**Google Data Analytics**

## **Process Data from Dirty to Clean**

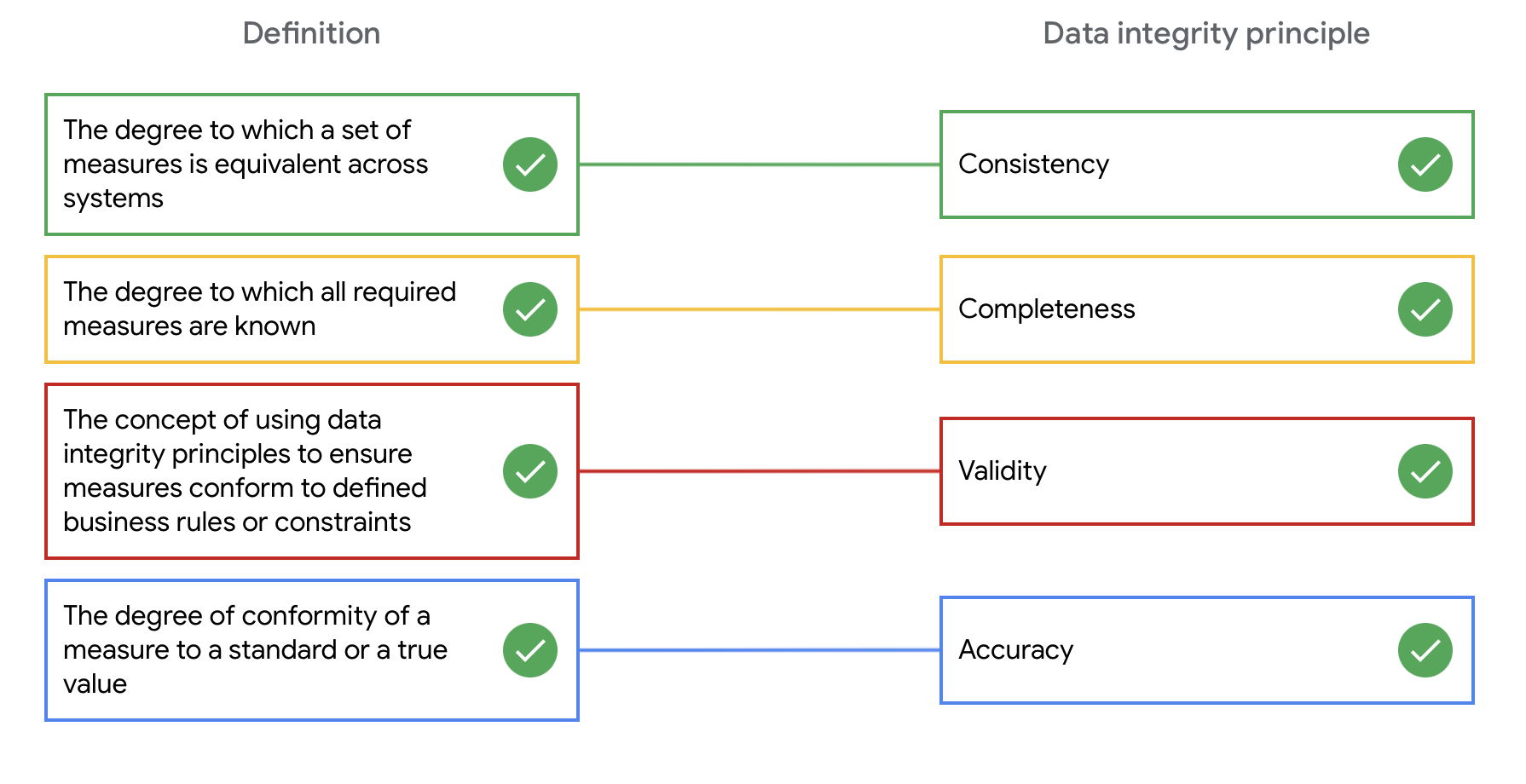
**Data Integrity:**

The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle

Threats to data integrity:

* Replicated
* Transferred
* Manipulated
* ﻿﻿Human error
* ﻿﻿Viruses
* ﻿﻿Malware
* ﻿﻿Hacking
* ﻿﻿System failures

| **Data constraint** | **Definition** | **Examples** |
| --- | --- | --- |
| **Data type** | **Values must be of a certain type: date, number, percentage, Boolean, etc.** | **If the data type is a date, a single number like 30 would fail the constraint and be invalid** |
| **Data range** | **Values must fall between predefined maximum and minimum values** | **If the data range is 10-20, a value of 30 would fail the constraint and be invalid** |
| **Mandatory** | **Values can’t be left blank or empty** | **If age is mandatory, that value must be filled in** |
| **Unique** | **Values can’t have a duplicate** | **Two people can’t have the same mobile phone number within the same service area** |
| **Regular expression (regex) patterns** | **Values must match a prescribed pattern** | **A phone number must match ###-###-#### (no other characters allowed)** |
| **Cross-field validation** | **Certain conditions for multiple fields must be satisfied** | **Values are percentages and values from multiple fields must add up to 100%** |
| **Primary-key** | **(Databases only) value must be unique per column** | **A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.** |
| **Set-membership** | **(Databases only) values for a column must come from a set of discrete values** | **Value for a column must be set to Yes, No, or Not Applicable** |
| **Foreign-key** | **(Databases only) values for a column must be unique values coming from a column in another table** | **In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table** |
| **Accuracy** | **The degree to which the data conforms to the actual entity being measured or described** | **If values for zip codes are validated by street location, the accuracy of the data goes up.** |
| **Completeness** | **The degree to which the data contains all desired components or measures** | **If data for personal profiles required hair and eye color, and both are collected, the data is complete.** |
| **Consistency** | **The degree to which the data is repeatable from different points of entry or collection** | **If a customer has the same address in the sales and repair databases, the data is consistent.** |

****

# 

# **Well-aligned objectives and data:**

It's important to check that the data you use aligns with the business objective.

Good alignment means that the data is relevant and can help you solve a business problem or determine a course of action to achieve a given business objective.

## **Clean data + alignment to business objective = accurate conclusions**

## **Alignment to business objective + additional data cleaning = accurate conclusions**

## **Alignment to business objective + newly discovered variables + constraints = accurate conclusions**

**Insufficient Data:**

Types of insufficient data:

* ﻿﻿Data from only one source
* ﻿﻿Data that keeps updating
* ﻿﻿Outdated data
* ﻿﻿Geographically-limited data

Ways to address insufficient data

* ﻿﻿Identify trends with the available data
* ﻿﻿Wait for more data if time allows
* ﻿﻿Talk with stakeholders and adjust your objective
* ﻿﻿Look for a new dataset

## **Data issue 1: no data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data. | If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees. |
| If there isn’t time to collect data, perform the analysis using proxy data from other datasets.  *This is the most common workaround.* | If you are analysing peak travel times for commuters but don’t have the data for a particular city, use the data from another city with a similar size and demographic. |

## 

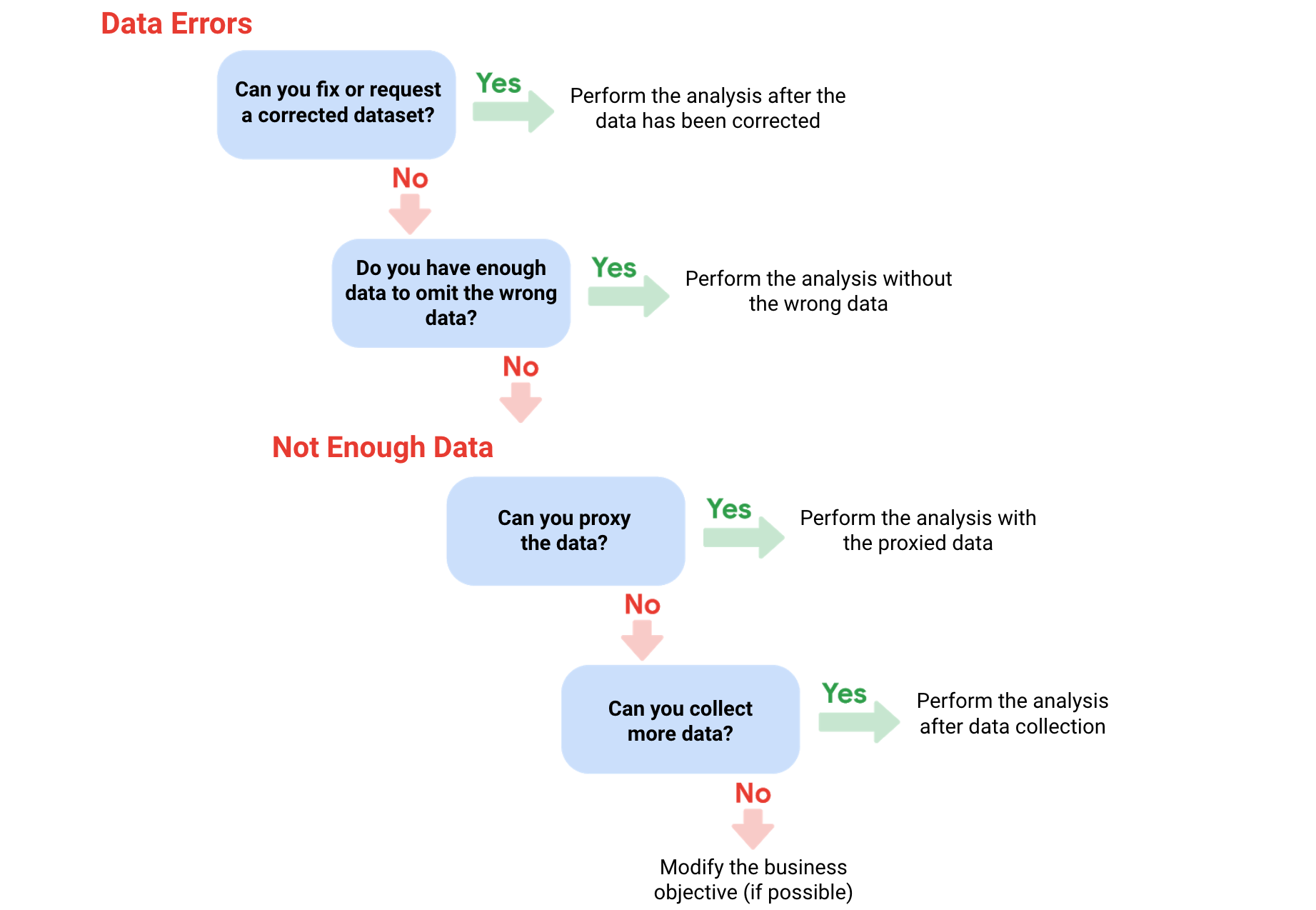
## **Data issue 2: too little data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis to align with the data you already have. | If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older* *only*. |

## **Data issue 3: wrong data, including data with errors\***

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| If you have the wrong data because requirements were misunderstood, communicate the requirements again. | If you need the data for female voters and received the data for male voters, restate your needs. |
| Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors. | If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values. |
| If you can’t correct data errors yourself, you can ignore the  wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won’t cause systematic bias. | If your dataset was translated from a different language and some of the translations don’t make sense, ignore the data with bad translation and go ahead with the analysis of the other data. |

***Important note****: sometimes data with errors can be a warning sign that the data isn’t reliable. Use your best judgement.*

****

| **Terminology** | **Definitions** |
| --- | --- |
| **Population** | The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company. |
| **Sample** | A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population. |
| **Margin of error** | Since a sample is used to represent a population, the sample’s results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population. |
| **Confidence level** | How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study. |
| **Confidence interval** | The range of possible values that the population’s result would be at the confidence level of the study. This range is the sample result +/- the margin of error. |
| **Statistical significance** | The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance. |

## Things to remember when determining the size of your sample

When figuring out a sample size, here are things to keep in mind:

* Don’t use a sample size less than 30 ([Reference](https://docs.google.com/document/d/1qs9JwakDpOL4LuRj6pY8KjjgdMN5BO5OUc8eRv-yWe4/edit?usp=sharing)). It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
* The confidence level most commonly used is 95%, but 90% can work in some cases.

Increase the sample size to meet specific needs of your project:

* For a higher confidence level, use a larger sample size
* To decrease the margin of error, use a larger sample size
* For greater statistical significance, use a larger sample size

**Statistics:**

**Statistical power:**

The probability of getting meaningful results from a test

Statistical power can be calculated and reported for a completed experiment to comment on the confidence one might have in the conclusions drawn from the results of the study. It can also be used as a tool to estimate the number of observations or sample size required in order to detect an effect in an experiment.

**Hypothesis testing:**

A way to see if a survey or experiment has meaningful results

**Statistically Significant:**

If a test is **statistically significant**, it means the results of the test are real and not an error caused by random chance

Usually, you need a statistical power of at least **0.8 or 80%** to consider your results statistically significant

* [Reference](https://docs.google.com/document/d/1qs9JwakDpOL4LuRj6pY8KjjgdMN5BO5OUc8eRv-yWe4/edit?usp=sharing)

## **Proxy data examples**

| **Business scenario** | **How proxy data can be used** |
| --- | --- |
| A new car model was just launched a few days ago and the auto dealership can’t wait until the end of the month for sales data to come in. They want sales projections now. | The analyst proxies the number of clicks to the car specifications on the dealership’s website as an estimate of potential sales at the dealership. |
| A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years. | The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years. |
| The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren’t publicly available yet. | The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier. |

**Margin of error**

The maximum amount that the sample results are expected to differ from those of the actual population

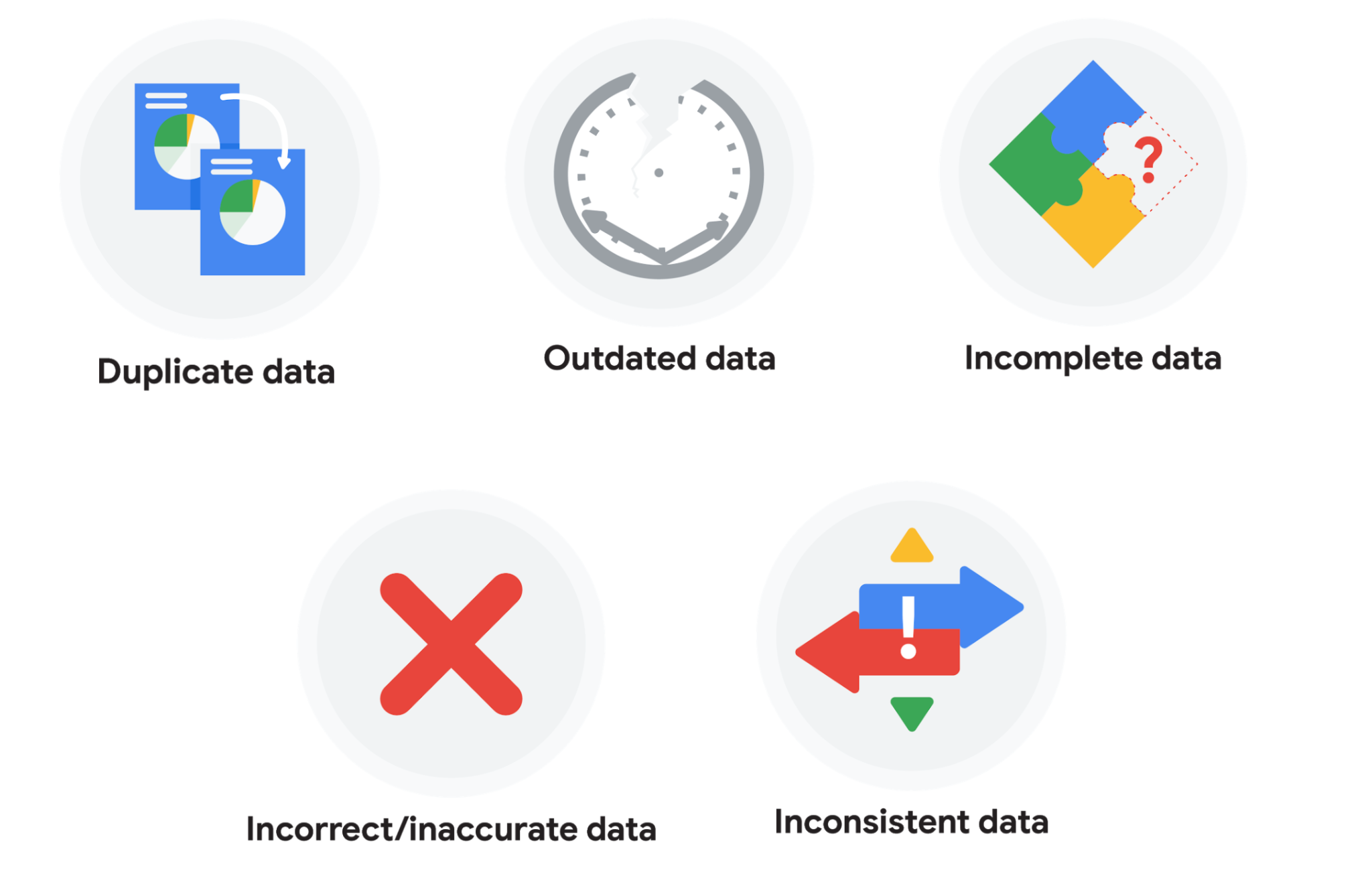
To calculate margin of error you need:

* ﻿﻿Population size
* ﻿﻿Sample size
* ﻿﻿Confidence level

**Note**: There are **Sample size** and **Margin of Error** **calculators** available online.

**Dirty and Clean data:**

**Types of Dirty Data:**

****

### 

### 

### **Duplicate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| **Any data record that shows up more than once** | **Manual data entry, batch data imports, or data migration** | **Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval** |

### **Outdated data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is old which should be replaced with newer and more accurate information | People changing roles or companies, or software and systems becoming obsolete | Inaccurate insights, decision-making, and analytics |

### **Incomplete data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is missing important fields | Improper data collection or incorrect data entry | Decreased productivity, inaccurate insights, or inability to complete essential services |

### **Incorrect/inaccurate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is complete but inaccurate | Human error inserted during data input, fake information, or mock data | Inaccurate insights or decision-making based on bad information resulting in revenue loss |

### **Inconsistent data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| **Any data that uses different formats to represent the same thing** | **Data stored incorrectly or errors inserted during data transfer** | **Contradictory data points leading to confusion or inability to classify or segment customers** |

**Some of the errors you might come across while cleaning your data could include:**

# 

# **Workflow automation:**

workflow automation is the process of automating parts of your work. That could mean creating an event trigger that sends a notification when a system is updated. Or it could mean automating parts of the data cleaning process.

| **Task** | **Can it be automated?** | **Why?** |
| --- | --- | --- |
| Communicating with your team and stakeholders | No | Communication is key to understanding the needs of your team and stakeholders as you complete the tasks you are working on. There is no replacement for person-to-person communications. |
| Presenting your findings | No | Presenting your data is a big part of your job as a data analyst. Making data accessible and understandable to stakeholders and creating data visualisations can’t be automated for the same reasons that communications can’t be automated. |
| Preparing and cleaning data | Partially | Some tasks in data preparation and cleaning can be automated by setting up specific processes, like using a programming script to automatically detect missing values. |
| Data exploration | Partially | Sometimes the best way to understand data is to see it. Luckily, there are plenty of tools available that can help automate the process of visualising data. These tools can speed up the process of visualising and understanding the data, but the exploration itself still needs to be done by a data analyst. |
| Modelling the data | Yes | Data modelling is a difficult process that involves lots of different factors; luckily there are tools that can completely automate the different stages. |

**SQL:**

| **Features of Spreadsheets** | **Features of SQL Databases** |
| --- | --- |
| **Smaller data sets** | **Larger datasets** |
| **Enter data manually** | **Access tables across a database** |
| **Create graphs and visualisations in the same program** | **Prepare data for further analysis in another software** |
| **Built-in spell check and other useful functions** | **Fast and powerful functionality** |
| **Best when working solo on a project** | **Great for collaborative work and tracking queries run by all users** |

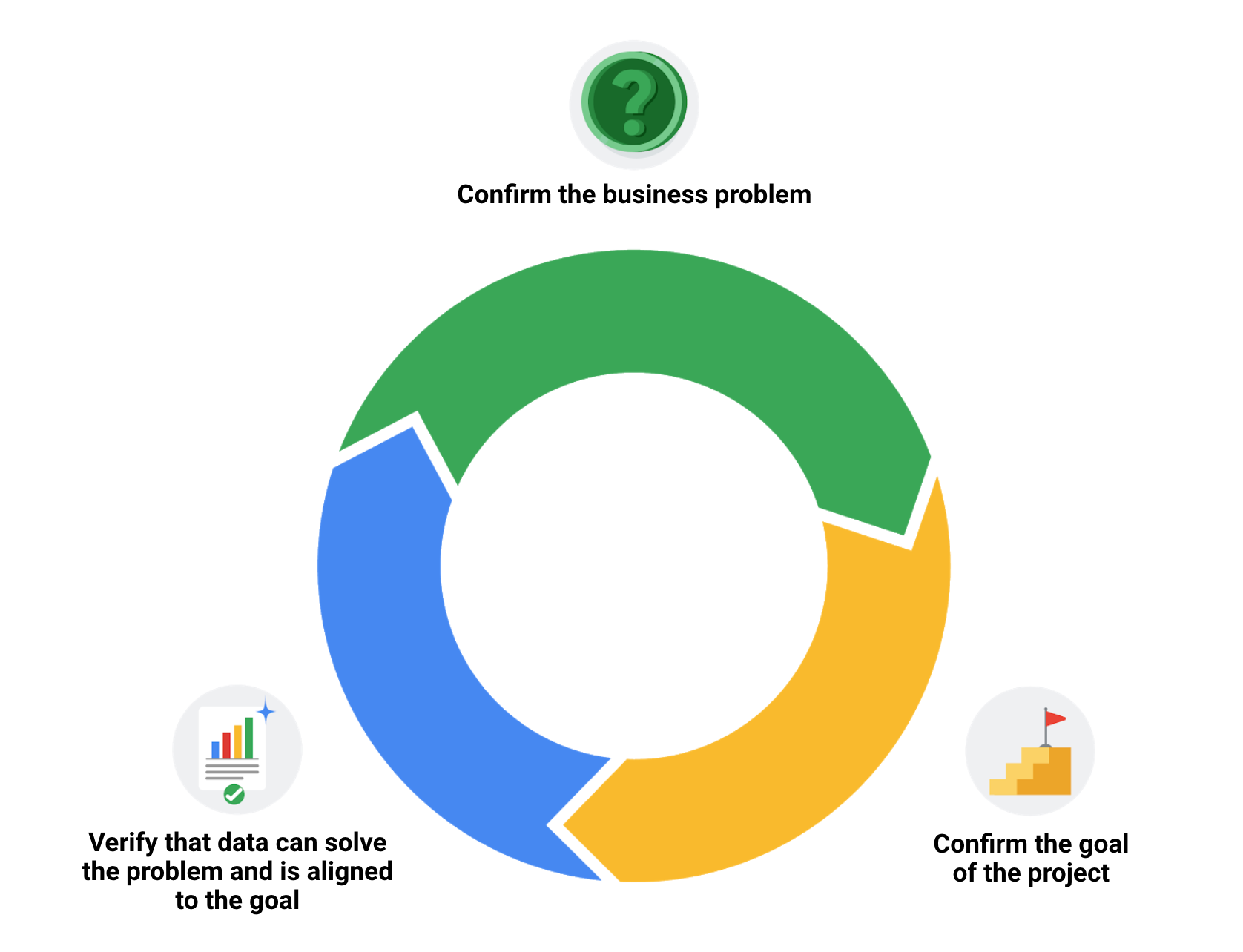
**Verifying & Reporting the cleaning report:**

**Verification:**

A process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable

See the big picture when verifying data-cleaning

1. ﻿﻿﻿Consider the business problem
2. ﻿﻿﻿Consider the goal
3. ﻿﻿﻿Consider the data



**Documentation:**

The process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

Why documentation:

* ﻿﻿Recover data-cleaning errors
* ﻿﻿Inform other users of changes
* ﻿﻿Determine quality of data

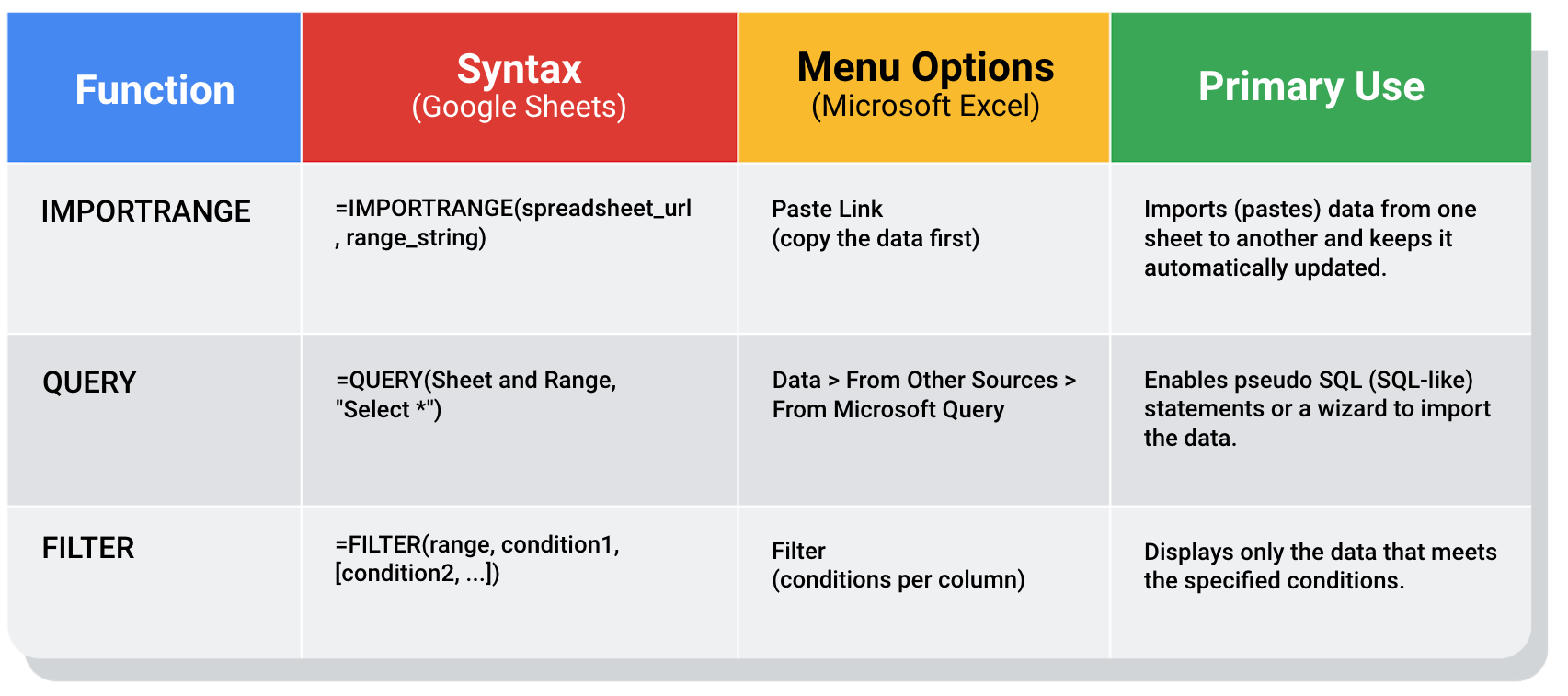
**Changelog**

A file containing a chronologically ordered list of modifications made to a project

Typically, a changelog records this type of information:

* Data, file, formula, query, or any other component that changed
* Description of what changed
* Date of the change
* Person who made the change
* Person who approved the change
* Version number
* Reason for the change
* **Added:** new features introduced
* **Changed:** changes in existing functionality
* **Deprecated**: features about to be removed
* **Removed:** features that have been removed
* **Fixed:** bug fixes
* **Security**: lowering vulnerabilities

# **Advanced functions for speedy data cleaning:**

****

**Resume tips(also refer to Resources):**

Accomplished [X]

As measured by [Y]

By doing [Z]

**PAR strategy:**

**P**roblem

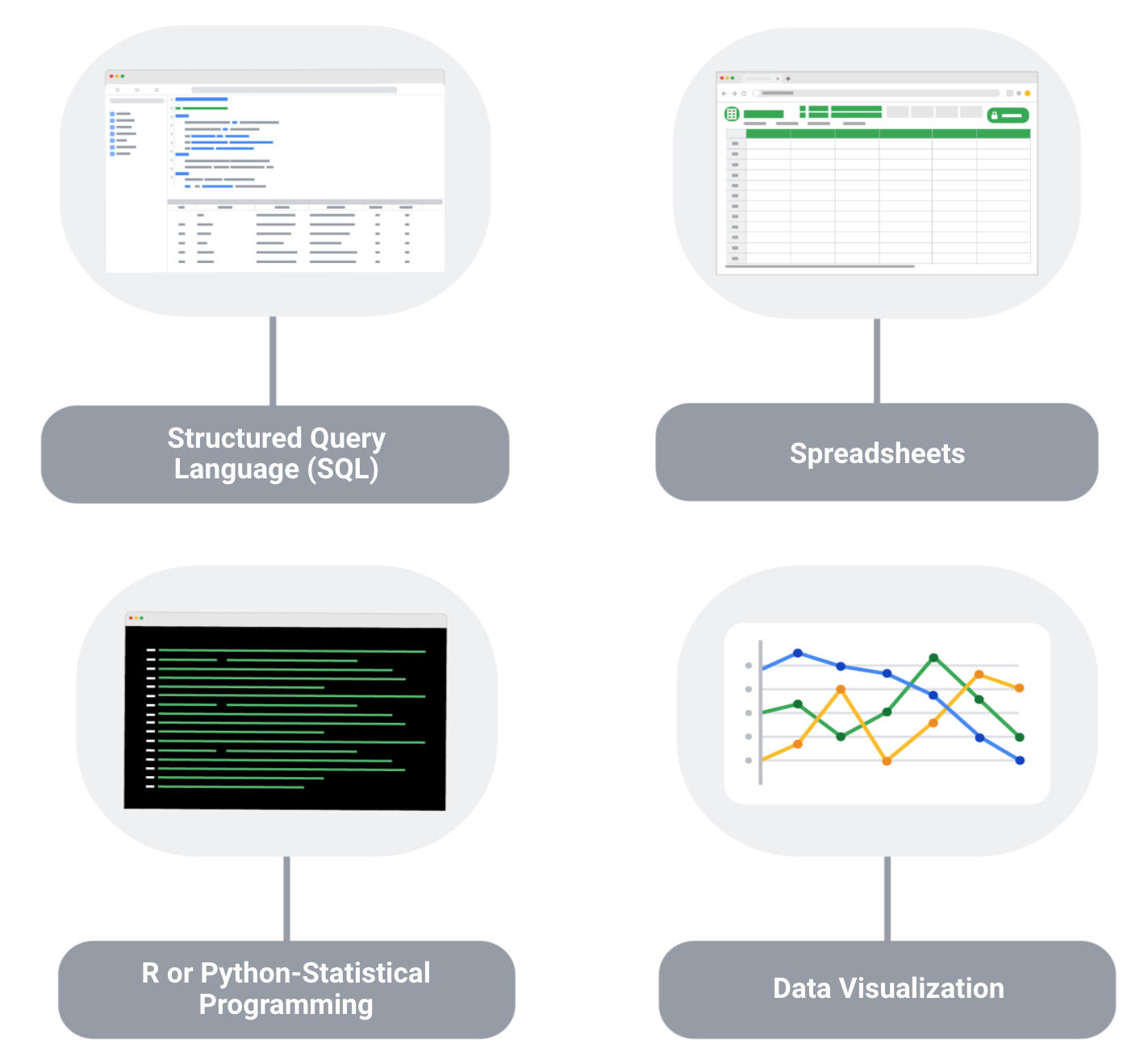
**A**ction

**R**esult

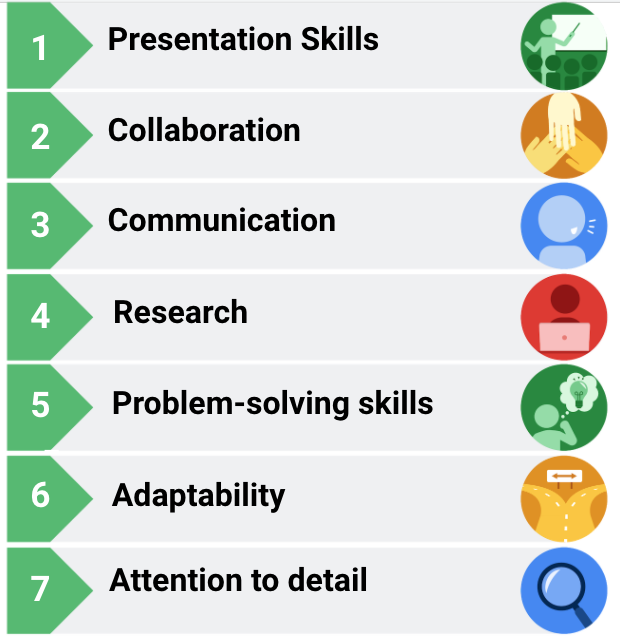
Ex: Effectively implemented and communicated daily workflow procedures to fellow team members, resulting in a 15% increase in productivity

* **Problem:** Previously-absent workflow procedures
* ﻿﻿**Action:** Implemented and communicated daily workflow procedures
* **﻿﻿Result:** 15% increase in productivity

## **Common professional skills for entry-level data analysts**

****

**Soft Skills required as an entry-level Data analyst:**

****