

# Unit III

## Applications, Evaluations, and Methods

# Solving Data Problems

- Solving Data Problems

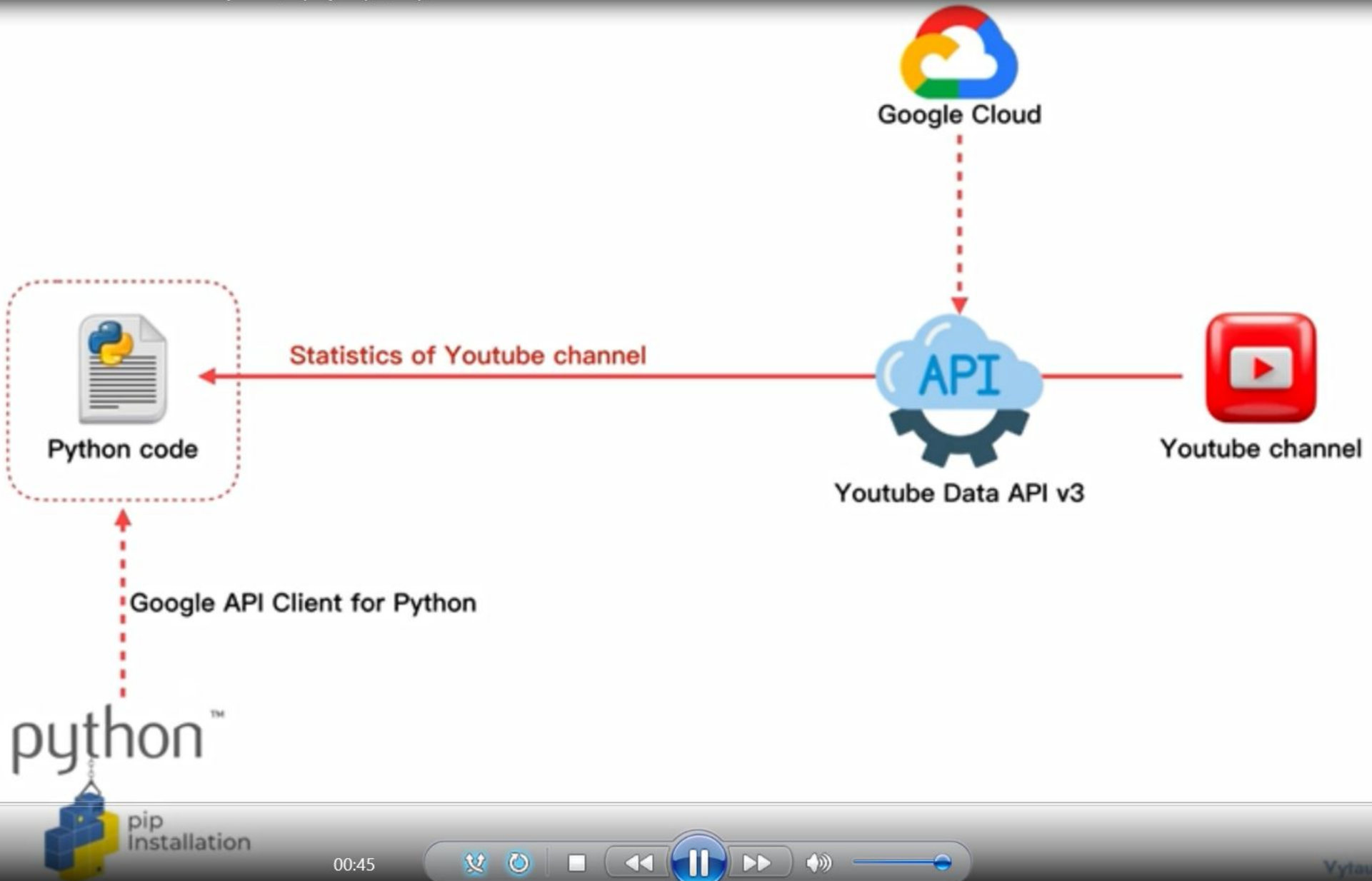
# Collecting and Analyzing Twitter Data

- Step 1: Signing Up for Using Twitter APIs
- Step 2: Get Your Keys and Tokens
- Step 3: Get the Twitter Search Python Script

# Collecting and Analyzing YouTube Data

- Step 1: Signing Up for Using YouTube APIs
- Step 2: Select the Correct API and Get Your API Key
- Step 3: Install the Packages
- **Micro level steps:**
  - Go to Google Cloud Console.
  - Click on the project drop-down at the top, then “New Project”. ( log in using gmail account )
  - Enter a project name and click “Create”.
  - In the Google Cloud Console, navigate to “APIs & Services” > “Library”. ( or type APIS and services in search text box) then, enable APIs
  - Search for “YouTube Data API v3” and click on it.
  - Click “Enable”.
  - Go to “APIs & Services” > “Credentials”.
  - Click “+ CREATE CREDENTIALS” and select “API key”.

# Collecting and Analyzing YouTube Data



# Collecting and Analyzing YouTube Data



The YouTube Analytics API allows you to create your own custom dashboard from the key performance indicator data you want from your YouTube channel. These are -

1. Likes count - Number of times your video was liked
2. Dislikes count - Total number of times the dislike button was hit
3. Shares count - How often your video was shared
4. Views count - How many times your YouTube video was watched
5. Comments count - The number of comments your video has gained under it. This includes reply threads
6. View duration - The number of seconds or minutes your video was watched
7. Subscriber count - The total number of subscribers your channel has
8. New subscriber count - How many new subscribers you have gained

# Collecting and Analyzing Yelp reviews

- Founded in 2004, Yelp is a popular online directory for discovering local businesses ranging from bars, restaurants, and cafes to hairdressers, spas, and gas stations.
- Yelp's website, Yelp.com, is a crowd-sourced local business review and social networking site.
- The site has pages devoted to individual locations, such as restaurants or schools, where Yelp users can submit a review of their products or services using a one to five stars rating scale.

# Key Services Provided by Yelp

## 1. Business Reviews and Ratings:

1. **User Reviews:** The most popular feature of Yelp is its review system. Users can write reviews and provide ratings (from 1 to 5 stars) about businesses, which helps other users make informed decisions.
2. **Ratings Aggregation:** Yelp aggregates the reviews and gives businesses an overall rating, which often influences a business's reputation.

## 2. Business Information:

1. **Business Profiles:** Each business listed on Yelp has its own profile where essential information like address, phone number, operating hours, and amenities (e.g., "outdoor seating", "Wi-Fi available") is displayed.
2. **Photos and Menus:** Users can upload photos of businesses and their offerings. Restaurants often display their menus.

## 3. Yelp Reservations and Waitlist:

1. **Online Reservations:** Many restaurants partner with Yelp to allow customers to book reservations through the platform.
2. **Waitlist Management:** Some restaurants use Yelp's waitlist feature to manage and notify customers about their table availability.



# Key Industries Covered by Yelp

- **Restaurants, Bars, and Cafes:** Yelp is particularly popular for food-related businesses, and it's one of the most used tools for finding restaurants.
- **Home Services:** Many users turn to Yelp to find home-related services like electricians, plumbers, and contractors.
- **Health and Beauty:** Listings for spas, salons, dentists, and doctors are also common on Yelp.
- **Retail and Shopping:** Local shops, boutiques, and larger retailers are listed and reviewed on Yelp.
- **Travel and Tourism:** Users frequently review hotels, attractions, and tour services.

# Data Collection Methods

- prior data may not be available, or what is available is not as useful for a given application.
- Need for designing new experiments and doing various quantitative and qualitative analyses to address emerging problems.

# Data Collection Methods

## 1. surveys

- To collect data specific to a problem or a question you are asking about how people think about something, for which data does not exist.
- census, Nielsen ratings, or exit polls on an election night, these are all variations of surveys.
- A survey can be conducted in person, on paper, or online.
- Reasons to conduct a survey:
  - 1. Uncover answers
  - 2. Evoke discussion
  - 3. Base decisions on objective information
  - 4. Compare results
- A survey can provide valuable, targeted information, but it's important to first identify the specific question or problem being addressed.
- The type of questions to be included should align with goals. If you need numerical data (e.g., how many people think "ABC" or "XYZ"), closed-ended or multiple-choice questions are effective. However, if you're seeking public opinion, open-ended questions may be more appropriate.

# Data Collection Methods

## 2. Survey Question Types

- Multiple-choice-type questions : useful when it is tiresome for the respondent to read through and they may lose patience
- Examples : questions related to gender (Male/Female), age groups (e.g., 18–25, 26–30, 31–40, 41–50, >51), and employment (Employed, Not employed, Self-employed).
- It is often a good idea to offer a “Prefer not to answer” option with each of these questions.
- multiple-choice questions are also used to get responses for other kinds of inquiries that are more complex.
- Example :
- ask customers/users about what was the most dissatisfying thing about a new service/interface using a multiple choice question that has five possible answers: (1) input dialogue; (2) transaction confirmation screen; (3) speed of the interaction; (4) audible and visual feedbacks with each interaction; and (5) readability of the messages.

# Data Collection Methods

## 2. Survey Question Types

- Rank-order-type questions : ask respondents to choose one thing over another in preference.
- These are usually things that come in paired comparisons, like price versus quality.
- These are fairly easy to answer but may not be very useful unless you are testing a hypothesis.
- **Example:** Rank the following factors in order of importance when choosing a smartphone (1 being the most important, 5 being the least important):
  - Battery life
  - Camera quality
  - Price
  - Brand reputation
  - Storage capacity
- In this type of question, respondents are required to assign a rank to each item, helping to assess priorities or preferences.

# Data Collection Methods

## 2. Survey Question Types

- Rating or open-ended questions may work for your survey purpose. A Likert scale (of 1 to 5) – widely used in scaling responses in surveys –
- Gives respondents five statements and asks them to choose one that comes closest to how they feel, ranging from (1) strongly agree to (5) strongly disagree.
- Respondents are typically asked to pick one statement

## 2. Survey Question Types

### d. Open ended questions

- An open-ended question is constructed so the respondent cannot answer “yes” or “no” but must give more information.
- Open-ended questioning is useful to master if you are interviewing people on camera and you need a good “sound bite.” (brief, memorable snippets of dialogue that can be used in media or presentations.)
- Example : What do you think would make this website more user friendly?”
- more difficult to quantify but it may be the most fruitful in giving you the reason behind the answer.
- few examples of open-ended questions:
  - Can you describe a challenge you faced in your work and how you overcame it?
  - What inspired you to pursue this career path?
  - Instead of asking How was your experience? Tell me your favorite thing on vacation?

## 2. Survey Question Types

### e. Dichotomous (closed-ended) questions

- ask for “yes” or “no,” or “true” or “false” answers.
- For instance, “In the last 30 days, have you seen any ads on Facebook for shoes?”
  - \_\_\_\_\_ yes
  - \_\_\_\_\_ no
- or
- “Five plus three equals eight.” \_\_\_\_\_ true \_\_\_\_\_ false.
- Program the survey, depending on the answer, to jump to another question.
- If a person answered “yes” to seeing ads for shoes, then we could ask if they remembered the brand.
- If they say “no,” then we could put some brand names in front of them and see if one might be the answer



### 3. Survey Audience

- People who will be taking the survey
- Using an online survey service and broadcasting the survey, we may not be able to control who sees and responds to the survey.
- And since we do not see these people in person, it may be hard to assess if the responses received were from the right kind of audience.
- not be easy to figure out if even a human took the survey ( BOTs are taking)
- Not able to track if the same individual responds to the survey more than once.
- Not able to check are they just randomly clicking around the multiple-choice questions?
- To address these issues :
- Use a service such as CAPTCHA to verify that indeed a human is taking the survey.
- Use a question or two somewhere in between real questions that tests if the respondent is paying attention. An example is “What is  $1 + 1$ ?”

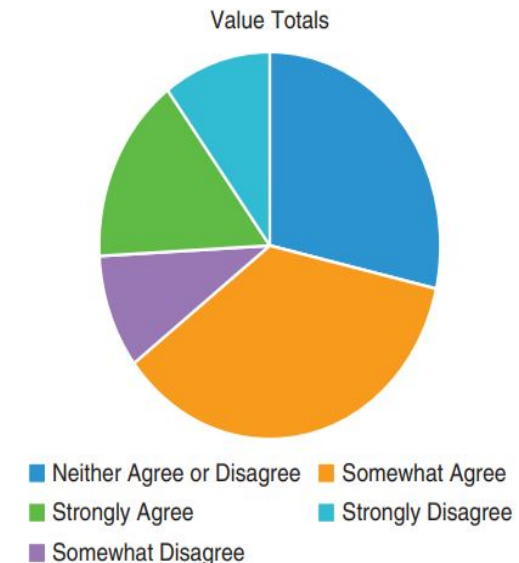
# 4. Survey Services

- do the survey using mobile social media, email, or, even more old-fashioned, by handing out a piece of paper at the exit of an event and asking people to answer on the spot.
- **Free service:**
  - **Microsoft's Office 365** offers a mobile-only option to conduct a survey using Excel
  - **SmartSurvey**<sup>4</sup> is out of the UK and offers free plans but also student discounts
    - There are options to create your own new survey
    - Use a survey template, for which they have more than 50 varieties for questions related to realms as broad as client services, educational experiences, hospitality, and marketing.
    - And there is a section of Web-ready surveys that we could tailor and shoot out via social network or post to a website to gather information from visitors
  - Conduct a survey through Google Forms and we can use the template gallery that can be tailored for personal, work-, or education-related surveys and assessments and results will stay in your Google Doc.
- **Paid surveys:** If able to afford to pay for services or if we have a subscription, we could look into more refined and sophisticated tools available from **SurveyMonkey and Qualtrics**
- Many social media services also allow one to create an instant poll

# 5. Analyzing Survey Data

- Spreadsheet program such as Excel or Google Sheets.
- can easily import your survey data (most tools allow this) into one of such programs.
- If used Google Forms to conduct the survey, the data is already in Google Sheets
- Or write a report that summarizes the results.
- Include a chart that visualizes the range of responses;

Row labels	Count for "I think this website is user friendly."	
Neither Agree or Disagree	4	<i>Note the values are the same for "somewhat" and "strongly"</i>
Somewhat Agree	2	
Somewhat Disagree	2	
Strongly Agree	1	
Strongly Disagree	1	
<b>Total</b>	10	



## 6. Pros and cons of surveys

- Pros :
  - 1.Results could indicate a direction to pursue.
  - 2. are relatively inexpensive to administer, especially online, and can quickly reach a large audience, allowing for the collection of data in a short period of time.
  - 3. Respondents may feel more comfortable sharing their honest opinions if they can complete the survey anonymously,
- Cons :
  - 1. may be skewed due to the mood the respondents.
  - 2. if you get results you did not expect? You may want to do another survey to see if the results were indicative. People may not answer the questions honestly.
  - 3. boost the probability of honesty by telling them no one will know who answered what – that it is a blind survey

To overcome the cons,

- Give respondents a variety of ways to answer.
- Try to ask more or less the same question or a couple of questions in different ways, from different angles

## 7. Interviews and Focus Groups

- Interviews and focus groups can deliver rich, targeted information with the benefit of the exact words of the interviewee.
- In person and in their own words, respondents may offer nuggets you will not get from a survey.
- And a direct quotation can be a compelling way to conclude a paper if it supports your findings.
- Interviews are generally done one on one and focus groups are just what they sound like – a group of people responding to focused questions or stimuli. I
- Interviews may be better suited to more intimate-type questions, and focus groups may be ideal for testing something that is easy to talk about in public.

## 8. Why Do an Interview?

- What is it that makes them good at bringing out revealing truths or stories.
- Example : If you have a thesis about what you are hoping to answer, confirm, or explore, you should prepare questions that ask for answers in various way
- As an interviewee you prepare and put your best foot forward.
- Not only a variety of types of questions, but ask the same question from another angle.
- You do not want the person you are interviewing to second guess your goal so that they give you predictable answers
- some of your questions are “fillers” and some are what you are really hoping to answer.

## 9. Why Focus Groups?

- a brand manager wants to test a new cereal that is still in development with an established brand behind it.
- Now try it out on a group of willing people in a focus group.
- Get at exactly what it is that would make the cereal great, not just good.
- Ask if it reminds them of any other cereal.
- Want to know if they like it and if they would buy it or what they think might make it better
- Focus groups can also be useful for brainstorming.
- Need to get feedback on a developing logo for a new brand or a brand update, then show variations and get reaction

# 10. Interview or Focus Group Procedure

- want to attract people to interview who fit the profile and are willing to talk about the subject.
- Once crafted the questions, do a test-run on a friend or two to see if the questions generate answers that can work with.
- Then start with :
  1. Agreement. Have them write out and sign their name and the date as well as their contact information under a statement that says they are competent and willing to divulge their thoughts/opinions today for the purpose of the interview. Also, add that their identity and privacy will be protected
  2. Ice breaker. Help your interviewee relax by offering them water and asking how was their trip getting there.
  3. Honest opinion. Tell them upfront that you hope they are willing to tell the truth. Ask if it is alright if tape-record the interview. Remind them that their identity will be protected and that ask them for permission to quote them anonymously after the process of interviewing others is complete if they have made a useful statement.
  4. Plan. Tell them that you will be asking different kinds of questions from a prepared list and they can take their time answering, but that the entire interview will last no more than (whatever you initially told them when you asked for the interview), say, half an hour. If gone over time because a worthwhile conversation happened, ask for permission to continue.



# 11. Analyzing Interview Data

- Once conducted the interviews or focus groups, hopefully we now have a rich resource in a recorded format.
- By way of illustrating the next step, let us consider the job of the court reporter.
- If you are a court reporter, your job is to sit through courtroom proceedings and capture the exact verbosity that transpires – from attorneys, witnesses, defendant, judge – on a stenotype machine.
- Transcribe the paper from shorthand symbols back into words.
- Sit down at a computer and type, so a document with words can be emailed or printed out.
- This process needs to happen with interview recordings.
- can methodically listen to the recordings and type what you hear yourself, though it is time-consuming. Or You could ask a student or a friend (and pay them) to listen to the
- recording and type it up. Or, you could use a professional transcription service that can turn your job into the written word in hours.
- They will charge you by the minute, but it will be worth it in time saved and attention to detail. Once you have your transcription (your content), then you can analyze it

# 12. Pros and Cons of Interviews and Focus Groups

- Is the interview the most honest of conversations?
- So how would you get an honest answer from someone you are interviewing for a research project
- project that you are genuinely interested in their answers
- Prepare and then practice on your classmates or a friend before the real thing.
- In focus groups, going to have the natural dynamic of some people dominating and some being more passive.
- Some will be tempted to piggyback on what someone else said.
- But, there will be interesting things that develop as a conversation ensues, and sometimes the stimulation of people thinking together, out loud (sometimes called brainstorming), can bring rich fodder for your project.

# 13. Log and Diary Data

- The experience of your study participants might best be captured in a log or a diary for the flexibility it gives the user-participant.
- They could do the log on an app on their smartphone or the diary from their laptop at home.
- Difference between a log and a diary:
- Examples : Logs
  - One traditional use of a log is by a ship's captain to note things like weather, geospatial coordinates, and anything unusual at predictable times of day so that he or someone else can look back at the record for why they hit an iceberg!
  - Or an athlete might use a log to note his/her progress while training for the Olympics.
  - For a research study, you might use an activity log to capture specific data at regular check-ins to assess a quality over time.
- Diary : use a diary if you want someone's experience or perceptions daily over a set time period. You might ask that a diary entry be made once a day, at the end of each day, over the course of your study or experiment.

# 13. Log and Diary Data

- Samples :

**Table 12.2** A generic activity log template.

Name:

Date:

Time	Activity/ Task	Feeling/ Observation	Duration/ Measurement
------	-------------------	-------------------------	--------------------------

10:00 am

1:00 pm

4:00 pm

8:00 pm

**Table 12.3** A generic diary template.

Daily reflection for (Name):

(Date): / /

What was your experience of doing “XYZ” today?

If you had to summarize it in one word, what would it be?

How might this work more efficiently/effectively if you did something differently? Share any other thoughts. Thank you!

# 14. User Studies in Lab and Field

- Example : Testing UI □ It is the designed relationship between the user and the product – essentially, a mechanism through which a user could interact with a system.
- If you want to test the performance of a new car or a website, you need some willing participants, or users. User studies can be done in a lab setting, where you have the most control and opportunity to observe, or “out in the field,” where less control may be a more realistic gauge of the product.
- Use a lab-based setting for something that needs supervision or requires confidentiality
- there are also computer programs that track a user’s experience, so you do not necessarily need to be present.
- Use hardware equipment such as an eye-tracker for collecting data about where a user is looking on the screen, heart-rate monitors, electroconductance monitors to measure emotional responses, and EEG monitors to collect data about brain activity.
- They come with their own software that you can use for analysis. Or, they let you export the data in a format (such as XML and CSV) that can be understood by other analysis software.
- A field-based study, where the user is out and about in their daily life – their normal context – interacting with the product, such as an app, and giving self-reported feedback, is a more “natural” situation that might be useful when you are looking for a big-picture experience.
- User studies in both the lab setting or out in the field are likely to yield feedback on the quality of the user experience and not numerical data.

# Picking Data Collection and Analysis Methods

# Quantitative Methods

## Definition:

- Involve collecting data that can be **measured and expressed numerically**.
- Commonly used in scientific research and include techniques like surveys, experiments, and statistical analysis.
- **Approach:**
  - **Deductive reasoning:** Start with a hypothesis, collect data, and use the results to confirm or refine the theory.
  - Focus on what *is* (descriptive statistics: mean, median, mode, variance, etc.) and then move to **inferential statistics** (correlation, regression, classification).
- **Examples:**
  - Calculating volume using geometry.
  - Lab studies counting behaviors or quantifying opinions (e.g., using a Likert scale).
- **Advantages:**
  - Objective, can generalize from samples to populations.
  - Easier to scale to large datasets.

# Qualitative Methods

- **Definition:**

Involve collecting data that captures **qualities, experiences, opinions, or behaviors**—usually in **non-numeric** form.

- **Approach:**

**Inductive reasoning:** Start with observations, identify patterns, and develop theories from the ground up.

Methods include **ethnography, interviews, focus groups, and observations**.

- **Challenges:**

- Harder to measure or generalize.
- Time-consuming.

- Analysis involves **coding** and identifying **themes** through methods like **constant comparison** and **grounded theory**.

- **Advantages:**

- Captures rich, contextual information.
- Valuable in early stages of research to define what to measure quantitatively later.



# Mixed Methods

- **Definition:**

Combines **quantitative and qualitative** approaches in one study to **offset the weaknesses** of each method.

- **Why Use It:**

- Quantitative = structured and scalable but lacks context.


- Qualitative = rich in insight but hard to scale and analyze statistically.

- **Example:**

- In a web search behavior study:

- Use quantitative methods to measure time, query types, and frequency.
- Follow up with qualitative interviews to understand **why** users behaved a certain way.
- This combination gives both **behavioral patterns** (quantitative) and **motivational insights** (qualitative).

# Evaluation

- **Comparing Models**
-  **Why Compare?**
  - Different models can solve the same problem with different results.
  - We need **evaluation metrics** to determine which model performs best, especially on new, unseen data.

# Evaluation Metrics

Metric	Formula	Meaning
Accuracy	$(TP + TN) / (TP + FP + TN + FN)$	proportion of correctly predicted instances out of the total instances in the dataset.
Precision	$TP / (TP + FP)$	How many predicted positives are <i>actually</i> positive
Recall	$TP / (TP + FN)$	How many actual positives were <i>correctly</i> identified
F-measure	$2 * (Precision * Recall) / (Precision + Recall)$	Harmonic mean of precision and recall; balances both
ROC Curve	TPR vs FPR graph	Visualizes trade-off between true positives and false positives
AUC	Area under the ROC curve	Higher = better overall performance
AIC	Lower AIC = better	Penalizes model complexity while rewarding good fit
BIC	Lower BIC = better	Like AIC, but <b>penalizes complexity more heavily</b>

# Precision

- **Formula:**  $TP / (TP + FP)$
- **TP (True Positives):** Correctly predicted positive cases.
- **FP (False Positives):** Incorrectly predicted positive cases.
- **Meaning:** Out of everything the model *predicted* as positive, how many were *actually* positive?
- **Use Case:** Precision is important when **false positives are costly**.
  - Example: A spam filter. You don't want to mark important emails as spam (false positives).

## Recall

- **Formula:**  $TP / (TP + FN)$
- **FN (False Negatives):** Actual positives missed by the model.
- **Meaning:** Out of everything that *was actually positive*, how many did the model correctly identify?
- **Use Case:** Recall is important when **missing a positive case is costly**.
  - Example: Diagnosing cancer. Better to catch all possible cases (even with some false alarms).

# F-measure (F1 Score)

**Formula:**  $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

- **Meaning:** It's the harmonic mean of precision and recall — gives a **balanced score** when you need to consider both.
- **Use Case:** Useful when there is an **imbalance between classes** (e.g., rare events like fraud detection).

## ROC Curve (Receiver Operating Characteristic Curve)

- **What it plots:**
  - **True Positive Rate (TPR)** = Recall
  - **False Positive Rate (FPR)** =  $FP / (FP + TN)$
- **Meaning:** Shows the **trade-off** between sensitivity (recall) and fall-out (false positive rate) at various thresholds.
- **Use Case:** Helps visualize model performance and decide on a threshold for classification.

# Example

		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

- Suppose we test **100 people**. The results:
- **Actual COVID-positive = 40 people**
- **Actual COVID-negative = 60 people**
- Test outcomes:
- **True Positive (TP) = 35** → 35 sick people correctly detected.
- **False Negative (FN) = 5** → 5 sick people wrongly declared healthy.
- **False Positive (FP) = 10** → 10 healthy people wrongly flagged as COVID.
- **True Negative (TN) = 50** → 50 healthy people correctly cleared.

Confusion Matrix - COVID Test Example

Actual Class	Predicted Class	
	Predicted Positive	Predicted Negative
Actual Positive	TP = 35	FN = 5
Actual Negative	FP = 10	TN = 50

# Calculation

$$\begin{aligned}\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{35 + 50}{35 + 50 + 10 + 5} = \frac{85}{100} = 0.85\end{aligned}$$

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP + FP} \\ &= \frac{35}{35 + 10} = \frac{35}{45} \approx 0.778\end{aligned}$$

$$\begin{aligned}\text{Recall} &= \frac{TP}{TP + FN} \\ &= \frac{35}{35 + 5} = \frac{35}{40} = 0.875\end{aligned}$$

$$\begin{aligned}F1 &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \cdot 0.778 \cdot 0.875}{0.778 + 0.875} = \frac{1.362}{1.653} \approx 0.824\end{aligned}$$

- **ROC Curve (Receiver Operating Characteristic Curve)**
- **What it plots:**
  - **True Positive Rate (TPR)** = Recall
  - **False Positive Rate (FPR)** =  $FP / (FP + TN)$
- **Meaning:** Shows the **trade-off** between sensitivity (recall) and fall-out (false positive rate) at various thresholds.
- **Use Case:** Helps visualize model performance and decide on a threshold for classification.

- **AUC (Area Under the ROC Curve)**
- **What it represents:** The **area under the ROC curve**.
- **Value range:** Between 0 and 1.
  - Closer to 1 = better model performance.
  - 0.5 = no better than random guessing.
- **Use Case:** Overall performance indicator — a higher AUC means the model is **better at distinguishing** between positive and negative classes.



- **Why AIC and BIC?**
- When building models, there's a **trade-off**:
- **Simple models** → easy to explain, faster, less costly.
- **Complex models** → may fit data better, but harder to interpret, more computationally heavy, and risk **overfitting**.
- We need a way to **balance goodness of fit with simplicity**.  
This is where **AIC and BIC** come in.

- **AIC (Akaike Information Criterion)**
  - Developed from **information theory**.
  - Measures **how much information is lost** when a model explains the process behind the data.
  - Formula (simplified idea):  $2k - 2\ln(L)$
  - Where:
    - $k$  = number of parameters in the model.
    - $L$  = likelihood (how well the model fits the data).
- 👉 Interpretation:
- Lower AIC = **less information loss = better model**.
- BUT: AIC is **relative only**.
  - It tells you which model is better **compared to others**, but not whether any model is absolutely “good.”
  - Example: If all candidate models fit poorly, AIC still picks the “least bad.”
- ✓ Example:

Suppose we fit two regression models:

  - Model A: 3 predictors  $\rightarrow$  AIC = 120
  - Model B: 5 predictors  $\rightarrow$  AIC = 118

Here, Model B is slightly better (lower AIC), but the difference is small, so Model A might be chosen if interpretability is important.

- **BIC (Bayesian Information Criterion)**

- Similar idea to AIC but derived from **Bayesian probability**.
- Formula (simplified):  $\ln(n)k - 2\ln(L)$
- Where:
  - $n$  = number of data points.
  - $k$  = number of parameters.

- 🙌 Interpretation:

- Lower BIC = better model.

- **Penalizes complexity more strongly than AIC** (especially when  $n$  is large).

- So, BIC tends to prefer **simpler models** compared to AIC.

- ✅ Example:

- Model A: BIC = 130

- Model B: BIC = 125

Model B is better under BIC.

But if Model B is much more complex, AIC may favor it, while BIC penalizes it heavily.

### **AIC (Akaike Information Criterion)**

- **Concept:** Balances **model fit** and **complexity**.
- **Formula (conceptually):**  $2k - 2\ln(L)$ 
  - **k:** number of parameters in the model.
  - **L:** likelihood of the model (how well it fits the data).
- **Lower AIC is better.**
- **Use Case:** Helps choose between **multiple models** — especially when trying to avoid overfitting.

### **BIC (Bayesian Information Criterion)**

- **Similar to AIC**, but it **penalizes complexity more heavily**.
- **Formula (conceptually):**  $\ln(n)k - 2\ln(L)$ 
  - **n:** number of observations (sample size).
- **Lower BIC is better.**
- **Use Case:** More conservative than AIC; better for avoiding overly complex models when the sample size is large.

# Training–Testing & A/B Testing

- **Training–Testing Split**
  - Train on **large portion** of data.
  - Test on **unseen data**.
  - Purpose: Check **generalizability** (how well model works beyond training data).
  - Must be **randomly split** to avoid bias.
- **Validation Set**
  - Separate dataset used during training.
  - Helps prevent **overfitting**.
  - Example: In decision trees, you use validation data to check if accuracy improvements are genuine or just memorization of training data.

## A/B Testing

- Controlled experiment to compare **two (or more) variants**.
- Common in **business & web analytics**.
- ✓ Example:
  - Food delivery app tests ad campaigns:
    - Version A (social media ad) → Promo code AX
    - Version B (email ad) → Promo code AY
- Measure redemption rates → Pick best strategy.
- 👉 Can extend to A/B/C/D... testing (more variants).




# Cross-Validation

- Instead of fixed train–test splits, cross-validation provides more reliable evaluation.

## Holdout Method

- One-time split into training & testing.
- Simple but results can depend heavily on how data is split.

## K-Fold Cross-Validation

- Dataset split into **k parts**.
- Each part gets a turn as test set.
- Average error = overall evaluation.
-  Example: 10-fold CV
- Dataset divided into 10 equal parts.
- Model trained 10 times → each time leaving out 1 part.
- Accuracy averaged across 10 runs.
-  Advantage: Every data point is used for both training and testing.
-  Disadvantage: More computationally expensive.

## Leave-One-Out Cross-Validation (LOO-CV)

Extreme case of k-fold, where  $k=N$  (number of data points).

Train on all except 1 point, test on that point. Repeat for all points.

Very accurate but expensive for large datasets.

Works well for **small datasets** or **locally weighted models**.