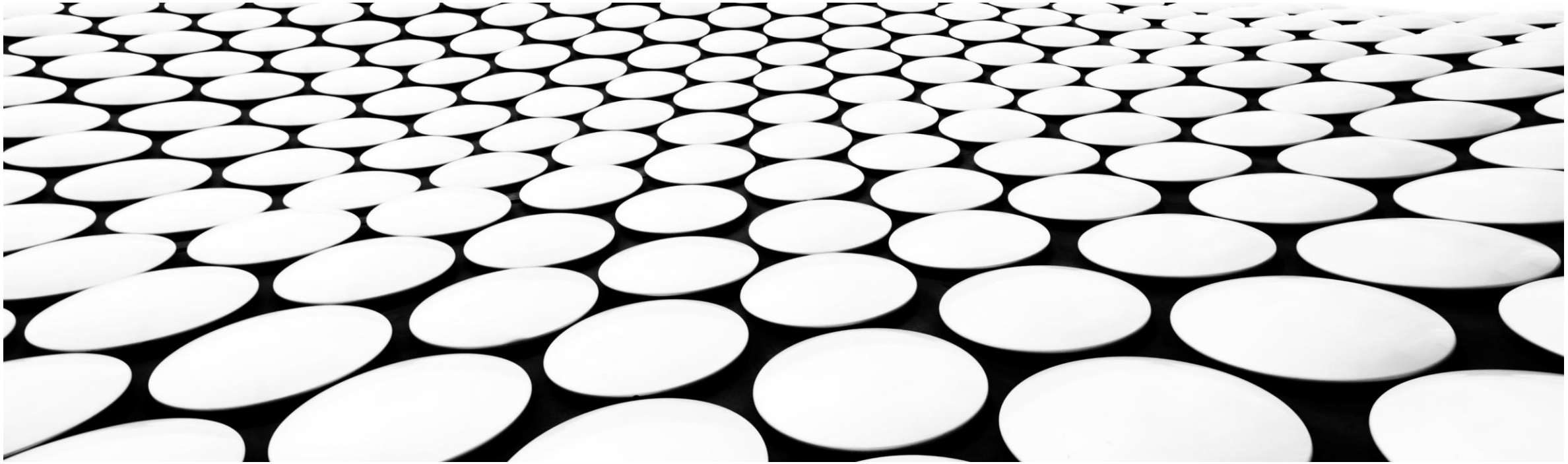
DATA COLLECTION

Exploration of diverse approaches



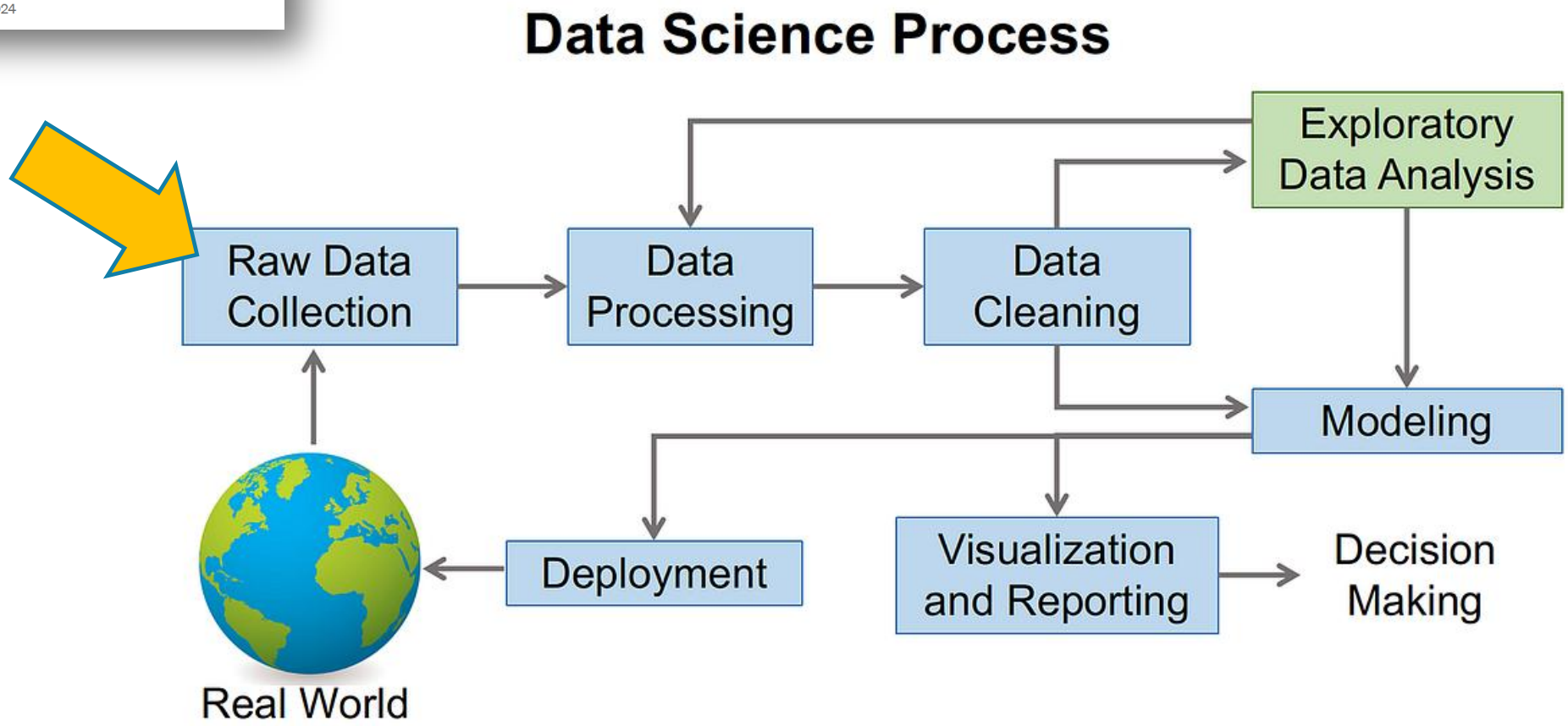
DATA COLLECTION

- Data Science Process
- Types of questions (insights)
- Overview of data collection methods

Data Science Process

Data Science Process: A Comprehensive Guide

Abhijit · Follow
6 min read · Jan 15, 2024



Source

7 Steps to a Successful Data Science Project

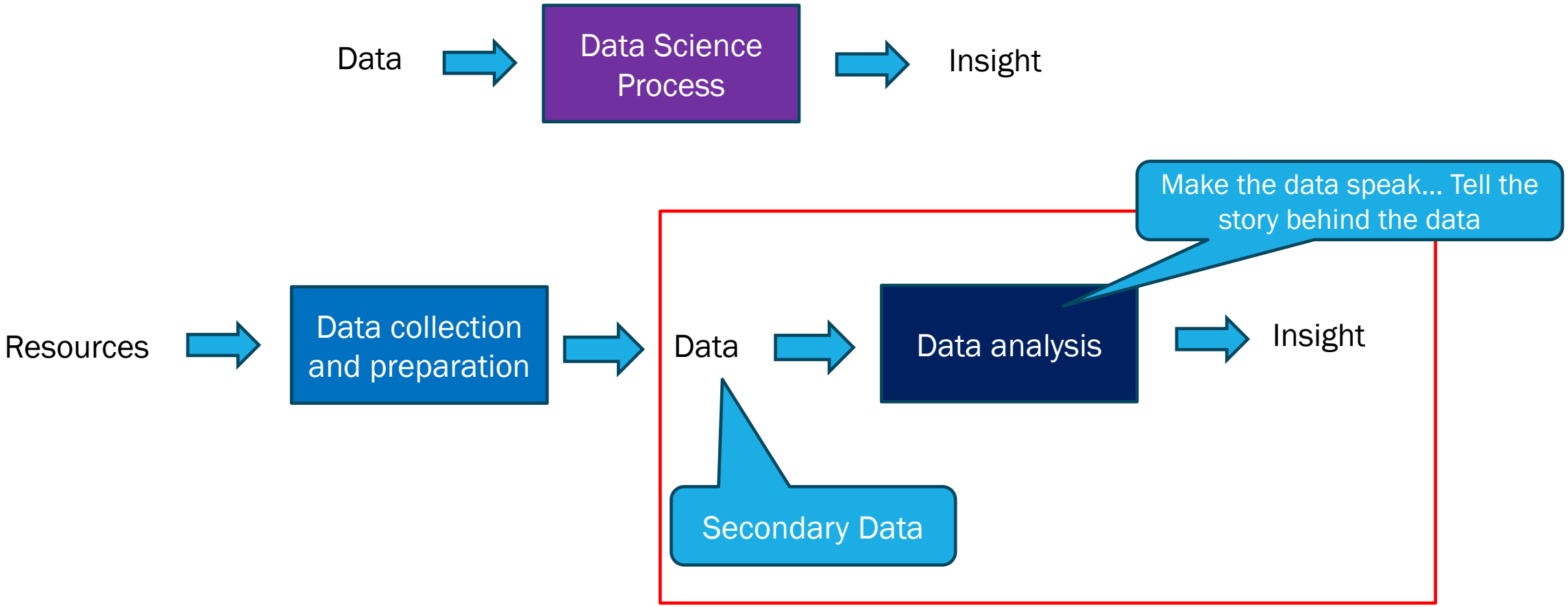
Beginners Guide on Completing a Data Science Project from Scratch

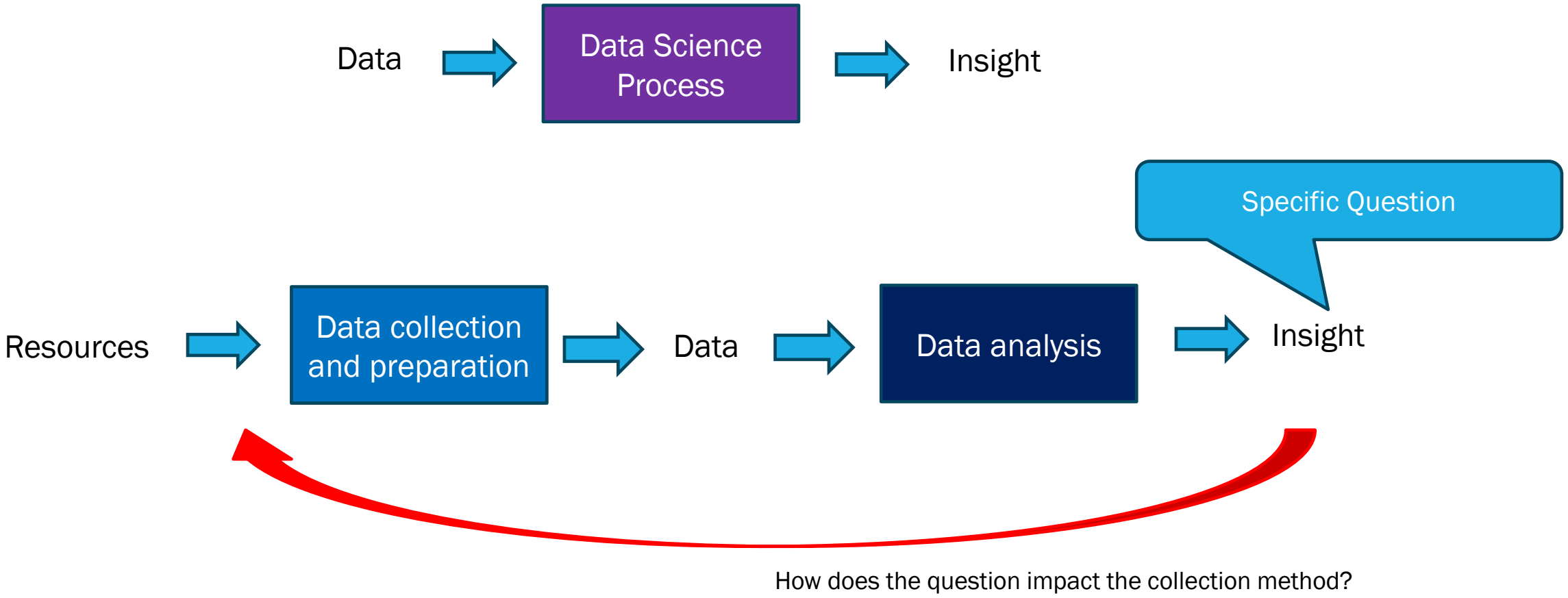
Amit Bharadwa · Follow
Published in Towards Data Science · 7 min read · Feb 6, 2021

The method is as follows:

1. **Problem Statement**
2. **Data Collection**
3. **Data Cleaning**
4. **Exploratory Data Analysis (EDA)**
5. **Feature Engineering**
6. **Modelling**
7. **Communication**

Source


Data
Data Science Process
Insight
Make the data speak… Tell the story behind the data
Resources
Data collection and preparation
Data
Data analysis
Insight
Secondary Data


Data
Data Science Process
Insight
Specific Question
Resources
Data collection and preparation
Data
Data analysis
Insight
How does the question impact the collection method?

Questions

Example 1 – Retail Domain

How much money are younger people spending on clothing compared to older people?

Example 2 – Fashion Domain

How do consumers feel about the ethical sourcing of products in the fashion industry?

Example 3 – Digital Marketing Domain

What is the expected bounce rate if the company adopts the suggested website redesign?

Example 4 – Environment Domain

How do air quality levels vary across different regions of Ottawa during peak traffic hours?

Example 5 – Urban Planning Domain

What are the most common complaints about public transportation services in Gatineau?

Example 6 – Public Health Domain

How do social media users perceive the new health policy announced by the government?

Example 7 – E-commerce Domain

What is the average number of hours a user spends on an e-commerce site per week?

Example 8 – Human Resources Domain

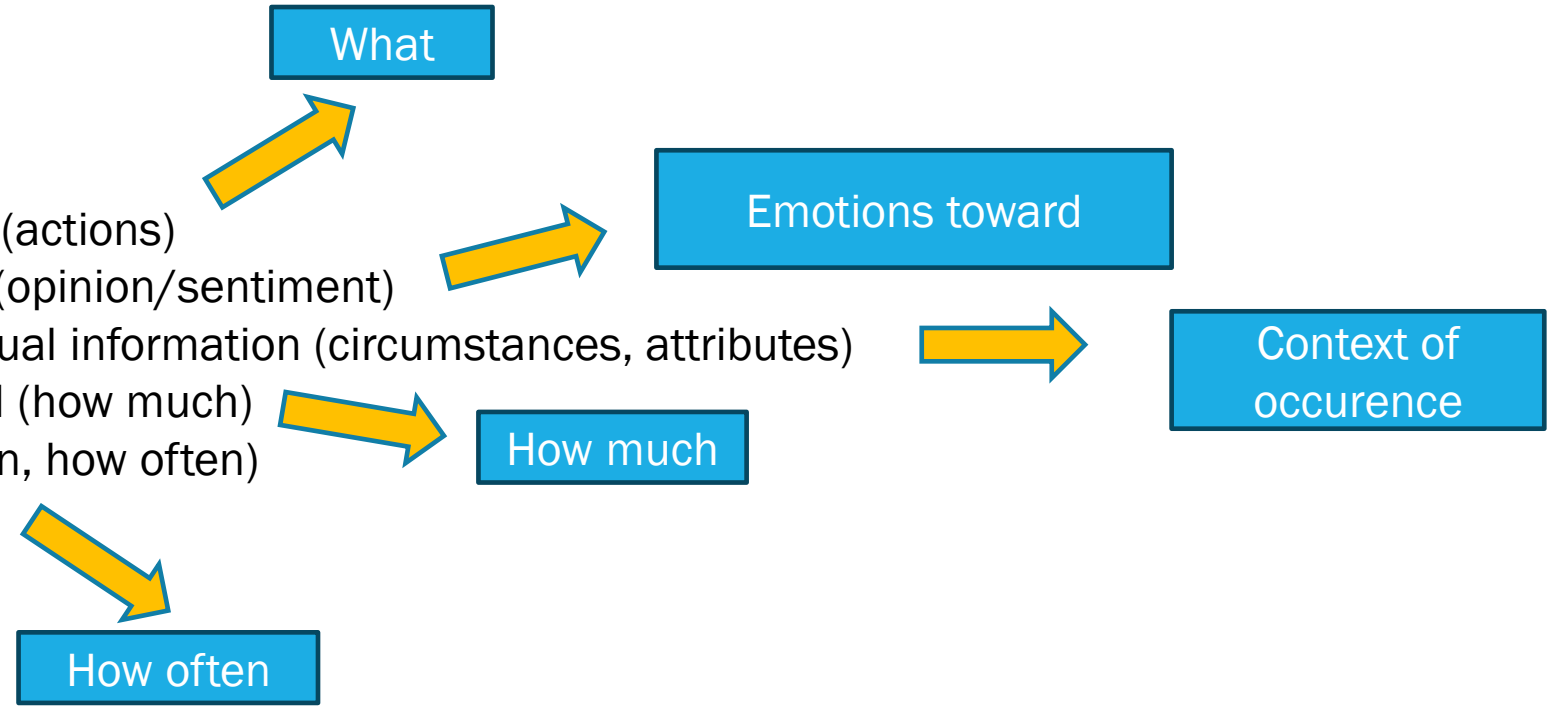
What are the primary motivations of employees for choosing remote work over on-site work in a company?

Example 9 – Healthcare Domain

What are the effects of a new drug treatment on the recovery rate of patients with a specific condition?

What are we trying to capture?

1. Behavioral information (actions)
2. Attitudinal information (opinion/sentiment)
3. Environmental/Contextual information (circumstances, attributes)
4. Financial/Transactional (how much)
5. Frequency/Usage (when, how often)



Example 1 – Retail Domain

How much money are younger people spending on clothing compared to older people?

transactional

context

Example 2 – Fashion Domain

How do consumers feel about the ethical sourcing of products in the fashion industry?

opinion

context

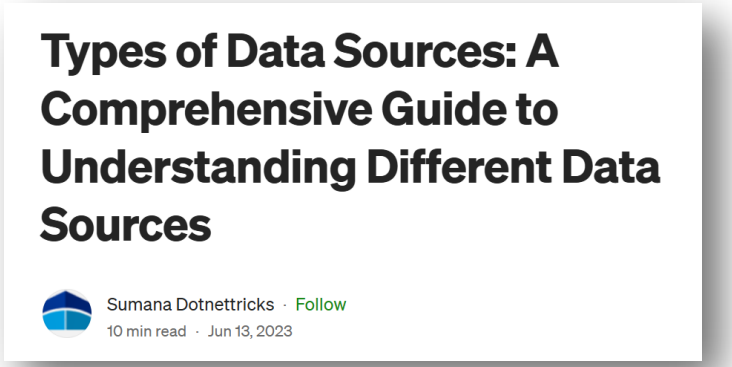
Example 3 – Digital Marketing Domain

What is the expected bounce rate if the company adopts the suggested website redesign?

behavior

context

Collection Methods



Source

Really nice resource for listing types of data sources. I'll just organize them a bit differently

Types of Data Collection Methods

1. Self-Reported Data
2. Behavioral and Observational Data
3. Automated Data
4. Digital and Platform-based Data

1. Self-Reported Data
2. Behavioral and Observational Data
3. Automated Data
4. Digital and Platform-based Data

Self-Reported Data	Surveys Interviews Focus groups
Data directly provided by individuals	

Surveys

Purpose:
To systematically collect quantifiable data from a sample of individuals to understand their opinions, preferences and satisfaction levels.

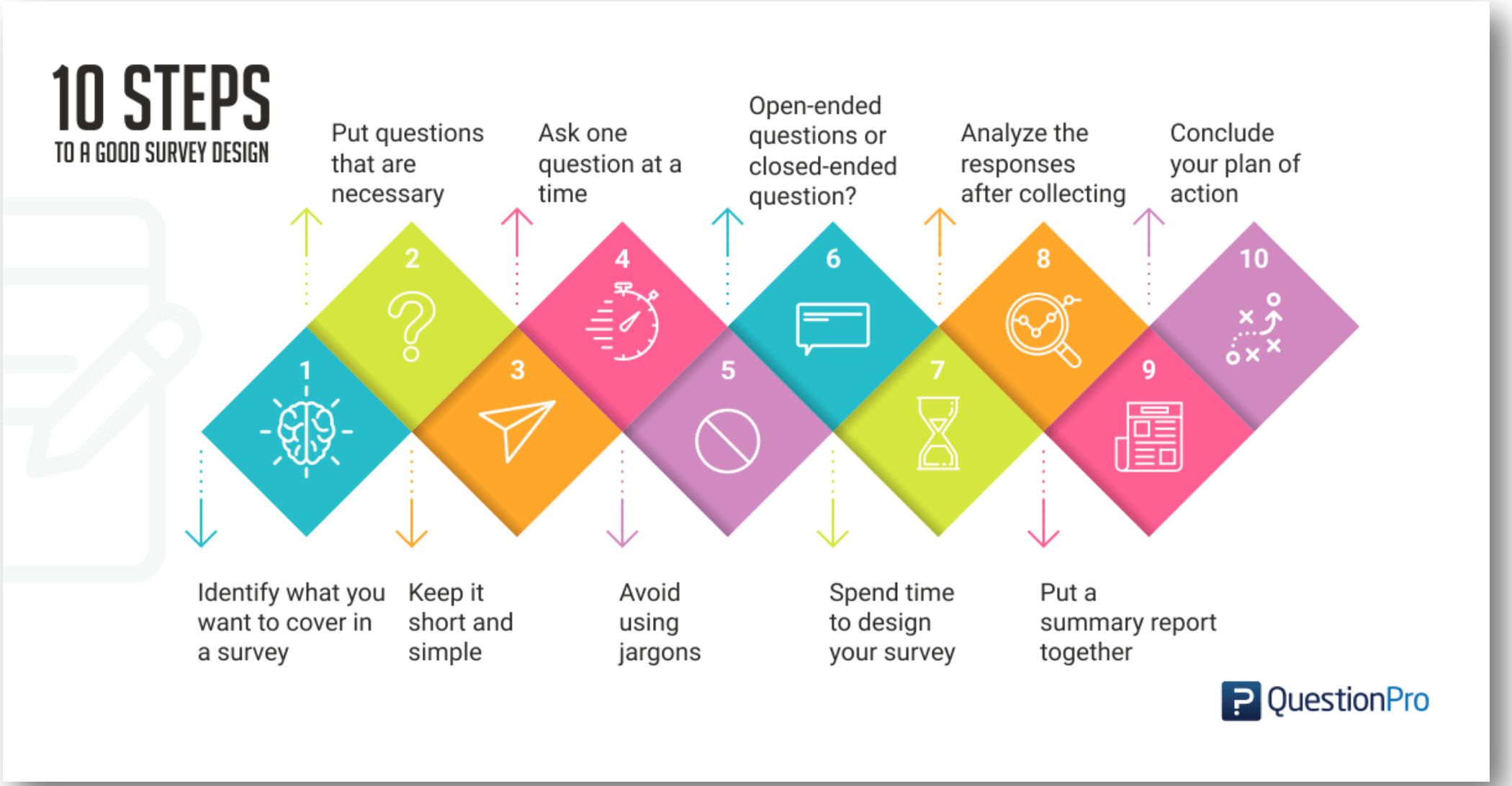
Example:
A tourism board wants to understand the preferences of international tourists visiting their country.

Statistics Canada performs many surveys

Source


70%
of consumers say they have changed their food purchasing habits in the last year because of increasing food prices
COST OF FOOD
of these
70%
66%
56%
50%
Purchasing more discounted or on sale food items
Purchasing less expensive brands
Shopping more often at low-cost retailers
Cutting back on the amount of food purchased

Source


10 STEPS
TO A GOOD SURVEY DESIGN
1
Identify what you want to cover in a survey
2
Put questions that are necessary
3
Keep it short and simple
4
Ask one question at a time
5
Avoid using jargons
6
Open-ended questions or closed-ended question?
7
Spend time to design your survey
8
Analyze the responses after collecting
9
Put a summary report together
10
Conclude your plan of action
QuestionPro

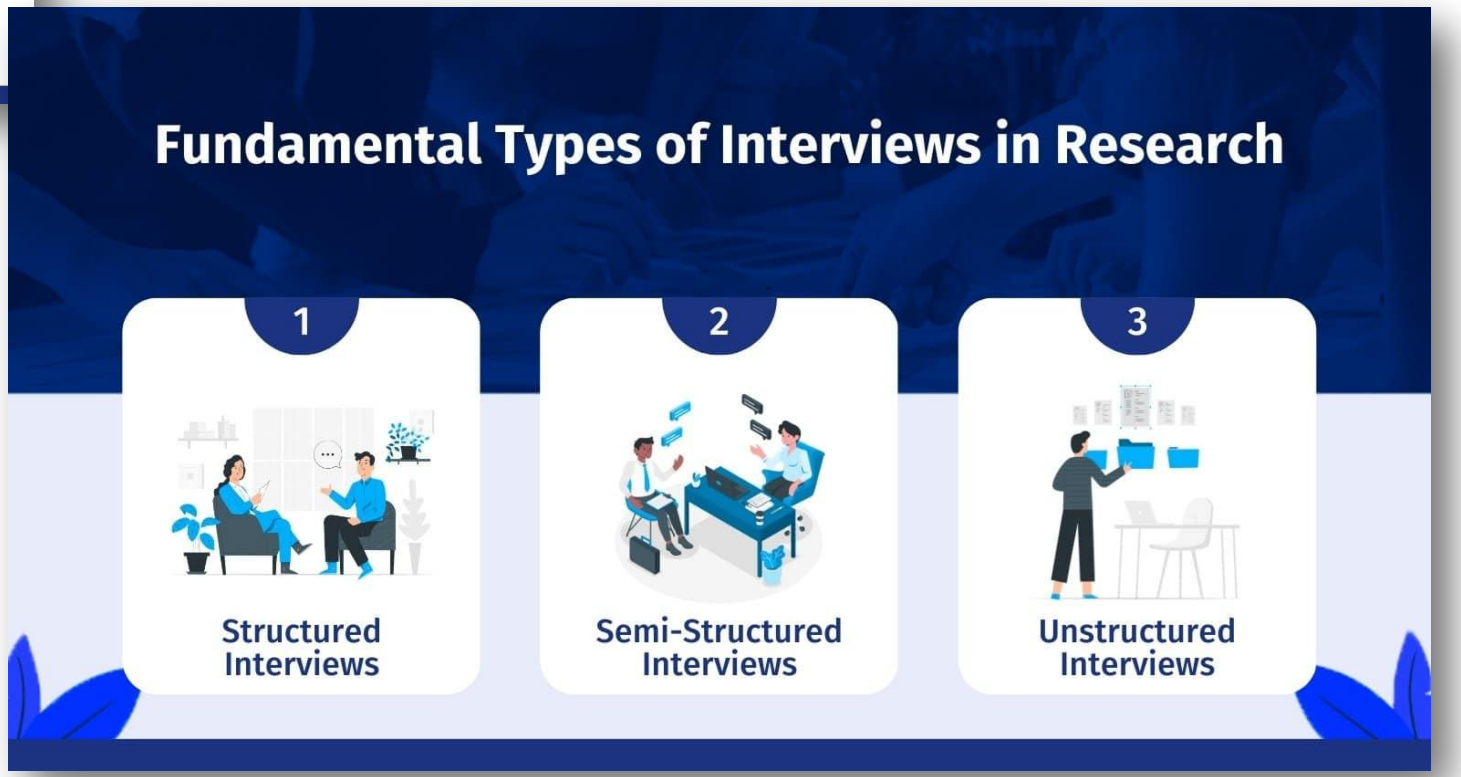
Source

Interviews

Purpose:
To gather detailed, qualitative insights by exploring an individual's experiences, perspectives, or knowledge through one-on-one conversation.

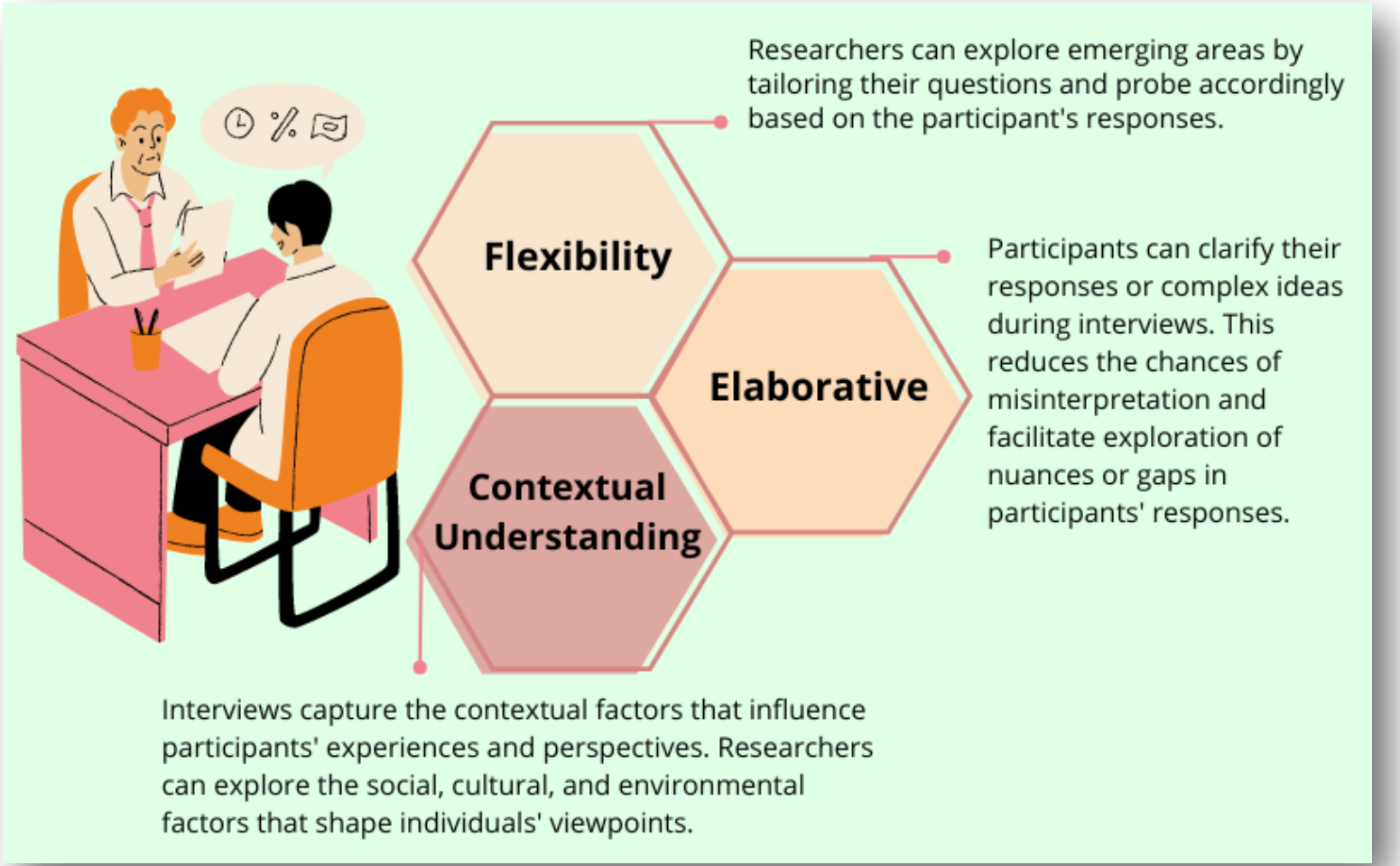
Example:
Conducting interviews with patients to understand their personal experiences with a new medical treatment.


TYPES OF
INTERVIEWS
QuestionPro


Fundamental Types of Interviews in Research
1
Structured Interviews
2
Semi-Structured Interviews
3
Unstructured Interviews

Source


2
Semi-Structured Interviews

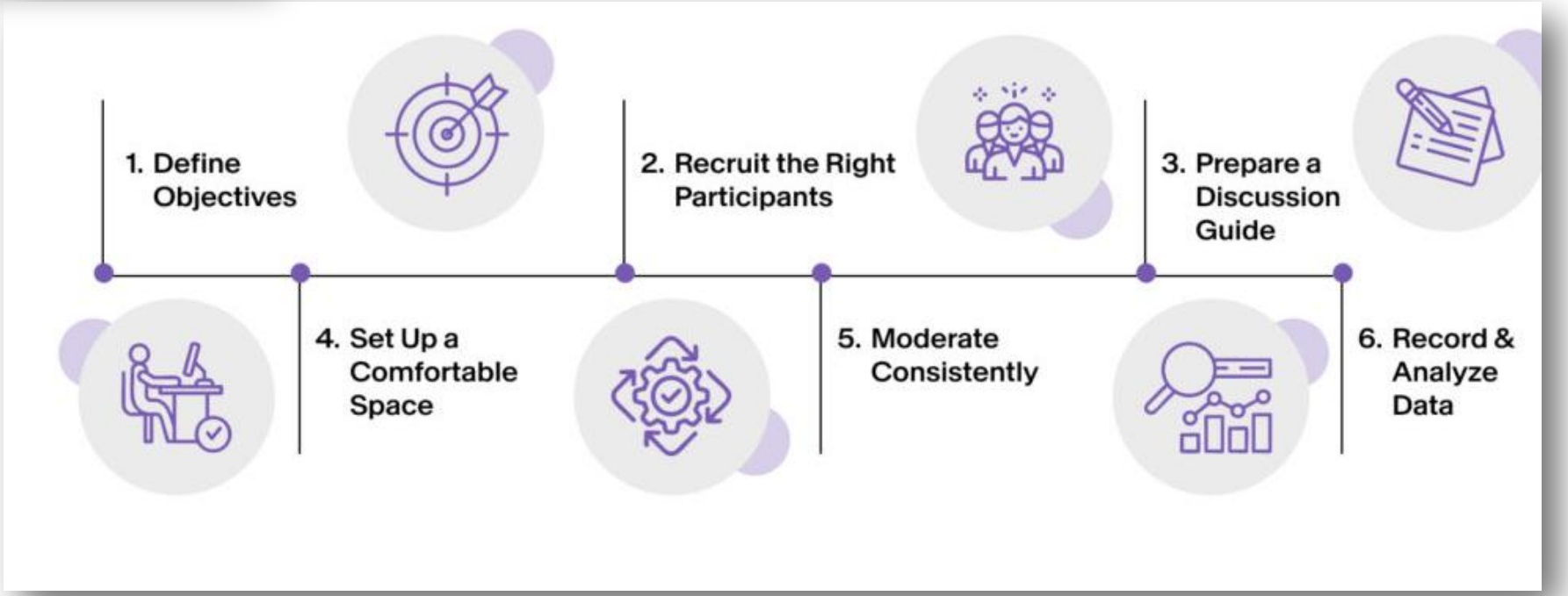

Researchers can explore emerging areas by tailoring their questions and probe accordingly based on the participant's responses.
Flexibility
Elaborative
Contextual Understanding
Participants can clarify their responses or complex ideas during interviews. This reduces the chances of misinterpretation and facilitate exploration of nuances or gaps in participants' responses.
Interviews capture the contextual factors that influence participants' experiences and perspectives. Researchers can explore the social, cultural, and environmental factors that shape individuals' viewpoints.

Source

Focus Groups

Purpose:
To gather diverse perspectives, ideas, or feedback through guided discussions among a small group of participants on a specific topic.

Example:
Organizing a focus group of parents to explore opinions on proposed changes to school nutrition programs.


1. Define Objectives
2. Recruit the Right Participants
3. Prepare a Discussion Guide
4. Set Up a Comfortable Space
5. Moderate Consistently
6. Record & Analyze Data

Source

1. Self-Reported Data
2. Behavioral and Observational Data
3. Automated Data
4. Digital and Platform-based Data

Behavioral and Observational Data	Observations Experiments
Data collected through monitoring or observing human behavior.	

Observations

Purpose:
To collect data by systematically watching and recording behaviors, actions, or events in their natural or controlled settings.

Examples:
Observing customers in a retail store to analyze shopping behaviors and product preferences.

Observing traffic flow at an intersection to study the effects of a newly implemented traffic light system.


OBSERVATIONAL
RESEARCH
QuestionPro


Observational Research Methods
Have a Clear Objective
Get Permission
OK
Unbiased Observation
Hide Your Observers
Documentation
Data Analysis
QuestionPro

Source

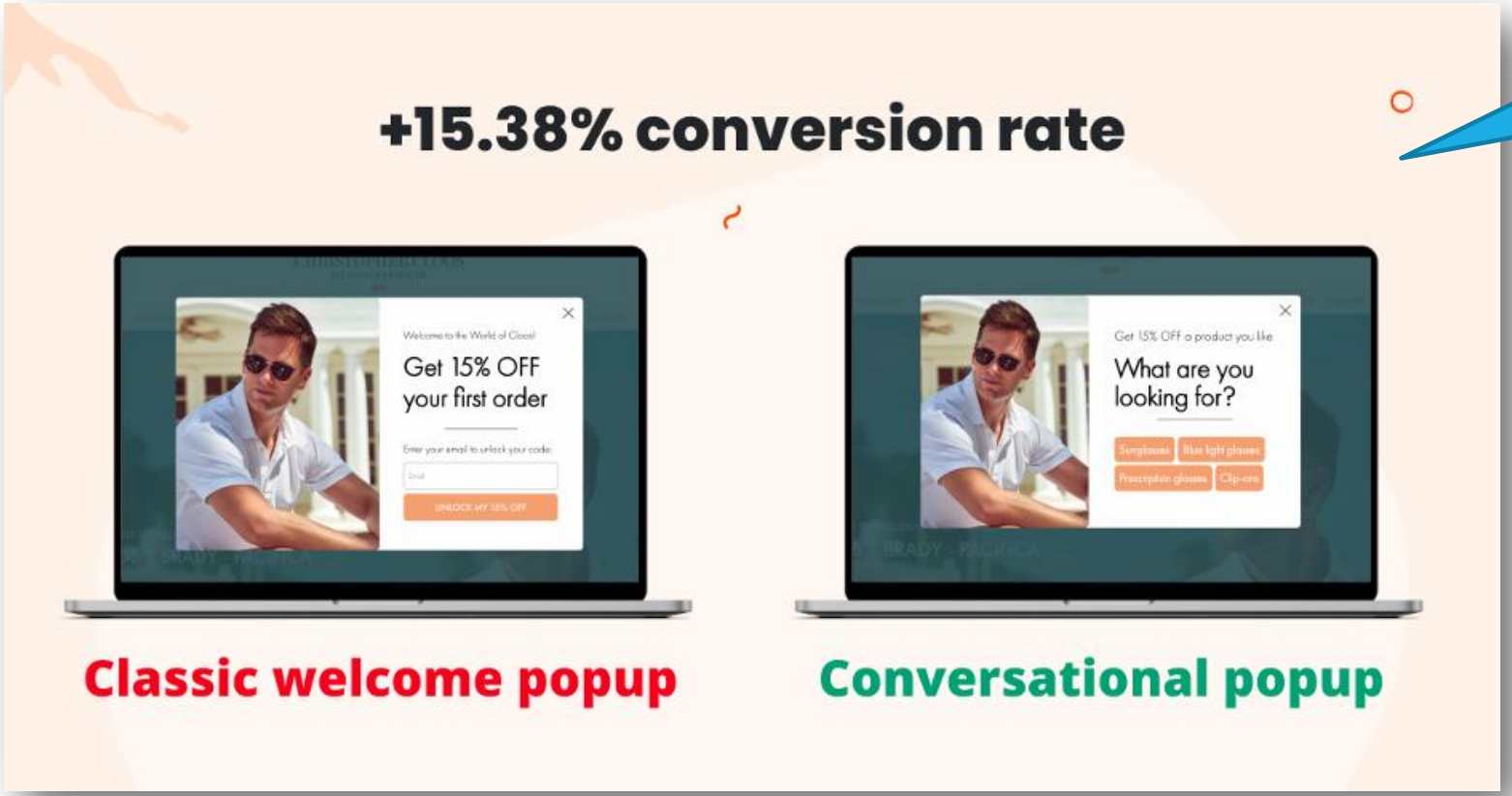
Experiments

Purpose:
Data collected by manipulating variables in a controlled setup (e.g., clinical trials, A/B testing).

Examples:
A travel agency wants to test whether offering a limited-time discount on adventure packages increases bookings.

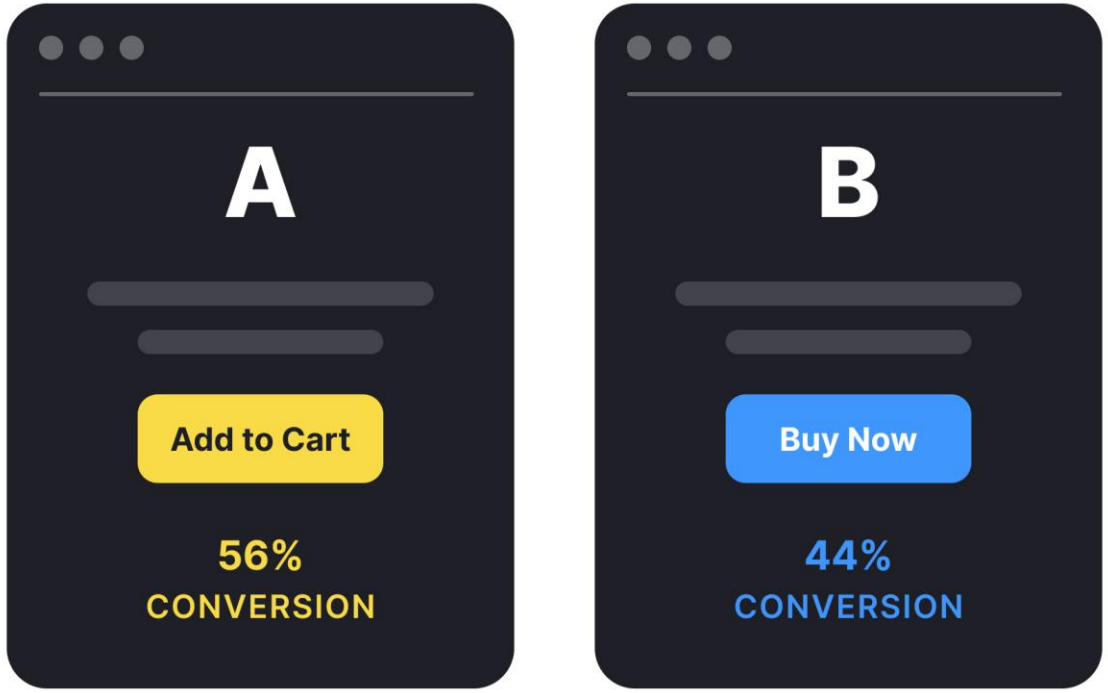
Conducting an experiment to test the effectiveness of a new drug by comparing patient recovery rates between a treatment group and a control group.

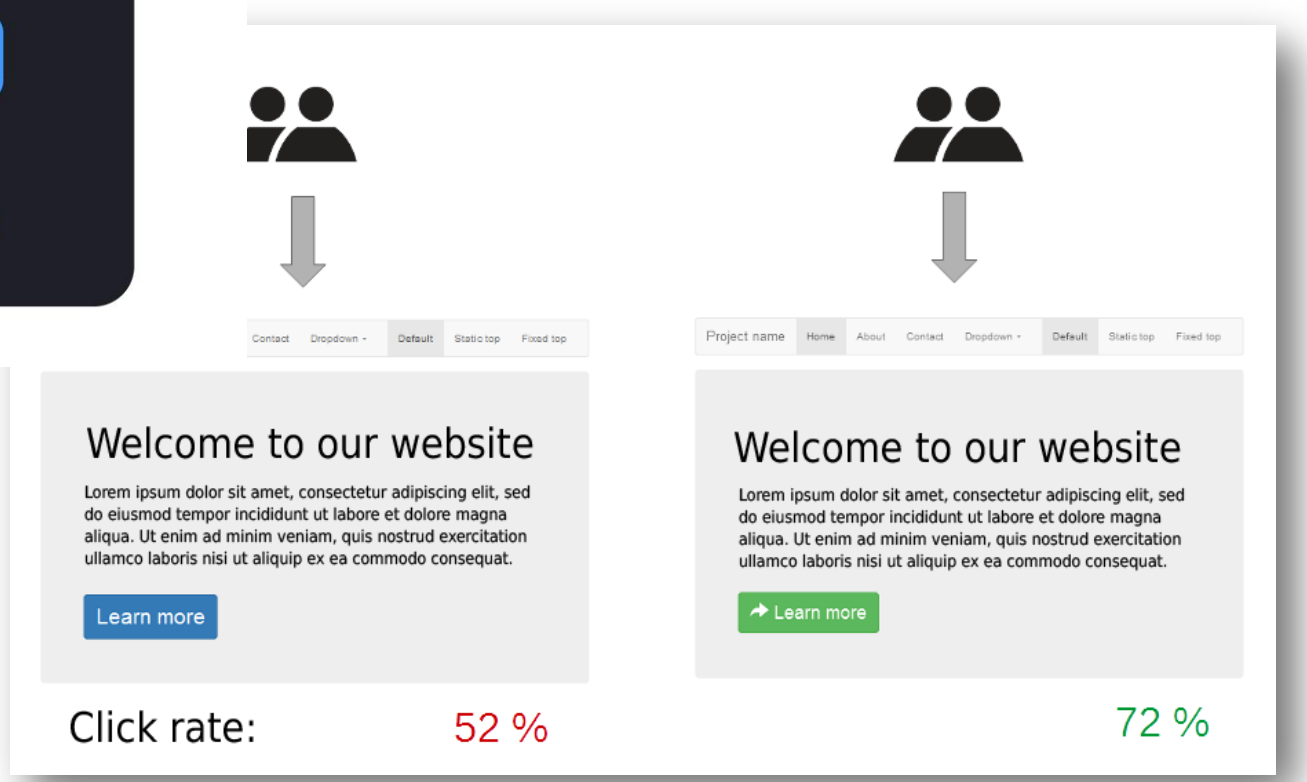
A conversion rate records the percentage of users who have completed a desired action. Conversion rates are calculated by taking the total number of users who 'convert' (for example, by clicking on an advertisement), dividing it by the overall size of the audience and converting that figure into a percentage.



A/B testing

Source


A
Add to Cart
56%
CONVERSION
B
Buy Now
44%
CONVERSION


Contact
Dropdown
Default
Static top
Fixed top
Welcome to our website
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.
Learn more
Project name
Home
About
Contact
Dropdown
Default
Static top
Fixed top
Welcome to our website
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.
Learn more
Click rate:
52 %
72 %

1. Self-Reported Data
2. Behavioral and Observational Data
3. Automated Data
4. Digital and Platform-based Data

Automated Data	Logs Sensor Transactions
Data generated obtained through an automated system.	

Logs

Purpose:
Analyse automatically generated records of system or user activities (e.g., app usage, website clicks). Log files capture information about system errors, user interactions, and more.

Example:
Analyzing website log files to identify popular pages and navigation patterns.

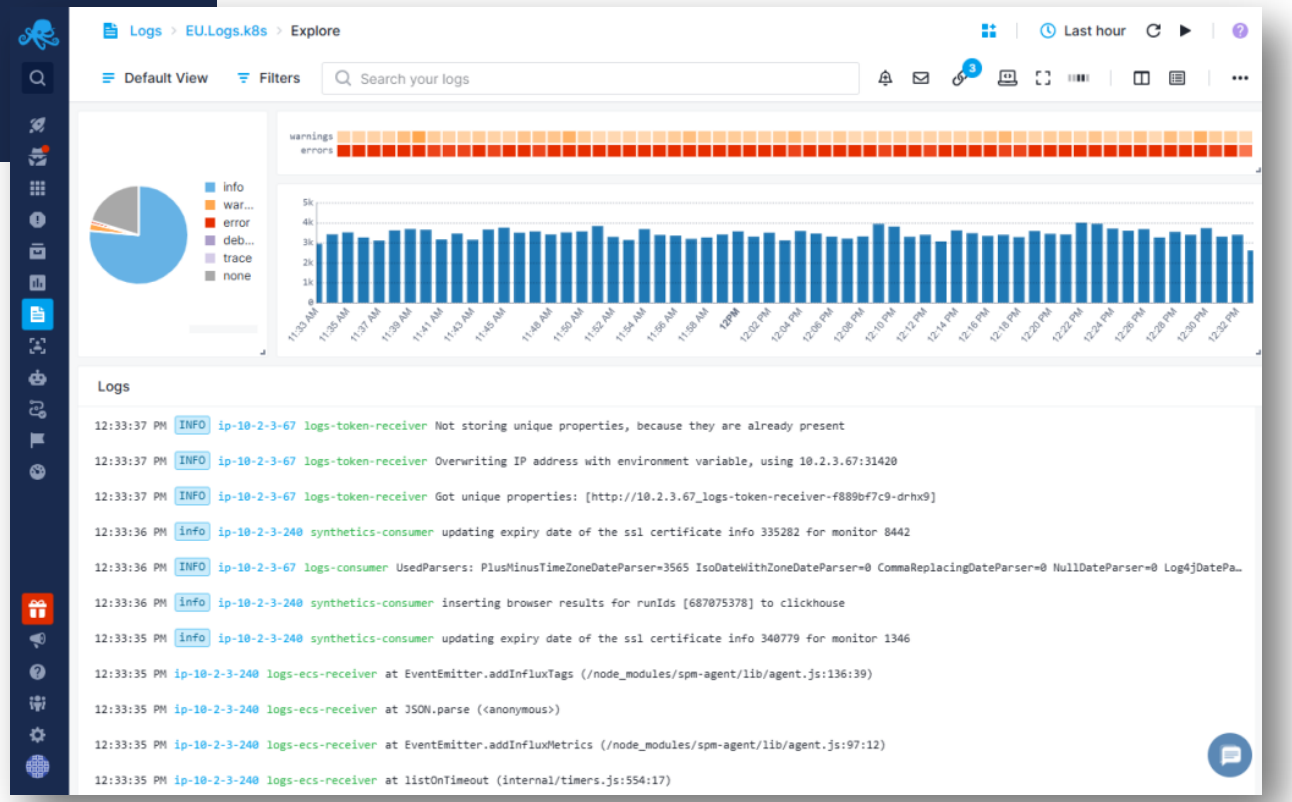
13 Best Log Analysis Tools of 2025. Top Paid, Free & Open-Source Log Analyzers Reviewed

TOOLS & COMPARISONS

Updated on: September 29, 2025

Best Log Analyzers

1. Sematext Logs
2. Elastic Stack
3. Graylog
5. Loggly
5. Splunk
6. Logz.io
7. Sumo Logic
8. SolarWinds
9. ManageEngine
10. Papertrail
11. Mezmo
12. Datadog
13. LogicMonitor



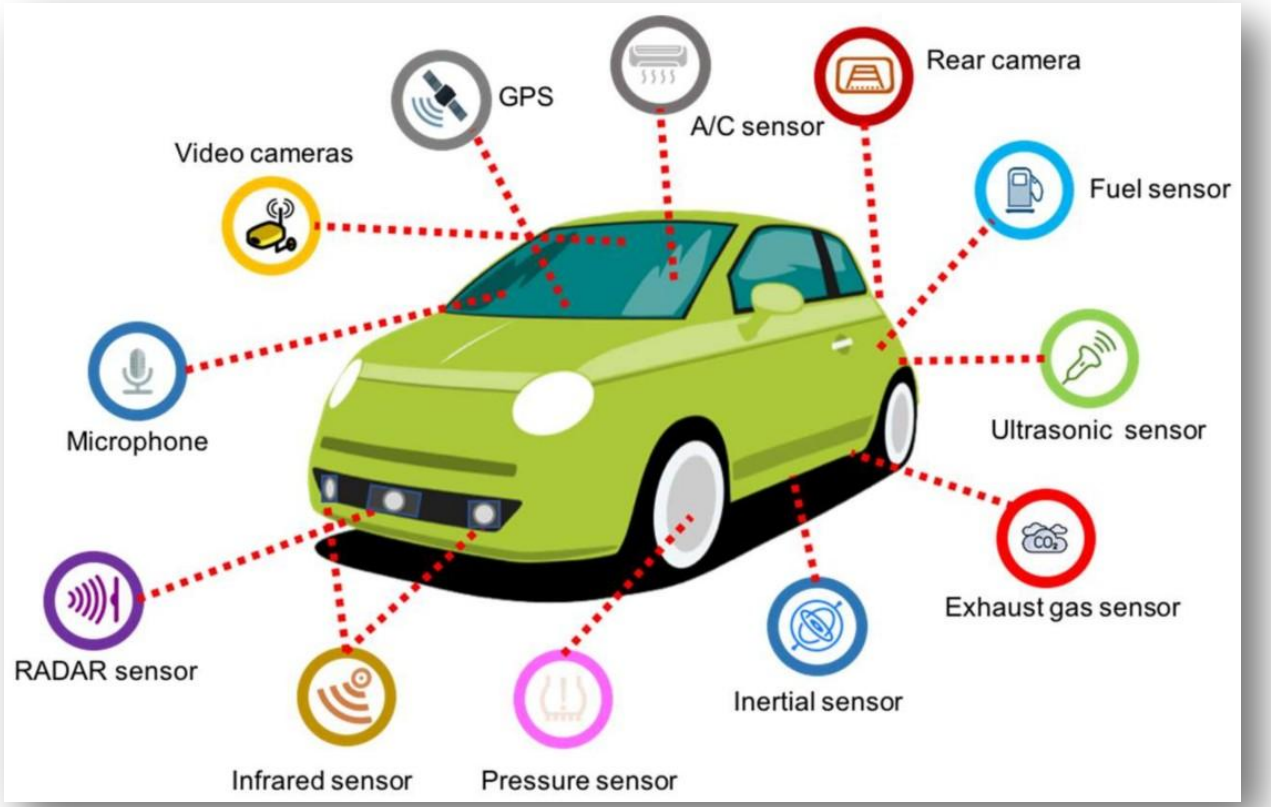
Source

Sensors

Purpose:
Analyse measurements coming from devices monitoring physical or environmental conditions (e.g., motion, temperature, air quality). Sensors are ubiquitous in today's world, collecting data on various physical parameters such as temperature, pressure, motion, and more.

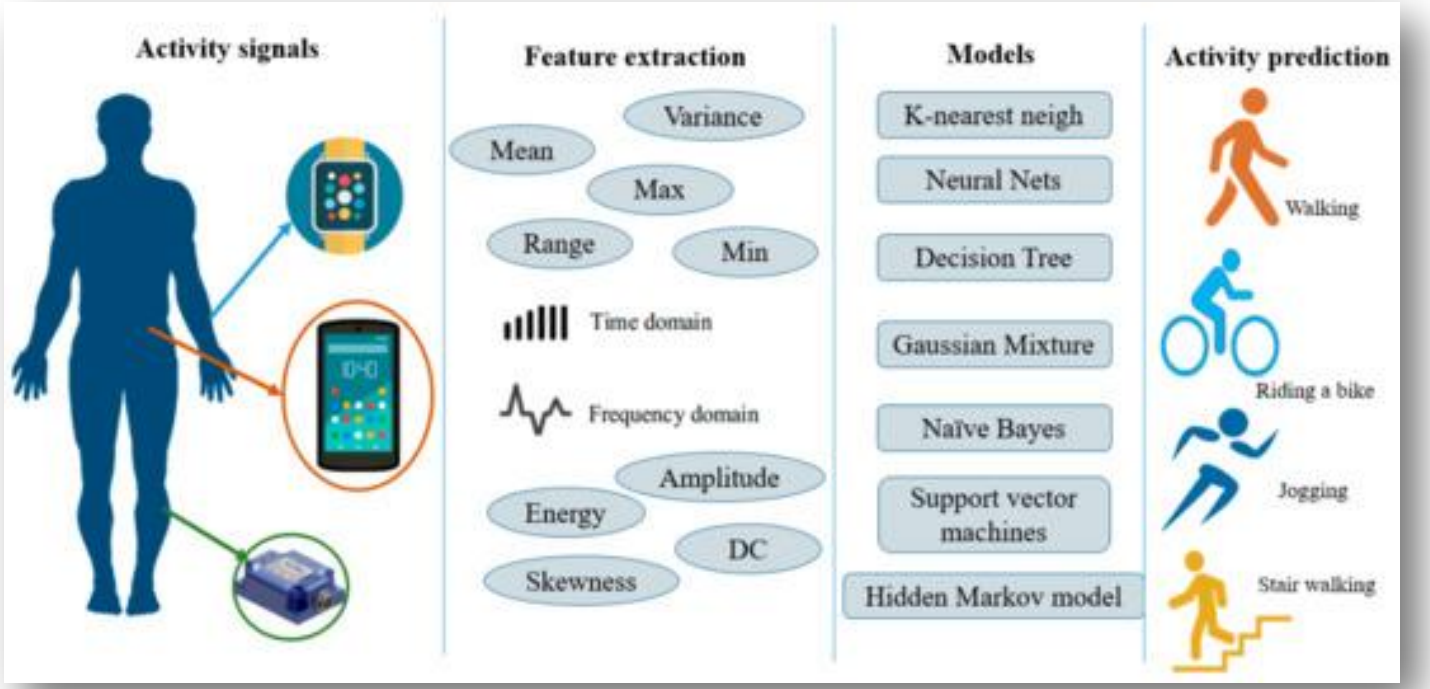
Example:
Using air quality sensors to monitor pollution levels in urban areas over time.

Sensor Data Analytics - Automobile



Source

New Sensor Data Structuring for Deeper Feature Extraction in Human Activity Recognition †



Source

ScienceSoft
Data Services

Industry-Specific Uses of Sensor Data Analytics

Manufacturing

- Monitoring equipment & machinery (e.g., temperature, vibration, lubrication, pressure), environmental conditions, resource utilization, and employees' activity.

Agriculture

- Monitoring weather conditions, crop growth and health.
- Monitoring the chemical composition (acidity, moisture, nutrient levels) of soil and agricultural waste.
- Real-time livestock monitoring (e.g., location, body temperature).
- Monitoring and automatically adjusting greenhouse conditions based on pre-set parameters and real-time sensor data.

Source

ScienceSoft
Data Services

Industry-Specific Uses of Sensor Data Analytics

Smart cities

- Monitoring air quality, energy consumption, driving patterns, vehicle and foot traffic flow, etc.
- Enabling predictive city infrastructure maintenance.
- Forecasting energy and water consumption.
- Root cause analysis and intelligent suggestions to improve traffic management, resource consumption, waste management, and public health.

Healthcare

- Remote patient monitoring and IoT for medical devices.
- Identifying trends in patient symptoms to determine the factors influencing patient health and help improve care outcomes.

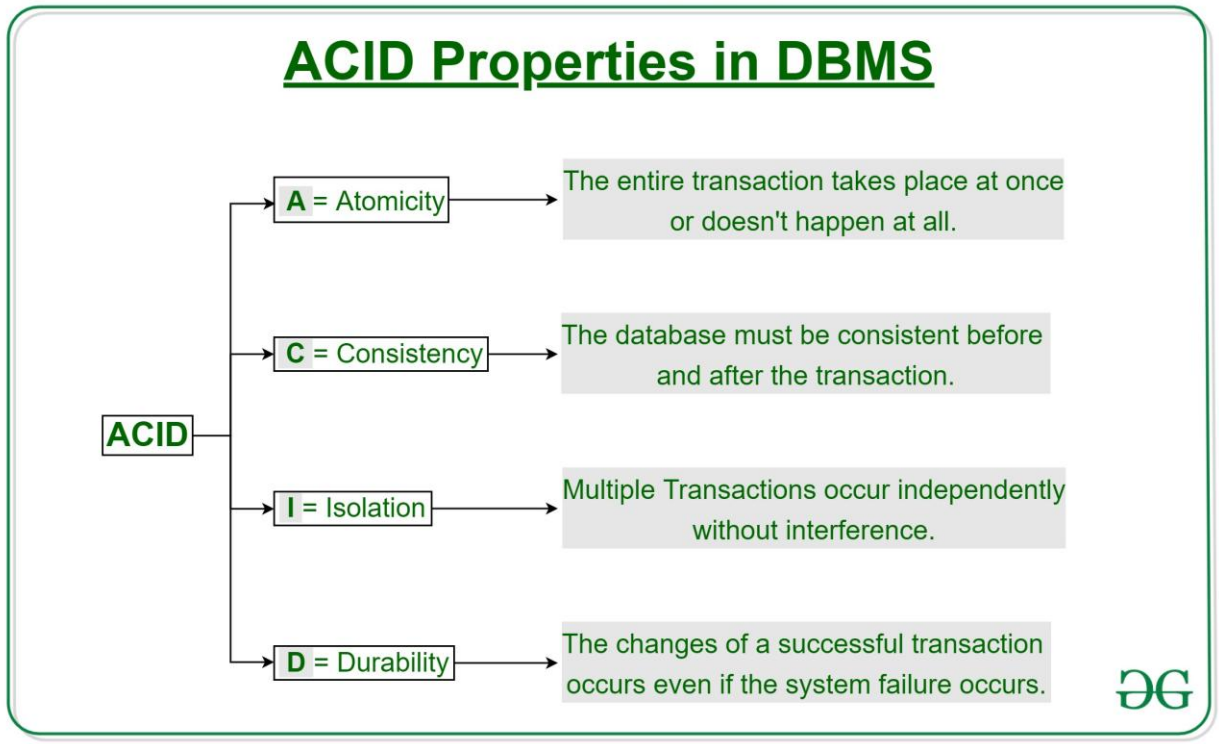
Source

Transactions

Purpose:
To collect detailed records of transactions, such as purchases or exchanges, to analyze patterns, behaviors, and financial activity. Transactional data typically captures the details of exchanges or interactions between entities (e.g., customers, organizations, or systems) and provides a record of a specific event or action.

Examples:
Review purchase records of renewable energy equipment from manufacturers, distributors, or financial institutions providing green loans.

	A	B	C	D	E	F
1	Date	Transaction ID	Customer Name	Category	Amount	Payment Method
2	2023-01-01 00:00:00	T00001	John Doe	Forex	54697.34	Check
3	2023-01-01 11:40:32	T00002	Alice Brown	Equity	-6230.63	Cash
4	2023-01-01 23:21:05	T00003	Jane Smith	Derivative	99837.92	Wire Transfer
5	2023-01-02 11:01:37	T00004	Jane Smith	Commodity	4473.36	Credit Card
6	2023-01-02 22:42:10	T00005	Jane Smith	Fixed Income	71438.19	Cash
7	2023-01-03 10:22:42	T00006	Bob Johnson	Forex	80311.67	Credit Card
8	2023-01-03 22:03:15	T00007	Jane Smith	Forex	31036	Check
9	2023-01-04 09:43:47	T00008	Charlie Green	Forex	11653.73	Wire Transfer
10	2023-01-04 21:24:19	T00009	John Doe	Derivative	863.59	Wire Transfer
11	2023-01-05 09:04:52	T00010	Alice Brown	Commodity	72346.66	Credit Card
12	2023-01-05 20:45:24	T00011	Alice Brown	Forex	39791.89	Credit Card
13	2023-01-06 08:25:57	T00012	John Doe	Commodity	68508.95	Wire Transfer
14	2023-01-06 20:06:29	T00013	Charlie Green	Fixed Income	90694.84	Credit Card
15	2023-01-07 07:47:02	T00014	Bob Johnson	Equity	6124.21	Wire Transfer
16	2023-01-07 19:27:34	T00015	Alice Brown	Equity	91108.81	Online Banking
17	2023-01-08 07:08:06	T00016	Bob Johnson	Fixed Income	35278.91	Check
18	2023-01-08 18:48:39	T00017	Bob Johnson	Equity	23579.37	Wire Transfer
19	2023-01-09 06:29:11	T00018	Alice Brown	Forex	93736.85	Check
20	2023-01-09 18:09:44	T00019	Bob Johnson	Derivative	98971.69	Check
21	2023-01-10 05:50:16	T00020	John Doe	Equity	11878.14	Online Banking
22	2023-01-10 17:30:49	T00021	Jane Smith	Fixed Income	62252.22	Wire Transfer
23	2023-01-11 05:11:21	T00022	Alice Brown	Derivative	1714.48	Credit Card
24	2023-01-11 16:51:54	T00023	Jane Smith	Fixed Income	61600.54	Credit Card
25	2023-01-12 04:32:26	T00024	Alice Brown	Forex	81004.46	Wire Transfer
26	2023-01-12 16:12:58	T00025	Charlie Green	Fixed Income	65294.84	Wire Transfer
27	2023-01-13 03:53:31	T00026	Jane Smith	Forex	35906.65	Check
28	2023-01-13 15:34:03	T00027	Alice Brown	Equity	32137.3	Check
29	2023-01-14 03:14:36	T00028	Bob Johnson	Derivative	33243.47	Cash
30	2023-01-14 14:55:08	T00029	Jane Smith	Commodity	54868.3	Check



1. Self-Reported Data
2. Behavioral and Observational Data
3. Automated Data
4. Digital and Platform-based Data

Digital and Platform-Based Data	Social Media Data Web Scraping APIs
Data originating from digital platforms, systems, or online activity.	

Social Media Data

Purpose:
To collect data from social media platforms to analyze user sentiment, trends, or interactions.

Examples:
Analyzing Twitter data to understand public sentiment regarding a political candidate during an election campaign.

Analyze social media post to understand people awareness and sentiment about climate change are increasing in response to global weather events.

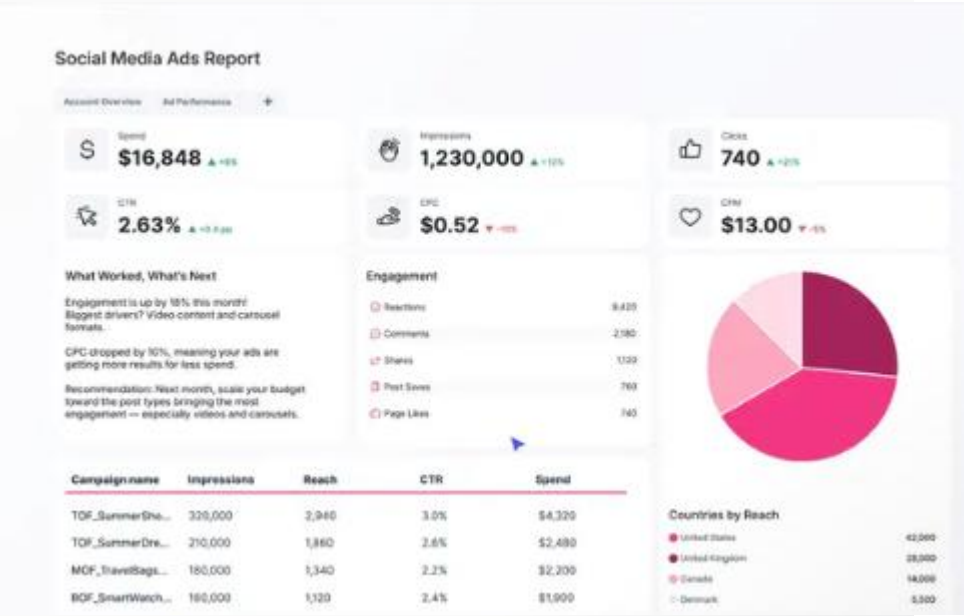
Each element needs to be analyzed differently

Source

12 Best Social Media Analytics Tools in 2026

Here's a summary of the top social media analytics tools we'll be reviewing:

1. Whatagraph
2. Metrics Watch
3. Keyhole
4. Google Analytics 4 (GA4)
5. Sprout Social
6. Buffer
7. Hootsuite
8. Brandwatch
9. Rival IQ
10. Zoho Social
11. Mentionlytics
12. BuzzSumo

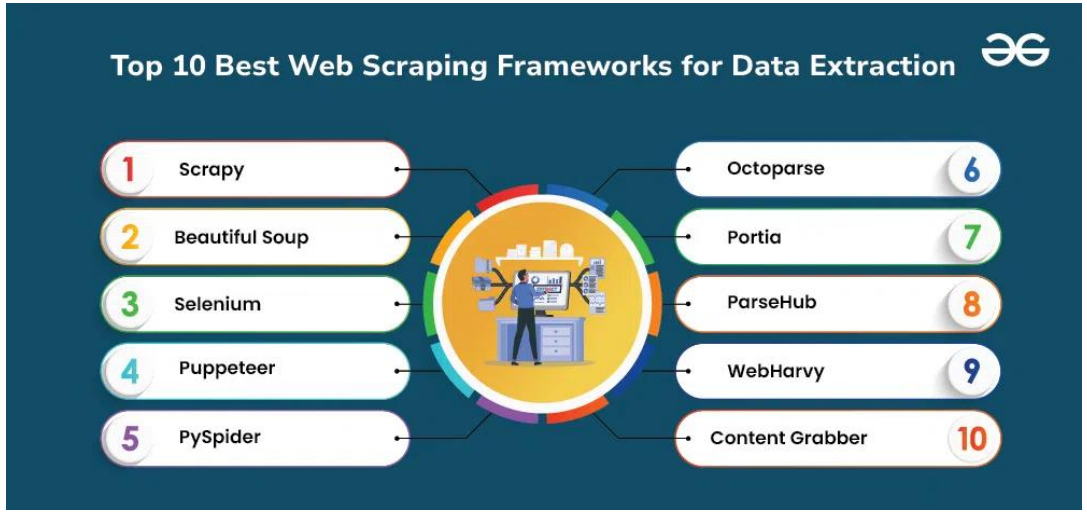


Web Scraping

Purpose:
To extract data from websites automatically, often for collecting large amounts of structured information from online sources. Web scraping is particularly useful for gathering information from websites that do not offer structured APIs or downloadable datasets.

Example:
Using web scraping to gather product pricing data from multiple e-commerce websites for competitive analysis.

Source

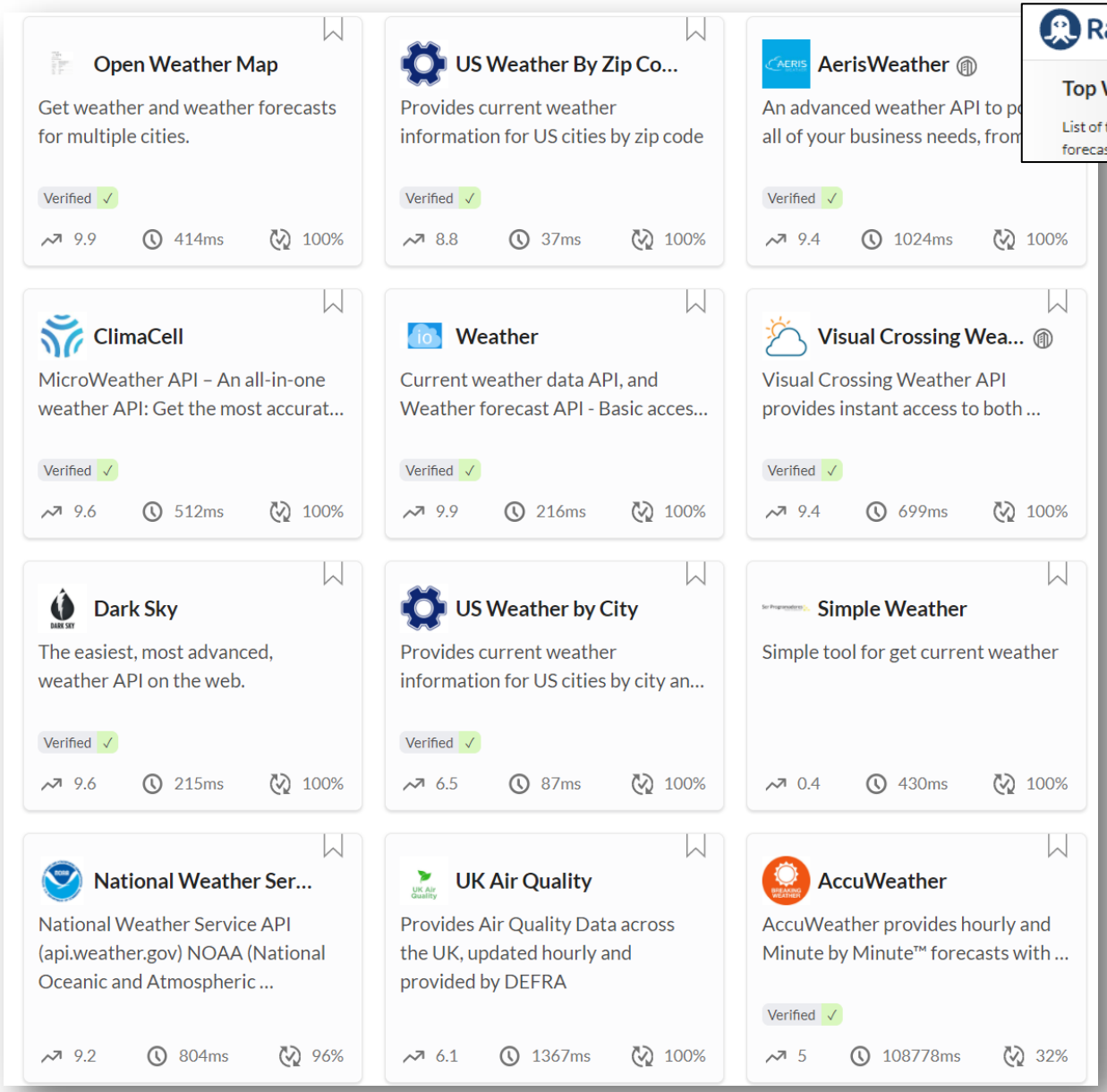


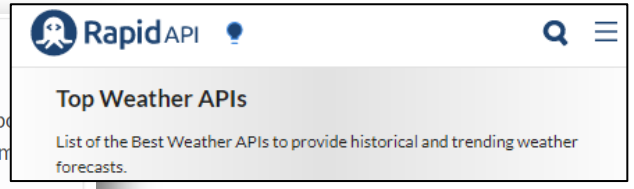
Source

APIs

Purpose:
To collect data by accessing and interacting with external platforms or services through predefined interfaces for specific data requests. APIs and web services enable developers to access and interact with data from various online platforms and databases.

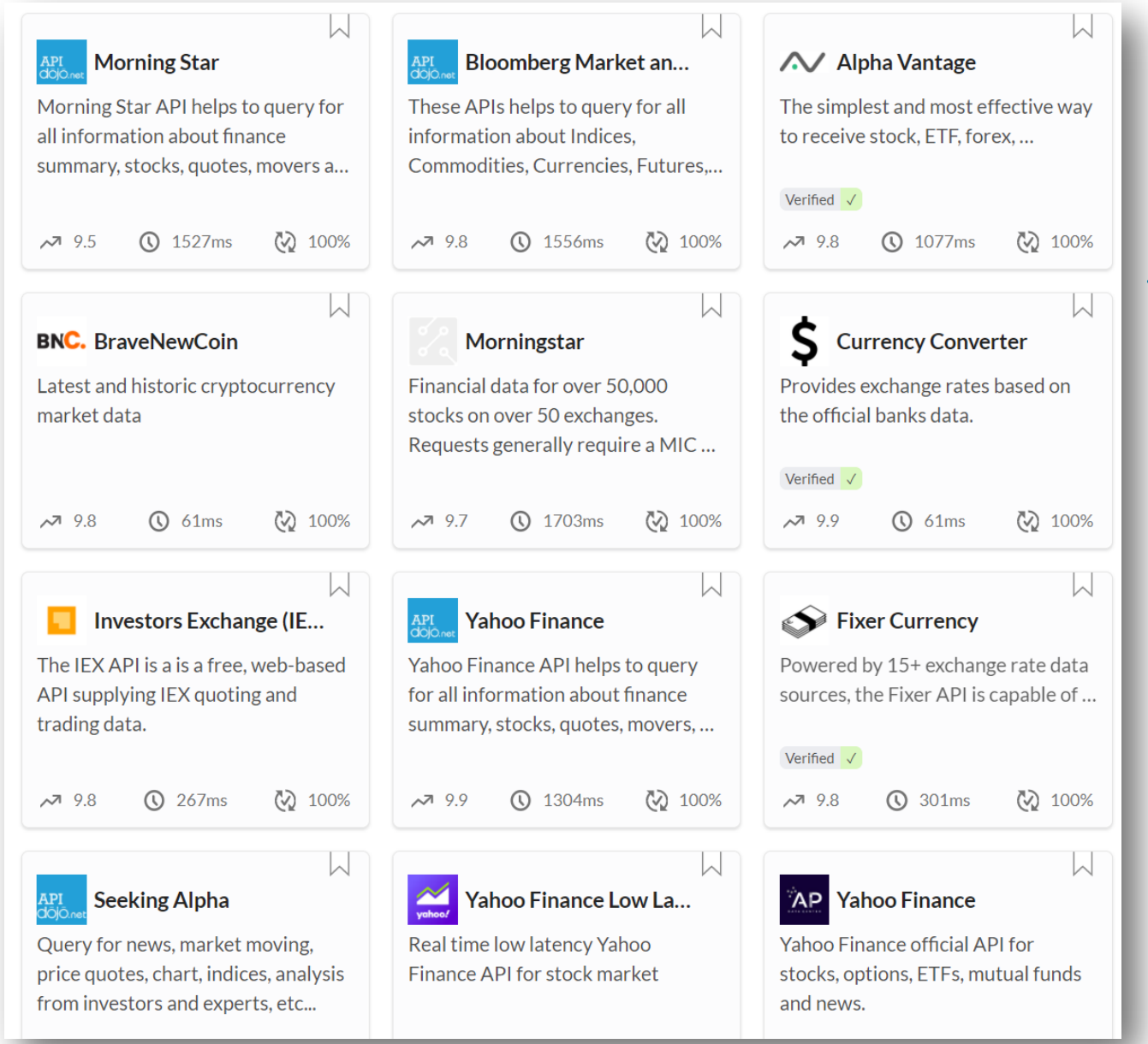
Examples:
Using a weather API to collect real-time temperature and humidity data for a mobile app.





The « problem » is which to choose?

Source



Not just for weather, for finance and other areas.

Source

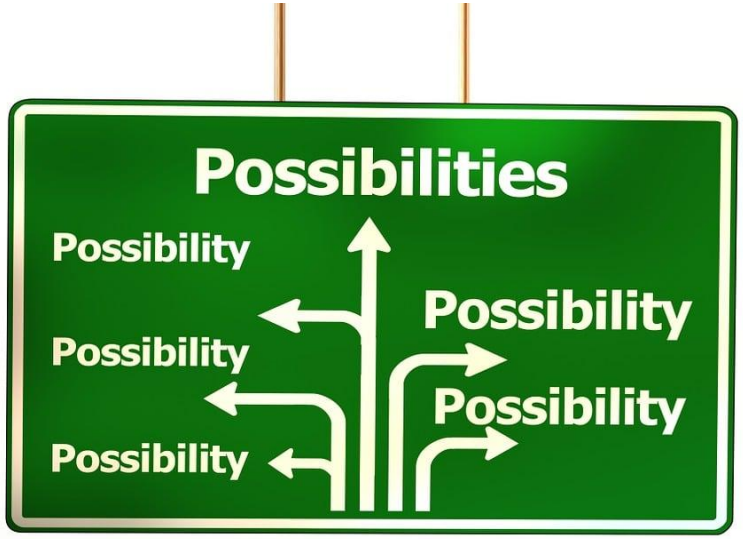
DECISION FACTORS

Self-Reported Data	Surveys Interviews Focus groups
Data directly provided by individuals	

Behavioral and Observational Data	Observations Experiments
Data collected through monitoring or observing human behavior.	

Automated Data	Logs Sensor Transactions
Data generated obtained through an automated system.	

Digital and Platform-Based Data	Social Media Data Web Scraping APIs
Data originating from digital platforms, systems, or online activity.	



Quite overwhelming… How to choose/decide on data collection method?

- Resource constraints (Effort/Cost)
- Time constraints
- Goal of the study
- Availability of already collected data
- Domain expertise
- Technical expertise
- Data storage possibilities
- Ease of use and integration
- Reliability
- Ethical and Privacy concerns

DATA COLLECTION

- Data Science Process
- Types of questions (insights)
- Overview of data collection methods

Continue with case studies on Friday