

Data Science and Data Analytics

Introduction to Data Science

Julian Amon, PhD

Charlotte Fresenius Privatuniversität

March 14, 2025

What is Data Science?

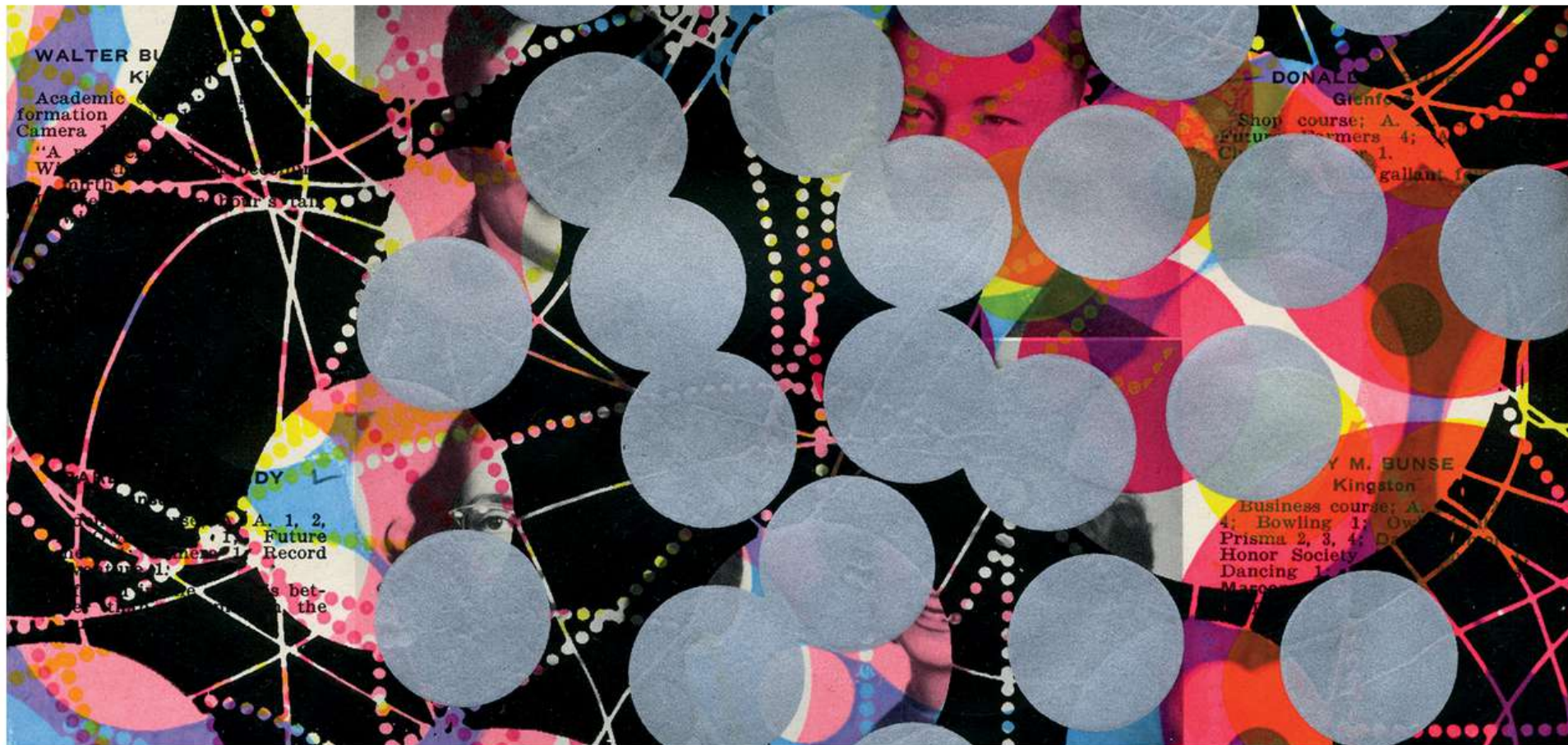
Data Science – A sexy profession in 2012?

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)



Andrew J Buboltz, silk screen on a page from a high school yearbook, 8.5" x 12", 2011. Tamar Cohen

Summary. Back in the 1990s, computer engineer and Wall Street "quant" were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for... [more](#)

Source: [Harvard Business Review](#) in October 2012

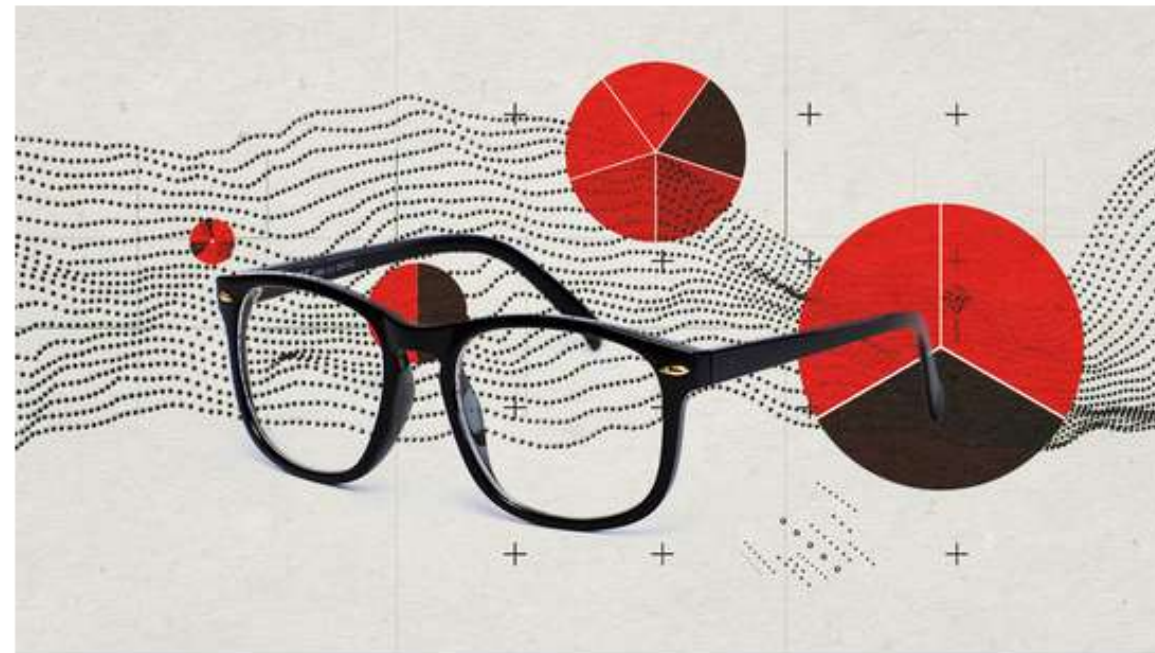
Data Science – A sexy profession today?

Analytics And Data Science

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022



HBR Staff/StudioM1/Moritz Otto/Getty Images

Summary. Ten years ago, the authors posited that being a data scientist was the “sexiest job of the 21st century.” A decade later, does the claim stand up? The job has grown in popularity and is generally well-paid, and the field is projected to experience more growth than almost... [more](#)

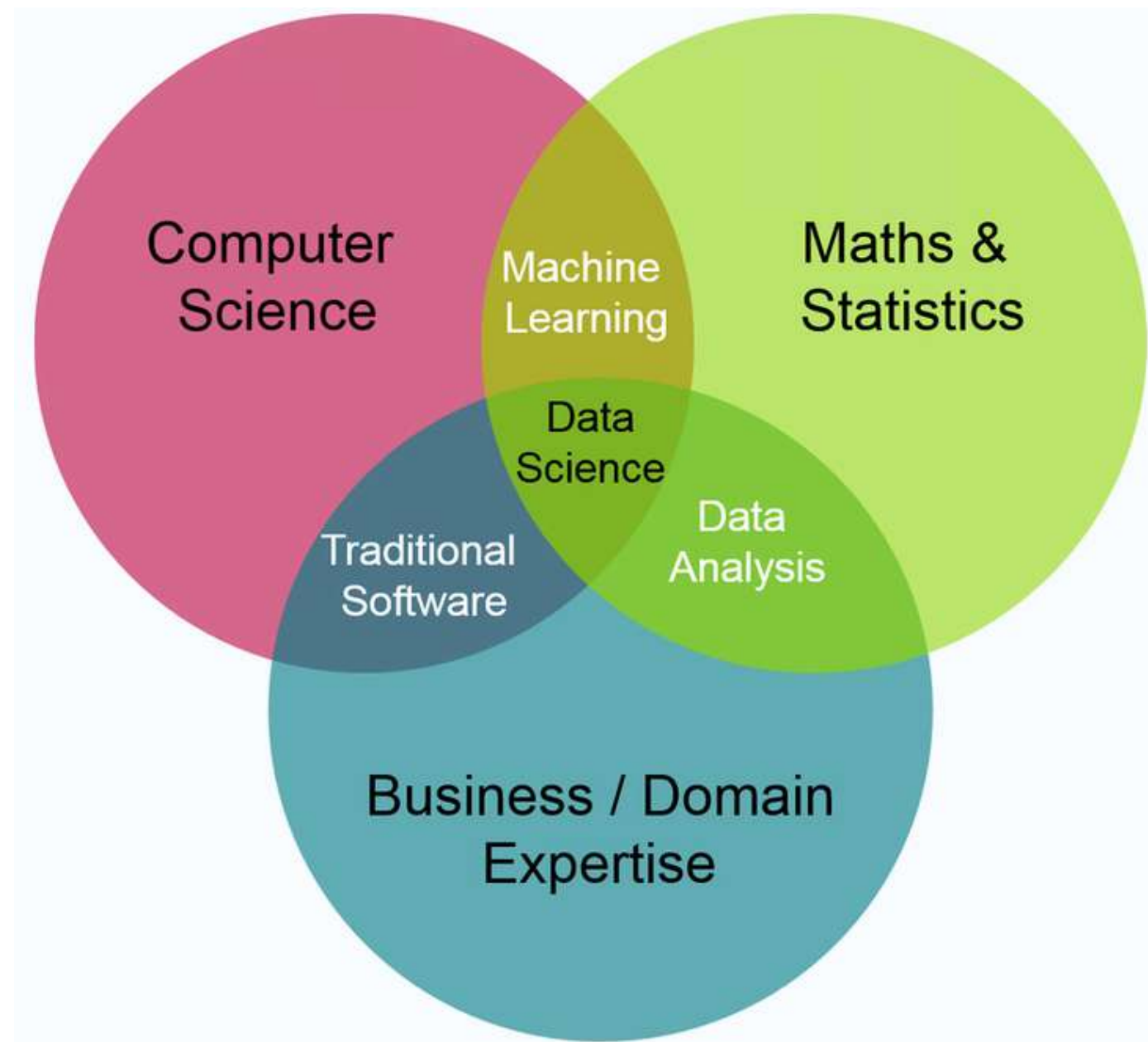
Source: [Harvard Business Review](#) in July 2022

What is Data Science?

Any ideas?

What is Data Science?

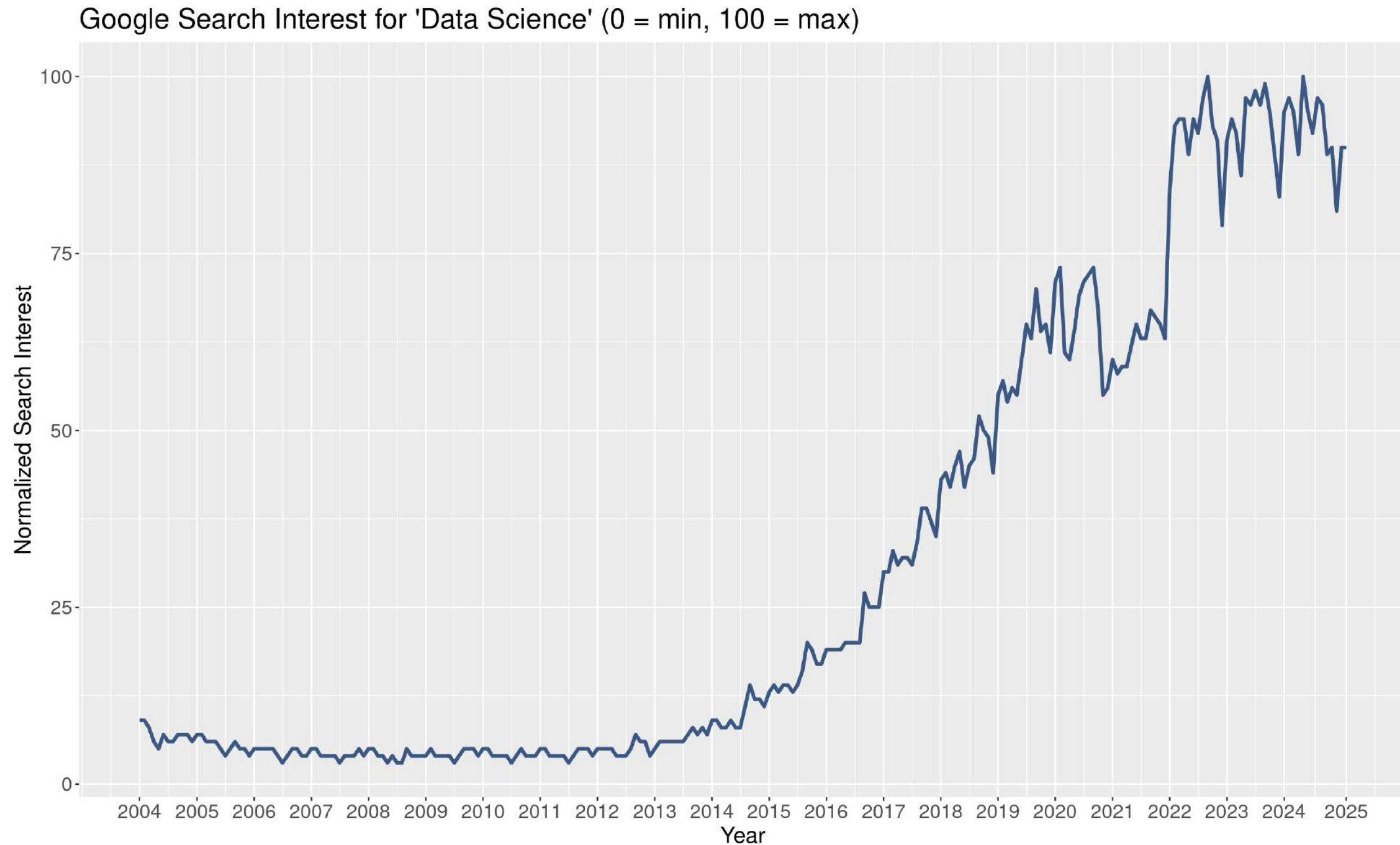
Data science is an interdisciplinary academic field that combines statistics, mathematics and computing with specific subject matter expertise to uncover actionable insights hidden in an organization's data.



A brief history of Data Science

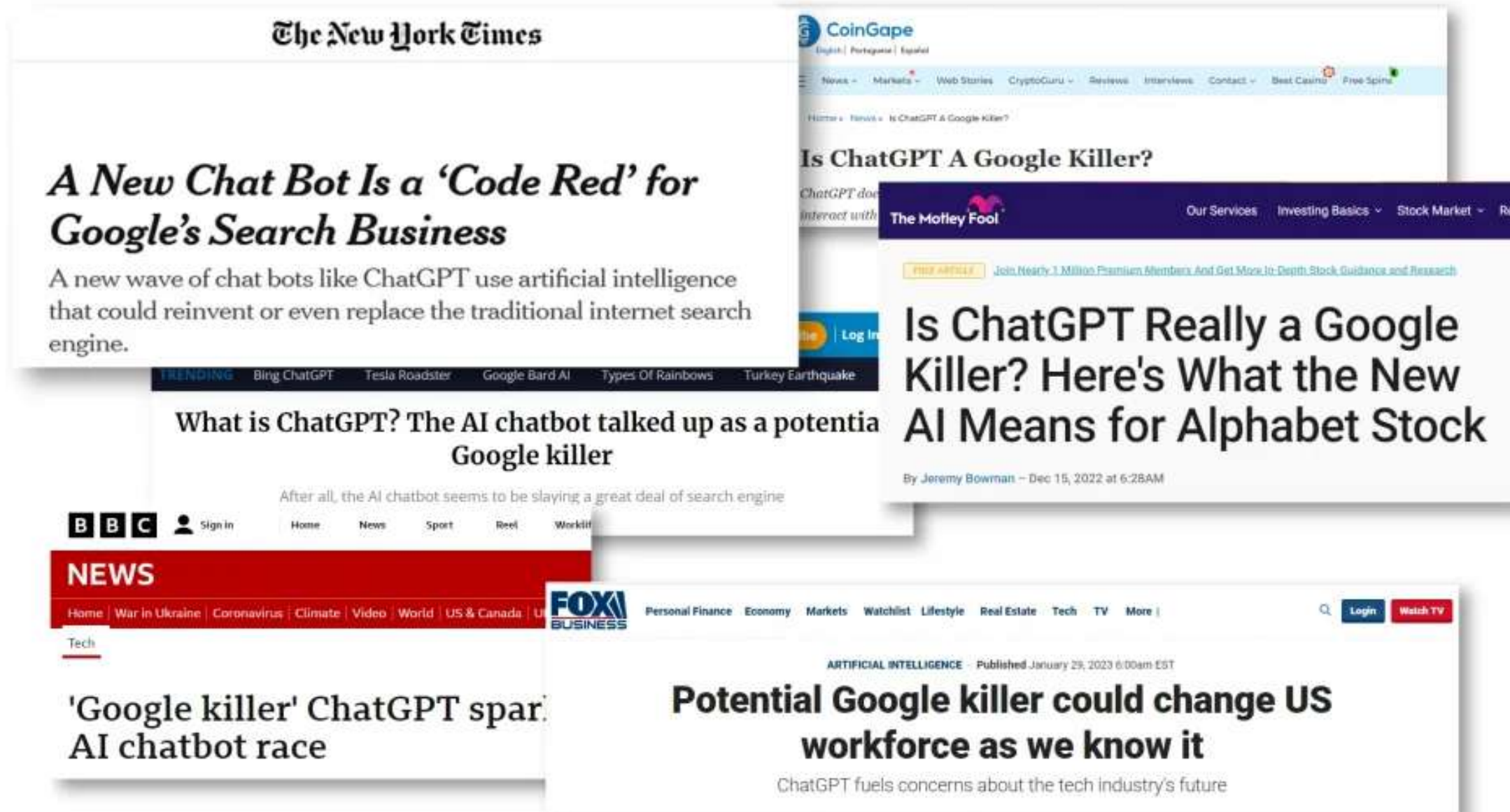
- **1962:** Statistician John Tukey describes a field he calls *data analysis*, similar to what we now understand to be data science.
- **1974:** Computer scientist Peter Naur proposes the term *data science*, but in a different sense to today, namely as an alternative name for computer science.
- **1992:** Conference participants at the University of Montpellier II acknowledge the emergence of a new, inherently interdisciplinary field called *data science* focused on gaining insight from very diverse types of data.
- **2001:** William S. Cleveland introduces data science as its own discipline in his article [Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics](#).
- **2003:** Columbia University starts publishing *The Journal of Data Science*.
- **2012:** With the HBR article on data scientist as the “sexiest job of the 21st century”, the term finds its way from the scientific realm into the mainstream.

A brief history of Data Science



Data Science and Artificial Intelligence (AI)

Since the arrival of ChatGPT, everyone is talking about AI...

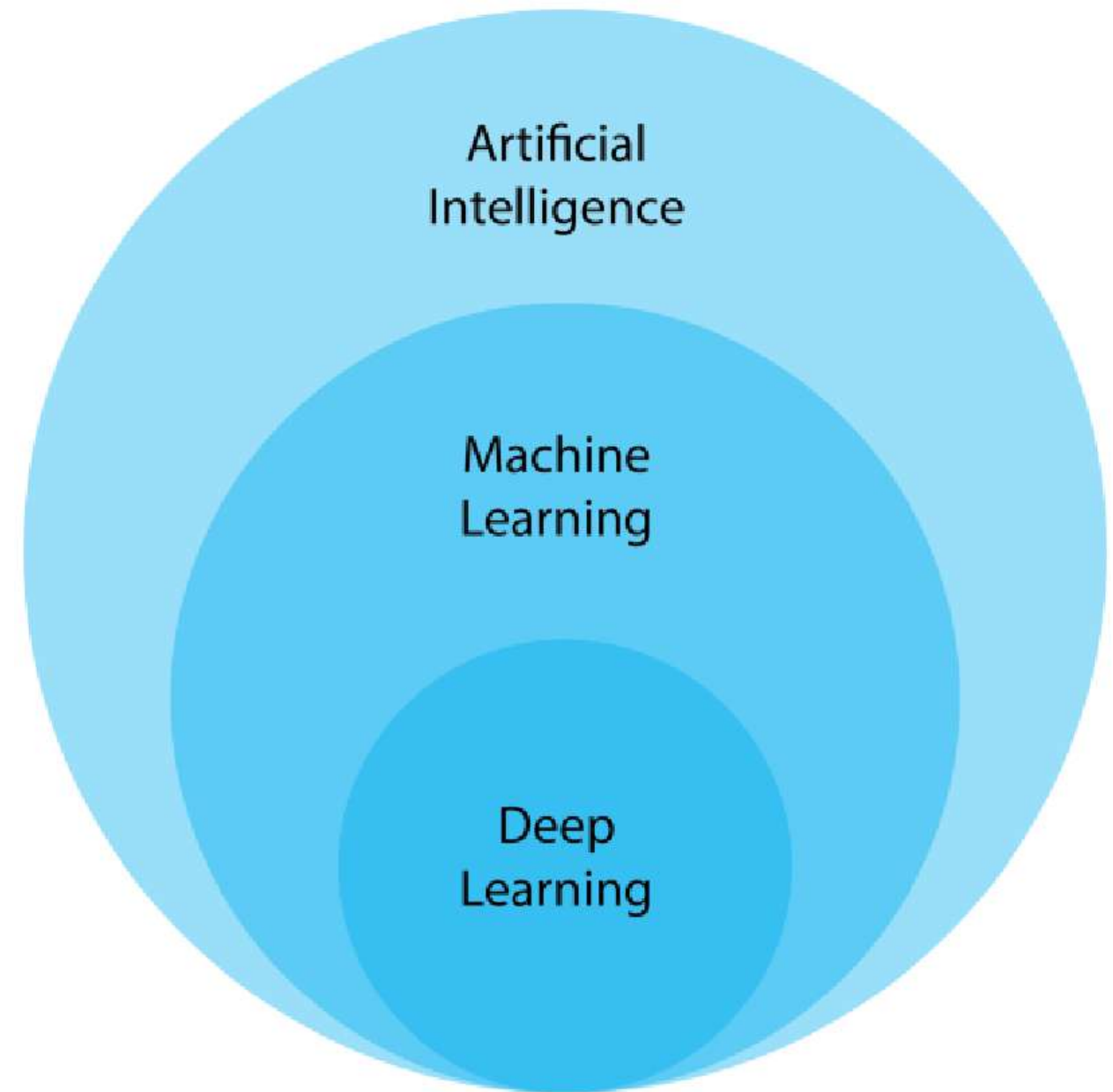


Is Data Science = AI? How are the two related?

Data Science and Artificial Intelligence (AI)

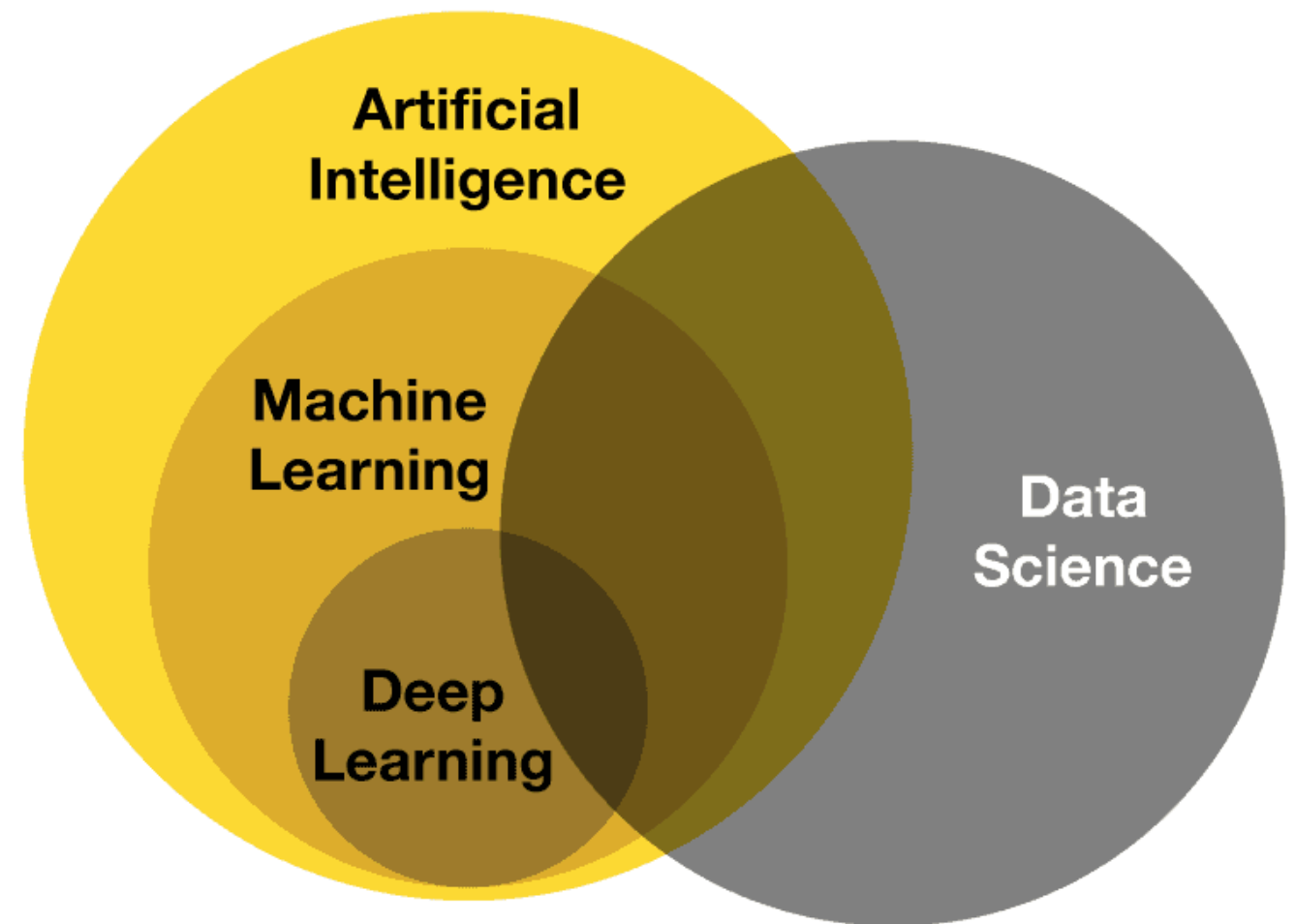
To answer these questions, we first need some further definitions:

- **Artificial Intelligence (AI)** is the broad field of developing machines that can replicate human behaviour, including tasks related to perception, reasoning, learning and problem-solving.
- One way of achieving this is through **Machine Learning (ML)** algorithms that detect patterns in large data sets and learn to make predictions from them.
- **Deep Learning (DL)** is an advanced branch of ML based on so-called neural networks. Innovations in this field are responsible for recent AI breakthroughs, such as ChatGPT.



Data Science at the heart of AI

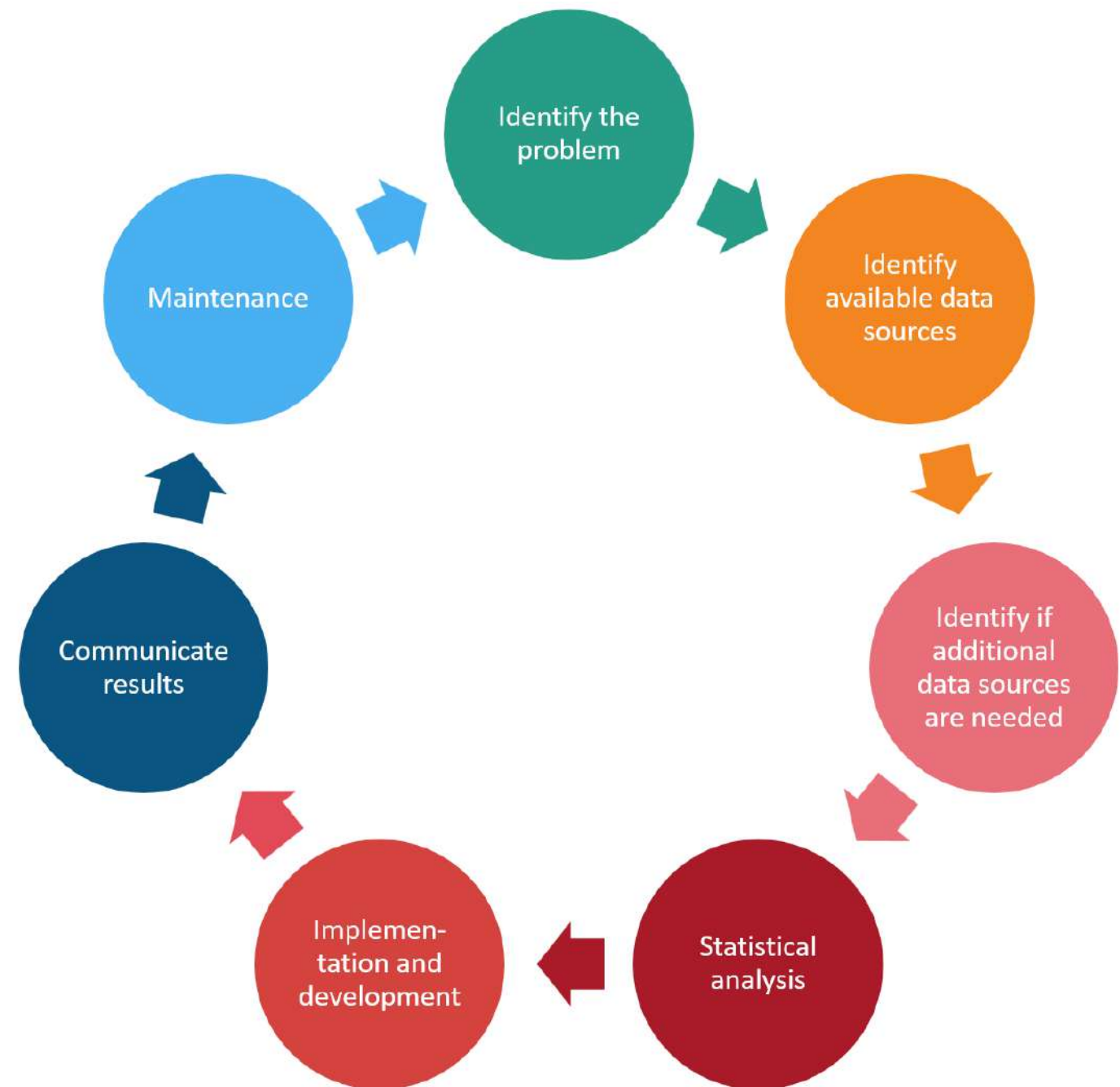
- With all of these fields, data science has significant **overlap**, but nonetheless, constitutes its own discipline.
- Analogously, there are branches of AI (e.g. robotics or symbolic reasoning) that do *not* revolve around learning patterns from various data sources.
- Many of the advances in AI require precisely the **interdisciplinary** combination of computer science, maths, statistics and domain expertise that data science provides.
- And yet, a data scientist does much more than develop and apply algorithms for AI...



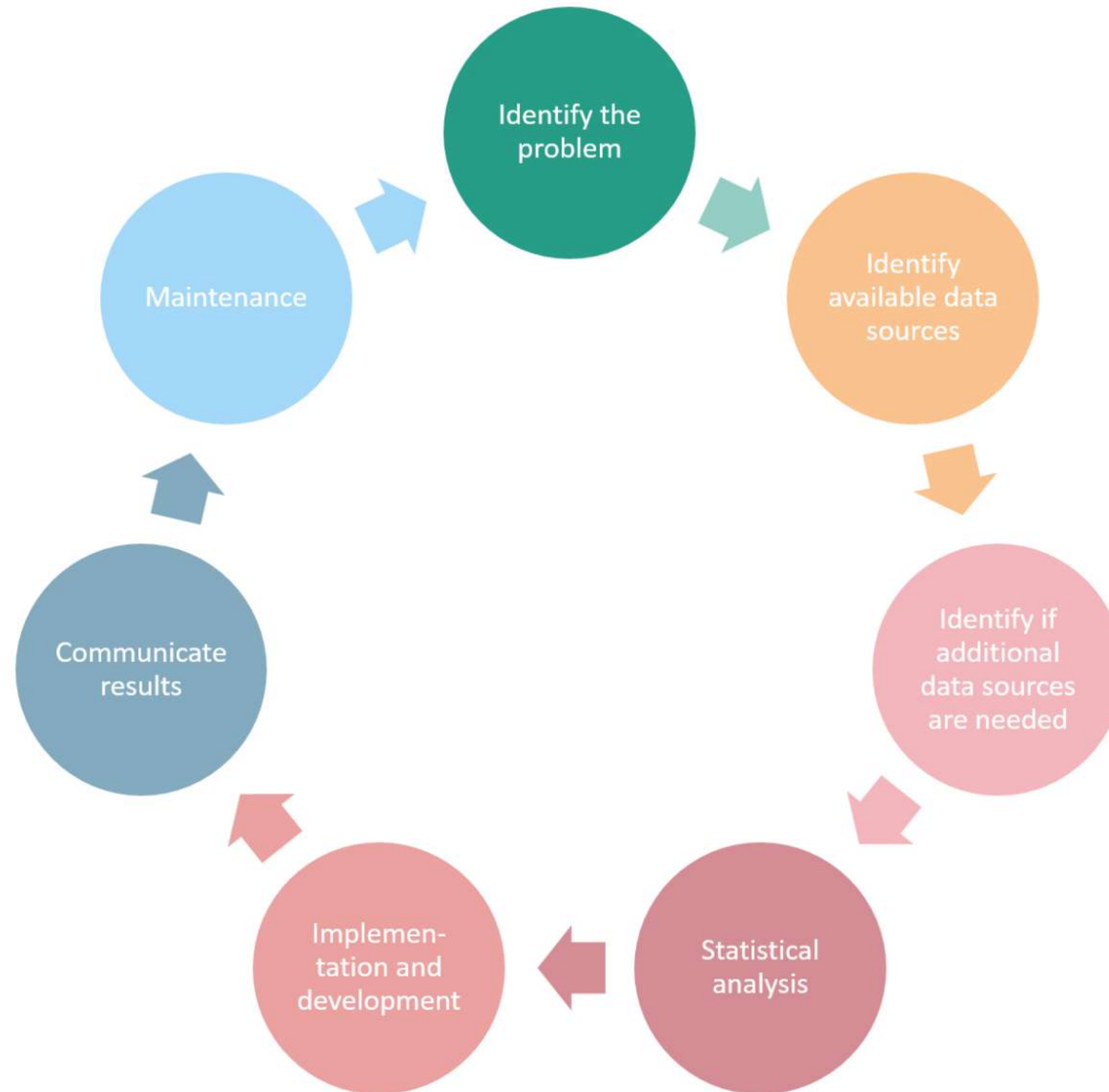
What does a Data Scientist do?

So what does a Data Scientist do?

- Now that we know roughly, what **data science** is, what does this discipline actually entail more concretely?
- In other words, what is it that a data scientist does every day?
- There are many ways to structure the data science, let's first look at the **seven stages** of a data science project.



Step 1 – Identify the problem



Step 1 – Identify the problem – Key aspects

- **Understand the business context:**
 - Familiarize yourself with the industry, market trends, and internal operations.
 - **Example:** As a data scientist in a struggling e-commerce company, study industry benchmarks and analyze internal sales, user behavior, and engagement data.
- **Define the problem statement:**
 - Together with the business, clearly articulate what issue needs solving.
 - **Example:** “Our customer churn rate has increased by 20% over the last six months, impacting revenue.”
- **Identify stakeholders:**
 - Determine who is impacted and who can provide insights.
 - **Example:** *marketing* to explain and refine their existing customer engagement strategies or *customer service* to point out existing issues in post-purchase support.

Step 1 – Identify the problem – Key aspects

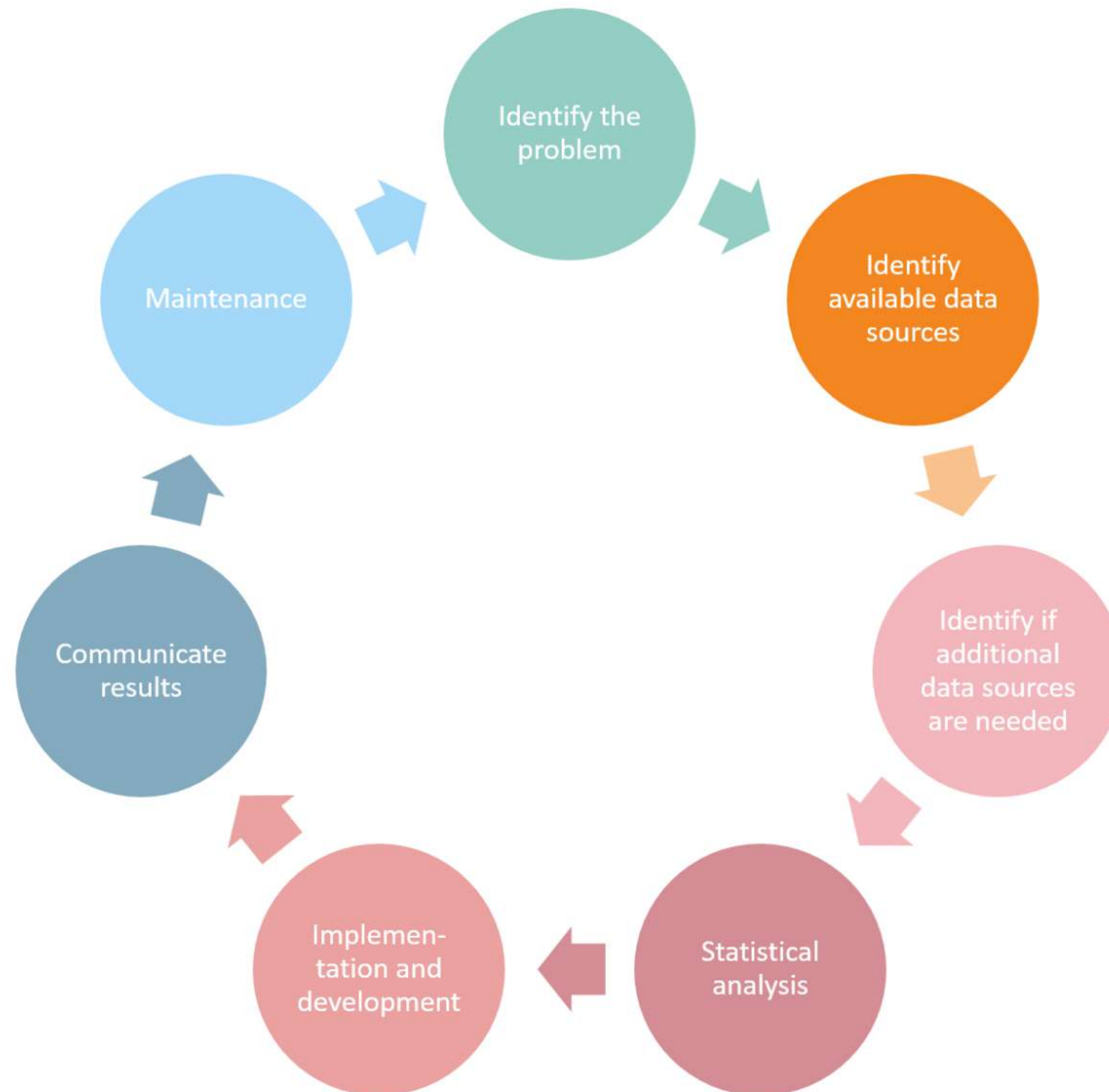
- **Outline goals & objectives:**
 - In collaboration with business, specify what success looks like and define measurable KPIs.
 - **Example:** “Reduce the churn rate by 10% within the next six months.”
- **Make the goals actionable:**
 - Identify potential strategies to achieve the goals as well as the data required to implement them.
 - **Example:** train a **machine learning model** to predict customer churn and use this model to tailor new customer engagement strategies. Requires data on past customer behaviour on the website and whether or not they churned.

Step 1 – Identify the problem – DS contributions

Importance of data science



Step 2 – Identify available data sources



Step 2 – Identify available data sources

- **Definition:** A **data source** refers to the physical or digital system where data is generated, stored and managed as a data table, data object or any other storage format. It is also where, stakeholders of this data (e.g. a data scientist) can access data for further use.
- **Why identify data sources early?**
 - **Foundation for analysis:** Data sources determine the quality, scope, and reliability of your insights.
 - **Strategic planning:** Knowing where data comes from helps define project goals and design the analytics pipeline.
 - **Integration & innovation:** Combining diverse sources (internal and external) can reveal hidden trends and create competitive advantages.

Classifying data sources

Internal sources

- **Definition:** Data generated, stored, and maintained within the organization.
- **Examples:**
 - Enterprise systems (CRM, ERP)
 - In-house databases
 - Spreadsheets, documents, ...
- **Key point:** Proprietary data that reflects your organization's operations.

External sources

- **Definition:** Data sourced from outside the organization.
- **Examples:**
 - Public datasets (government statistics, market research)
 - APIs (social media, financial data feeds)
- **Key point:** Supplement internal data and provide broader market or industry insights.

Internal data sources – CRM system

A Customer Relationship Management (CRM) system helps manage **customer data**. It helps businesses keep customer contact details up to date, track every customer interaction, and manage customer accounts.



Internal data sources – ERP system

The CRM typically forms part of the Enterprise Resource Planning (ERP) system. This type of software integrates data about **all main business processes** in a single system.



Internal data sources – Databases

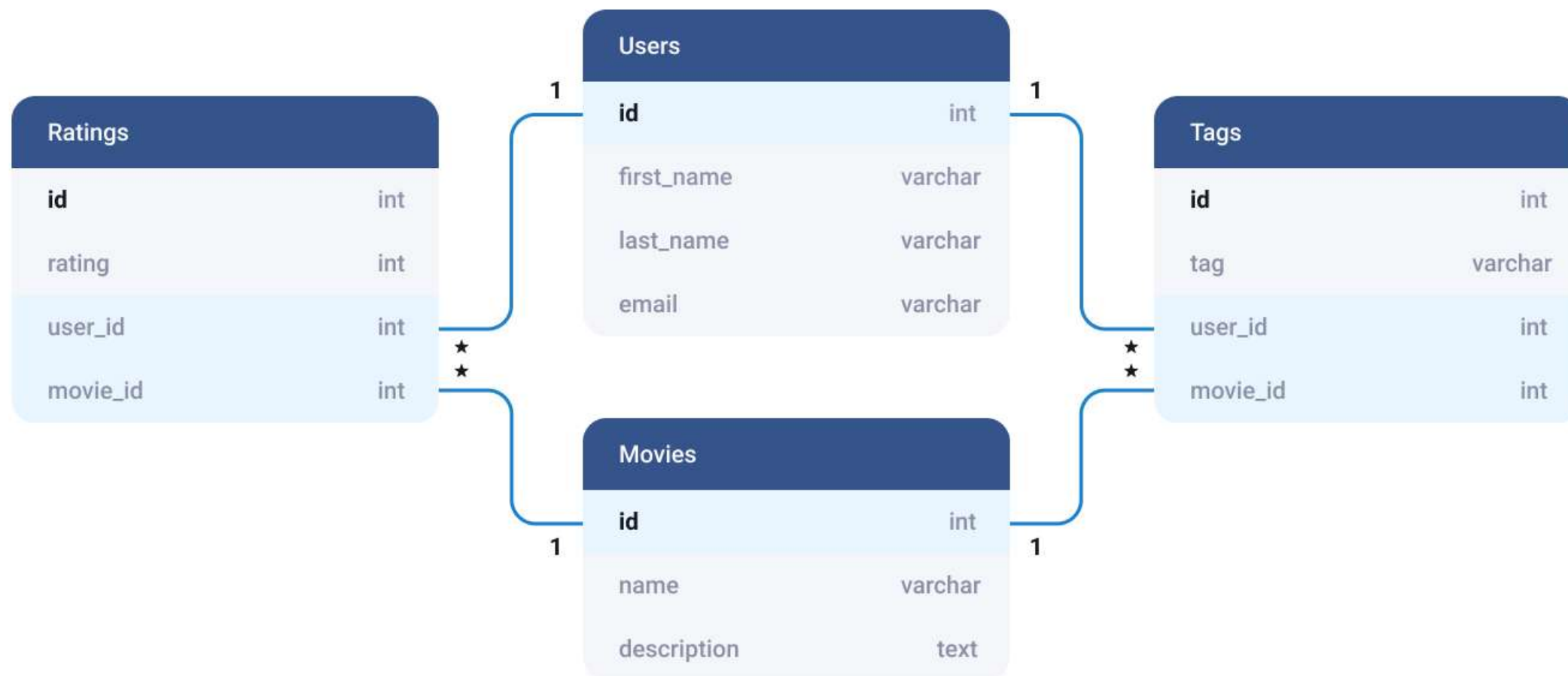
- A **database** is an organized collection of data stored and accessed electronically.
- Databases are managed using specialized software called a **Database Management System (DBMS)**, which allows users to store, retrieve, and manipulate data efficiently.
- Databases are the backbone of modern applications, supporting businesses, organizations, and systems across industries.
- Different types of databases can be classified in terms of their structure, usage, or storage methods.
Two main types are:

- Relational / SQL
- NoSQL



Internal data sources – Relational databases

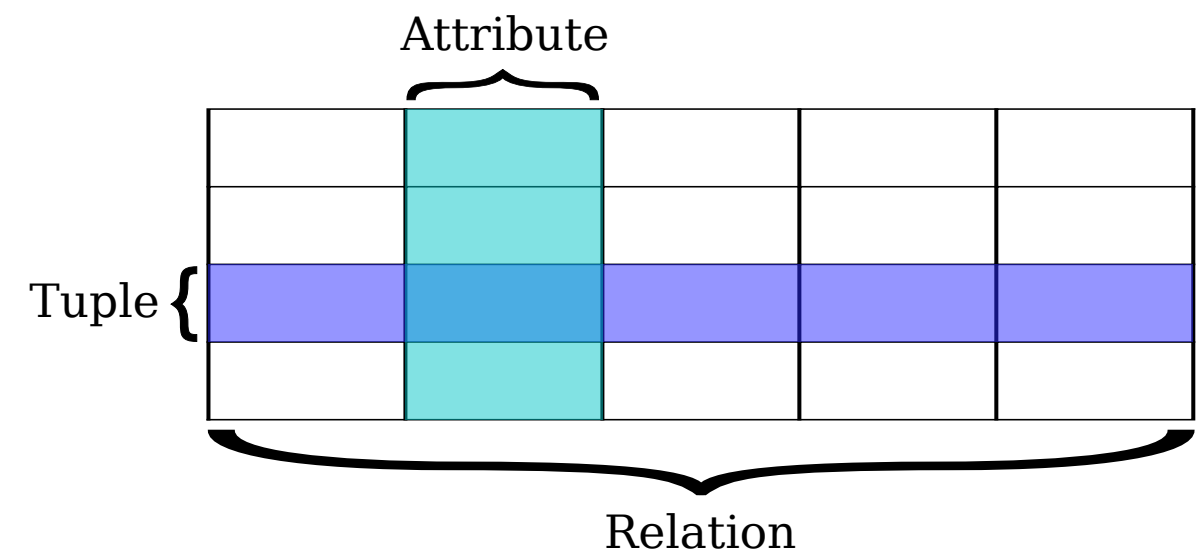
A **relational database** organizes data into **tables** with certain pre-defined relationships to each other. These relationships are logical connections, established on the basis of interaction among these tables.



Internal data sources – Relational databases

The **relational model** underlying such databases uses special lingo to refer to the entities that make up its structure:

Familiar term	Relational DB term
Table	Relation
Row	Tuple
Column	Attribute



- Unlike tables, relations are **unordered**, i.e. you can shuffle the order of rows or columns and still get the same relation.
- Attributes (columns) specify a **data type** (e.g. integer or text), and each tuple (or row) contains the value of that specific data type.

Internal data sources – Relational databases

- All tables in a relational database have an attribute known as the **primary key**, which is a unique identifier of a row.
- Each row can be used to create a **relationship** between different tables using a **foreign key**, i.e. a reference to a primary key of another existing table.

Let's see how this looks in practice: Suppose we have the following table:

book_id	Title	Author	Format	Publisher	Country	Price
1	Harry Potter	J.K. Rowling	Paperback	Banana Press	UK	\$20
2	Harry Potter	J.K. Rowling	E-book	Banana Press	UK	\$10
3	Sherlock Holmes	Conan Doyle	Paperback	Guava Press	US	\$30
4	The Hobbit	J.R.R. Tolkien	Paperback	Banana Press	UK	\$30
5	Sherlock Holmes	Conan Doyle	Paperback	Guava Press	US	\$15

Clearly, the primary key in this table is **book_id**.

Internal data sources – Relational databases

- What if “Banana Press” changed its name to “Pineapple Press”?
- We would have to change every single instance of “Banana Press” in the **Publisher** attribute. Very cumbersome...
- What if we instead organized our database like this?

book_id	Title	Author	Format	Publisher_ID	Price
1	Harry Potter	J.K. Rowling	Paperback	pub_1	\$20
2	Harry Potter	J.K. Rowling	E-book	pub_1	\$10
3	Sherlock Holmes	Conan Doyle	Paperback	pub_2	\$30
4	The Hobbit	J.R.R. Tolkien	Paperback	pub_1	\$30
5	Sherlock Holmes	Conan Doyle	Paperback	pub_2	\$15

Publisher_ID	Publisher	Country
pub_1	Banana Press	UK
pub_2	Guava Press	US

Internal data sources – Relational databases

- We introduced the **foreign key** “Publisher_ID” into book relation.
- Now, we would just have to change the name of the publisher once, namely in the publisher relation.
- This is an example of a process called **database normalization**, which helps reduce redundancy and improve data integrity.
- To retrieve data from a relational database, one uses a specialized computer language called **Structured Query Language (SQL)**, which gives relational databases an alternative name.
- To reconstruct our original book relation, the SQL statement would be:

```
1 SELECT b.book_id, b.Title, b.Author, b.Format,  
2         p.Publisher, p.Country, b.Price  
3 FROM Books AS b  
4 LEFT JOIN Publishers AS p ON b.Publisher_ID = p.Publisher_ID
```

Internal data sources – Relational databases

SQL will not be part of this course, but data scientists working in industry tend to write a lot of SQL...



Internal data sources – Types of data

- Relational databases are the de-facto standard for managing **structured data**, i.e. data that is highly organized in a table-like format.
- Data can also take other forms though:
 - **Unstructured data** is not organized in a specific way and does therefore not lend itself to the relational model.
 - **Examples:** text documents, images and video.
 - **Semi-structured data** is a hybrid of the two preceding forms, containing some organizational elements (like tags or markers), but allowing for more flexibility than data stored in a relational model.
 - **Examples:** Data stored in JSON, XML or HTML files.

Internal data sources – Types of data

For unstructured and semi-structured data, the relational model is typically not suitable, or at least suboptimal. For this purpose, **non-relational models** are preferred.

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

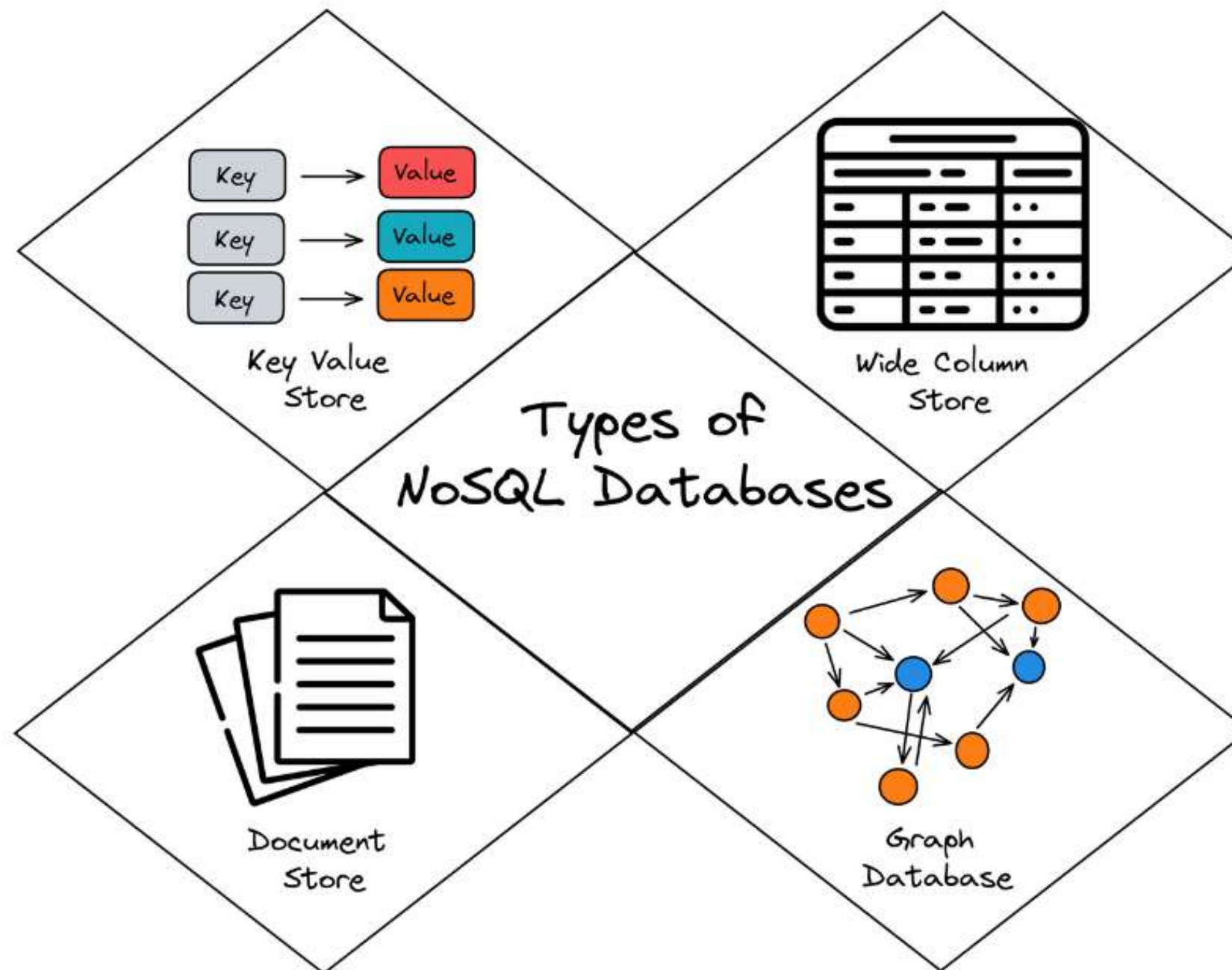
Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Internal data sources – NoSQL databases

- Despite some fuzzy terminology, non-relational databases are commonly referred to as **NoSQL**.
- Types of NoSQL databases include:
 - **Key-value store**: simplest form of a NoSQL database, typically of little use in data science.
 - **Wide column store**: uses tables, rows and columns similar to relational DBs, but names and format of columns can vary from row to row.
 - **Document store**: perfect for semi-structured data, can handle complex data structures.
 - **Graph database** prioritize relationships between data objects. They use nodes (data entities) and edges (relationships) to model data.
- As an illustration of the kind of benefits achievable through NoSQL databases, we consider an example for a document store.

Internal data sources – NoSQL databases



Internal data sources – NoSQL databases

- As the name suggests, a **document store** is centred around the concept of “document”, like a JSON or XML file.
- All documents are assumed to be encoded in the same format.
- Each document has a unique key that can be used to retrieve it.
- A collection of documents could be considered analogous to a **table** in a relational database, where each row is a document. However, a collection of documents is much more **flexible**:
 - In a table, all rows must have the same sequence of columns.
 - In a collection of documents however, each document can have completely different fields.

Internal data sources – NoSQL databases

```
{
  "Title": "Harry Potter",
  "Author": "J.K. Rowling",
  "Publisher": "Banana Press",
  "Country": "UK",
  "Sold as": [
    {"Format": "Paperback", "Price": "$20"},
    {"Format": "E-book", "Price": "$10"}
  ]
}
```

Document 1 – Harry Potter

```
{
  "Title": "Sherlock Holmes",
  "Author": "Conan Doyle",
  "Publisher": "Guava Press",
  "Country": "US",
  "Sold as": [
    {"Format": "Paperback", "Price": "$30"},
    {"Format": "E-book", "Price": "$15"}
  ]
}
```

Document 2 – Sherlock Holmes

```
{
  "Title": "The Hobbit",
  "Author": "J.R.R. Tolkien",
  "Publisher": "Banana Press",
  "Country": "UK",
  "Sold as": [
    {"Format": "Paperback", "Price": "$30"}
  ]
}
```

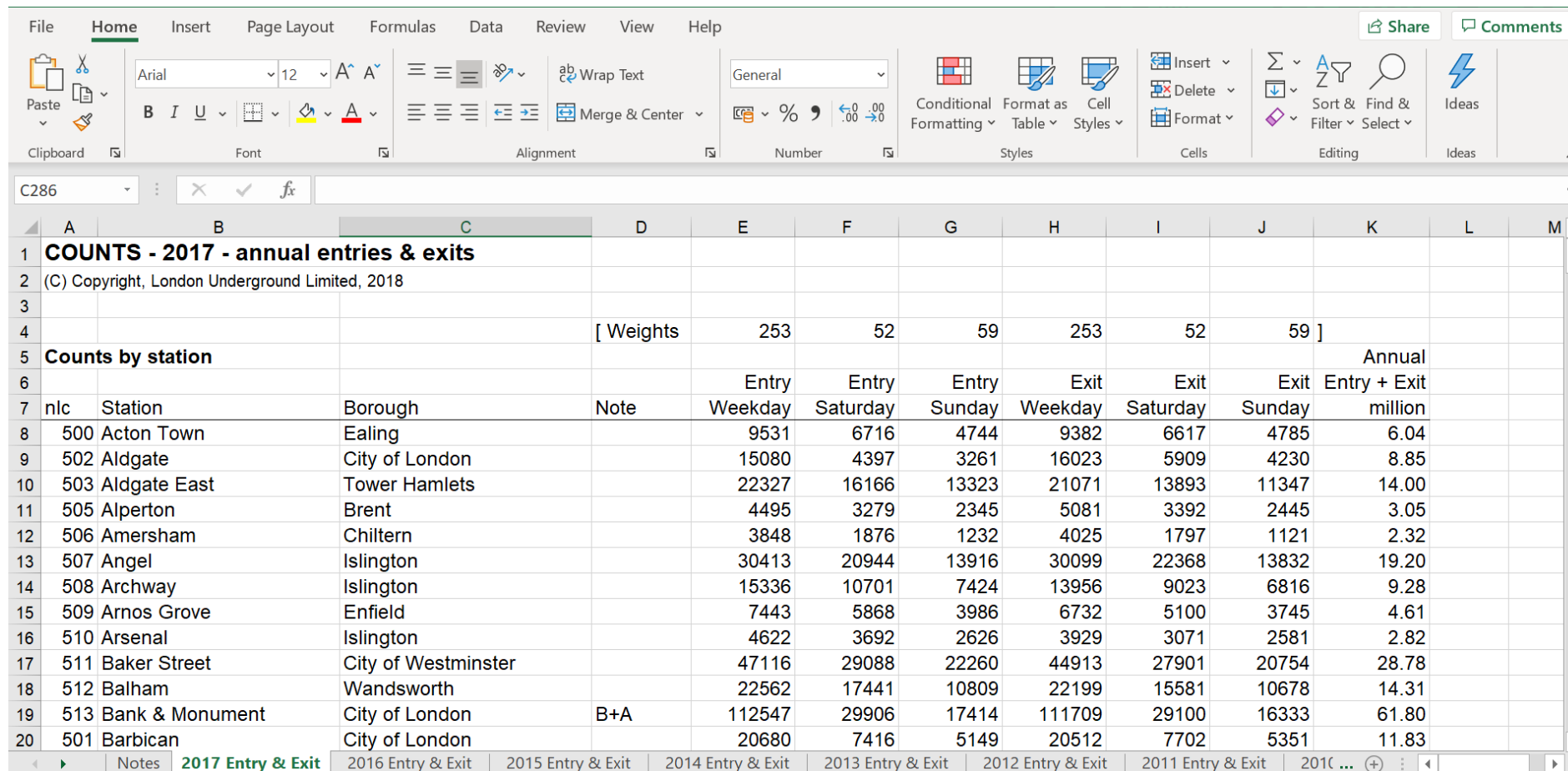
Document 3 – The Hobbit

Advantages of such a NoSQL database:

- No need to normalize
- Easier extensibility: changes in one document do not affect others.
- Better locality (all information on a document in one place instead of spread across several tables).

Internal data sources – Others

- Unfortunately, a lot of data in corporations is *not* in a nicely structured form in a database... Instead, big amounts of data are spread around in **messy spread sheets**, e-mails, presentations, csv-files, etc.
- Organizing and cleaning such data to gain meaningful insight is a key part of being a data scientist.



COUNTS - 2017 - annual entries & exits										
(C) Copyright, London Underground Limited, 2018										
[Weights 253 52 59 253 52 59]										
Counts by station										Annual
n/c	Station	Borough	Note	Weekday	Saturday	Sunday	Weekday	Saturday	Sunday	Entry + Exit million
500	Acton Town	Ealing		9531	6716	4744	9382	6617	4785	6.04
502	Aldgate	City of London		15080	4397	3261	16023	5909	4230	8.85
503	Aldgate East	Tower Hamlets		22327	16166	13323	21071	13893	11347	14.00
505	Alperton	Brent		4495	3279	2345	5081	3392	2445	3.05
506	Amersham	Chiltern		3848	1876	1232	4025	1797	1121	2.32
507	Angel	Islington		30413	20944	13916	30099	22368	13832	19.20
508	Archway	Islington		15336	10701	7424	13956	9023	6816	9.28
509	Arnos Grove	Enfield		7443	5868	3986	6732	5100	3745	4.61
510	Arsenal	Islington		4622	3692	2626	3929	3071	2581	2.82
511	Baker Street	City of Westminster		47116	29088	22260	44913	27901	20754	28.78
512	Balham	Wandsworth		22562	17441	10809	22199	15581	10678	14.31
513	Bank & Monument	City of London	B+A	112547	29906	17414	111709	29100	16333	61.80
501	Barbican	City of London		20680	7416	5149	20512	7702	5351	11.83

External data sources – Public data sets

- External data sources are often crucial to understand the general economic environment, market or industry trends, and much more.
- The first approach towards external data sources is usually to verify whether there are any **publicly available** data sets from reliable sources, such as:

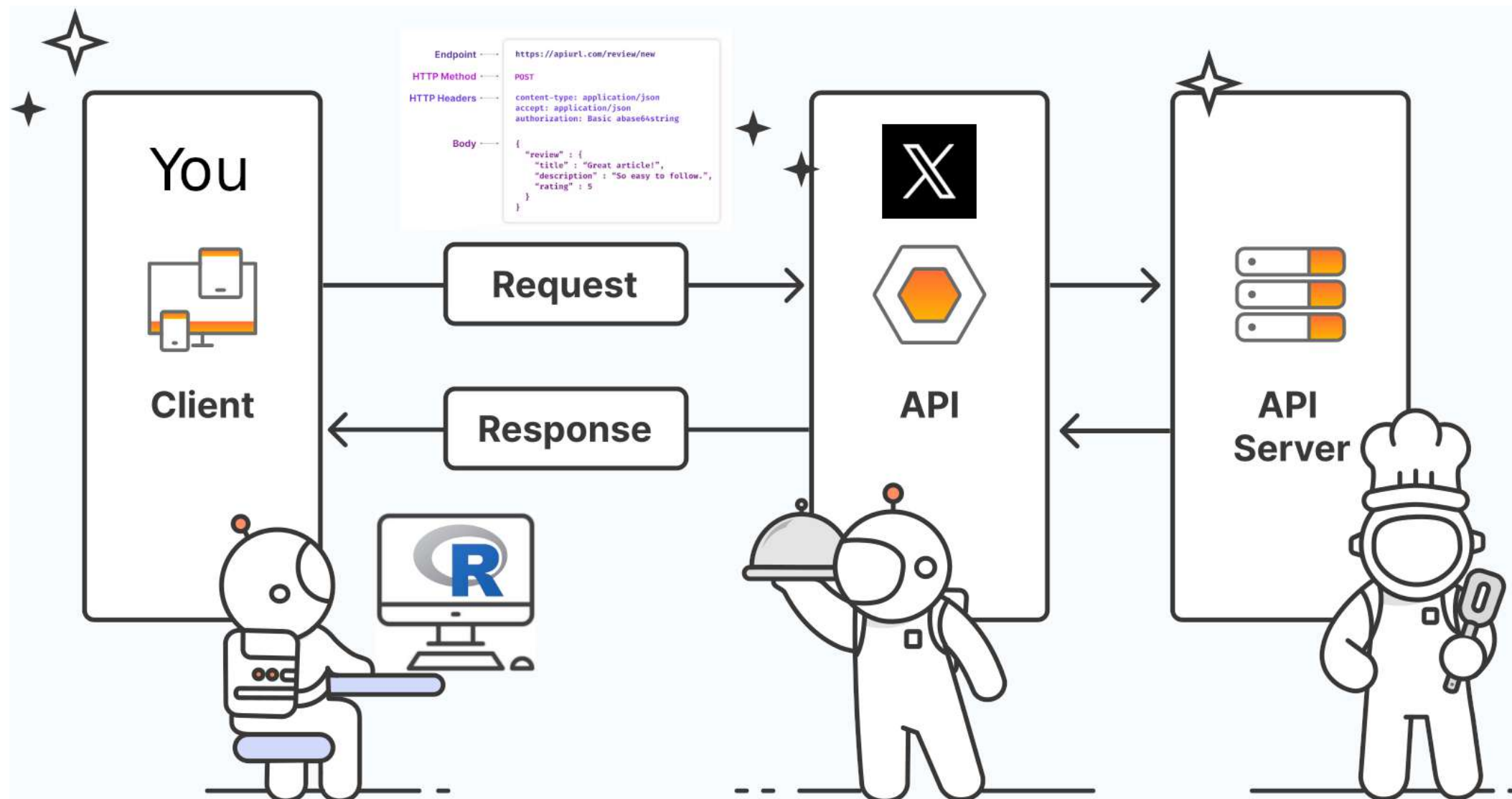


External data sources – APIs

- Often, rather than downloading a data set in form of a csv or xls file, data providers offer **Application Programming Interfaces (APIs)** to access their data directly.
- APIs serve as a bridge between different software applications, enabling them to communicate and share data in (near) real-time.
- As a data source, APIs can be used, for instance, to access data from...
 - ... social media to detect and leverage current viral trends.
 - ... AI-providers like OpenAI to incorporate information produced by ChatGPT.
 - ... tech companies like Google to identify highly-rated competitors in the area based on Google Maps.
 - ...

External data sources – APIs

Let's say you wanted to find out, how many views all of your company's tweets on X generated last week. If your company tweets a lot, that's a lot of manual labour... Instead, we could write a program (in R) that uses the [X API](#) to gather this information.



External data sources – APIs

- Communication with APIs works through a **request and response** cycle: the request is sent to the API, which retrieves the data and returns it to the user in form of the response.
- An API request will typically include the following components:
 - **Endpoint:** a dedicated URL declared by the API provider to access the desired resource, e.g. <https://api.x.com/2/insights/historical> (see [here](#)).
 - **Method:** the type of operation the client wants to perform. Data retrieval operations typically use the **GET method** (see [here](#) for more information on HTTP methods).
 - **Parameters:** the variables indicating the specific instructions for the API to process, e.g. the time period for which you want to analyze the views of all tweets.
 - **Request headers:** meta information about the request, e.g. authentication credentials, such as an **API key**.
 - **Request body:** contains further information to be passed to the API, e.g. parameters that are not passed as part of the endpoint.

External data sources – APIs

HTTP Request

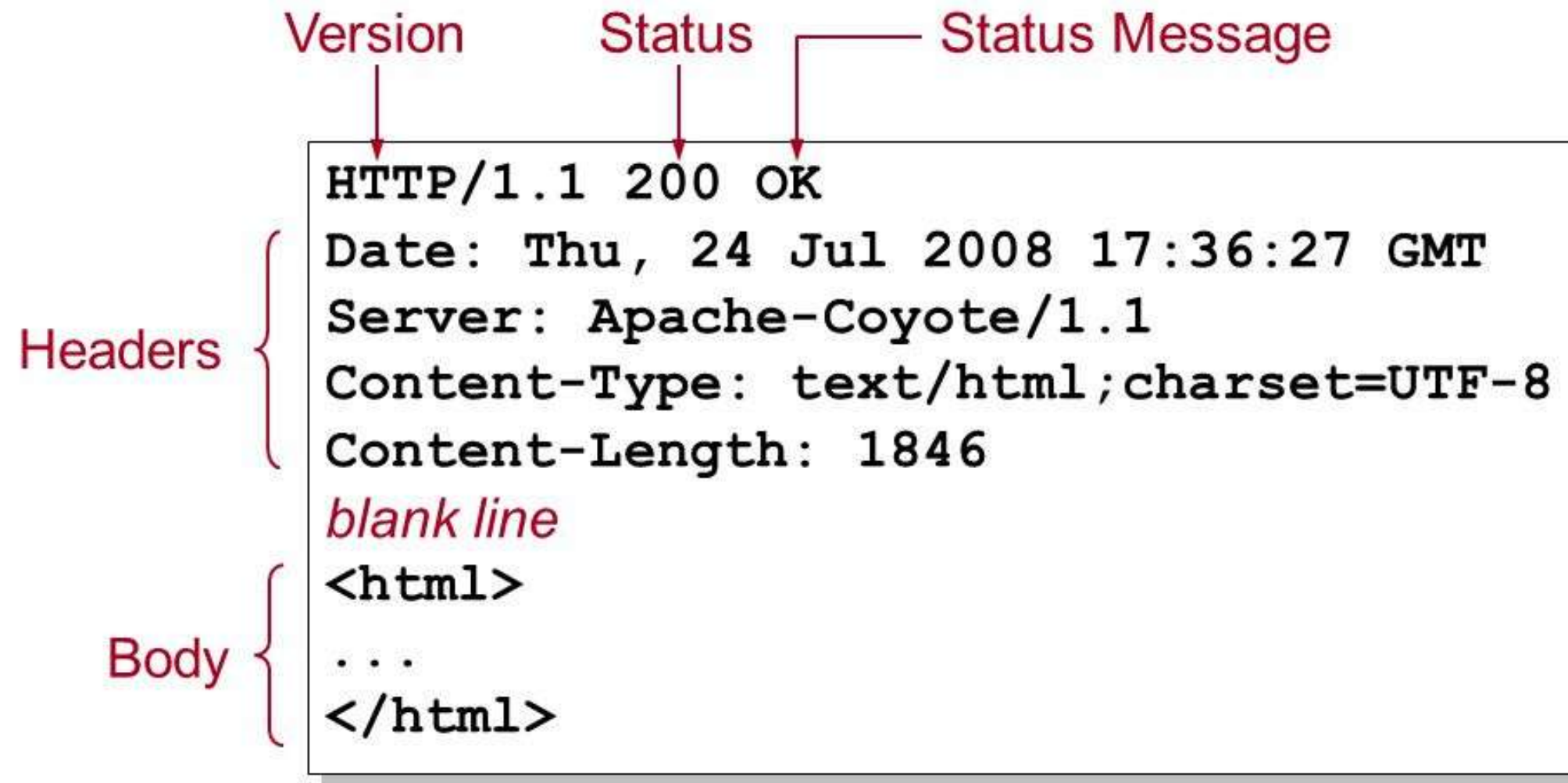


External data sources – APIs

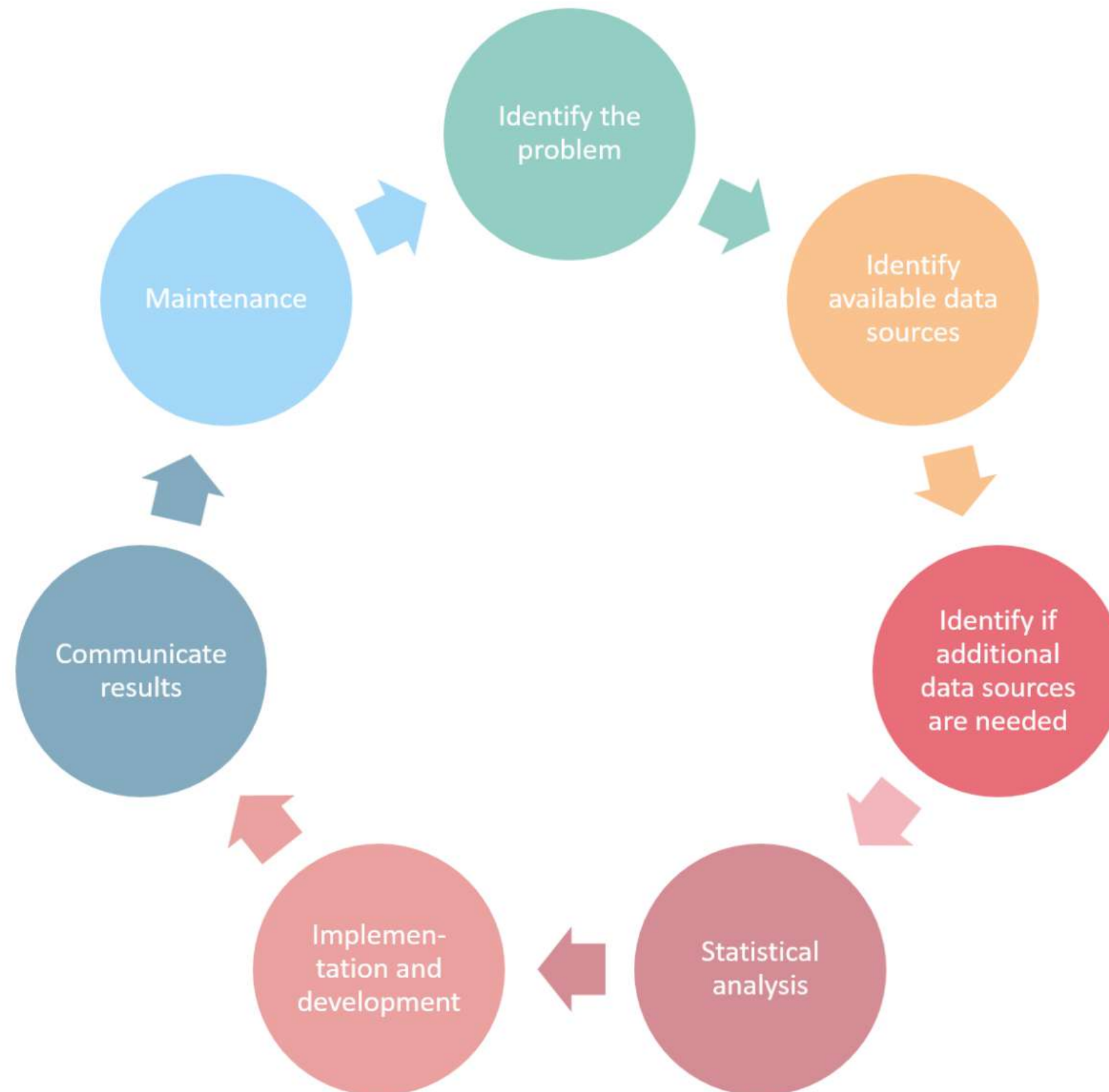
- The **response** returned by the API will then typically include the following components:
 - **Status code:** HTTP status codes are three-digit codes that indicate the outcome of an API request. Some common ones are:
 - 200: OK (server successfully returned the requested data)
 - 400: Bad request (e.g. malformed request syntax)
 - 404: Not found (server cannot find requested resource)
 - **Response headers:** very similar to request headers, but provide additional information about response.
 - **Response body:** includes actual requested data, e.g. the number of views for each of last week's tweets.

External data sources – APIs

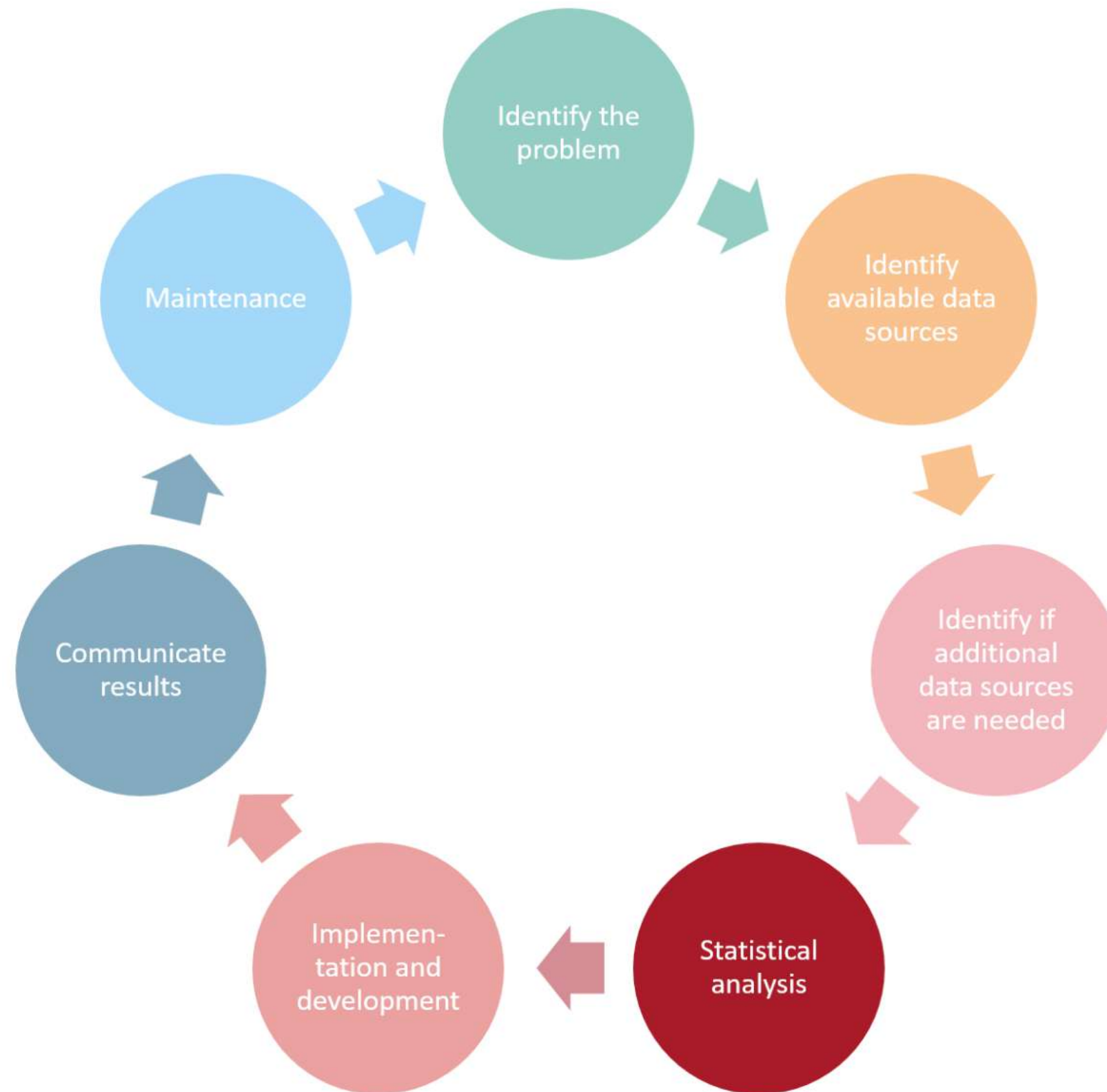
HTTP Response



Step 3 – Identify if additional sources are needed



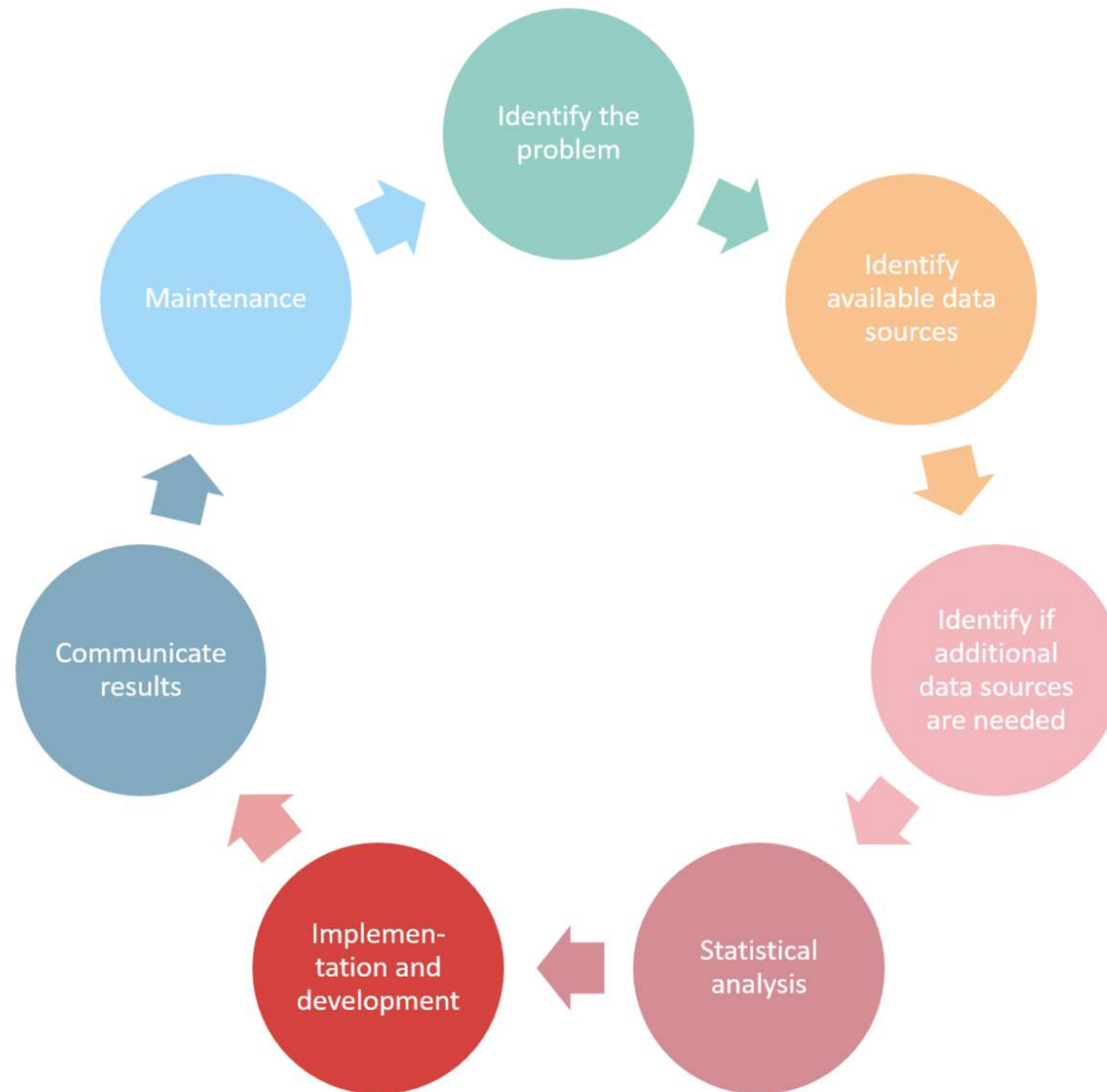
Step 4 – Statistical analysis



Step 4 – Statistical analysis

- After data has been obtained, the data scientist needs to clean, process and explore it. This can involve:
 - Handling missing values, outliers, and noise.
 - **Exploratory data analysis (EDA)**: Visualizations, correlation analysis, and summary statistics.
- Based on clean data, **statistical tests** can be employed to validate existing business hypotheses. Here, all the tests you have (hopefully) learned in your statistics classes come in, e.g. T-tests, ANOVA, χ^2 tests, etc.
- Through this process, the data scientist can start to develop an **understanding** of the data.

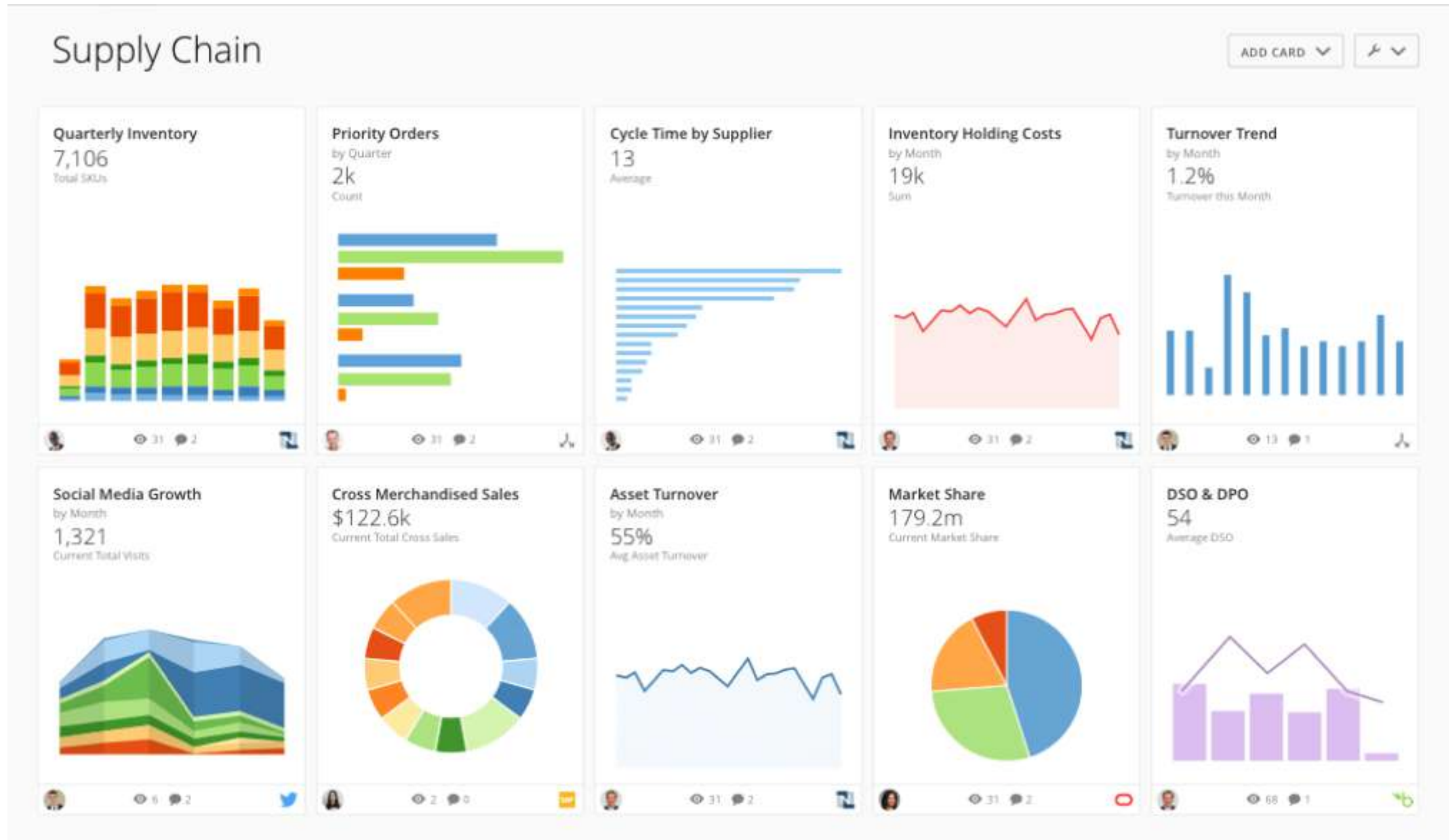
Step 5 – Implementation and development



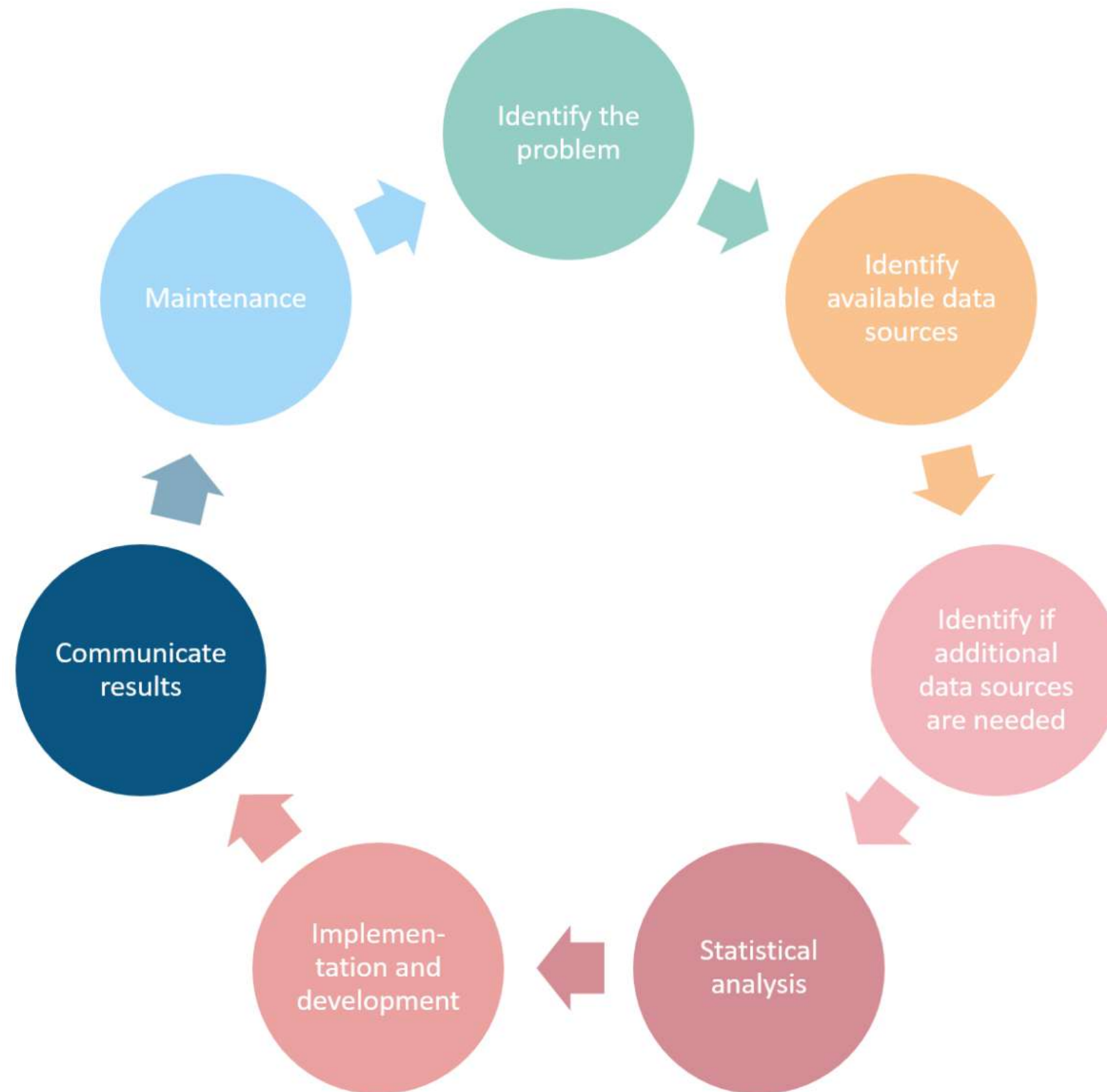
Step 5 – Implementation and development

- After familiarizing themselves with the data, data scientists then have to **translate insights into action**. This involves offering some kind of product back to the business.
- This could take the form of:
 - Training and deploying a **machine learning model**, e.g. offering the business a model that tells them the probability that a given customer will churn.
 - Building **dashboards** to make the data-driven insights accessible to non-technical people in business.
 - Support strategic decision making by providing appropriate statistics, figures, graphs and other data.

Step 5 – Implementation and development



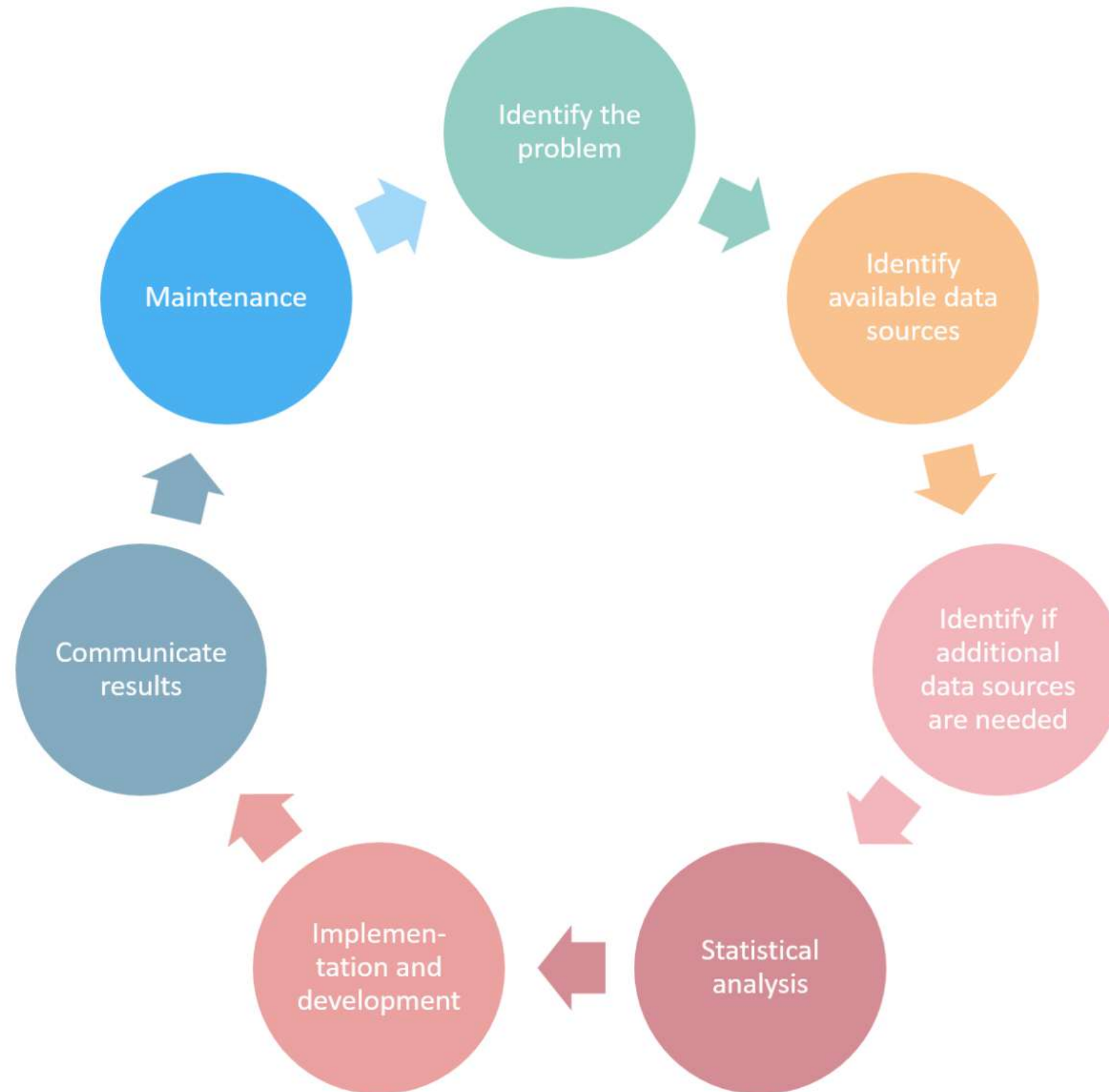
Step 6 – Communicate results



Step 6 – Communicate results

- Communication is an often underappreciated aspect of data science in business: any data insight can only be as good as it can be made understood by the people that (should) act on it.
- Key communication goals:
 - **Simplify** complex technical details without losing accuracy.
 - **Tailor the message to your audience** (executives, technical teams, clients).
- Communication strategies:
 - **Visual storytelling:** Use charts, graphs, and dashboards to highlight trends and key metrics.
 - **Narrative & context:** Frame results within a business story: What problem was solved? How do the results impact the business?
 - **Actionable insights:** Include clear recommendations and next steps.

Step 7 – Maintenance

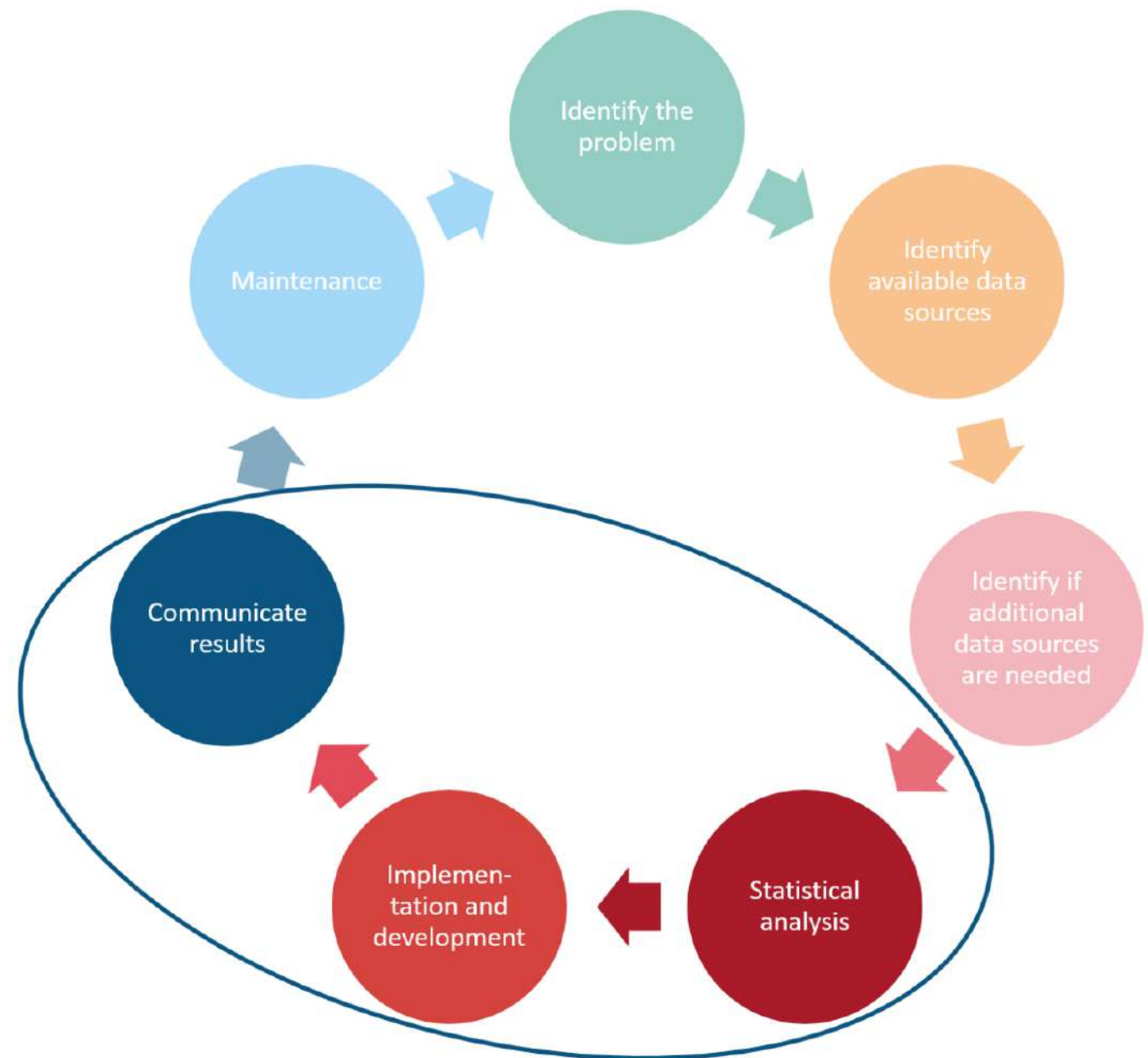


Step 7 – Maintenance

- Once a **data science product** like a machine learning model or a dashboard is “in production”, i.e. is used by business, it will not age like fine wine...
- Instead, changes in data over time and general **software rot** will quickly cause the product to be useless if left unchanged.
- Hence, data scientists need to subject their products to continuous **maintenance**:
 - **Monitor performance over time**: track metrics to detect data drift or deterioration in model performance.
 - **Update models and systems**: retrain models with new data when necessary.
 - Regularly engage stakeholders for **feedback**
 - Identify new problems and areas for improvement and **restart** the cycle.

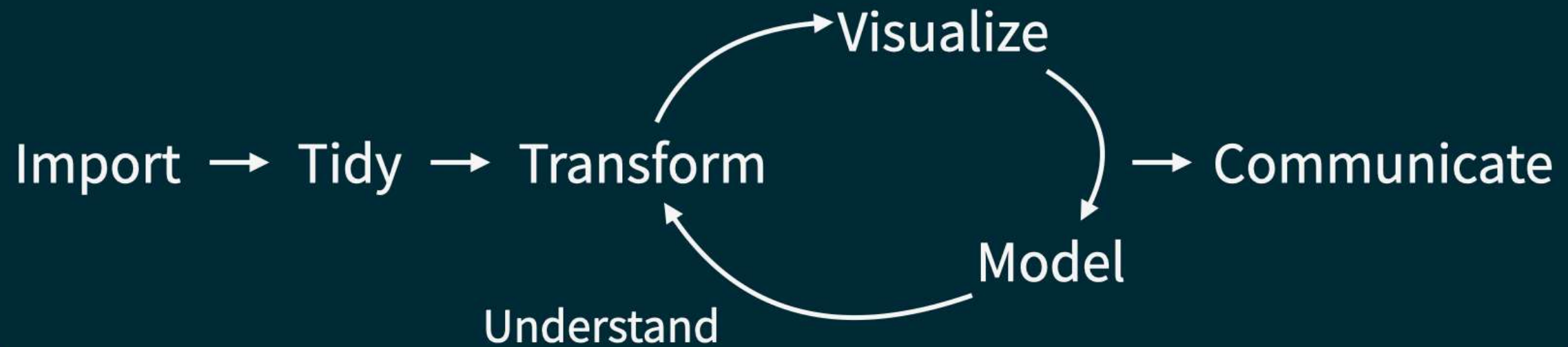
The essence of Data Science

- Data science is a demanding field that requires skills in maths, statistics and computer science, but also high levels of **domain expertise** and **excellent communication**.
- Thus, it is impossible to cover all areas of DS in a single course...
- This course focuses on what could be called the **essence of data science**:
 - Statistical analysis
 - Implementation and development
 - Communication of results
- We will further break these steps down in the **data science workflow**.



The Data Science workflow

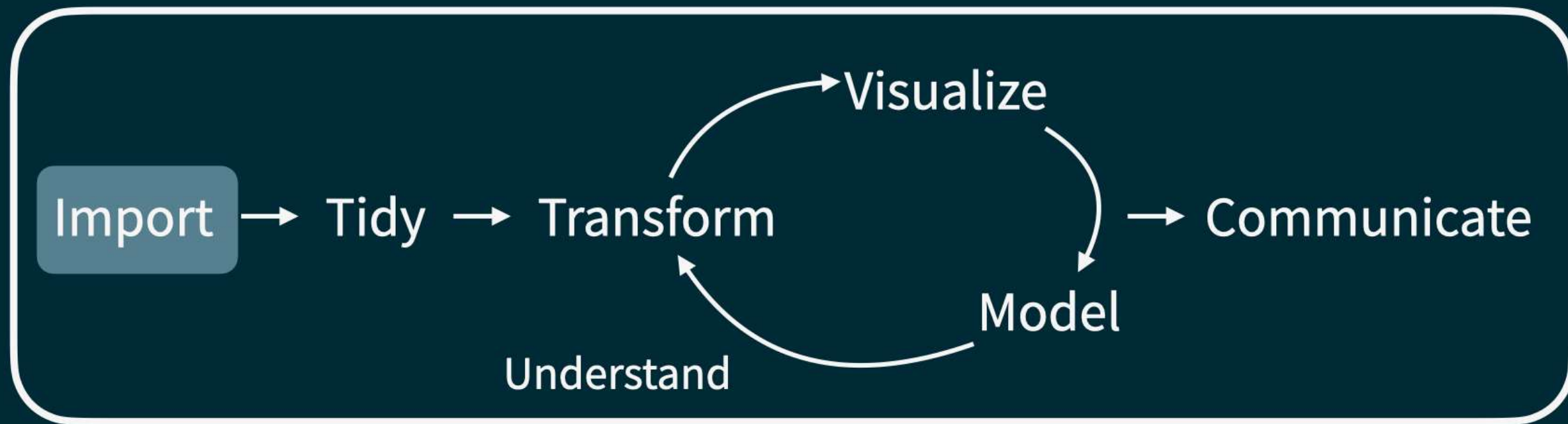
The Data Science workflow



Program

Source: [Wickham et al. \(2023\)](#)

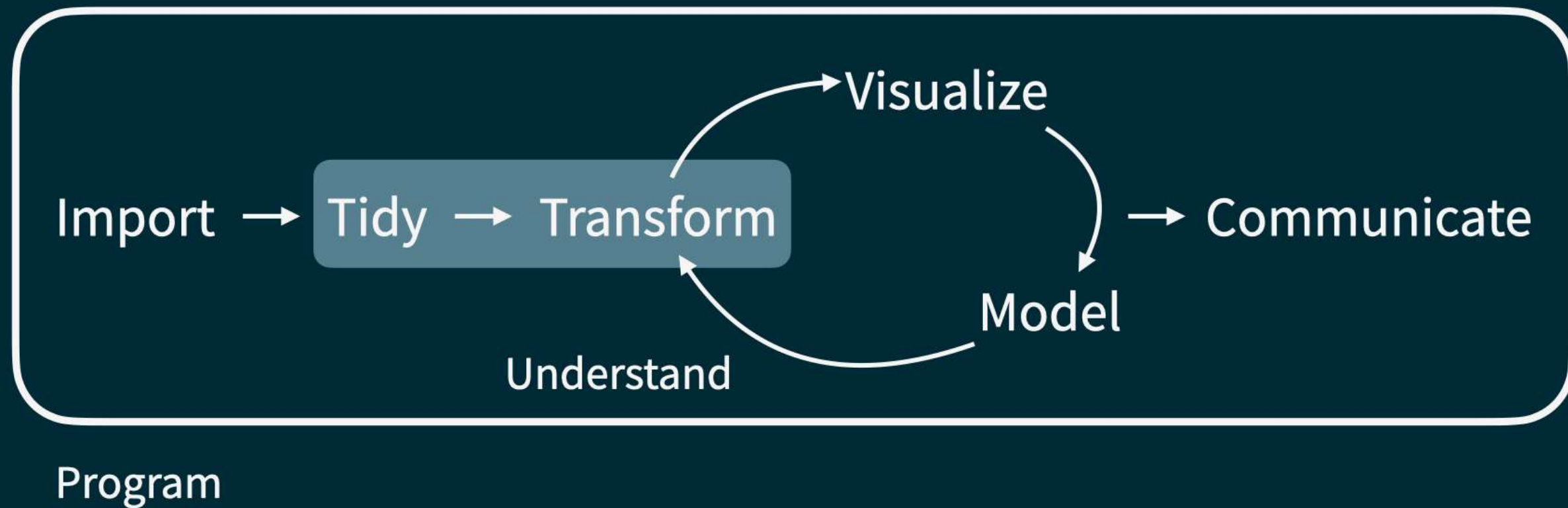
The Data Science workflow – Part I



Program

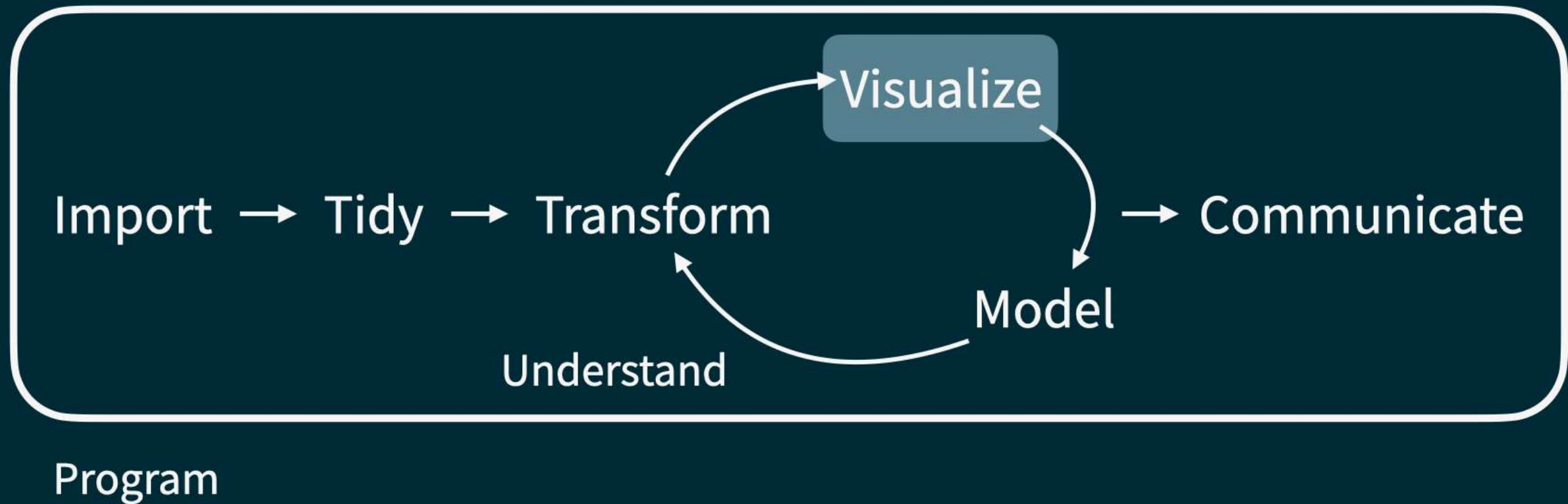
Source: [Wickham et al. \(2023\)](#)

The Data Science workflow – Part I



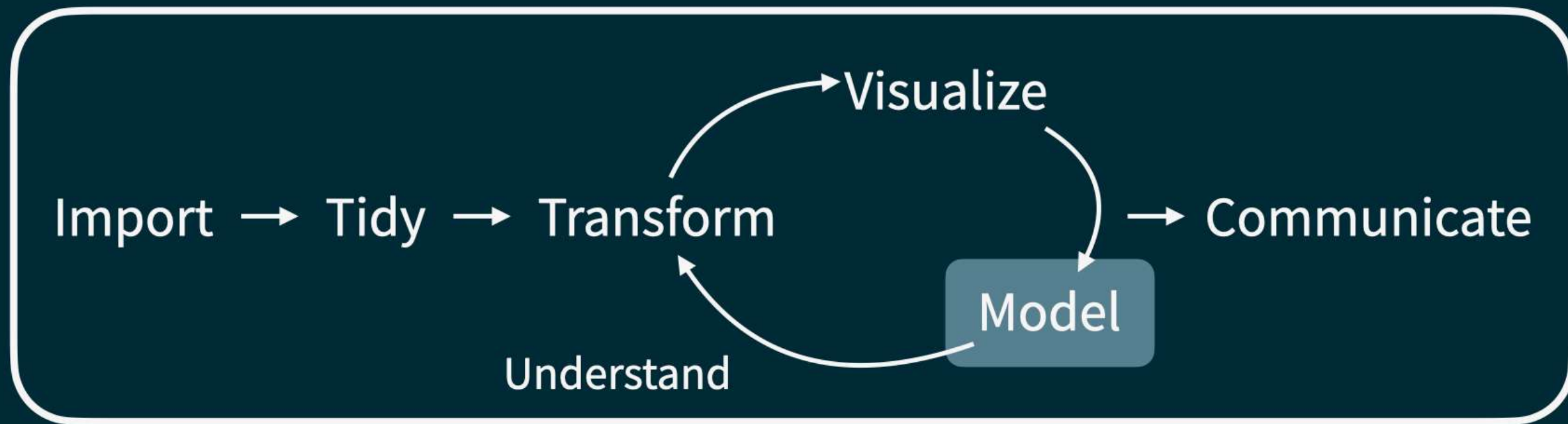
Source: [Wickham et al. \(2023\)](#)

The Data Science workflow – Part II



Source: [Wickham et al. \(2023\)](#)

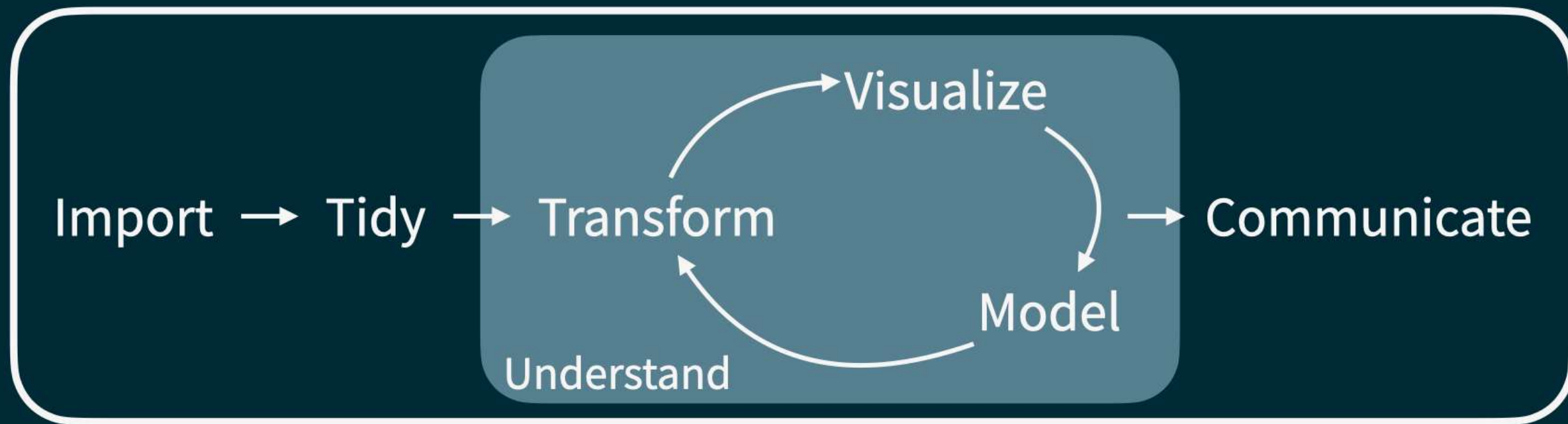
The Data Science workflow – Part III



Program

Source: [Wickham et al. \(2023\)](#)

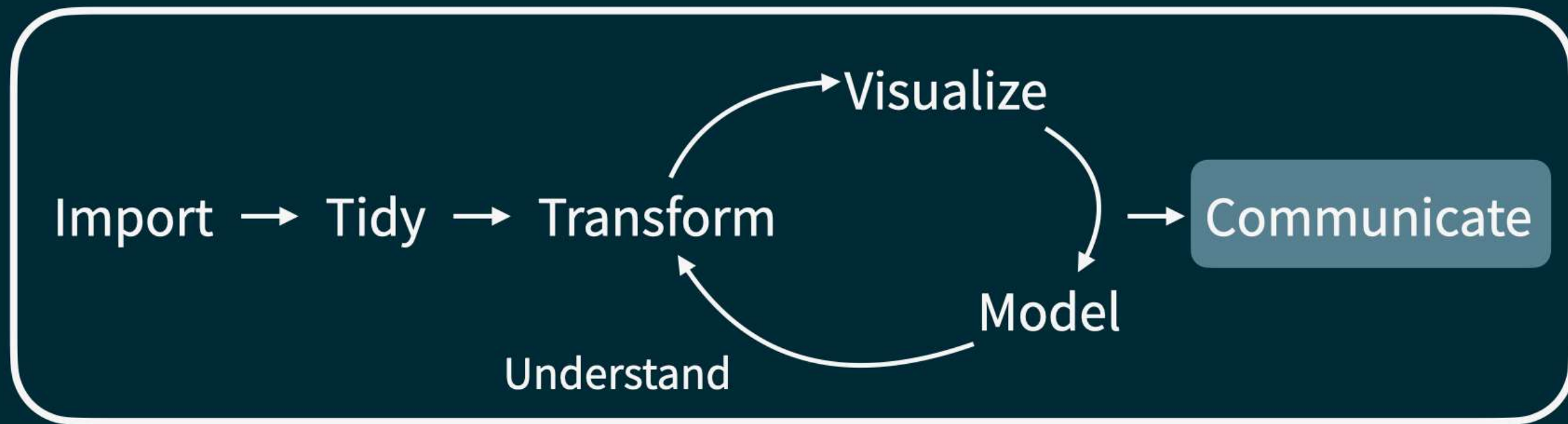
The Data Science workflow – Understanding data



Program

Source: [Wickham et al. \(2023\)](#)

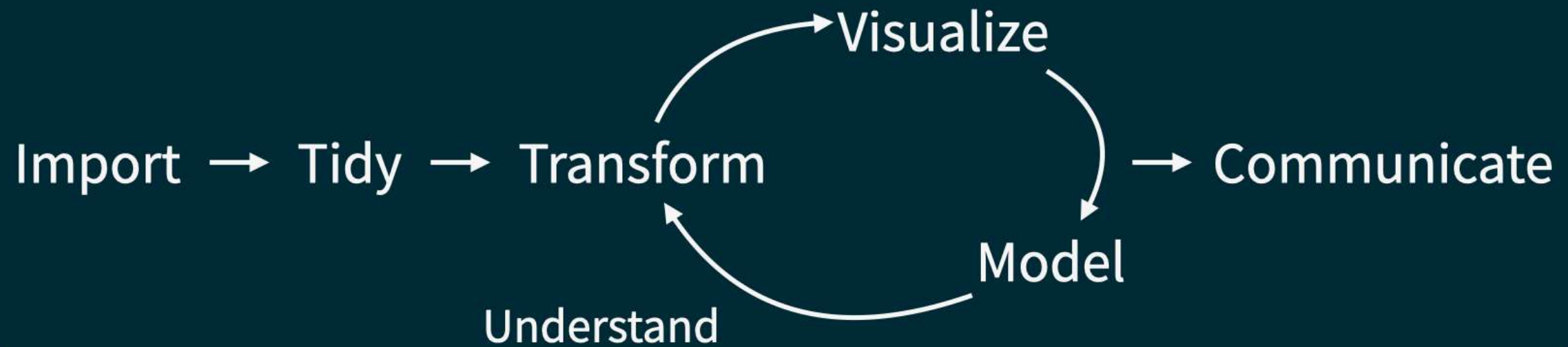
The Data Science workflow – Part IV



Program

Source: [Wickham et al. \(2023\)](#)

The Data Science workflow – How? Program!

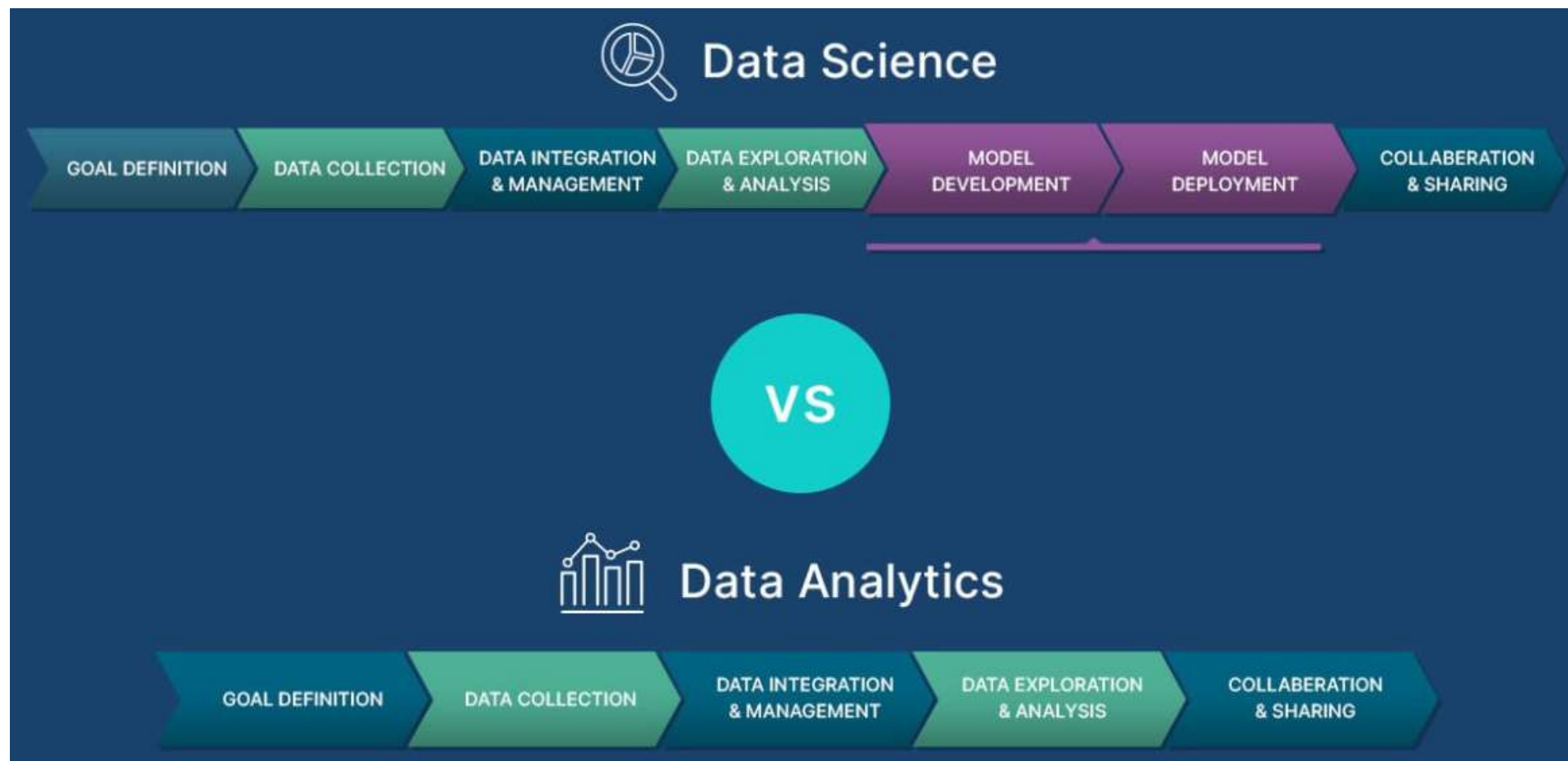


Program

Source: [Wickham et al. \(2023\)](#)

How about Data Analytics?

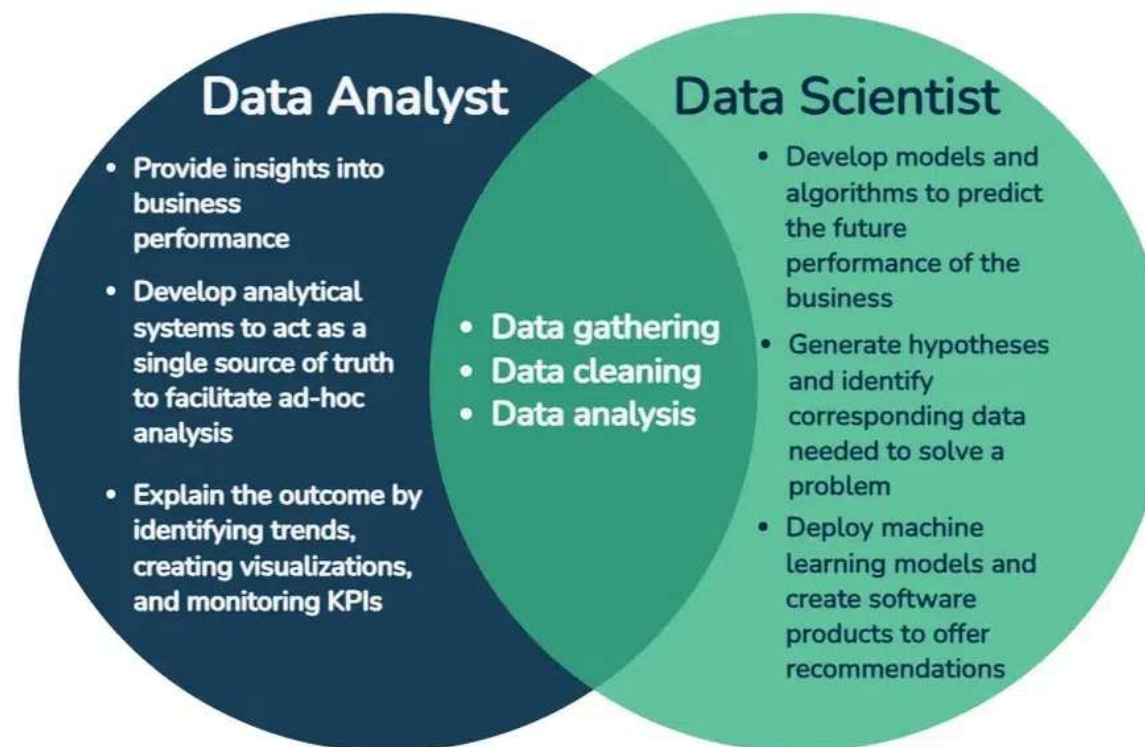
- The name of the course is **Data Science and Data Analytics**, but only data science has been mentioned so far. How about data analytics?
- Essentially, **data analytics** is data science without the strong computer science and machine learning focus. Therefore, the steps involved in data analytics are typically a subset of the steps in a data science project:



How about Data Analytics?

The two fields thus have a **large intersection**, which is also reflected in the tasks that data scientists and data analysts share:

Difference between a Data Analyst and a Data Scientist



As we are focused on tasks in the intersection, we will mostly be referring to **data science** as the overarching topic in the remainder of this course.