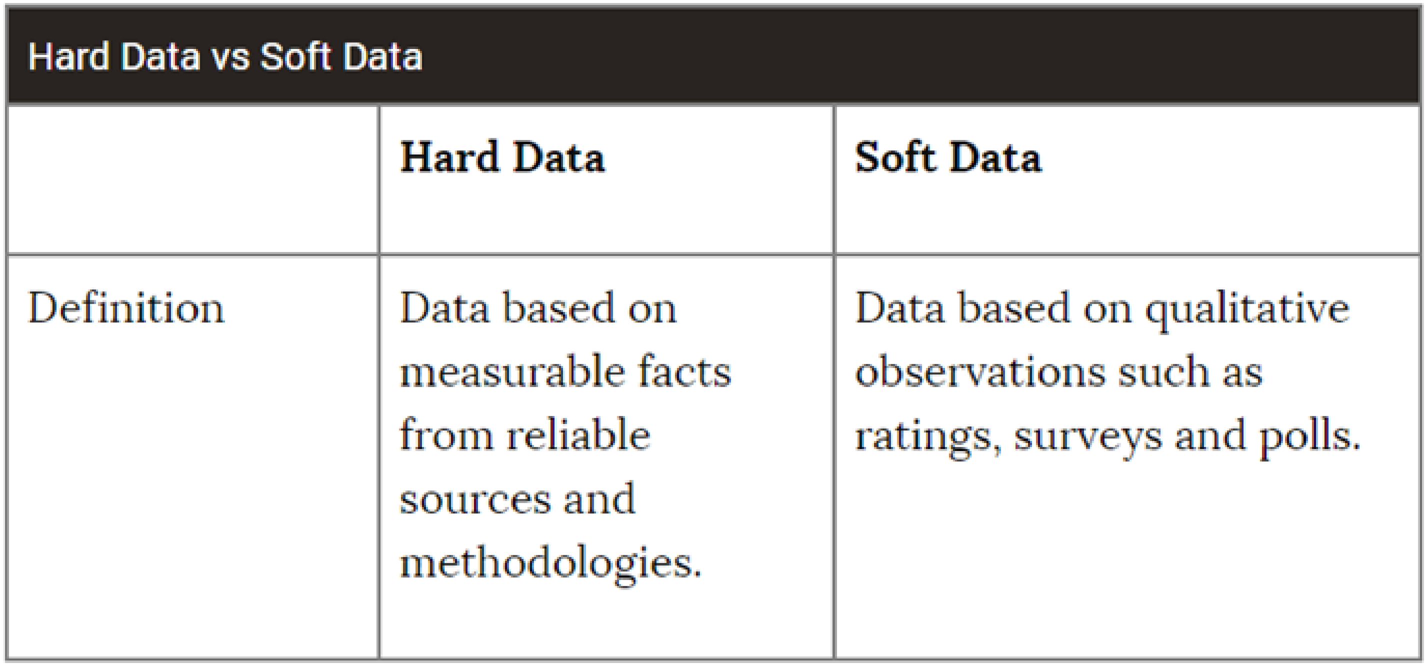
Business objects which are agreed on and shared across the enterprise

Data that hasn't be processed

Data Literacy – Data Types

- **Qualitative data** - everything that refers to the quality of something: a description of colors, the texture and feel of an object, a description of experiences, and an interview are all qualitative data.

- **Quantitative data** - data that refers to a number, e.g. the number of golf balls, the size, the price, a score on a test, etc.



Hard Data vs Soft Data

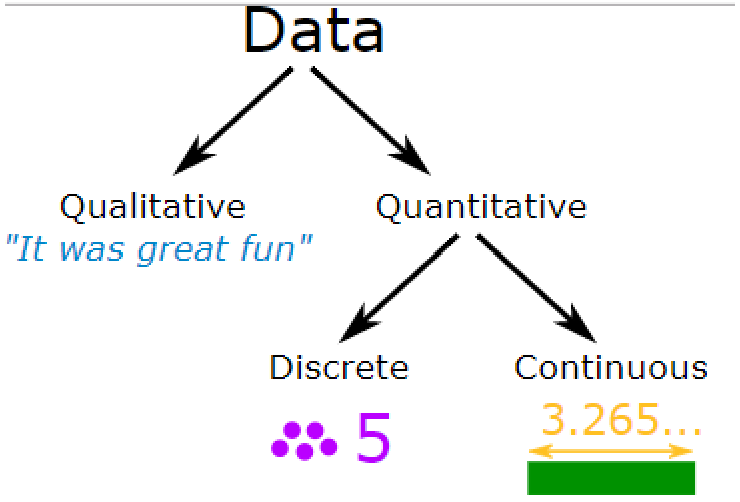
	Hard Data	**Soft Data**
Definition	Data based on measurable facts from reliable sources and methodologies.	Data based on qualitative observations such as ratings, surveys and polls.

Data Literacy – Data Types

- **Categorical data** puts the item we describe into a category.
  - For example, an item can categorized as “new”, "used" or “broken”.

- **Discrete data** is based on counts. Only a finite number of values is possible, and the values cannot be subdivided meaningfully.
  - Examples are scores in tests, head count, shoe sizes or number of languages a person speaks.

- **Continuous data** is numerical data with a continuous range. It can take any value (within a range).
  - Examples are a person's height (could be any value, within the range of human heights), weight of cars, speed of the train.



Counted Measured

Unstructured vs. Structured Data

Data for Humans

A plain sentence – “we have 5 white used golf balls with a diameter of 43mm at 50 cents each” – might be **easy for a human** to understand, but for a computer, it is very difficult.

- The above sentence is what we call **unstructured data**.
- **Unstructured data** has no transparent underlying structure—it's impossible to mechanically figure out exactly what refers to what.
- Likewise, **PDFs** and **scanned images** may contain information which is **pleasing to the human eye** as it is laid-out nicely, but they are **not machine-readable** or structured presentations of data.

Data for Computers

- Computers are inherently different from humans. It can be exceptionally hard to make computers extract information from certain sources. Some tasks that humans find easy are still difficult to automate. For example, interpreting text that is presented as an image is still a challenge for a computer.
- If you want your computer to process and analyze your data, it has to be able to read and process it. This means it needs to be **structured** and **machine-readable**.

CSV

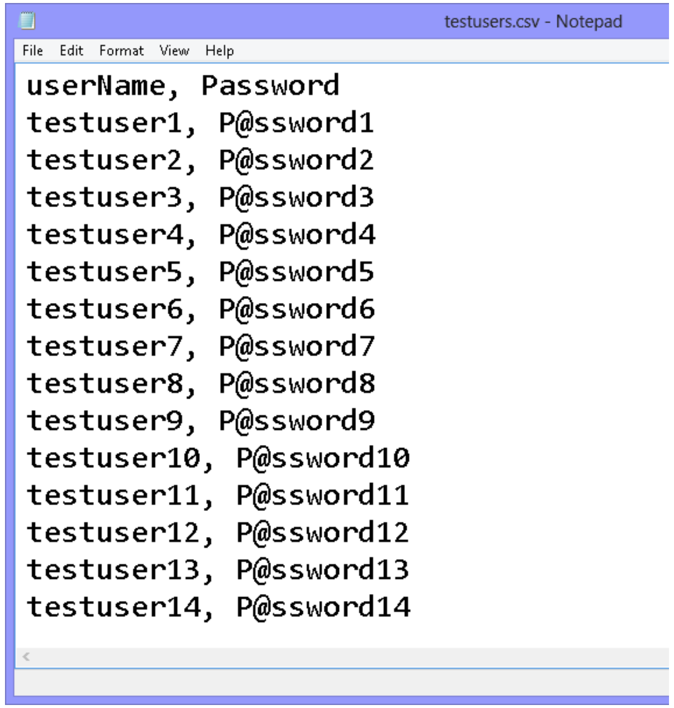
Most commonly used formats for exchanging data is CSV, which stands for "**comma separated values**".

It might look like this:

“quantity”, “color”, “condition”, “item”, “category”, “diameter (mm)”, “price per unit (AUD)”, 5,”white”,”used”,”ball”,”golf”,43,0.5

Data formatted like this is **tabular data**, data which forms a table consisting of **rows** and **columns**.

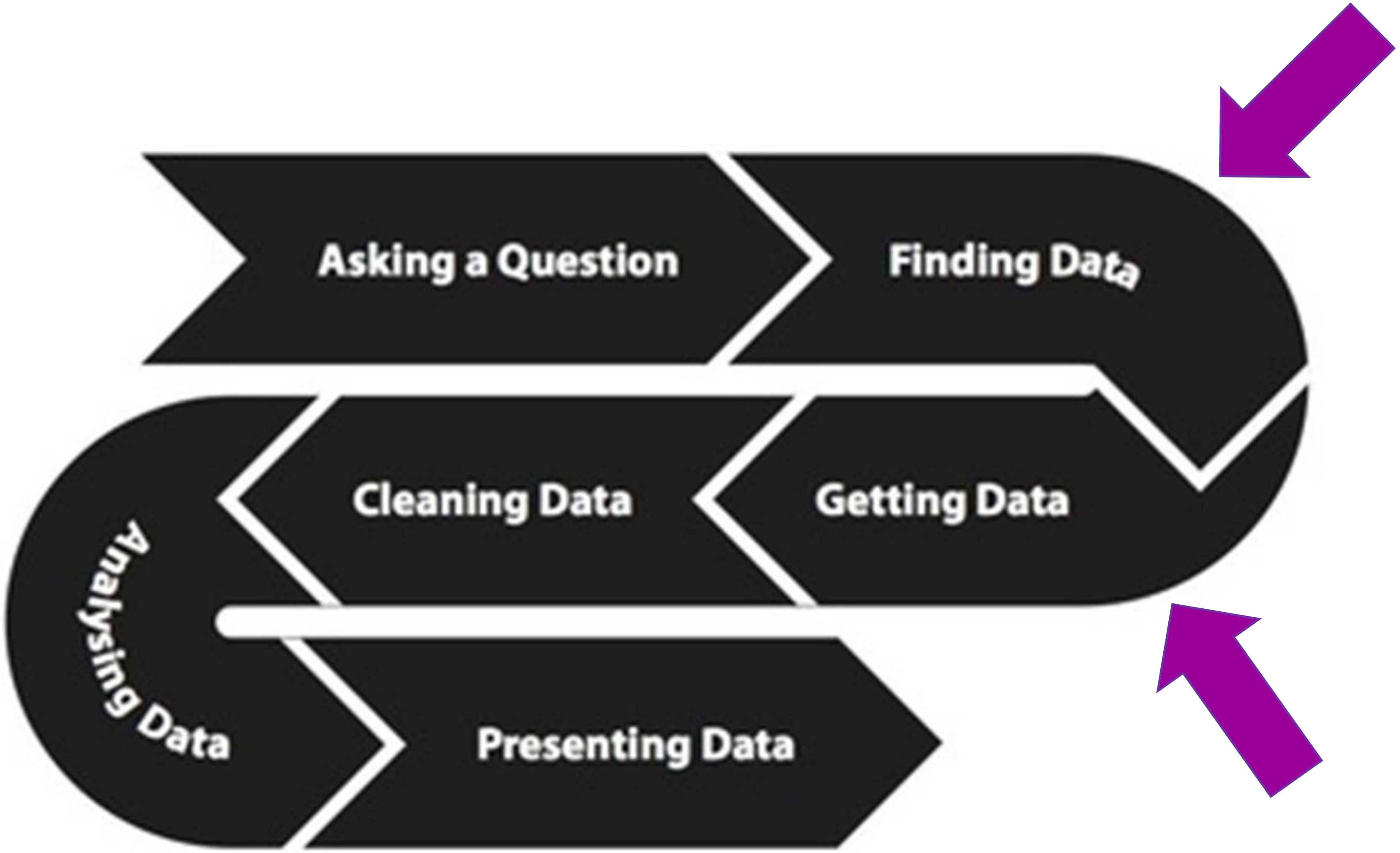
- Think of **each line** in the CSV as a **row**, and think of **each part of a line** separated by a comma as part of a **column**.
- Each **row** represents a **single data item**, and each **column** represents a **property**—the only exception is the first row, which gives the names of the columns.



```
userName, Password
testuser1, P@ssword1
testuser2, P@ssword2
testuser3, P@ssword3
testuser4, P@ssword4
testuser5, P@ssword5
testuser6, P@ssword6
testuser7, P@ssword7
testuser8, P@ssword8
testuser9, P@ssword9
testuser10, P@ssword10
testuser11, P@ssword11
testuser12, P@ssword12
testuser13, P@ssword13
testuser14, P@ssword14
```

Data Acquisition – Stage 2 and 3

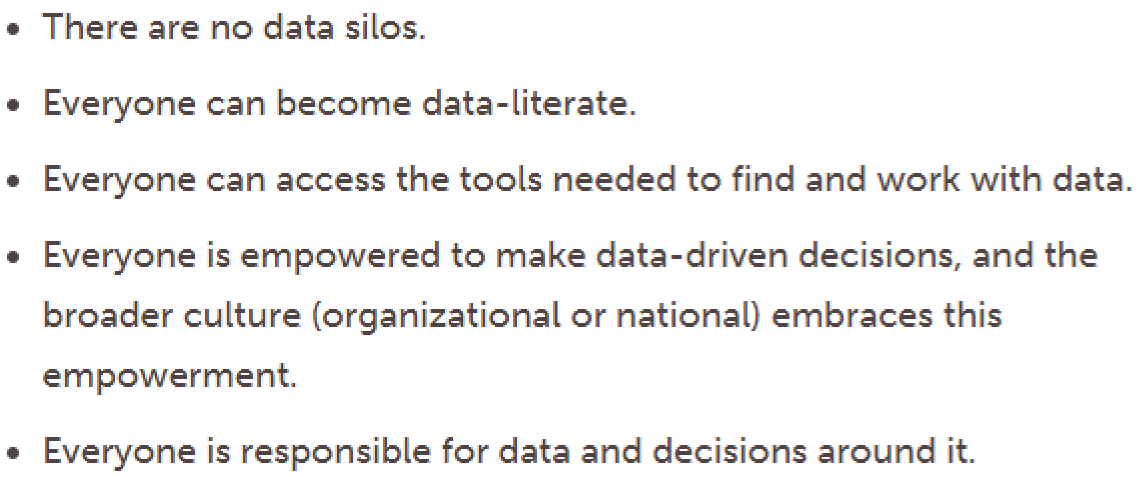
Data acquisition is the processes of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.



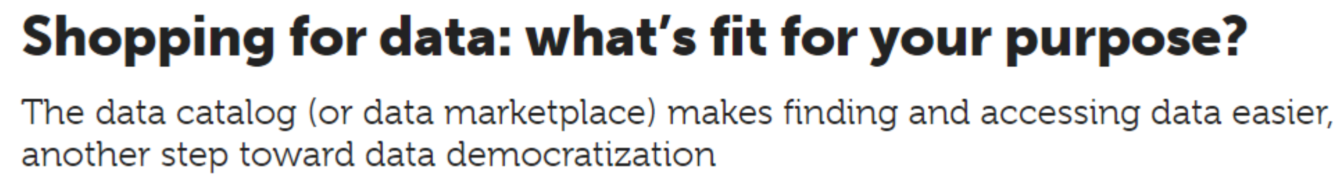
Finding Data

- Searching and finding data that is **readily available**.
- **Data democratization** is the idea that digital information should be accessible and understandable to the average end user as a basis for decision-making.

Aspirations behind data democratization



- There are no data silos.
- Everyone can become data-literate.
- Everyone can access the tools needed to find and work with data.
- Everyone is empowered to make data-driven decisions, and the broader culture (organizational or national) embraces this empowerment.
- Everyone is responsible for data and decisions around it.



Shopping for data: what's fit for your purpose?

The data catalog (or data marketplace) makes finding and accessing data easier, another step toward data democratization

https://www.infoworld.com/article/3246632/database/data-democratization-finally-living-up-to-the-name.html

Finding Data

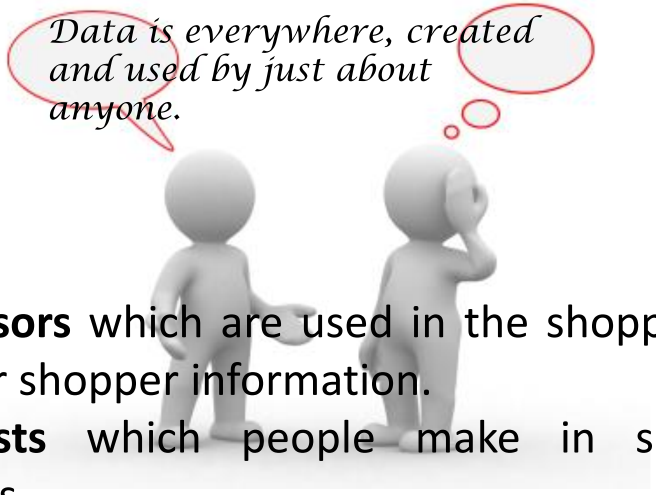
Think who might collect the data or who has the data.

- Could it have been collected by a government agency?
- An NGO or non-profit organization?
- A private business or industry group?
- Academic researchers?

Once you know that what you want **exists**, it's time to **hunt it down**.

- Is it freely available on the web? Check Google—you never know!
- Or part of a package to which the library already subscribes?
- Is it something we can buy? (And is it within the library's budget and can the purchase be made quickly enough to fit your timeframe?)
- Can it be requested directly from the researcher? There’s a reason articles usually include author contact information

Sources of Data



The **Sensors** which are used in the shopping complex to gather shopper information.

The **posts** which people make in social media platforms.

The **digital pictures and videos** we capture in our phone.

The **purchase transaction** which is made through e-commerce.

Artificial Knowledge

Information created by artificial intelligence such as predictions and interpretations.

Sensor

Data from sensors such as cameras, global positioning units, proximity sensors, thermometers, accelerometers, magnetometers and gyroscopes.

User Input

Input from users such as taps on a screen.

Interactions

Interactions such as a mobile device connecting to a website or a financial transaction.

Calculated

Data that is calculated from other data such as the revenue of a company that is calculated from accounting journal entries.

Metadata

Data about data such as control data in a database.

Raw Versus Processed Data

Raw Data

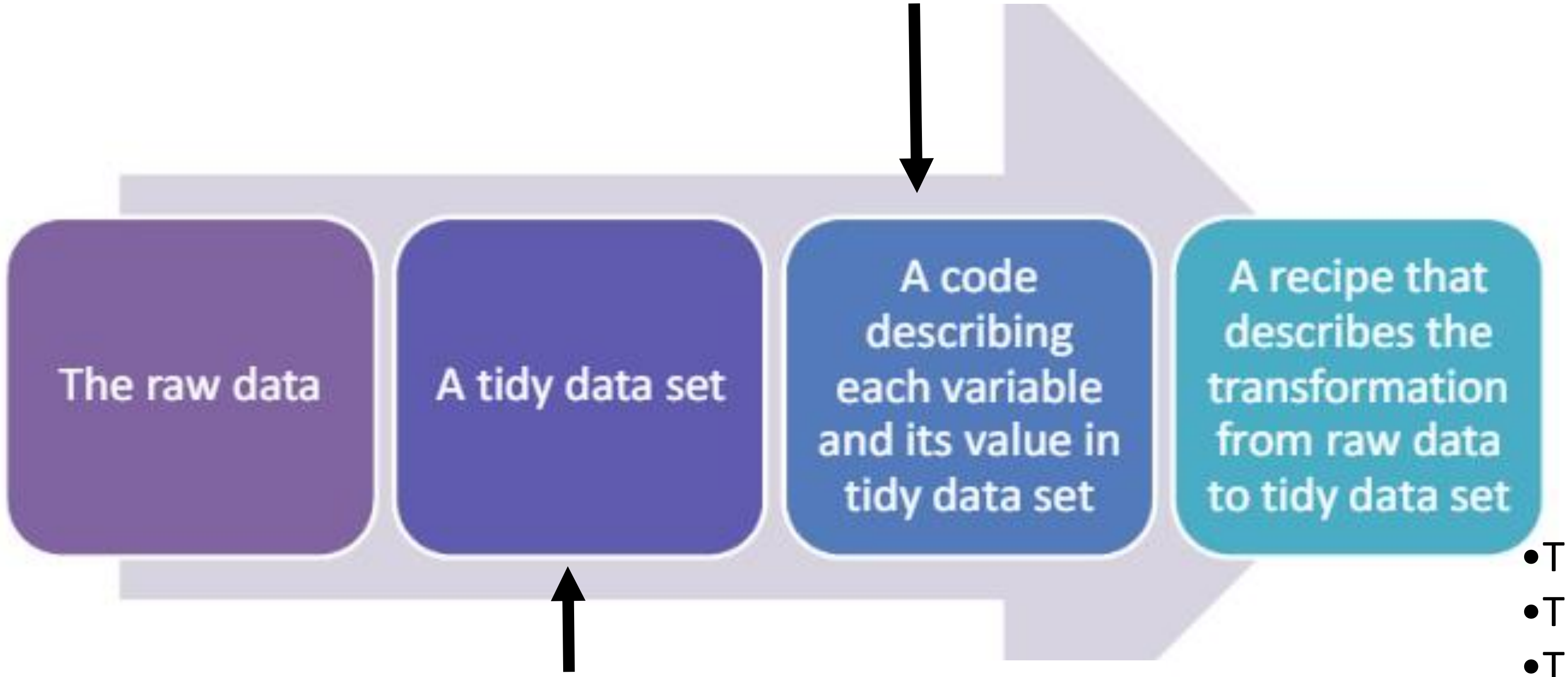
- Data from the original source
- Data that is often or difficult to be analyzed
- Data that needs to be processed before analyzing
- Usually raw data can be converted to processed data in one time.

Processed Data

- Data that is ready for analysis
- Processing can include merging, subsetting, transforming etc.
- There may be standards for processing
- All processing steps should be recorded.

Data Processing

- It is the meta data
- Information that describes the variables (including the measurement units)
- Information that describes the summary choices made
- Information that describes the experimental design used



The raw data

A tidy data set

A code describing each variable and its value in tidy data set

A recipe that describes the transformation from raw data to tidy data set

Explicit steps and exact recipe to get through 1 - 3 (**instruction list**)

- The programming script i.e. R script
- The input for the script –raw data
- The output will be the processed – tidy data set

- Each measured variable should be in one column
- Each different observation of that variable should be in a different row
- There should be one table for each "kind" of variable
- If you have multiple tables, they should include a column in the table that allows them to be linked

Getting Data

Before you can even think about putting your data to work, you have to figure out how you're going to obtain all the data you need.
Data is either collected (i) **slowly over time** as it becomes available to a database you set up, or (ii) **downloaded from a source** of curated datasets.

Data can be acquired for a data science project by many different means.

Some ways of acquiring data:

- ✓using data portals
- ✓submitting Freedom of Information requests
- ✓government cooperation
- ✓data scraping
- ✓use technology to bridge the gap between how information is shared and what is necessary for project.

Data Portals

Many countries have set up dedicated **open data** portals.

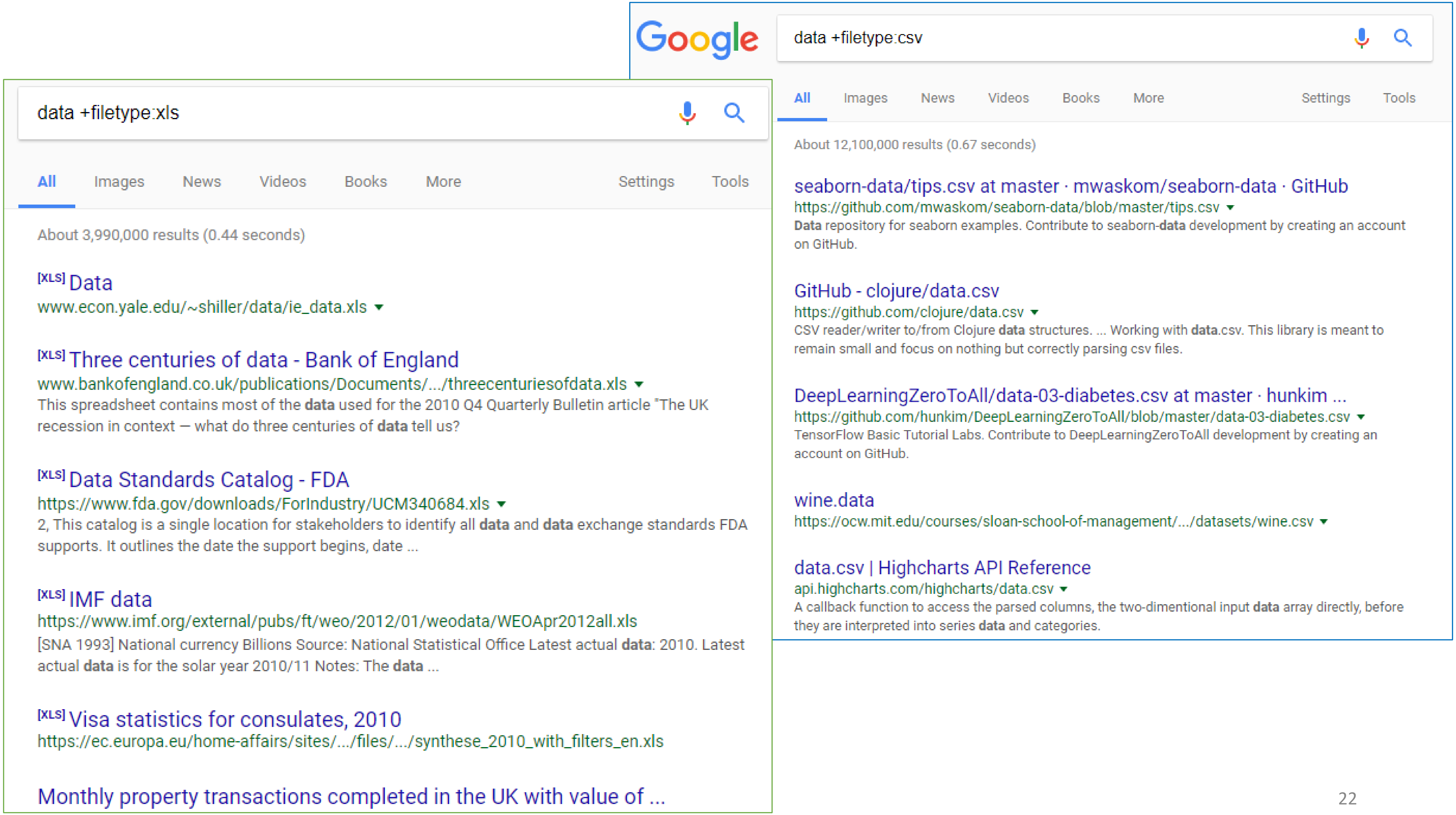
  - published datasets and relevant metadata

- Two indexes : **The Open Data Census** and **datacatalogs.org**
- The **Guardian** also provides a world government data search engine.
- Comparative data - available from the data portals of the **United Nations**, **World Bank**, and **World Health Organization**.

A lot of government data is indexed by ordinary web search engines.

- The trick to finding this data is anticipating its file format. If you limit your searches to **machine-readable file formats** specific to the type of data you want (e.g. CSV or XLS for tabular data, SQL or DB for databases), your search results are likely to be relevant data.
- Add "+filetype:extension" to your Google query to look for files with a specific extension, e.g. "+filetype:csv" to look for CSV files.

- Check the **Open Knowledge Foundation's Data Hub**, "a community-run catalogue of useful sets of data on the Internet", to see if anyone else has put up the data you're looking for.



data +filetype:xls

About 3,990,000 results (0.44 seconds)

[XLS] Data
www.econ.yale.edu/~shiller/data/ie_data.xls ▾

[XLS] Three centuries of data - Bank of England
www.bankofengland.co.uk/publications/Documents/.../threecenturiesofdata.xls ▾
This spreadsheet contains most of the **data** used for the 2010 Q4 Quarterly Bulletin article "The UK recession in context — what do three centuries of **data** tell us?

[XLS] Data Standards Catalog - FDA
https://www.fda.gov/downloads/ForIndustry/UCM340684.xls ▾
2, This catalog is a single location for stakeholders to identify all **data** and **data** exchange standards FDA supports. It outlines the date the support begins, date ...

[XLS] IMF data
https://www.imf.org/external/pubs/ft/weo/2012/01/weodata/WEOApr2012all.xls
[SNA 1993] National currency Billions Source: National Statistical Office Latest actual **data**: 2010. Latest actual **data** is for the solar year 2010/11 Notes: The **data** ...

[XLS] Visa statistics for consulates, 2010
https://ec.europa.eu/home-affairs/sites/.../files/.../synthese_2010_with_filters_en.xls

Monthly property transactions completed in the UK with value of ...

data +filetype:csv

About 12,100,000 results (0.67 seconds)

seaborn-data/tips.csv at master · mwaskom/seaborn-data · GitHub
https://github.com/mwaskom/seaborn-data/blob/master/tips.csv ▾
Data repository for seaborn examples. Contribute to seaborn-**data** development by creating an account on GitHub.

GitHub - clojure/data.csv
https://github.com/clojure/data.csv ▾
CSV reader/writer to/from Clojure **data** structures. ... Working with **data**.csv. This library is meant to remain small and focus on nothing but correctly parsing csv files.

DeepLearningZeroToAll/data-03-diabetes.csv at master · hunkim ...
https://github.com/hunkim/DeepLearningZeroToAll/blob/master/data-03-diabetes.csv ▾
TensorFlow Basic Tutorial Labs. Contribute to DeepLearningZeroToAll development by creating an account on GitHub.

wine.data
https://ocw.mit.edu/courses/sloan-school-of-management/.../datasets/wine.csv ▾

data.csv | Highcharts API Reference
api.highcharts.com/highcharts/data.csv ▾
A callback function to access the parsed columns, the two-dimentional input **data** array directly, before they are interpreted into series **data** and categories.

Open Data

- **Open data** is data that is made freely and easily available to anyone to use, reuse and distribute.
- The Open Knowledge Foundation, an organization dedicated to bringing "openness" to the mainstream, defines the following key factors that make data "open":
  - **Access & availability** - data is available to all in a convenient and modifiable form
  - **Re-use & redistribution** - terms of use allow for reusing, remixing and redistributing the data
  - **Universal participation** - there are no restrictions on who may do any of the above with the data
- Similarly, the FAIR research principles advocate for **Findable, Accessible, Interoperable, and Reusable** data.

https://mozillascience.github.io/open-data-primers/1-open-data-what.html

Free: Datasets List

- https://elitedatascience.com/datasets
- https://www.kaggle.com/datasets
- https://www.dataquest.io/blog/free-datasets-for-projects/
- https://archive.ics.uci.edu/ml/datasets.html
- https://towardsdatascience.com/top-10-great-sites-with-free-data-sets-581ac8f6334

APIs

- Open data is sometimes provided through an **application programming interface** (API).
- This is a web-based method for retrieving, searching, or even updating data dynamically from within a programming language environment.
- APIs provide up-to-date data in a granular and filtered form, removing the need to repeatedly process and update source files.
- A common use case for APIs is relatively time-sensitive information, such as procurement calls and contracts which are released every day.

Application Programming Interface (API): a specification allowing two pieces of software to interface with each other, without either having knowledge of the inner workings of the other.

Freedom of Information (FOI)

- Many countries decided to increase the transparency of their governments by introducing **Freedom of Information** (FoI) legislation.
- **Laws** - enable every citizen to request documents and other material from parts of the government which do not merit special protection (e.g. due to concerns over privacy, national security, or commercial confidentiality).
- FoI requests may be necessary when you want to **get more detail on the projects that government money is funding**.
- A good example of this process is the **Sunlight Foundation's** request for information on the **Airport Improvement Program** in the United States.
- If you see some interesting patterns in your high-level spending data, don't be afraid to dig deeper and ask for more detailed program information.

Preparing FOI request

- Want to submit an FoI request, but not sure where to start, who to address your request to, or how to write it? Access Info is an organization that works to help people obtain the information they require from the public bodies that hold it. They have also produced a toolkit for FoI requests. It's primarily aimed at journalists, but most of the tips are equally relevant for other researchers.

- Before submitting your FoI request, consider whether you could acquire the data by some other route. Journalists, activists, and CSOs have long had their own channels of acquiring information. Sometimes having a good relationship with a press officer or a civil servant is good enough, and making a formal request for information is unnecessary—your friendly press-officer may even feel offended if you don't ask them nicely first. FoIs generate a lot of paperwork (hence grumpy civil servants), so if you do have the contacts, it may be a good idea to ask nicely first!

- If an FoI request is your best option, make sure to invest some preparation in formulating your request. The documents or databases that are requested should be clearly identified, you should be aware of the department or unit in charge of the request, and you should address possible concerns over privacy or commercial confidentiality in your request.

- Don't count on receiving data in a machine-readable form. The FoI legislation in force in many countries was established before the need for structured data became apparent, and many laws do not allow the citizen to request a particular format. Many governments choose to release information on paper rather than in a structured digital form, which makes the data processing step more painful.

Data Scraping

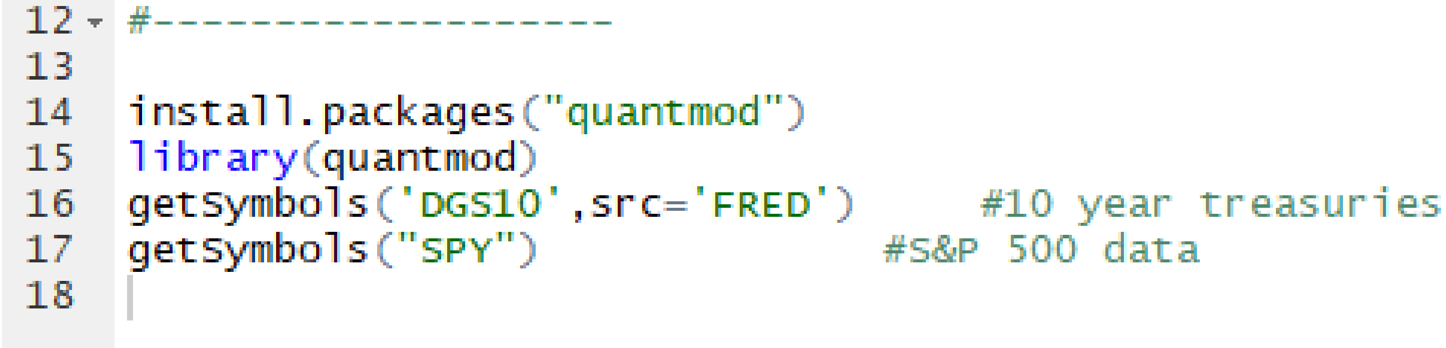
- **Scraping** refers to **transforming unstructured documents** (online database interfaces, PDF files, or even printed documents) into a **structured form** that is ready for computational processing and analysis.
- Although many easy-to-use scraping tools which do not require much technical know-how exist, many of the most rewarding data scraping tasks – such as the automated scraping of thousands or millions of web sites or the mass interpretation of PDF files – require some programming ability.
- Start learning about what data scraping is, with ScraperWiki (now renamed as QuickCode, https://quickcode.io/)

Web Scraping

- **Web scraping** is the process of pulling data from a website’s source code.
- It generally involves **writing a script** that will identify the information a user wants and pull it into a new file for later analysis.

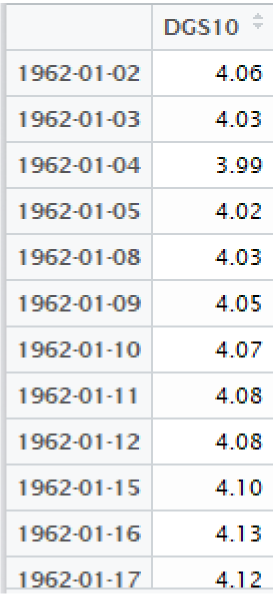
Examples of Web Scraping

Web scraping is a technique employed to extract data from a website.

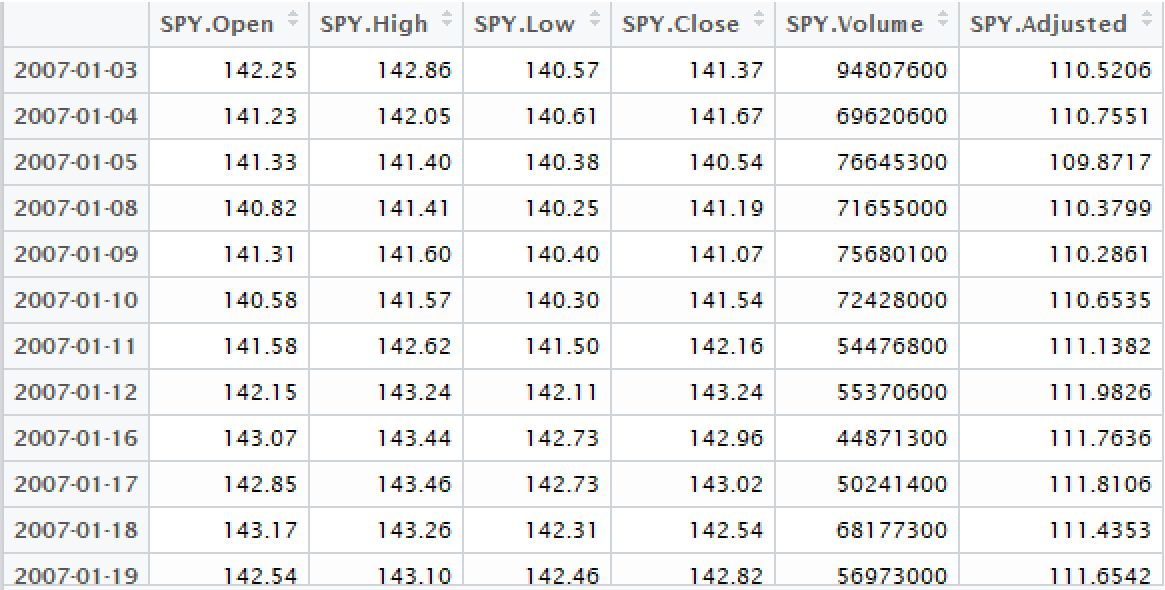


```
#--------------------

install.packages("quantmod")
library(quantmod)
getSymbols('DGS10',src='FRED')     #10 year treasuries
getSymbols("SPY")                  #S&P 500 data

```

Source: http://www.programmingr.com/examples/webscraping-stock-prices-economics-data-r/



	DGS10
1962-01-02	4.06
1962-01-03	4.03
1962-01-04	3.99
1962-01-05	4.02
1962-01-08	4.03
1962-01-09	4.05
1962-01-10	4.07
1962-01-11	4.08
1962-01-12	4.08
1962-01-15	4.10
1962-01-16	4.13
1962-01-17	4.12



	SPY.Open	SPY.High	SPY.Low	SPY.Close	SPY.Volume	SPY.Adjusted
2007-01-03	142.25	142.86	140.57	141.37	94807600	110.5206
2007-01-04	141.23	142.05	140.61	141.67	69620600	110.7551
2007-01-05	141.33	141.40	140.38	140.54	76645300	109.8717
2007-01-08	140.82	141.41	140.25	141.19	71655000	110.3799
2007-01-09	141.31	141.60	140.40	141.07	75680100	110.2861
2007-01-10	140.58	141.57	140.30	141.54	72428000	110.6535
2007-01-11	141.58	142.62	141.50	142.16	54476800	111.1382
2007-01-12	142.15	143.24	142.11	143.24	55370600	111.9826
2007-01-16	143.07	143.44	142.73	142.96	44871300	111.7636
2007-01-17	142.85	143.46	142.73	143.02	50241400	111.8106
2007-01-18	143.17	143.26	142.31	142.54	68177300	111.4353
2007-01-19	142.54	143.10	142.46	142.82	56973000	111.6542

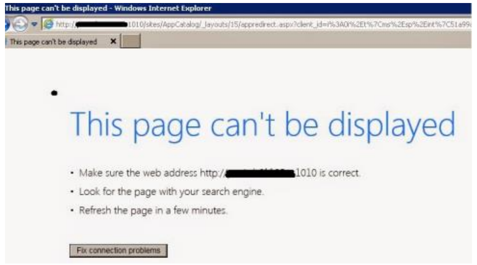
Getting Data Out of Scanned Documents

- In some cases, the only way to gain access to a dataset is through the **digitization of printed material**.
- How to leverage data which is trapped inside scanned documents and images?
  - **Optical character recognition (OCR)** software is built to do this, accepting scanned pictures and PDF documents as an input.
    - ABBYY FineReader, and some open-source software packages, such as Google's Tesseract.
- Automated data entry solutions do a great job of reading scanned documents and images and then transferring that data into a different format such as excel sheet or csv.

https://docparser.com/blog/extract-data-scanned-documents-images/

Keeping the Data Around

- Experience has shown that data does disappear, whether through the government redesigning its web sites, new policies that retract transparency rules, or simple system failures.
- Help prevent the disappearance of data by keeping your own archival copies.
- For data found on the web, this means downloading complete copies of web sites – a process called mirroring – which is a fairly well established technique that can easily be deployed by civil society organizations and other researchers.
  - **Mirroring** involves using a computer program called a **web crawler** to harvest all the web pages from a specified web page, e.g. a ministry home page.
  - Also possible to find old versions of web sites via the **Internet Archive's Wayback Machine**, a project that aims to create up-to-date copies of all public web sites and archive them forever.



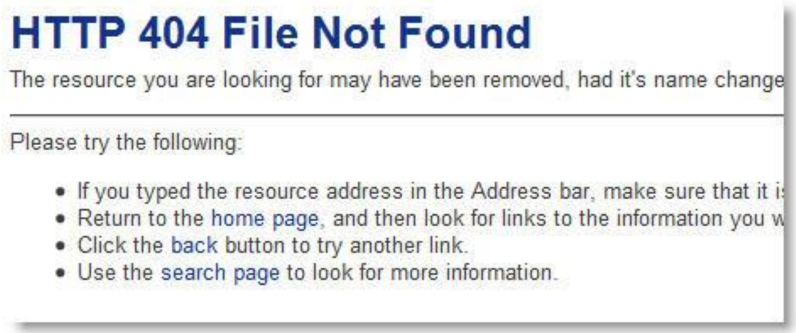
This page can't be displayed - Windows Internet Explorer

http://[redacted]1010/sites/AppCatalog/_layouts/15/appredirect.aspx?client_id=i%3A0i%2Et%7Cms%2Esp%2Eint%7C51a99

This page can't be displayed

This page can't be displayed

- Make sure the web address http://[redacted]1010 is correct.
- Look for the page with your search engine.
- Refresh the page in a few minutes.



HTTP 404 File Not Found

The resource you are looking for may have been removed, had it's name change

Please try the following:

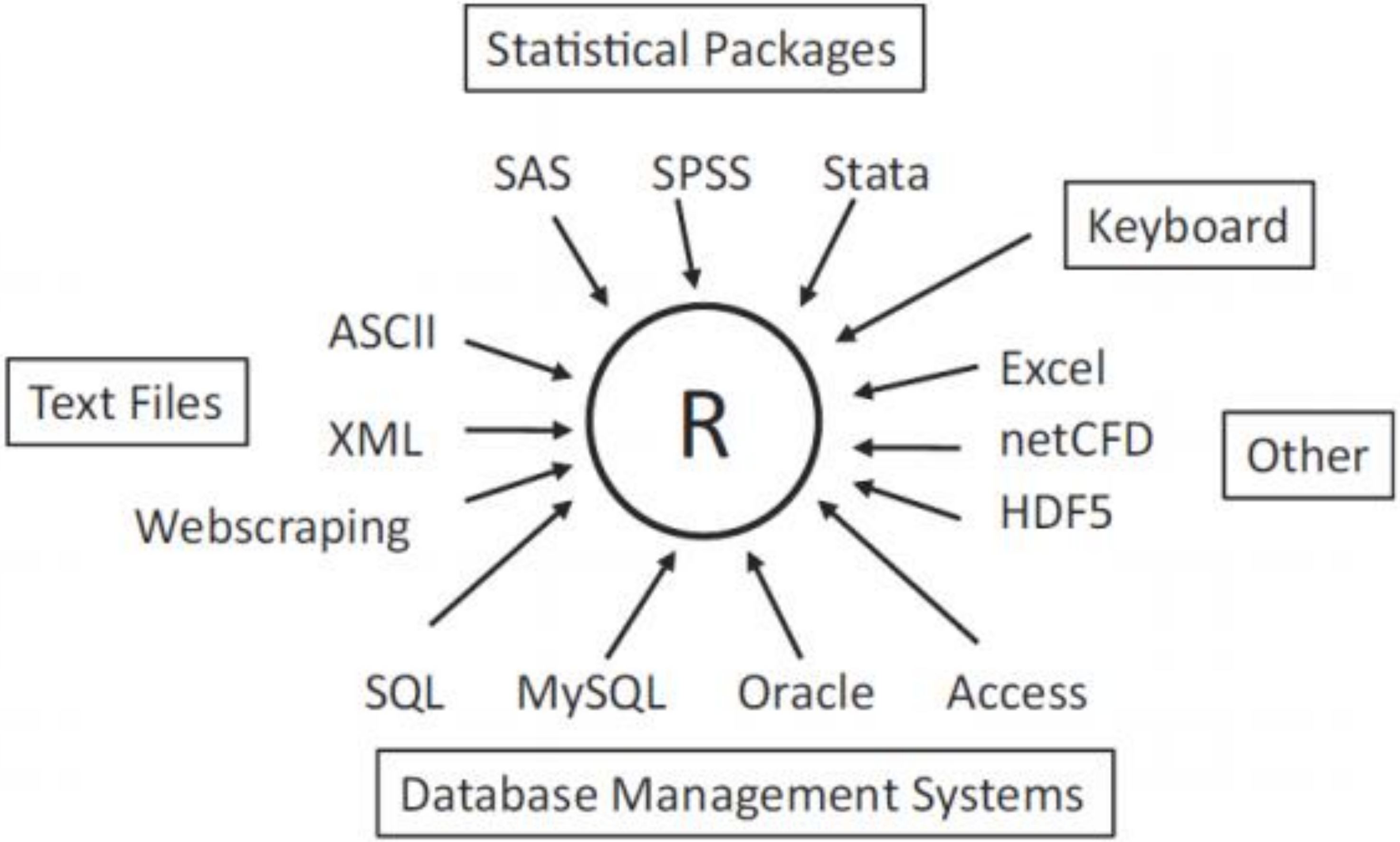
- If you typed the resource address in the Address bar, make sure that it i
- Return to the home page, and then look for links to the information you w
- Click the back button to try another link.
- Use the search page to look for more information.

Downloading files

- Always check and set your working directory using the **getwd( )** and **setwd( )** commands
- To check whether the “Data” directory has been created or not:

```
if (!file.exists("data")){
dir.create("data")
}
```

Sources of data for R



Tutorial 3

Web Scraping with R

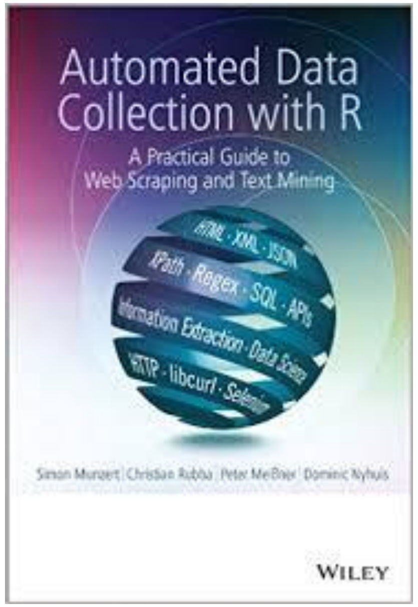
Try out the tutorial from ONE of the following link :

- ✓ https://slcladal.github.io/webcrawling.html
- ✓ https://www.geeksforgeeks.org/web-scraping-using-r-language/
- ✓ https://medium.freecodecamp.org/an-introduction-to-web-scraping-using-r-40284110c848

You may **choose any web page** to scrape.
Show that you had successfully done some web scraping.
Submit on Spectrum for Tutorial 3.

References

- John Spacey, Data Guide
- http://okfn.booktype.pro/spending-data-handbook/getting-cleaning/
- https://r-dir.com/reference/datasets.html
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- http://kek.ksu.ru/eos/WM/AutDataCollectR.pdf


Automated Data Collection with R
A Practical Guide to Web Scraping and Text Mining
HTML · XML · JSON
XPath · Regex · SQL · APIs
Information Extraction · Data Science
HTTP · libcurl · Selenium
Simon Munzert | Christian Rubba | Peter Meißner | Dominic Nyhuis
WILEY