Programming and Artificial Intelligence

An Introduction to
Information and Communication Technology

Egyptian Baccalaureate

2nd Year

Part Two

11

SPRIX

Programming and Artificial Intelligence

Egyptian Baccalaureate

2nd Year

The International Baccalaureate™

has reviewed the pedagogical and assessment approaches underpinning this textbook.

Ministry of Education and Technical Education

New Administrative Capital

Cairo, Egypt

Introduction

In light of Egypt's Vision 2030 and the state's ongoing efforts to develop education, the Ministry of Education and Technical Education places great emphasis on updating school curriculums — particularly in information and communication technology — due to its vital role in fostering computational thinking, problem-solving abilities, and responsible digital citizenship, with a focus on building a deep understanding of fundamental concepts rather than rote memorization. The content is organized in a gradual and systematic manner that enables students to progress smoothly from simple to complex concepts, linking every concept to real-life situations, and relying on interactive methods that include teamwork and discussion, making learning more enjoyable.

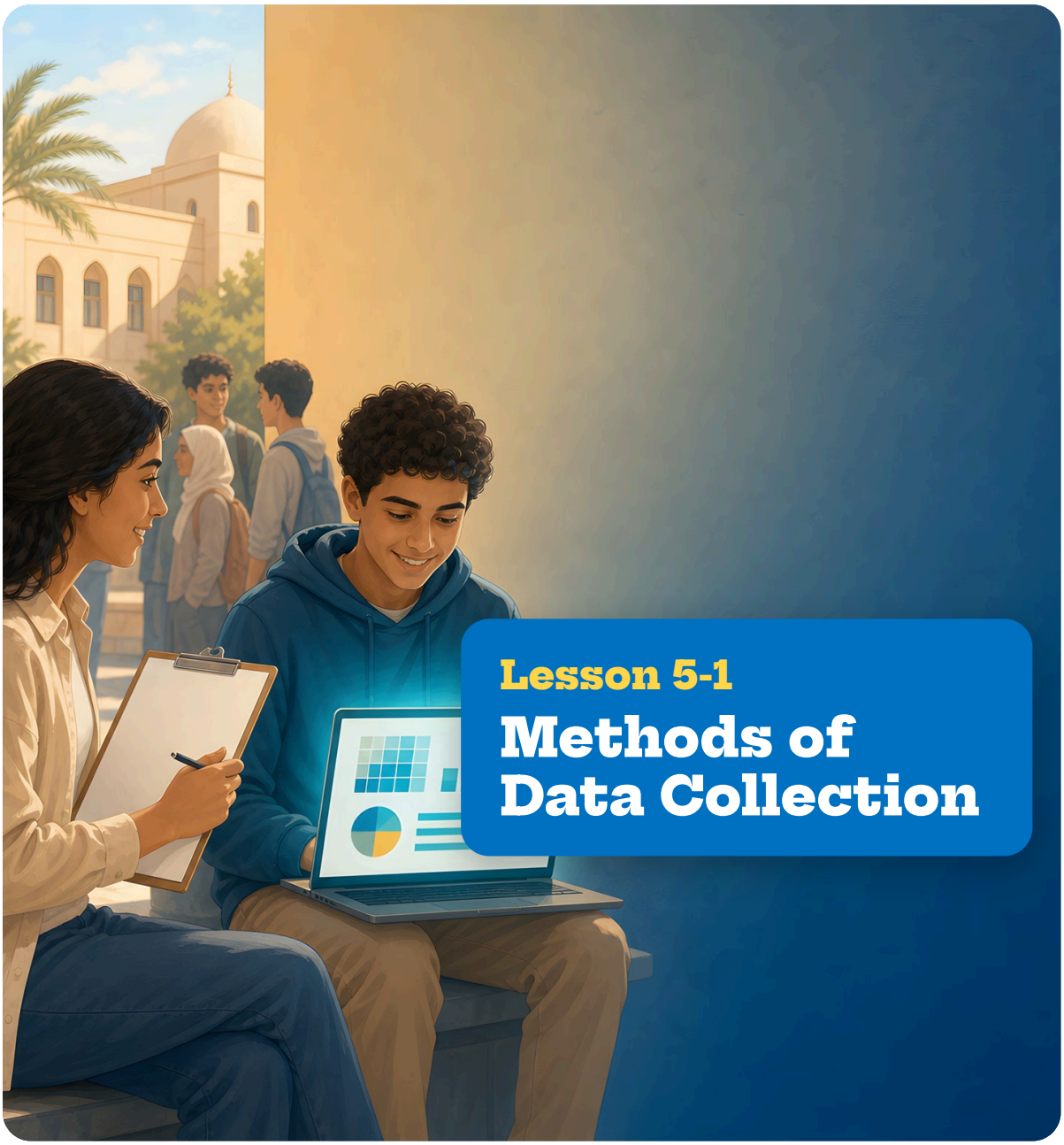
In this context, the Ministry has been keen to benefit from successful international experiences in teaching information technology, data analysis and artificial intelligence, given the speed at which these fields are reshaping work, communication and society, and the need for curriculums that keep pace with them while remaining clear and accessible to every student.

From this concern, this book offers an integrated introduction to information and communication technology. Part 1 addresses information technology and society, cybersecurity, web applications, and web and media design; Part 2 addresses data collection and cleaning, analysis and communication, and machine learning and artificial intelligence — so that the student meets data and AI as tools to be understood, questioned and used with judgement.

This book represents a step within a comprehensive plan to develop information and communication technology curriculums across all grade levels, with the aim of building an integrated educational system that combines the best international practices with Egypt's authentic identity.

The Ministry of Education and Technical Education firmly believes that developing education is a true investment in the nation's future, and that updating curriculums is not an end in itself, but rather a means to build an aware, creative generation, capable of keeping pace with scientific and technological changes in today's world.

Contents



Learning Objectives

- Distinguish primary from secondary data and explain when each is appropriate.
- Explain how sampling works, and describe how sampling bias can lead to confident but wrong conclusions.

Before you can trust what data tells you, you have to trust how it was gathered. Good analysis moves through three stages — collection, organization, then analysis — and if the wrong data is collected, no careful work later can rescue it. Some data you gather yourself for your exact purpose (primary data); some already exists, collected by others (secondary data). And when a whole group is too large to survey, you study a sample — but only a fair sample lets you speak for the whole. A biased sample gives a confident answer that is simply wrong.

? Guiding question: How do you collect data you can trust, and how can a small sample fairly represent a whole population?

Point!

1. The Data Collection Stage

(1) Data analysis progresses through the flow of "collection → organization → analysis," after the purpose has been clarified. In **(data collection)**, the first stage, it is necessary to plan what data will be gathered and how, in line with the purpose of the analysis.

(2) If data that does not match the purpose is collected, it is impossible to draw appropriate conclusions, no matter how carefully later organization and analysis are carried out. For this reason, data collection is an important stage that determines the quality of the entire analysis.

2. Primary Data and Secondary Data

(1) **(Primary data)**…data that is newly collected to fit a specific purpose. *e.g.,* surveys, interviews, observations, experiments. Strength: the items collected can be designed to match the purpose, providing excellent originality and reliability. Constraint: collection takes time, effort, and cost.

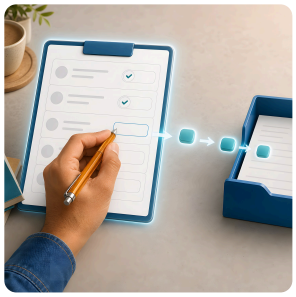
(2) **(Secondary data)**…data that has already been collected and made public by others, and is used in analysis. *e.g.,* open datasets, existing statistics, industry reports. Strength: large amounts of data can be obtained in a short time. Constraints: it may not perfectly match one's purpose; bias or omissions may have arisen during editing or summarizing.

(3) Primary data, since it can be collected to match one's own purpose, has excellent reliability and originality, making it a high-value type of data that can serve as the core of analysis. On the other hand, the value of secondary data lies in being able to easily make use of large-scale data and existing statistics that cannot be obtained from primary data alone. In actual analysis, the two are appropriately combined depending on the purpose and constraints.

Key Fact:
Good analysis starts with good collection: primary data is gathered fresh for a purpose; secondary data already exists; sampling studies part of a population to estimate the whole — and only an unbiased sample supports valid conclusions.

Key Concepts:

- data collection
- primary / secondary data
- population & sample
- sampling (simple random, stratified)
- sampling bias



Sound collection is the base of trustworthy data.

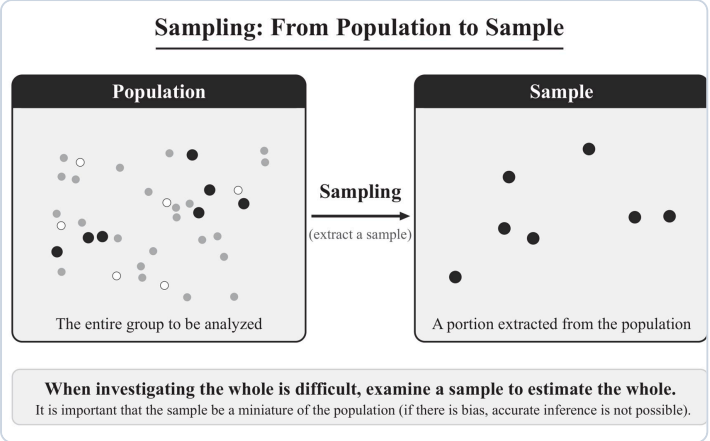
KEY TERMS
data collection — the first stage: plan what to gather and how.
primary data — collected fresh for a purpose.
secondary data — already collected by others.

Primary data is gathered fresh; secondary is reused.

3. Sampling

(1) The entire group that is the target of analysis is called the **(population)**, and the part of the population that is actually surveyed is called the **(sample)**. Surveying the entire population is called a **(complete survey)**, also called a census, and surveying a chosen sample is called a **(sample survey)**.

(2) **(Sampling)**…selecting a sample from the population. Depending on the sampling method, the reliability of the conclusions obtained varies greatly.



Sampling: from population to sample

(3) A typical sampling technique is **(random sampling)**, as described below. Random sampling makes it likely that the sample represents the population, allowing the analysis results to be applied to the population as a whole.

Method	Description
(Simple random sampling)	Choose a sample from the population purely by chance, without bias
(Stratified sampling)	Divide the population into strata such as gender or age group, and randomly choose from each stratum

(4) Simple random sampling is the most basic method and can be used even when little is known about the population. Stratified sampling can reflect the composition of the population (such as the ratios of gender and age group) in the sample, making it easier to reduce bias when differences between strata affect the analysis.

4. Bias in Surveys

(1) **(Sampling bias)**…this occurs when the method of selecting a sample is biased and does not correctly reflect the actual state of the population. When sampling bias is present, the conclusions obtained cannot be applied to the population as a whole.

(2) Typical examples of bias. Self-selection and time/location bias concern who is surveyed (selection bias); response bias concerns how those surveyed answer (measurement error).

PAUSE & THINK

A government census you download — is that primary or secondary data for you?

A fair sample mirrors the whole; a biased one misleads.

KEY TERMS

population — the whole group analyzed.
sample — the part actually surveyed.
sampling bias — an unfair sample that misrepresents the whole.

Type of bias	Description	Example
(Self-selection bias)	Bias caused by only those who want to respond actually responding	In a web survey, only people with high interest respond
(Response bias)	Not a selection problem: people already surveyed give answers that look desirable rather than accurate	Reporting a “study time” longer than it really was
(Time/location bias)	Bias caused by the timing or location of the survey being limited	Surveying in front of a station on weekday afternoons, missing office workers and students

(3) It is difficult to eliminate sampling bias completely, but bias can be reduced by carefully designing the survey: how subjects are chosen, how questions are worded, and how the survey is conducted.

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Primary data refers to data that has already been collected and made public by others.

B Selecting a sample from a population is called sampling.

C Stratified sampling is a method in which the population is divided into strata, and a sample is randomly chosen from each stratum.

D When sampling bias is present, the conclusions obtained may not be applicable to the population as a whole.

(2) Match each description (a, b) of a sampling method with the most appropriate option from the choices below (A, B).

a Choose a sample from the population by chance, without bias.

b Divide the population into strata such as gender or age group, and randomly choose from each stratum.

[Choices] **A** Stratified sampling **B** Simple random sampling

✓ Solution

(1)

A Primary data is data that is newly collected to fit a specific purpose. Data already collected and made public by others is secondary data. Therefore, ×.

B Correct definition of sampling. Therefore, ○.

C Correct definition of stratified sampling. Therefore, ○.

D Correct description of the effect of sampling bias. Therefore, ○.

💡 PAUSE & THINK

A web poll anyone can enter — whose voices are missing, and which bias is that?

(2)

a: Choosing without bias by chance is simple random sampling. **B**

b: Dividing into strata and choosing randomly from each is stratified sampling. **A**

Try

1 Answer the following questions.

1. What is the term for data that is newly collected to fit a specific purpose?
2. What is the term for data that has already been collected and made public by others, and is used in analysis?
3. What is the term for selecting a sample from a population?
4. What is the term for when the method of selecting a sample is biased and does not correctly reflect the actual state of the population?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that is most appropriate as a method of collecting primary data.

A Use statistical data published by a government agency.

B Create a survey to fit one's own purpose and collect responses.

C Cite figures from a report issued by an industry association.

D Use survey data from a study conducted earlier by someone else as is.

(2) For each survey situation (a, b), choose the most appropriate sampling method from the choices below (A, B).

a When investigating study time across all students in a school, you want to choose the sample taking into account differences in tendencies between grades.

b You want to grasp the opinions of all eligible voters in a city without bias.

[Choices] **A** Simple random sampling **B** Stratified sampling

EXAM-STYLE QUESTION

Explain why a biased sample can make an analysis worthless even when the calculations are correct. **[6] Refer to:** the sample as a miniature of the population.

PAUSE & THINK

Which option gathers new data *you* designed — and which just reuses others' data?

(3) From the following options (A–D), choose the one that is most appropriate as an example in which sampling bias has occurred.

A Surveying 100 students randomly chosen from all students in the school by drawing lots.

B Dividing all residents into age-group strata and surveying a randomly chosen sample from each stratum.

C Approaching passers-by in front of the station on weekday afternoons and reporting the results as “the opinions of all citizens.”

D Designing the wording of a survey question carefully so that respondents can answer truthfully.

Think as an Engineer — Investigate & Decide

Your school wants to know how long students typically spend commuting to school each day, but you cannot survey everyone. You design how to choose and question a sample.

☐ **Ask whoever is easiest to reach**

☐ **Plan a fair, representative sample**

1 · Collect data (plan the sample). Decide how many students you would ask and how you would pick them so the sample fairly represents the whole school. As a quick pilot, tally the commute of 8–10 classmates. State the population and the sample clearly.

2 · Analyze the risk of bias. Name one group your method might silently drop — for example, students who were absent that day, or students who never open an online survey link. Explain how missing that group would skew the estimated commute, and say which bias it is closest to (self-selection, response, or time/location).

3 · Decide. Choose simple random or stratified sampling for this survey and justify the choice from the risk you found in step 2 — think about whether commute really differs between identifiable subgroups here. **In pairs,** compare methods and agree on the single change that would most reduce bias.

Give evidence for your answer.

Stuck? Start from who might be left out. Ask yourself: could one part of the school answer very differently from the rest — a whole group, or students who never see your survey? If a subgroup might differ systematically, the real question is how to give every subgroup a fair chance *before* you pick your method. Work out which subgroups matter here, then let that guide your choice.

🌍 In a New Context

A shopping mall surveys only visitors on Saturday afternoon and reports the results as “what all our customers think.” Name the bias, explain who is missed, and describe one change that would make the sample fairer.

❓ Lesson Question — Answered

The lesson asked how to collect data you can trust and how a small sample can represent a whole population. Trustworthy analysis begins with collection: after the purpose is clear, you choose primary data (gathered fresh for that purpose, with high reliability but more cost) or secondary data (already collected by others, fast to obtain but perhaps mismatched or biased), often combining both. When the population — the whole group analyzed — is too large, you study a sample through sampling, aiming for a miniature of the population. Random sampling makes that likely: simple random sampling picks purely by chance, while stratified sampling splits the population into strata and picks from each to mirror its make-up. The danger is sampling bias, where a skewed selection cannot speak for the whole; it is reduced by careful design of who is chosen, how questions are worded, and how the survey is run.

KEY TERMS

simple random sampling — picked purely by chance.
stratified sampling — split into strata, then pick from each.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

In data analysis, it is important to plan what data to collect according to the purpose. The data to be collected can be classified into two types: (a), which is newly collected, and (b), which has already been collected and made public by others. Also, when investigating the entire population is difficult, (c) is performed to select a part of the population. At this time, if bias arises in the way of selection, (d) occurs, and the analysis results may not be applicable to the population as a whole.

1. Fill in the blanks (a)–(d).

⭐ REFLECT & CHALLENGE

Reflect: Think of an online poll you have seen. Whose voices were probably missing from it, and why? **Challenge:** You may survey only 40 of your school's 800 students about a new schedule. Describe exactly how you would choose those 40 to avoid bias, and name the one group you would make sure is represented.

3 Answer the following questions.

(1) From the following options (A–D), choose the one that is most appropriate as an example of secondary data.

A Survey responses collected from students in one's own class.

B Population statistics published by the government.

C Traffic on a school commute route observed and recorded by oneself.

D Records of an interview designed and conducted by oneself.

(2) From the following options (A–D), choose the one that most appropriately describes sampling methods.

A Simple random sampling is a method in which the surveyor preferentially chooses subjects that are easy to gather.

B Stratified sampling is a method in which the population is divided into strata, and a sample is randomly chosen from each stratum.

C Simple random sampling is a method in which the population is divided into strata, and a representative is chosen from each stratum.

D A complete survey is a method in which only part of the population is surveyed.

(3) From the following options (A–D), choose the one that is NOT appropriate as a way to reduce sampling bias.

A Choose subjects without bias from the entire population.

B Limit the timing and location of the survey, and gather responses only from people in specific conditions.

C Choose the sample taking into account the strata of the population (age group, gender, etc.).

D Design the wording of survey questions so that respondents can answer truthfully.

⭐ Key takeaway

Remember:
Collect the right data (primary vs secondary), study a fair sample of the population, use random sampling (simple or stratified) so the sample mirrors the whole, and guard against sampling bias.

Learning Objectives

- Explain how missing values, outliers, and duplicate data are handled when cleaning a dataset.
- Describe data type conversion, normalization and standardization, and the reshaping of data to fit a purpose.

Real data arrives messy: blank cells, a value ten times too large, the same row entered twice, dates written three different ways. Before any analysis can be trusted, the data must be cleaned and transformed to fit the question. There is a rule in computing — "garbage in, garbage out" — which means careful calculation on flawed data still gives a flawed answer. This lesson is about turning raw data into data you can actually rely on.

? Guiding question: Why must data be cleaned and transformed before analysis, and what choices does each fix involve?

Key Fact:
Raw data is rarely ready to use. Cleaning fixes missing values, outliers, and duplicates; transformation converts types, rescales, and reshapes data to fit the purpose. "Garbage in, garbage out" — the quality of any analysis depends on this preparation.

Key Concepts:

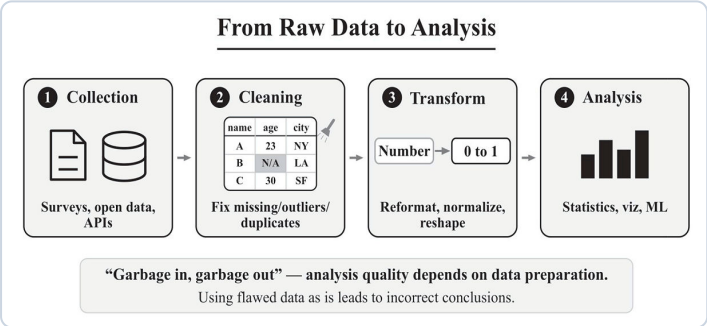
- data cleaning / transformation
- missing values, outliers, duplicates
- data type conversion
- normalization / standardization
- reshaping (pivot, join, aggregation)

KEY TERMS
data cleaning — fixing missing values, outliers, duplicates.
data transformation — reshaping/rescaling data to fit the purpose.

Point!

1. The Data Cleaning and Transformation Stages

(1) Collected data often cannot be used for analysis as is. Adjusting issues such as missing values, outliers, and duplicate data, and arranging the data into a format that fits the purpose of the analysis, is called **(data cleaning)** and **(data transformation)**.



From raw data to analysis: collect, clean, transform, analyze

(2) If issues are left unaddressed during analysis, incorrect conclusions may be drawn. For this reason, cleaning and transformation are important stages that support the accuracy and reliability of the analysis.

2. Handling Missing Values

(1) Using data containing missing values (values where data is absent) directly in analysis can prevent calculations from being performed correctly or cause biased results. For missing values, one of the following three responses is chosen.

PAUSE & THINK

A blank cell in a survey — delete the row, fill it in, or flag it? What decides?

Fix blanks and judge outliers before analysis.

Response	Description	When to apply
Deletion	Remove rows or columns containing missing values from the data	When there are few missing values and the impact on the whole is small
Imputation	Fill in the missing parts with values such as the mean, median, or surrounding values	When missing values are scattered without bias, and replacing them with representative values from other data does not significantly change the overall trend
Flagging	Leave a marker indicating that the value was missing, and analyze the data as is	When the missingness itself has meaning (e.g., the fact that someone refused to answer)

(2) Which response to choose is determined by the proportion and pattern of missing values, and the purpose of the analysis. For example, imputing the mean for an item where missing values account for a large proportion of all data may produce values that are far from reality. There is no single correct response; judgment according to the purpose is required.

3. Handling Outliers and Duplicate Data

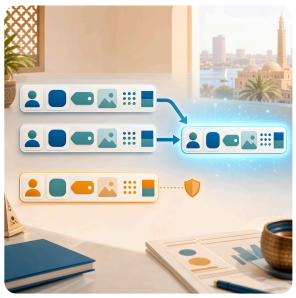
(1) Outliers (values that deviate significantly from other data) are not necessarily errors. They may be **(abnormal values)** caused by input or measurement errors, or they may be values that reflect reality. *e.g.,* If the sales data of a certain store shows that one day has remarkably high sales, it may be an input error, or it may genuinely reflect a sale or event on that day.

KEY TERMS
missing value — a value that is absent.
outlier — a value far from the others.
abnormal value — an outlier caused by an error.

(2) Handling outliers begins with identifying the cause.

① When the cause is an input or measurement error…correct or delete the value

② When it is a value that reflects reality…in principle, keep it and decide how to handle it based on the purpose of the analysis



Remove only genuinely duplicated records.

(3) Duplicate data (a state in which the same content exists in multiple records) affects the results of aggregation and is, in principle, removed. However, it is important to confirm that the data is genuinely duplicated. Even if records appear identical, they may include data that should be retained as separate records. *e.g.,* If the same customer purchases the same product twice on the same day, the records are separate transactions; combining them into one would make the sales aggregation incorrect.

4. Data Type Conversion

(1) Collected data is sometimes difficult to handle in its original format. Arranging the format to make data usable for analysis is called **(data type conversion)**. Typical examples include the following.

Type of conversion	Description	Example
Conversion from string to numeric	Convert numbers stored as strings into numeric values that can be used in calculations	“123” → 123
Unification of date format	Align dates with different notations into a single format	“2026/4/1,” “2026-04-01,” “01 Apr 2026” → 2026-04-01
Unification of notation	Combine values with the same meaning into a single notation	A mix of “Male/Female” and “M/F” → standardized to “Male/Female”

(2) If formats are not aligned, data cannot be aggregated correctly. For example, if numbers are stored as strings, sums and means cannot be calculated; if date notations are not unified, data from the same day will be treated as belonging to different days. For this reason, data type conversion is a step that must be checked before analysis.

5. Normalization and Standardization

(1) When the units or ranges (scales) of data differ greatly between variables, variables with larger values may strongly influence the analysis results. To prevent this, processing is performed to align the scale of data. Typical methods are **(normalization)** and **(standardization)**.

Method	Description	Main purpose
Normalization	Convert data to a range from 0 to 1	Align data with different units into a comparable form
Standardization	Convert data to a form with mean 0 and standard deviation 1	Position values relative to the spread of the data

(2) For example, when analyzing variables with different units and scales together, such as height (cm) and weight (kg), normalization can align both into the range from 0 to 1, making them easier to compare. Specifically, if height data ranges from 160 cm to 180 cm, normalizing converts 160 cm → 0, 170 cm → 0.5, 180 cm → 1, so the minimum becomes 0 and the maximum becomes 1.

(3) Standardization aligns data based on how far each value is from the mean. For example, when test scores from multiple subjects are standardized, even if the mean and spread differ between subjects, the position of each student can be compared.

6. Reshaping Data According to Purpose

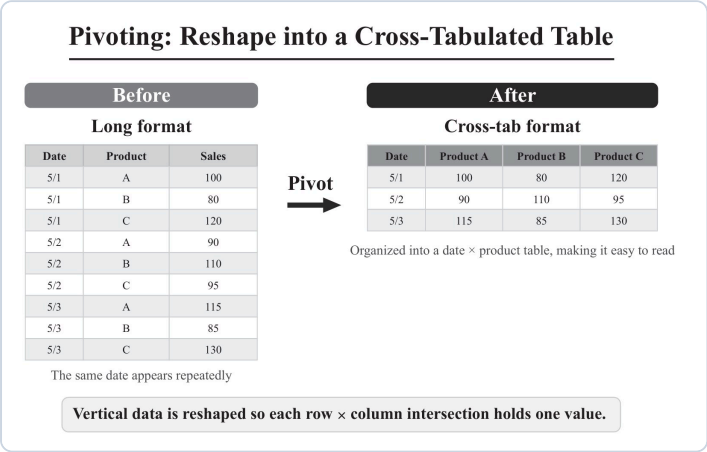
(1) Processing that changes the arrangement or grouping of data according to the purpose of the analysis is called **(data reshaping)**. The following three operations are typical.

PAUSE & THINK

Height in cm and weight in kg — why might the bigger numbers dominate an analysis?

Normalization rescales variables to a common range.

Operation	Description	Example
(Pivot)	Reshape vertically arranged data into a table of row × column combinations	Reshape data with date, product, and sales arranged vertically into a “date × product” table
(Join)	Connect multiple datasets via a common item	Connect a student roster and a grade sheet by “student ID” into a single table
(Aggregation)	Calculate sums, means, etc. for each group and summarize	Compute monthly totals from daily sales



Pivoting: reshaping long-format data into a cross-tabulated table

(2) Which operation to perform depends on the purpose of the analysis. For example, if you want to see “monthly sales trends,” you need to aggregate daily sales data by month. It is important to clarify the purpose of the analysis first, and then plan how to reshape the data.

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Data cleaning is work performed before data collection.

B The responses to missing values are deletion, imputation, and flagging—three in total.

C All outliers are caused by input or measurement errors, so they must always be deleted.

D Standardization is processing that converts data to a form with mean 0 and standard deviation 1.

(2) Match each description (a–c) with the most appropriate processing from the choices below (A–C).

a Conversion of data to a range from 0 to 1

b Processing that connects different datasets via a common item

c Processing that calculates sums, means, etc. for each group and summarizes

💡 PAUSE & THINK

Is a store's one huge sales day always an error — or could it be real?

[Choices] **A** Aggregation **B** Join **C** Normalization

✓ Solution

(1)

A Data cleaning is the work of arranging collected data, and is performed after collection. Therefore, ×.

B Correct as the responses to missing values. Therefore, ○.

C Outliers include both abnormal values caused by input errors and values that reflect reality. They are not always deleted. Therefore, ×.

D Correct definition of standardization. Therefore, ○.

(2)

a: Conversion to 0–1 is normalization. **C**

b: Connecting via a common item is join. **B**

c: Summarizing by group is aggregation. **A**

Try

1 Answer the following questions.

1. What is the term for the processing that fills in the missing parts of data with values such as the mean or median?
2. What is the term for the processing that converts data to a range from 0 to 1?
3. What is the term for the processing that converts data to a form with mean 0 and standard deviation 1?
4. What is the term for the operation that connects multiple datasets via a common item?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that is NOT an appropriate response to missing values.

A When there are few missing values and the impact on the whole is small, delete rows containing missing values.

B When missing values are scattered randomly, fill in the missing parts with values such as the mean.

C When the missingness itself has meaning, leave a marker indicating that the value was missing and analyze the data.

D Impute the mean and analyze data on items where most of the data is missing.

EXAM-STYLE QUESTION

Explain, using "garbage in, garbage out," why data preparation matters as much as the analysis itself. **[6] Refer to:** one cleaning fix and one transformation.

(2) For each of situations (a, b), choose the most appropriate data processing from the choices below (A, B).

a You want to align date data with mixed notations such as “2026/4/1,” “2026-04-01,” and “01 Apr 2026.”

b You want to compare variables with different units and scales together, such as height (cm) and weight (kg).

[Choices] **A** Data type conversion (standardization of date format)
B Normalization

(3) From the following options (A–D), choose the one that most appropriately describes the handling of outliers.

A Outliers are all errors in the data, so delete all of them as soon as they are found.

B For outliers, identify the cause first; correct or delete them if they are due to input errors, and in principle keep them if they reflect reality.

C Outliers do not affect analysis, so no special handling is needed.

D Outliers can always be resolved by replacing them with the mean.

Think as an Engineer — Investigate & Decide

Your class collected a survey on daily screen time, but the raw table is messy: some cells are blank, one entry reads “9999” hours, dates are written three different ways, and two rows look identical. You prepare the data for analysis.

☐ **Analyze the raw table as is** ☐ **Clean and transform it first**

1 · Inspect the data. List each problem you can see find; for each, say which category it is (missing value, outlier, duplicate, or format issue) and note in one phrase how it would distort the analysis if left unfixed.

2 · Decide a fix and its risk. For the blank cells, choose deletion, imputation, or flagging and justify it from the proportion of blanks. For “9999,” say how you would check its cause — a typo, a wrong unit (minutes entered as hours), or a code entered for a refused answer — before removing it. Name one respondent whose true answer a careless fix might erase.

3 · Plan the transformation. State the one format conversion you must run so the dates aggregate correctly, and whether you would normalize the screen-time column. **In pairs,** compare cleaning plans and agree on which single fix carries the greatest risk of distorting the results — and why.

Give evidence for your answer.

PAUSE & THINK

If most of a column is blank, does filling it with the mean help — or invent data?

KEY TERMS

pivot — long rows → row × column table.
join — link datasets by a common item.
aggregation — summarize by group.

Stuck? Handle one problem type at a time in order: unify formats first, then check the outlier's cause, then decide the blanks, then remove only true duplicates — and never delete a value before you know why it is there.

🌍 In a New Context

A shop's sales table lists each day's sales for products A, B, and C on separate rows, but the manager wants one row per day with a column per product. Name the reshaping operation needed, and name the operation instead required to get each product's monthly total.

❓ Lesson Question — Answered

The lesson asked why data must be cleaned and transformed before analysis and what choices each fix involves. Collected data is rarely usable as is, and flawed data yields flawed conclusions no matter how careful the later calculation — garbage in, garbage out. Cleaning handles missing values (by deletion, imputation, or flagging, chosen from their proportion and meaning), outliers (judged by cause: correct or delete errors, keep real values), and duplicates (removed only when genuinely duplicated). Transformation makes data comparable and analyzable: data type conversion unifies formats (string to numeric, date formats, notation), normalization rescales to 0–1 and standardization to mean 0 and standard deviation 1 so no variable dominates by size, and reshaping — pivot, join, aggregation — arranges data to fit the question. Each step is a judgment guided by the purpose, not a fixed rule.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

If collected data is used directly in analysis, incorrect conclusions may be drawn. For this reason, (a), which addresses issues, and (b), which arranges the format to fit the purpose of the analysis, are necessary. For missing values, an appropriate response is chosen from deletion, (c), and flagging, depending on the proportion of missing values and the purpose of the analysis. Also, when

⭐ REFLECT & CHALLENGE

Reflect: Why can “filling in” missing data sometimes mislead more than leaving it out? **Challenge:** You have one survey column where 70% of answers are blank. Argue whether to impute, flag, or drop the column, and state the one condition under which your choice would change.

units or ranges differ greatly between variables, the scale is aligned through (d), which converts data to a range from 0 to 1, or (e), which converts data to a form with mean 0 and standard deviation 1.

1. Fill in the blanks (a)–(e).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that corresponds to data type conversion.

A Filling in missing values with the mean.

B Standardizing notations such as “2026/4/1” and “01 Apr 2026” into “2026-04-01.”

C Aligning the scales of height and weight to between 0 and 1.

D Aggregating daily sales into monthly totals.

(2) From the following options (A–D), choose the one that most appropriately describes normalization and standardization.

A Normalization is processing that converts data to a form with mean 0 and standard deviation 1.

B Standardization is processing that converts data to a range from 0 to 1.

C Normalization is used to align data with different units into a comparable form.

D Standardization is used to enlarge the scale of variables.

(3) For each processing (a–c), indicate which corresponds to pivot, join, or aggregation.

a Connecting a student roster and a grade sheet by “student ID” into a single table.

b Calculating monthly totals from daily sales data.

c Reshaping daily product sales into the form of a “date × product” table.

[Choices] **A** Pivot **B** Join **C** Aggregation

⭐ **Key takeaway**

Remember:
Clean first (missing values, outliers, duplicates), then transform (convert types, normalize or standardize, reshape by pivot, join, aggregation). Every fix is a judgment set by the purpose — garbage in, garbage out.

Learning Objectives

- Explain what open data is and identify major reliable sources of it.
- Describe how data is obtained through an API, and explain the importance of terms of use and source citation.

Some of the most powerful data in the world is free to use — population figures from the World Bank, global statistics from the UN, competition datasets on Kaggle. But "free" is not the same as "no rules." To rely on open data you need to know where to get it, how to fetch it (a download in a browser, or an API that returns data directly to your program), and how you are allowed to use it. Checking the terms of use and citing your source are what turn borrowed data into trustworthy analysis.

? Guiding question: Where does reliable open data come from, how do programs fetch it, and what rules govern its use?

📌 Point!

1. Using Open Data

(1) To use open data in analysis, it is necessary to understand where to obtain it, how to obtain it, and how to use it. By correctly observing these points, reliable analysis can be performed.

(2) Open data comes from a variety of sources, such as data from international organizations, data from government agencies, and datasets published by researchers and companies. It is important to choose data sources with high reliability that match the purpose of the analysis.

2. Major Open Data Sources

(1) The following are typical, internationally widely used providers of open data.

Provider	Description	Main uses
(World Bank Open Data)	Statistics on economy, population, education, and health for countries around the world	International comparisons and analysis of social/economic trends in each country
(UN Data)	International statistics published by the United Nations and its related agencies	Analysis of global issues such as population, development, and environment
(Kaggle)	A platform that runs data analysis competitions; datasets in various fields are made public	Acquisition of data for machine learning study and research

(2) On these platforms, data can be searched for and downloaded from a browser, and in many cases data can also be obtained through APIs.

3. Obtaining Data via APIs

(1) **(API (Application Programming Interface))**…rules for how pieces of software interact with each other. In web services, APIs are also widely used as a

Key Fact:
Open data comes from official providers (World Bank Open Data, UN Data) and platforms such as Kaggle, whose reliability varies; get it from a browser or through an API (often returning JSON); every dataset has terms of use to follow and a source to cite.

Key Concepts:

- open data
- World Bank Open Data / UN Data / Kaggle
- API / endpoint / JSON
- terms of use
- source citation

KEY TERMS

open data — data made public for anyone to use.
World Bank Open Data — country statistics on economy, population, and more.

Open data is public and free to reuse.

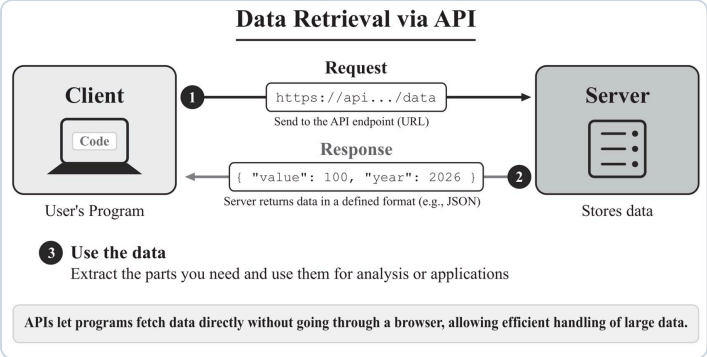
means of obtaining data. By using APIs, necessary data can be obtained directly without going through a browser, allowing large amounts of data to be handled efficiently.

(2) Obtaining data via API proceeds according to the following flow.

① The client (the user's program) sends a request to the API endpoint (the URL that is the destination of the request).

② The server processes the request and returns the requested data in a predetermined format.

③ The client receives the response, extracts the necessary parts, and uses them.



Data retrieval via API: request, response, and use

(3) The data format most widely used for data returned via APIs today is **(JSON (JavaScript Object Notation))**. JSON is a notation that represents data as combinations of names and values; it is easy for humans to read and easy to handle in various programming languages.

(4) When using an API, it is important to **(follow the terms of use)**, not send excessive requests in a short time, and confirm the scope of allowed use of the obtained data.

4. Data Use and Source Citation

(1) Even with open data, the scope of free use is set for each dataset. Always confirm the **(terms of use)** of the data before using it.

(2) When using data, it is required to **(cite the source)**. The source generally includes the data provider, the name of the data, the date of acquisition, and the URL.

(3) Citing the source is important both as an indication of the reliability of the data and to allow others to verify the analysis.

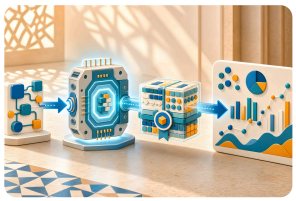
✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A World Bank Open Data publishes statistics on economy, population, education, and so on for countries around the world.

KEY TERMS

API — rules for how software interacts, used to fetch data.
endpoint — the URL a request is sent to.
JSON — a common name/value data format.



An API delivers fresh data straight into analysis.

PAUSE & THINK

"Open" data still has rules — what two things must you check before you use it?

Use data responsibly and credit its source.

B Using an API, necessary data can be obtained directly without going through a browser.

C As long as data is open data, it can be used freely without checking the terms of use.

D When using data, it is required to cite the provider, the name of the data, the date of acquisition, the URL, and so on as the source.

(2) Match each description (a–c) with the most appropriate open data provider from the choices below (A–C).

a Publishes statistics on economy, population, education, and so on for countries around the world.

b Provides international statistics published by the United Nations and its related agencies.

c A platform that runs data analysis competitions.

[Choices] **A** UN Data **B** World Bank Open Data **C** Kaggle

PAUSE & THINK

In an API call, who sends the request and who returns the data?

✓ Solution

(1)

A Correct description of World Bank Open Data. Therefore, ○.

B Correct description of an API. Therefore, ○.

C Even for open data, the terms of use differ for each dataset, so they must always be checked before use. Therefore, ×.

D Correct content of source citation. Therefore, ○.

(2)

a: Statistics on countries around the world is World Bank Open Data. **B**

b: Data from the United Nations and its related agencies is UN Data. **A**

c: The platform that runs data analysis competitions is Kaggle. **C**

EXAM-STYLE QUESTION

Explain why an analyst must check the terms of use and cite the source even when the data is free and open. **[6] Refer to:** your right to use the data, reliability, and letting others verify.

Try

1 Answer the following questions.

1. What is the name of the World Bank's open data portal that publishes statistics on economy, population, education, and so on for countries around the world?
2. What is the term for the rules for how pieces of software interact with each other, which is also widely used in web services for obtaining data?
3. What is the name of the data format widely used when data is returned from an API?
4. What is the term for what should be cited when using data, both to indicate reliability and to allow verification by others?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the flow of obtaining data via an API.

A The client sends a request to the server, and the server returns the data in a format such as JSON.

B The server one-sidedly continues to send data to the client.

C Clients exchange data directly with each other.

D The server only stores data; clients cannot retrieve it.

(2) For each of purposes (a, b), choose the most appropriate open data provider from the choices below (A–C).

a You want to compare GDP across countries.

b You want to find datasets from various fields for machine learning study.

[Choices] **A** UN Data **B** World Bank Open Data **C** Kaggle

(3) From the following options (A–D), choose the one that most appropriately describes the use of open data.

A Open data can be used freely, so the terms of use do not need to be checked.

B Source citation is required only for data published for a fee.

C Even when using open data, it is required to cite the provider, the name of the data, the date of acquisition, the URL, and so on as the source.

D For data obtained via an API, there is no need to follow the terms of use.

PAUSE & THINK

Which provider fits which purpose — and would you use a browser or an API?

Think as an Engineer — Investigate & Decide

For a school project you want to show how Internet access in Egypt has changed over the past ten years compared with two neighboring countries.

You must choose an open data source and a way to obtain it.

☐ **Take the first dataset you find** ☐ **Choose a reliable source and follow its rules**

1 · Inspect two sources. Look up two candidate providers for this exact question (for example World Bank Open Data and UN Data). For each, note whether it actually holds Internet-access figures by country and year, and how current the latest year is.

2 · Analyze fit and rules. Compare the two on a point the lesson table does not list — such as how far back the data goes, or whether its terms of use allow you to republish a chart in a public report. Name one risk of picking the wrong source here.

3 · Decide. Choose one provider and one fetch method (browser download or API) and justify both from what you found. Write the source citation you would include (provider, data name, date obtained, URL). **In pairs,** compare choices and agree on the more defensible source.

Give evidence for your answer.

Stuck? Start from the question, not the source: list exactly what fields you need (country, year, Internet-access rate), then keep only the providers that actually carry all three — and check each one's terms before you commit.

KEY TERMS

terms of use — the rules set for each dataset.
source citation — provider, name, date, URL.

🌍 In a New Context

A weather app on a phone shows today's forecast by fetching it from a public weather service. Describe the request-and-response steps between the app and the server, and name the data format the server most likely returns.

❓ Lesson Question — Answered

The lesson asked where reliable open data comes from, how programs fetch it, and what rules govern its use. Reliable open data comes from trusted providers chosen to fit the purpose — World Bank Open Data for country economic and social statistics, UN Data for United Nations international statistics, and Kaggle for datasets used in machine learning study. It can be downloaded through a browser or fetched by a program through an API: the client sends a request to an endpoint, and the server returns the data in a set format, most often JSON, which the client then uses. Because open data still comes with rules, every dataset carries terms of use that must be followed — including not overloading an API with requests — and every use requires citing the source (provider, data name, date of acquisition, URL), which signals reliability and lets others verify the analysis.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

To use open data in analysis, it is necessary to understand where to obtain it, how to obtain it, and how to use it. Statistics on economy, population, education, and so on for countries around the world can be obtained from sources such as (a), operated by the World Bank. Also, in addition to downloading individually from a browser, by using (b), necessary data can be obtained directly in formats such as (c) in response to a request. When using open data, the (d) determined for each dataset must always be checked, and when data is used, the provider, the name of the data, the date of acquisition, the URL, and so on must be cited as the (e).

1. Fill in the blanks (a)–(e).

⭐ REFLECT & CHALLENGE

Reflect: Think of a chart you saw online with no source. Did you trust it — and what one line would have changed that?
Challenge: You may cite only one source line for a chart built from a World Bank dataset you downloaded today. Write that citation line, and say which single element you would never drop and why.

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes Kaggle.

A A portal operated by the World Bank that publishes statistics for countries around the world.

B A site for international statistics operated by the United Nations.

C A platform that runs data analysis competitions, on which datasets in various fields are made public.

D A system that automatically records browsing histories of websites.

(2) From the following options (A–D), choose the one that is NOT an appropriate description of source citation.

A Source citation is important as an indication of the reliability of the data.

B Source citation is important to allow others to verify the analysis.

C Sources include the data provider, the name of the data, the date of acquisition, the URL, and so on.

D When using open data, it is not necessary to cite the source.

(3) From the following options (A–D), choose the one that is NOT an appropriate point of caution when using an API.

A Use the API in accordance with the terms of use determined by the provider.

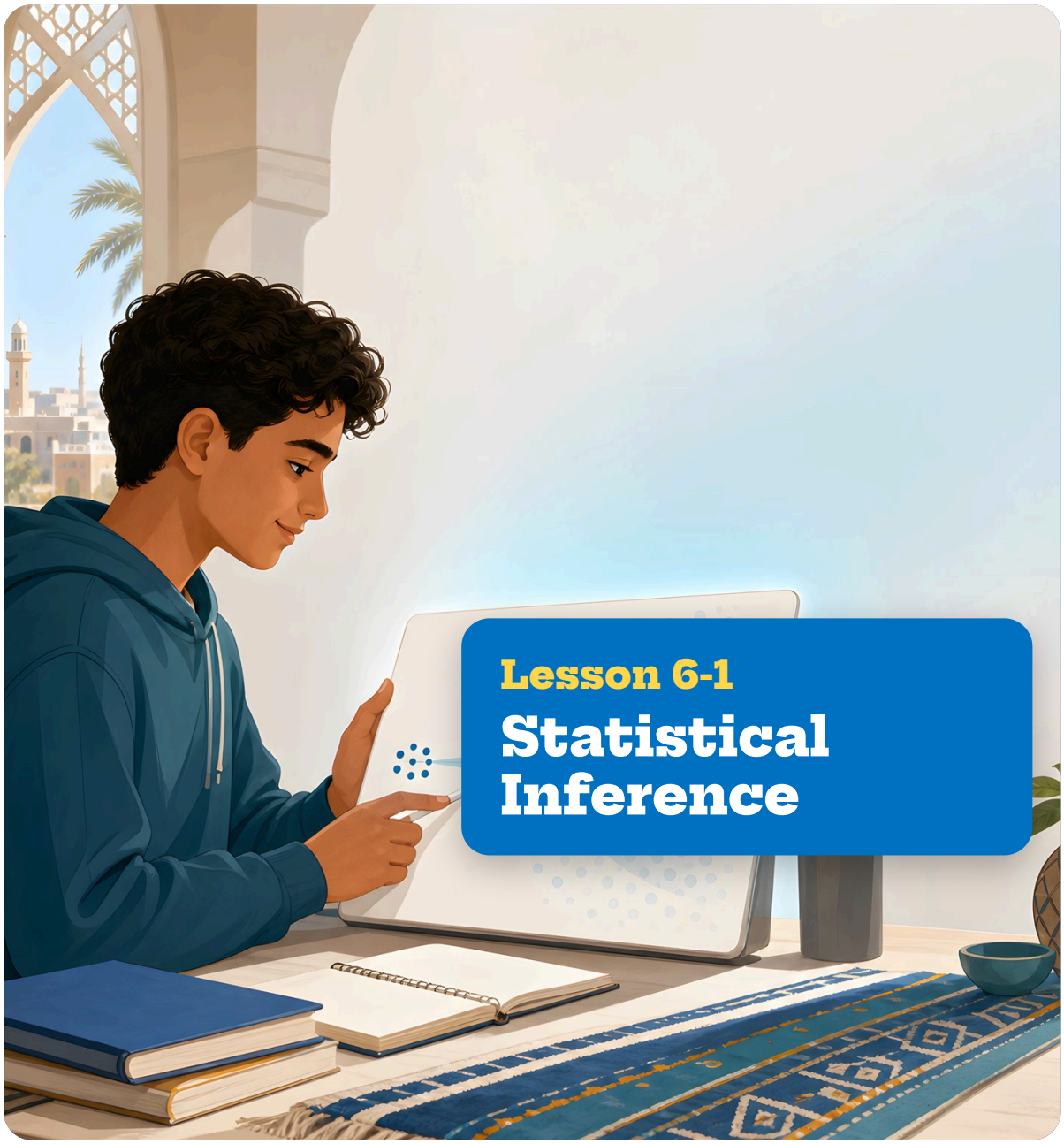
B Avoid sending excessive requests in a short time.

C Confirm the scope of allowed use of the obtained data in advance.

D Because of how APIs work, they may be used without checking the terms of use.

⭐ Key takeaway

Remember:
Pick a reliable open data source for your purpose (World Bank Open Data, UN Data, Kaggle), fetch by browser or API (client requests an endpoint; server returns JSON), and always follow the terms of use and cite the source.



Learning Objectives

- Explain the difference between point estimation and interval estimation, including the meaning of a confidence level.
- Describe the procedure of hypothesis testing, and explain the cautions needed when interpreting its results.

You can never survey everyone, so you study a sample and reason back to the whole population — that is statistical inference. It answers two kinds of question: "about how big?" (estimation) and "is this real, or just chance?" (hypothesis testing). But because a sample is only part of the whole, every answer carries uncertainty. The skill is not to hide that uncertainty but to measure it — with confidence intervals, p-values, and significance levels — and to remember that "statistically significant" is not the same as "important."

? Guiding question: How can we reason from a sample to a whole population, and how do we measure the uncertainty in that reasoning?

📌 Point!

1. The Statistical Inference Stage

(1) A sample is drawn from the entire group that is the target of analysis (the population) and surveyed, and the properties of the population are then inferred from the results of the sample. Inferring the properties of the population from the sample in this way is called **(statistical inference)**.

(2) Statistical inference can be broadly divided into the following two approaches.

Approach	Description	Example
(Estimation)	Infer the properties of the population in numerical terms	Infer the population mean height from the sample mean height
(Hypothesis testing)	Determine whether a hypothesis about the population is correct	Determine whether a new learning method is more effective than the conventional one

(3) In either case, since the sample is only part of the population, inference always involves uncertainty. In statistical inference, judgment is made while considering the degree of that uncertainty.

2. Point Estimation and Interval Estimation

(1) **(Point estimation)**…estimating the properties of the population as a single value calculated from the sample. *e.g.,* If the sample mean height is 168 cm, then the population mean height is also estimated to be 168 cm.

(2) **(Interval estimation)**…estimating that the properties of the population fall within a certain range (interval). *e.g.,* The population mean height is estimated to be in the range of 166 cm to 170 cm.

(3) The interval used in interval estimation is called the **(confidence interval)**, and how often such intervals contain the population value over repeated surveys is called the **(confidence level)**. Levels of 95% or 99% are common.

Key Fact:
Statistical inference reasons from a sample to a population. Estimation gives a value (point) or a range (interval, with a confidence level); hypothesis testing weighs a null against an alternative using a p-value and a significance level. All of it carries uncertainty — and significance is not the same as practical importance.

Key Concepts:

- statistical inference
- point / interval estimation
- confidence interval & level
- null / alternative hypothesis
- p-value & significance level
- Type I / II errors

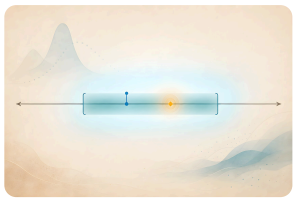
KEY TERMS
statistical inference — reasoning from a sample to the population.
estimation — inferring the population in numbers.
hypothesis testing — judging a claim about the population.

💡 PAUSE & THINK

A single value or a range — which estimate shows you how uncertain it is?

e.g., Suppose the population mean height is estimated by sample survey, and a confidence interval of “166 cm to 170 cm” is obtained at a 95% confidence level. The “95%” means that if the same survey method is repeated 100 times, about 95 of the resulting confidence intervals will contain the true population mean height.

(4) Point estimation is easy to understand because it is shown as a single value, but the uncertainty of the estimate is invisible. Interval estimation has width, so care is needed in interpreting the result, but it has the advantage of also showing the degree of uncertainty.



A point estimate is one value; an interval is a range.

3. The Idea of Hypothesis Testing

(1) Hypothesis testing…a procedure for determining whether a hypothesis about the population is acceptable based on the data at hand.

(2) In hypothesis testing, the following two hypotheses are set up.

Hypothesis	Description
(Null hypothesis) (H0)	A hypothesis serving as the starting point of the test, such as “no difference” or “no relationship”
(Alternative hypothesis) (H1)	The hypothesis adopted when the null hypothesis is rejected, such as “there is a difference” or “there is a relationship”

(3) In hypothesis testing, the probability that data collection could result in the data at hand is evaluated under the assumption that “there is no difference or relationship.” The reason for first assuming the hypothesis one wants to deny (the null hypothesis) is that, rather than directly showing “there is a difference or relationship,” it is logically easier to judge this by showing that, under the assumption “there is no difference or relationship,” the data at hand would be unlikely to be obtained. *e.g.,* It is difficult to directly show “the new learning method is effective.” Instead, one assumes “there is no effect” and then examines how low the probability is of obtaining a difference like the experimental result at hand, and if that probability is sufficiently small, one judges that “no effect” is unlikely.

(4) Based on the evaluation, the null hypothesis is judged to be either **(rejected (denied))** or **(not rejected)**.

Testing weighs evidence against the null hypothesis.

KEY TERMS
null hypothesis (H0) — the “no difference” starting point.
alternative hypothesis (H1) — adopted if H0 is rejected.
p-value — chance of data at least this extreme if H0 holds.

4. The Procedure of Hypothesis Testing

(1) Hypothesis testing follows this flow.

① Set the null hypothesis H0 and the alternative hypothesis H1

② Decide the **(significance level)** (5% = 0.05 and 1% = 0.01 are commonly used)

③ Compute the **(p-value)** from the data at hand

④ Compare the p-value with the significance level and make a judgment

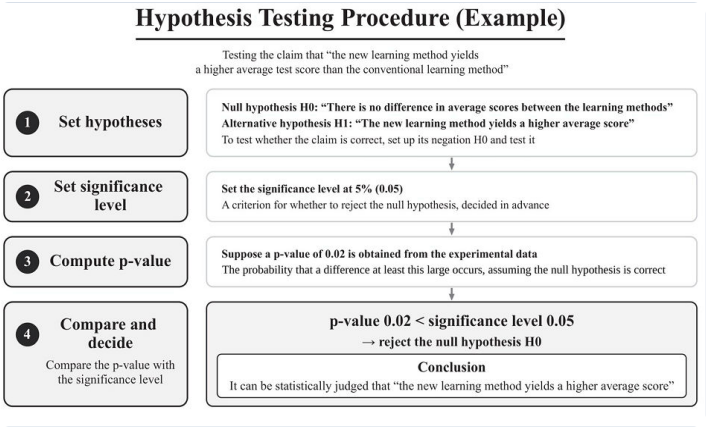
(2) p-value…assuming the null hypothesis is correct, the probability of obtaining a result as extreme as, or more extreme than, the data at hand. The

smaller the p-value, the less likely it is that the data would be obtained under the null hypothesis.

(3) Significance level…the threshold value used to decide whether to reject the null hypothesis.

(4) Criteria for judgment

Condition	Judgment
p-value < significance level	Reject the null hypothesis and adopt the alternative hypothesis (statistically significant)
p-value ≥ significance level	Do not reject the null hypothesis (cannot be said to be statistically significant)



The hypothesis testing procedure worked through an example

5. Major Hypothesis Testing Methods

(1) The test method used differs depending on the purpose of the analysis and the type of data. Major examples include the following.

Test	Main use	Example
(t-test)	Examine whether there is a difference in the means of two groups	Examine whether students' average test scores differ between a certain learning method and the conventional method
(Chi-squared test)	Examine whether there is a relationship between two categorical variables (data expressed by classification, such as gender, occupation, or preference)	Examine, from survey results, whether there is a relationship between gender and subject preference

6. Cautions Regarding Inference

(1) In statistical inference, there is always a possibility of judgment errors. There are two types of errors.

Type of error	Description
(Type I error)	Rejecting the null hypothesis when it is actually correct
(Type II error)	Not rejecting the null hypothesis when it is actually incorrect

PAUSE & THINK

If the p-value is 0.02 and the significance level is 0.05, do you reject the null or not?

PAUSE & THINK

Which test compares two group means, and which checks a relationship between categories?

KEY TERMS

Type I error — rejecting a true null hypothesis.
Type II error — keeping a false null hypothesis.
trade-off — lowering one error raises the other.

(2) Lowering the significance level reduces the Type I error, but tends to increase the Type II error. The two have a **(trade-off relationship)**.

(3) Care is needed in interpreting the p-value. The following misunderstandings must be avoided.

① The p-value is not “the probability that the null hypothesis is correct.”

② A small p-value does not mean “the effect is large.”

③ The comparison of the p-value with the significance level is not an absolute correctness criterion but a guide for judgment. The significance level (e.g., 0.05 or 0.01) is a value chosen according to the purpose of the analysis or the conventions of the field; falling below this value does not mean the result is certainly correct.

(4) Being statistically significant and being practically important are different. *e.g.,* Suppose an experiment compares two designs of a website, with click-rate data from 100,000 people split evenly. The old design is clicked by 0.4% of visitors, the new one by 0.5% — only 0.1 percentage points higher. Because the data set is large, this difference is likely to be statistically significant. However, whether that 0.1-point difference leads to actual profit or better user experience needs to be judged from a separate perspective.

(5) Statistical inference is a powerful means of drawing conclusions from data, but it is important to understand that it is a tool for judgment that includes uncertainty, and the results should be interpreted appropriately.

Type I is a false alarm; Type II is a miss.

PAUSE & THINK

With 100,000 people, a tiny difference can be “significant.” Does that make it important?

Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Inferring the properties of the population from the sample is called statistical inference.

B Point estimation is a method that estimates the properties of the population within a certain range.

C In hypothesis testing, the probability that data collection could result in the data at hand is evaluated under the assumption that the null hypothesis is correct.

D If a result is statistically significant, the difference is also necessarily important in practice.

(2) Match each description (a–c) with the most appropriate term from the choices below (A–C).

a A hypothesis serving as the starting point of the test, such as “no difference” or “no relationship”

b The probability of obtaining a result as extreme as, or more extreme than, the data at hand, assuming the null hypothesis is correct

c The threshold value for deciding whether to reject the null hypothesis

[Choices] **A** p-value **B** Significance level **C** Null hypothesis

✓ Solution

(1)

A Correct definition of statistical inference. Therefore, ○.

B Estimating within a range is interval estimation. Point estimation estimates a single value. Therefore, ×.

C Correct description of the idea of hypothesis testing. Therefore, ○.

D Statistical significance and practical importance are different; with a large amount of data, even a tiny difference can become significant. Therefore, ×.

(2)

a: A hypothesis serving as the starting point of the test is the null hypothesis. **C**

b: The probability of obtaining the data under the null hypothesis is the p-value. **A**

c: The threshold value for rejection is the significance level. **B**

Try

1 Answer the following questions.

1. What is the term for inferring the properties of the population from the sample?
2. What is the term for the method of estimating that the properties of the population fall within a certain range?
3. What is the term for the interval used in interval estimation?
4. What is the term for a typical hypothesis testing method that examines whether there is a relationship between two categorical variables?

EXAM-STYLE QUESTION

Explain why researchers test a null hypothesis they hope to disprove, rather than proving the effect directly. **[6] Refer to:** what a small p-value shows.

2 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the procedure of hypothesis testing.

A Set the null and alternative hypotheses → decide the significance level → compute the p-value → make a judgment

B Compute the p-value → set the null hypothesis → decide the significance level → make a judgment

C Decide the significance level → adopt the alternative hypothesis → compute the p-value → set the null hypothesis

D Make a judgment → set the null hypothesis → decide the significance level → compute the p-value

(2) Suppose that, in a hypothesis test with the significance level set at 0.05 (5%), the p-value is 0.03. From the following options (A–D), choose the one that most appropriately describes the judgment in this case.

A Reject the null hypothesis and adopt the alternative hypothesis.

B Do not reject the null hypothesis.

C Reject the alternative hypothesis.

D Change the significance level to 0.01 and recalculate.

(3) For each of analysis purposes (a, b), choose the most appropriate test method from the choices below (A, B).

a You want to examine whether students' average test scores differ between two teaching methods.

b You want to examine whether there is a relationship between gender and the category of products purchased.

[Choices] **A** Chi-squared test **B** t-test

(4) From the following options (A–D), choose the one that most appropriately describes the p-value.

A The p-value is the probability that the null hypothesis is correct.

B The smaller the p-value, the larger the effect is.

C The p-value is the probability of obtaining a result as extreme as, or more extreme than, the data at hand, assuming the null hypothesis is correct.

D p-value < 0.05 is an absolute criterion of correctness.

Think as an Engineer — Investigate & Decide

A school club claims its new online quiz tool raises students' average score on weekly tests. You are asked to judge the claim fairly using a hypothesis test.

☐ **Trust the club's own summary** ☐ **Frame and test the claim properly**

PAUSE & THINK

p-value 0.03 against a 0.05 threshold — below or above? So what do you do?

KEY TERMS

significance level — the threshold for rejecting H0.
statistically significant — p below the threshold.
practically important — big enough to matter in real life.

1 · Frame the hypotheses. The club reports its tool group averaged 74 on the last quiz versus 70 for a control group of the same size. Write the null and alternative hypotheses for this exact claim in your own words, state what data you would still need to actually run the test (beyond these two means), and say which significance level you would set (0.05 or 0.01) and why.

2 · Analyze the risks. Describe, in this context, what a Type I error and a Type II error would each mean for the club, and say which one you would most want to avoid here — a judgment the lesson's tables do not make for you.

3 · Decide and interpret. Suppose the test returns a p-value of 0.04. State your decision against the significance level you chose in step 1 (note it would differ at 0.05 versus 0.01). Then, if the result is significant, explain why even a significant result might still not justify switching every class to the tool. **In pairs,** compare interpretations and agree on what evidence beyond the p-value you would still want.

Give evidence for your answer.

Stuck? Start from what you are trying to disprove: the null is always “no difference.” Then ask what it would cost to reject it wrongly (Type I) versus to miss a real effect (Type II) — that cost is what should guide your significance level.

🌍 In a New Context

A shop tests two checkout-button colors with data from 200,000 visitors and finds the new color's click rate is 0.05 percentage points higher, with a p-value of 0.01 at the 5% significance level. State whether the result is statistically significant, and explain why the shop might still decide not to change the button.

❓ Lesson Question — Answered

The lesson asked how we reason from a sample to a whole population and how we measure the uncertainty in that reasoning. Because a sample is only part of the population, we use statistical inference — either estimation, which reports the population as a single value (point estimation) or as a range (interval estimation) whose confidence level says how often such intervals would contain the true value, or hypothesis testing, which sets a null hypothesis ("no difference") against an alternative and compares a p-value with a chosen significance level to decide whether to reject the null. The uncertainty is never removed: it is measured and reported. Judgment can fail as a Type I error (rejecting a true null) or a Type II error (keeping a false null), and lowering one raises the other. And a result that is statistically significant — especially with very large samples — is not automatically important in practice, so results must be interpreted, not just computed.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

In statistical inference, the properties of the population are inferred from the sample. Estimating the properties of the population as a single value is called (a), and estimating within a certain range is called (b). Also, the procedure of judging a hypothesis about the population is called (c). In hypothesis testing, the (d)—such as "no difference" or "no relationship"—and the opposite (e) are set up, and a judgment is made by comparing the (f) computed from the data at hand with the predetermined (g).

1. Fill in the blanks (a)–(g).

⭐ REFLECT & CHALLENGE

Reflect: Why is a 95% confidence level a statement about the method, not about one particular interval?
Challenge: A study of 500,000 users reports a "highly significant" effect ($p < 0.001$) whose size is 0.02 percentage points. Argue whether you would act on it, and name the one extra piece of information that would change your mind.

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes confidence intervals and confidence levels.

A A confidence interval with a 95% confidence level means that the true population value is contained in that interval with a probability of 95%.

B The higher the confidence level, the narrower the confidence interval.

C A confidence interval with a 95% confidence level means that, if the same survey method is repeated, about 95% of the resulting confidence intervals will contain the true population value.

D The confidence level is a value used as the criterion for judgment in hypothesis testing.

(2) From the following options (A–D), choose the one that is NOT an appropriate description of errors in hypothesis testing.

A A Type I error is rejecting the null hypothesis when it is actually correct.

B A Type II error is not rejecting the null hypothesis when it is actually incorrect.

C Lowering the significance level can simultaneously reduce both Type I and Type II errors.

D Type I and Type II errors have a trade-off relationship.

(3) From the following options (A–D), choose the one that most appropriately describes the interpretation of statistical inference.

A A statistically significant difference is necessarily an important difference in practice.

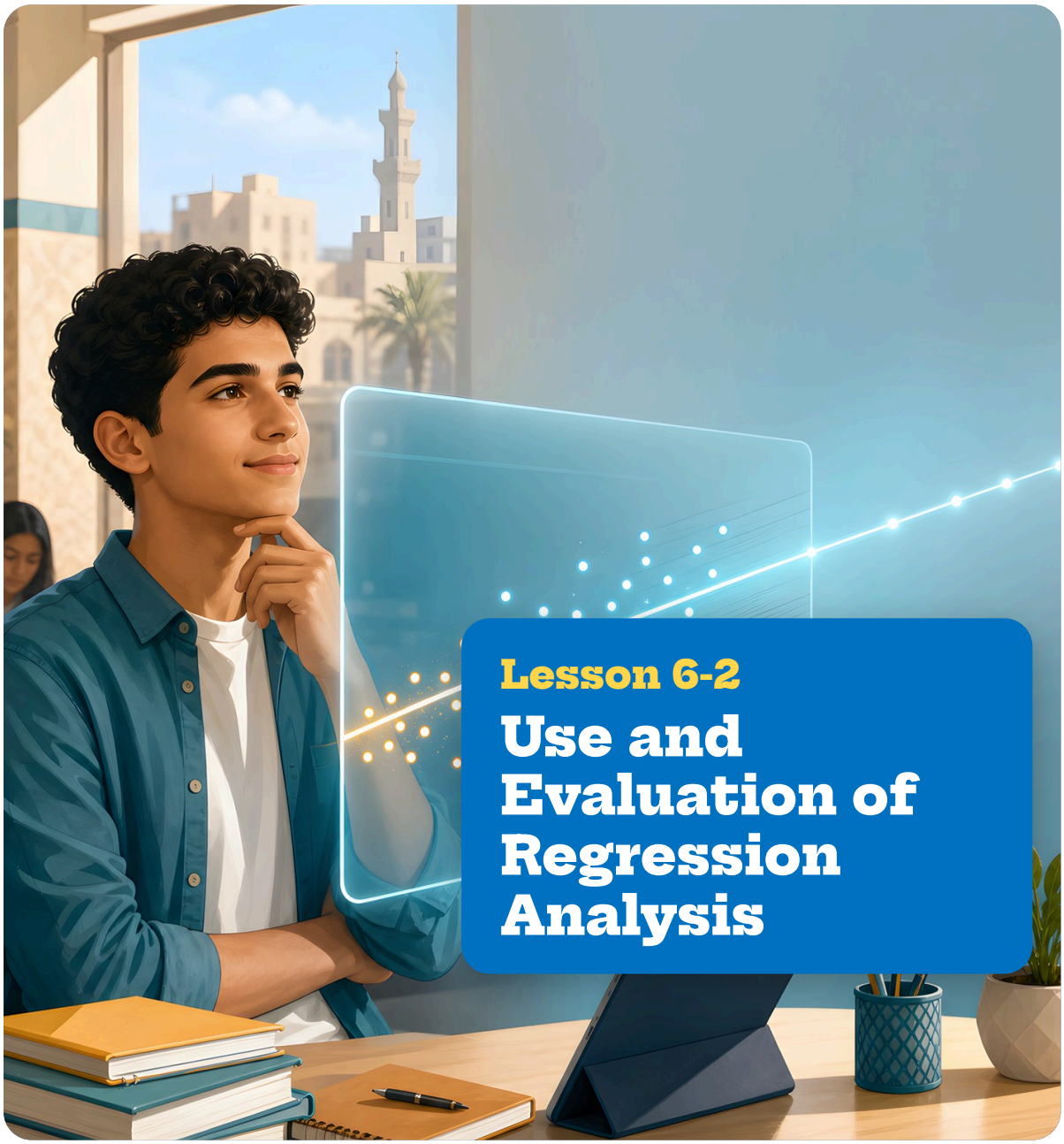
B With a large amount of data, even a tiny difference can become statistically significant.

C If the p-value falls below 0.05, the conclusion is certainly correct.

D Statistical inference can completely remove uncertainty.

⭐ Key takeaway

Remember:
Estimate with a point or an interval (confidence level = how often the method captures the truth); test by comparing a p-value with a significance level to reject or keep the null. Watch Type I / II errors, and never confuse statistical significance with practical importance.



Lesson 6-2

Use and Evaluation of Regression Analysis

Learning Objectives

- Explain the difference between simple and multiple regression, and what a regression model represents.
- Evaluate a regression model using R^2, RMSE, and MAE, and explain why a good fit does not prove causation.

Data often hides a relationship you can turn into a prediction: the warmer the day, the more ice cream sells. Regression analysis writes that relationship as an equation, so you can put in tomorrow's forecast temperature and predict tomorrow's sales. Yet a prediction equation is only a model — a simplified picture — and a careful analyst asks two more questions: how accurate is it (using R^2, RMSE, and MAE), and does "fits well" really mean "causes"? A line that fits the data is not proof of cause.

? Guiding question: How does regression turn data into a prediction, and how do we judge whether that prediction can be trusted?

Point!

1. The Regression Analysis Stage

(1) **(Regression analysis)**…a method that expresses the relationship between variables in the data as an equation, and uses it for prediction and simulation. For example, from past data on temperature and sales, one can express how sales change when temperature rises as an equation, and put tomorrow's forecast temperature into that equation to predict sales.

(2) In regression analysis, the variable that is the result is called the **(response variable) (dependent variable)**, and the variable used to explain or predict it is called the **(explanatory variable)**. In the temperature-and-sales example, sales are the response variable and temperature is the explanatory variable.

(3) The equation obtained from data in regression analysis is called the **(model)**. A model is a simplified representation of a real relationship—for example, putting the trend "the higher the temperature, the higher the sales" into a format that can be calculated.

(4) Both creating a model and evaluating whether the model can actually be used are important. A model that is only created is not useful; only by confirming its accuracy and limitations does it lead to practical use.

2. Simple and Multiple Regression

(1) **(Simple regression analysis)**…a method of predicting the response variable using one explanatory variable.

e.g., Predict ice cream sales (y) from temperature (x) alone. The prediction equation is expressed as a straight line in the format **$y = ax + b$**, and this line is called the **(regression line)**. a and b are determined by the **(least-squares method)** (a method that determines the line so as to minimize the sum of the squares of the discrepancies between the actual values and the regression line, each called a **(residual)**).

(2) **(Multiple regression analysis)**…a method of predicting the response variable using multiple explanatory variables. In real data, results are rarely determined by a single cause, and multiple regression analysis, which can consider several factors simultaneously, is suited to more realistic prediction.

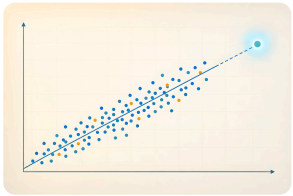
Key Fact:
Regression analysis expresses the relationship between an explanatory variable (the input used to make the prediction) and a response variable (the result) as a model — a line $y = ax + b$ for simple regression, or a sum of weighted variables for multiple regression. Judge the model with R^2, RMSE, and MAE, and never mistake a good fit for proof of causation.

Key Concepts:

- regression analysis
- response / explanatory variable
- simple / multiple regression
- regression line, least-squares
- R^2, RMSE, MAE
- correlation versus causation

KEY TERMS

regression analysis — expressing a relationship as an equation for prediction.
response variable — the result (dependent).
explanatory variable — the input used to make the prediction.



Fit a line to data, then use it to predict within that range.

e.g., Predict housing prices from a house's size, age, and distance from a transit station (multiple factors). The prediction equation is expressed as the sum of multiple explanatory variables, each multiplied by a coefficient (a numerical value representing the strength of its influence), plus an intercept. For example, when there are two explanatory variables, the format becomes $\boldsymbol{y = a_1x_1 + a_2x_2 + b}$ (x_1 and x_2 are the explanatory variables, a_1 and a_2 are their coefficients, and b is the intercept). When there are three or more explanatory variables, terms are added in the same way.

(3) The following differences exist between simple and multiple regression.

Aspect	Simple regression	Multiple regression
Number of explanatory variables	One	Two or more
Characteristics of the model	Simple and easy to interpret	Can express complex relationships
Suitable situations	When you want to capture the relationship between two variables concisely	When the response variable is influenced by multiple factors

3. Evaluation Indicators for Models

(1) To judge how useful a model created by regression analysis is, **(evaluation indicators)** are used. Three typical indicators are as follows.

Indicator	Description (concept of calculation)	Interpretation
(R^2 (coefficient of determination))	Expresses, on a 0–1 scale, how well the model explains the variability in the data	The closer to 1, the better the variation in the data is explained
(RMSE (root mean square error))	The square root of the mean of the squared discrepancies between predicted and actual values	The smaller the value, the higher the prediction accuracy; expressed in the same units as the response variable
(MAE (mean absolute error))	The mean of the absolute values of the discrepancies between predicted and actual values	The smaller the value, the higher the prediction accuracy; less affected by extremely large discrepancies

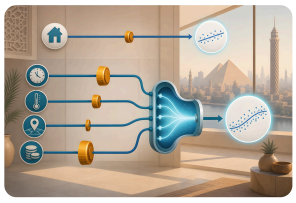
(2) Both RMSE and MAE indicate the size of prediction errors, but RMSE tends to weight large discrepancies more heavily. On the other hand, MAE evaluates discrepancies on average, so it is less affected by outliers. They are used differently according to the purpose of the analysis.

4. Cautions When Using Regression Models

(1) Regression analysis is a widely applicable method, but misuse can lead to incorrect conclusions. Pay particular attention to the following points.

PAUSE & THINK

One cause or several? Which kind of regression would house prices need?



Multiple regression combines several inputs.

KEY TERMS

R^2 — how well the model explains the variability in the data (0–1).
RMSE — error size, weights big misses more.
MAE — average error, less swayed by outliers.

PAUSE & THINK

A prediction line fits the data well — does that prove one variable causes the other?

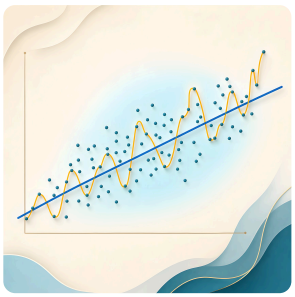
① Do not confuse correlation with causation: a prediction equation that fits does not establish it. Causation needs temporal precedence, exclusion of other factors' influence, and theoretical support.

② A model is not finished once created: Confirm accuracy with evaluation indicators, verify validity in light of the purpose, and do not apply it outside the range of the data it was fitted to.

③ Do not judge a model by R^2 alone: Make judgments together with other indicators.

④ Adding more explanatory variables does not necessarily improve a model: Appropriate variables must be chosen according to the purpose.

(2) The results of regression analysis should not be treated as absolute truth as is, but as one perspective that the data offers. Interpreting results appropriately and using them in real-world judgment is important.



An overfit model chases noise, not the trend.

PAUSE & THINK

Which error measure would you trust more when a few huge outliers exist?

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Multiple regression analysis is a method of predicting the response variable using one explanatory variable.

B R^2 (coefficient of determination) is an indicator showing how well the model explains the variability in the data.

C The equation obtained from data in regression analysis is called the model, and is a simplified representation of a real relationship.

D If the prediction equation fits, it can be said that there is a causal relationship between the explanatory variable and the response variable.

(2) Match each description (a–c) with the most appropriate indicator from the choices below (A–C).

a The mean of the absolute values of the discrepancies between predicted and actual values

b The square root of the mean of the squared discrepancies between predicted and actual values

c How well the model explains the variability in the data

[Choices] **A** RMSE **B** R^2 **C** MAE

✓ Solution

(1)

A Predicting with one explanatory variable is simple regression analysis. Multiple regression analysis uses multiple explanatory variables. Therefore, ×.

B Correct description of R^2. Therefore, ○.

C Correct definition of a model. Therefore, ○.

D Even if the prediction equation fits, that alone does not establish causation. Therefore, ×.

(2)

a: The mean of absolute discrepancies is MAE. **C**

b: The square root of the mean of squared discrepancies is RMSE. **A**

c: An indicator that explains the variability in the data is R^2. **B**

Try

1 Answer the following questions.

1. What is the term for the regression analysis that predicts the response variable using multiple explanatory variables?
2. What is the term for the indicator showing how well the model explains the variability in the data?
3. In regression analysis, what are the variables called that are the result and the input used to predict it, respectively?
4. What is the term for "the discrepancy between actual values and the regression line" used in simple regression analysis?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the difference between simple and multiple regression.

A Simple regression can express complex relationships, while multiple regression is simple.

B Simple regression uses multiple explanatory variables, and multiple regression uses one explanatory variable.

C Simple regression uses one explanatory variable, and multiple regression uses multiple explanatory variables.

D There is no difference between simple and multiple regression in the number of explanatory variables; only the calculation method differs.

(2) For each of analysis purposes (a, b), choose the most appropriate regression analysis method from the choices below (A, B).

a You want to predict ice cream sales from temperature (one variable).

b You want to predict housing prices from a house's size, age, and distance from a transit station (multiple variables).

[Choices] **A** Simple regression analysis **B** Multiple regression analysis

EXAM-STYLE QUESTION

Explain why a regression model with a high R^2 can still be a poor basis for claiming that one variable causes another. **[6] Refer to:** correlation versus causation.

PAUSE & THINK

One variable in, or several? That is the whole difference.

(3) From the following options (A–D), choose the one that most appropriately describes the interpretation of regression analysis results.

A If a prediction equation is obtained from regression analysis, the equation necessarily indicates a causal relationship.

B The results of regression analysis should be treated as one perspective that the data offers, and whether causation has been demonstrated must be carefully evaluated.

C If R^2 is high, the model is necessarily a good model.

D If the prediction equation fits, one may conclude that the explanatory variable is the cause of the response variable.

(4) From the following options (A–D), choose the one that is NOT an appropriate description of evaluation indicators for regression models.

A R^2 (coefficient of determination) takes values between 0 and 1; the closer to 1, the better the variation in the data is explained.

B RMSE is the square root of the mean of squared discrepancies between predicted and actual values, expressed in the same units as the response variable.

C MAE is the mean of the absolute values of the discrepancies between predicted and actual values, and is less affected by extremely large discrepancies.

D If the value of R^2 is close to 1, the model is necessarily a good model without needing to check other indicators.

PAUSE & THINK

A high R^2 alone — enough to trust a model, or do you need RMSE and MAE too?

Think as an Engineer — Investigate & Decide

Your school canteen wants to predict its daily cold-drink sales so it can order the right amount. You plan a regression model.

☐ **Trust one obvious cause** ☐ **Model it and test the model**

1 · Set up the model. Name the response variable, then list the explanatory variables you would try (for example the day's temperature, the number of students present, and whether it is an exam day). Decide whether this is a simple or a multiple regression, and say why.

2 · Judge and question it. Choose which evaluation indicator you would report (R^2, RMSE, or MAE) and justify it for this purpose. Then name one confounder (a confounding variable) — a hidden third factor that could make two variables move together without one causing the other — that could mislead you here.

3 · Decide and interpret. Suppose your model gives a high R^2. State whether that alone would let you claim "hot days cause higher sales," and what else you would need. **In pairs,** compare models and agree on the single most useful explanatory variable to keep.

Give evidence for your answer.

KEY TERMS

coefficient — the weight on an explanatory variable.
intercept — the constant *b* in the equation.
confounder — a hidden third cause behind a correlation.

Stuck? Start from the result you want to predict (sales), then ask which measurable things plausibly move it — and for each, ask "could something else be driving both?" A high R^2 tells you the line fits, not that the cause is real.

🌍 In a New Context

A city finds that months with more ice cream sales also have more swimming accidents, and a regression line fits the two closely. Explain why this does not mean ice cream causes accidents, and name the hidden factor most likely behind both.

❓ Lesson Question — Answered

The lesson asked how regression turns data into a prediction and how we judge whether it can be trusted. Regression analysis expresses the relationship between an explanatory variable (the input used to make the prediction) and a response variable (the result) as a model: a regression line $y = ax + b$ in simple regression (one explanatory variable), or a weighted sum such as $y = a_1x_1 + a_2x_2 + b$ in multiple regression (several variables), fitted by least squares to minimize the sum of squared residuals. Trust is judged with evaluation indicators used together — R^2 for how much variability is explained, and RMSE and MAE for the size of prediction errors (RMSE weighting large misses more, MAE less swayed by outliers) — never R^2 alone, and never by adding variables for their own sake. Above all, a good fit is only a correlation: causation requires temporal precedence, ruling out other factors, and theory. Regression offers one perspective from the data, to be interpreted, not obeyed.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

In regression analysis, the variable that is the result is called the (a), and the variable used to predict it is called the (b). The method that predicts using one

explanatory variable is called (c), and the method that predicts using multiple explanatory variables is called (d). To judge whether a model is good, evaluation indicators are used. Indicators include (e), which shows how well the model explains the variability in the data, and (f) and (g), which show the size of the discrepancies between predicted and actual values. To claim causation from the results of regression analysis, temporal precedence, exclusion of other factors' influence, theoretical support, and so on are necessary, and one must not conclude that there is (h) just because the prediction equation fits.

1. Fill in the blanks (a)–(h).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes correlation and causation.

A If there is correlation, causation necessarily holds.

B To claim causation, temporal precedence, exclusion of other factors' influence, theoretical support, and so on are necessary.

C If a prediction equation is obtained from regression analysis, the equation completely demonstrates causation.

D If there is no correlation, it can be said that there is no relationship at all between the two variables.

(2) From the following options (A–D), choose the one that most appropriately describes the interpretation of evaluation indicators.

A R^2 takes values between 0 and 1; the closer to 1, the better the variation in the data is explained.

B RMSE and MAE use the same calculation method, and their values always coincide.

C A model with a high R^2 is always a good model in any situation.

D MAE is an indicator that weights large discrepancies more heavily.

(3) From the following options (A–D), choose the one that is NOT an appropriate point of caution when using regression models.

A Do not conclude that there is causation just because the prediction equation fits.

B Confirm accuracy with evaluation indicators and verify the validity of the model in light of the purpose.

C If R^2 is close to 1, one may judge the goodness of the model without checking other indicators.

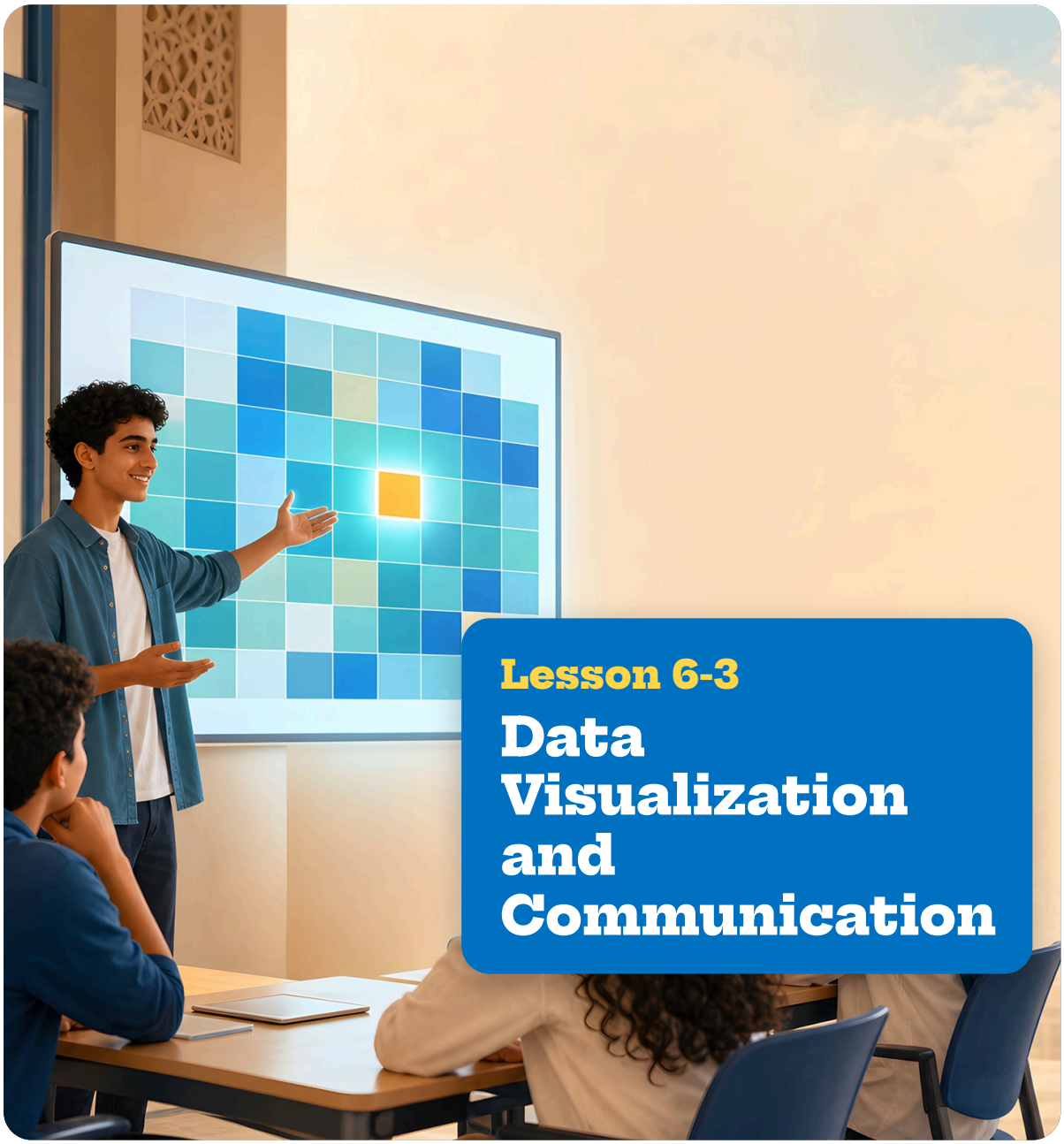
D Adding more explanatory variables does not necessarily improve accuracy. Appropriate variables must be chosen according to the purpose.

REFLECT & CHALLENGE

Reflect: Think of two things you have seen rise together (say, screen time and poor sleep). What hidden factor could drive both? **Challenge:** You may add only one more explanatory variable to a model that already has a high R^2. Argue whether you would add it at all, and state the one check you would run before trusting the bigger model.

⭐ Key takeaway

Remember:
Regression predicts a response variable from one (simple, $y = ax + b$) or several (multiple) explanatory variables, fitted by least-squares. Judge it with R^2, RMSE, and MAE together — and never read a good fit as proof of cause.



Lesson 6-3

Data Visualization and Communication

Learning Objectives

- Explain the role of data visualization and select suitable graphs, including for multiple variables.
- Apply data storytelling to communicate a clear and honest message from data.

A table of numbers can hide the very thing you most need to see. Turn it into the right picture and the pattern jumps out — the busiest hours, the strongest link, the odd outlier. Visualization does two jobs: it helps you understand your own data, and it helps you explain it to others. When one or two variables are involved, a bar, line, or scatter plot is enough; when three or more variables, or many combinations of information, must be shown at once, heatmaps and mosaic plots are used instead. And a chart alone is not a message: telling a clear data story is what makes an audience actually understand.

? Guiding question: How do we choose a visualization that reveals a pattern honestly, and how do we turn a chart into a message others understand?

Key Fact:
Visualization turns numbers into patterns the eye can grasp — to help the analyst think and to communicate. Common graphs suit one or two variables; heatmaps and mosaic plots handle more combinations. Good communication uses data storytelling: a clear message, only the elements that serve it, and honest labels.

Key Concepts:

- data visualization
- heatmap
- mosaic plot
- overplotting
- data storytelling

Point!

1. The Role of Data Visualization

(1) **(Data visualization)**…representing data as graphs or figures so that features that are hard to read from numbers alone can be grasped visually.

(2) Data visualization broadly has two roles.

Role	Description	Example
Deepen the analyst's own understanding	Visually grasp the distribution, trends, outliers, etc. of data and gain insight	When training records or study time are shown as a line graph, changes and trends that are hard to notice when numbers are merely listed in a table become easier to see
Communicate data to others	Convey the analysis results or claims to others in an easy-to-understand way	When the average test score of each class is shown as a bar graph, even someone seeing the data for the first time can compare differences between classes at a glance

(3) In either role, it is important to choose a graph that fits the content to be conveyed and to use representations that do not invite misunderstanding. The appropriate visualization method differs depending on the number and type of variables and the purpose of the analysis. When information that cannot be handled by familiar graphs needs to be shown, more advanced visualization techniques are used.

KEY TERMS
data visualization — showing data as graphs/figures to grasp features visually.

A clear chart speaks; a cluttered one hides.

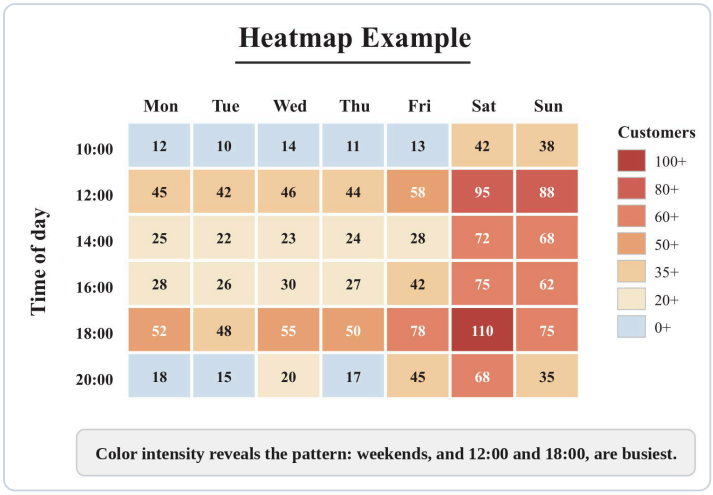
2. Visualization Handling Multiple Variables

Common graphs such as bar graphs, line graphs, and scatter plots are suited to representing the relationship between one or two variables. On the other hand, in analysis, situations often arise in which three or more variables, or many

KEY TERMS
heatmap — a color grid on two axes for a third value.
overplotting — points overlapping so density is hidden.

combinations of information, need to be handled at once. Heatmaps and mosaic plots are advanced visualization methods used for such cases.

(1) **(Heatmap)**…a graph in which items are arranged on two axes, and the size of a **(third value)** — a numerical value — corresponding to each combination is represented by the gradation of color or differences in hue. *e.g.,* Grasp the number of customers visiting at a glance by day of week × time slot; compare trends in average temperature for city × month.



A heatmap: color intensity shows the value for each row × column combination

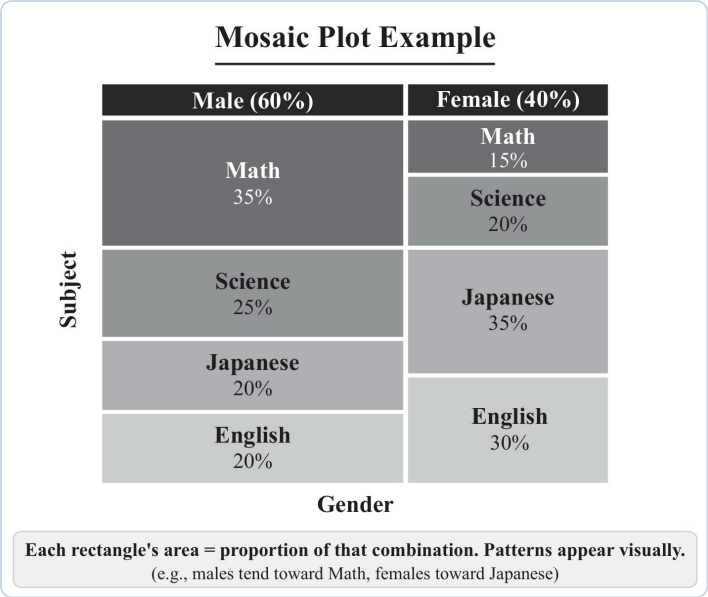
In a scatter plot, since the relationship between two quantitative variables is represented by points, when the number of points becomes large, they overlap and the density of data becomes hard to see (overplotting). Also, scatter plots are not suited to handling combinations of categorical variables. In such cases, a heatmap has the advantage of allowing the density and patterns of the data to be grasped at a glance.

(2) **(Mosaic plot)**…a graph that represents the relationship between two categorical variables by the size of rectangular areas. *e.g.,* Compare the distribution of gender × subject preference; grasp differences in purchase ratios for age group × product category.

One view can encode several variables at once.

PAUSE & THINK

Two categories compared by area, or a third value by color — which graph is which?



A mosaic plot: each rectangle's area shows the proportion of that category combination

The widths of row and column divisions correspond to the proportions of frequencies, and the area of each rectangle represents the proportion of frequency for that combination of categories. Even when the overall picture is hard to grasp from a numerical table (cross-tabulation), a mosaic plot visually shows differences in proportions as areas, making it easier to grasp relationships between categories intuitively.

3. Data Storytelling

(1) **(Data storytelling)**…using data and visualization to convey the message you want to convey, structured like a story.

(2) The following are the three basic elements of storytelling.

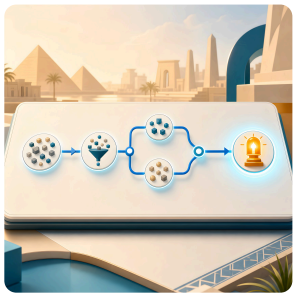
Element	Description	Example
Clarification of the message	Make it possible to express in one sentence what you want to convey through the graph	Make the title of a graph showing class test results contain the claim, such as “The class scored highly overall on the math test”
Selection of elements	Remove elements unrelated to the message and emphasize the parts you want to draw attention to	Make only the elements you want to draw attention to stand out in color, while keeping other elements in faded colors
Provision of supporting information	Properly arrange the title, axis labels, units, and notes	Make the graph title, such as “Average math test scores by class,” express the content at a glance, and clearly indicate units and item names like “points” on the vertical axis and “class” on the horizontal axis

(3) The depth of an audience's understanding can change greatly with how the same analysis result is conveyed. After visualizing data, it is important to put yourself in the audience's position and verify whether the message is conveyed.

KEY TERMS

mosaic plot — rectangle areas for two categorical variables.

data storytelling — structuring a data message like a story.



Storytelling guides the eye to one insight.

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Data visualization has both the role of deepening the analyst's own understanding and the role of communicating data to others.

B A heatmap is a graph that represents the size of a third value corresponding to combinations on two axes by the gradation of color or differences in hue.

C A mosaic plot is a graph that represents the relationship between two categorical variables by the size of rectangular areas.

D In data storytelling, it is important to include as many elements related to the graph as possible.

(2) Match each description (a–c) with the most appropriate graph from the choices below (A–C).

a A graph that represents the size of a numerical value corresponding to combinations on two axes by the gradation of color or differences in hue

b A graph that represents the relationship between two categorical variables by the size of rectangular areas

c A graph that represents the relationship between two quantitative variables by the position of points

[Choices] **A** Scatter plot **B** Heatmap **C** Mosaic plot

✓ Solution

(1)

A Correct as a role of data visualization. Therefore, ○.

B Correct definition of a heatmap. Therefore, ○.

C Correct definition of a mosaic plot. Therefore, ○.

D In data storytelling, it is important to remove elements unrelated to the message. Therefore, ×.

(2)

a: Showing values by gradation of color or hue is a heatmap. **B**

b: Showing relationships by area is a mosaic plot. **C**

c: Showing relationships by points is a scatter plot. **A**

PAUSE & THINK

Does adding more elements to a graph make the message clearer, or bury it?

Try

1 Answer the following questions.

1. What is the term for the graph that represents the size of a third value corresponding to combinations on two axes by the gradation of color or differences in hue?
2. What is the term for the graph that represents the relationship between two categorical variables by the size of rectangular areas?
3. In data visualization, of “deepening the analyst's own understanding” and “communicating data to others,” for what purpose is the latter performed?
4. What is the term for using data and visualization to convey the message you want to convey, structured like a story?

2 Answer the following questions.

(1) For each of situations (a, b), choose the most appropriate graph from the choices below (A, B).

a You want to grasp the pattern of customer visits by day of week × time slot at a glance.

b You want to intuitively show differences in proportions for gender × subject preference using areas.

[Choices] **A** Heatmap **B** Mosaic plot

(2) From the following options (A–D), choose the one that most appropriately describes a situation in which a heatmap is used.

A A situation in which one wants to simply compare a representative value for one variable.

B A situation in which the relationship between two quantitative variables can be sufficiently grasped with a scatter plot because the number of points is small.

C A situation in which one wants to grasp the distribution and patterns of numerical data over combinations on two axes at a glance.

D A situation in which one wants to show only one composition ratio of a categorical variable.

EXAM-STYLE QUESTION

Explain why choosing the wrong graph, or hiding a chart's labels, can mislead an audience even when the data is correct. **[6] Refer to:** matching the graph to the data and honest supporting information.

PAUSE & THINK

Two axes with a value by color, or two categories by area? Name each situation's graph.

PAUSE & THINK

Time-series of one value, or proportions of two categories — which suits a mosaic plot?

(3) From the following options (A–D), choose the one that most appropriately describes a situation in which a mosaic plot is used.

A A situation in which one wants to follow the time-series changes of one quantitative variable.

B A situation in which one wants to intuitively grasp the proportions of frequencies for combinations of two categorical variables using areas.

C A situation in which one wants to represent the correlation between two quantitative variables by points.

D A situation in which one wants to show the composition ratio of categories at a particular time using sectors.

(4) From the following options (A–D), choose the one that most appropriately describes data storytelling.

A Even if the message you want to convey is unclear, it will be communicated to the audience as long as the graph is beautiful.

B Clarification of the message, selection of elements, and provision of supporting information are the basic elements.

C It is better to include as many related elements as possible in a graph.

D Titles and axis labels detract from the appearance of a graph and should preferably be omitted.

⚙ Think as an Engineer — Investigate & Decide

Your class collected data on students' home governorate and their favorite sport, and also monthly rainfall for several Egyptian cities across the year. You must present each dataset clearly.

☐ **Use a bar graph for everything**

☐ **Match each dataset to the right graph**

1 · Choose a graph for each. For the governorate × favorite-sport data and for the city × month rainfall data, name the graph you would use. For at least one of them, justify your choice — could the other graph type also work here, and why is yours better?

2 · Guard against misleading. Name one design choice that could make your chosen chart mislead the reader (for example a color scale with no legend, or truncated categories), and say how you would prevent it — a judgment the lesson's examples do not hand you.

3 · Tell the story. For one of the two charts, write the one-sentence message you want the audience to take away, and name the single element you would emphasize to carry it. **In pairs,** compare messages and agree on the clearer one.

Give evidence for your answer.

KEY TERMS

cross-tabulation — a table of counts for category combinations.

misleading graph — one whose design distorts the true pattern.

Stuck? Ask what each dataset is made of: two categories → areas (mosaic); a value across two axes → color grid (heatmap). Then ask what one sentence you want the reader to leave with, and strip away everything that does not serve it.

🌍 In a New Context

A researcher plots study time against test score for every student in a large school as a scatter plot, but so many points overlap into a solid blob that the density cannot be read. Name a visualization that would reveal where most students actually fall, and explain why it works better here than the crowded scatter plot.

❓ Lesson Question — Answered

The lesson asked how to choose a visualization that reveals a pattern honestly and how to turn a chart into a message others understand. Data visualization serves two roles — deepening the analyst's own understanding and communicating to others — and the right method depends on the number and type of variables. Bar, line, and scatter plots suit one or two variables; when a scatter plot's points overlap (overplotting) or categories must be combined, a heatmap shows a value across two axes by color, and a mosaic plot shows two categorical variables by rectangular area. Turning a chart into a message is data storytelling, built from three elements: clarify the message in one sentence, select only the elements that serve it (emphasizing the key part, fading the rest), and provide honest supporting information (title, axis labels, units, notes). Because the same result can land very differently depending on how it is shown, the final step is to stand in the audience's position and check that the message truly arrives.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

Data visualization has both the role of deepening the analyst's own understanding and the role of (a) data to others. When information cannot be

handled by common graphs such as bar graphs and line graphs, advanced visualization techniques are used. The graph that represents the size of a third value corresponding to combinations on two axes by the gradation of color or differences in hue is called (b), and the graph that represents the relationship between two categorical variables by the size of rectangular areas is called (c). When communicating data, the concept of (d), in which the message is structured like a story, is important; the basic elements are clarification of the message, selection of elements, and provision of (e).

1. Fill in the blanks (a)–(e).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes a heatmap.

A A graph that represents the relationship between two quantitative variables by the position of points.

B A graph that represents the size of a third value corresponding to combinations on two axes by the gradation of color or differences in hue.

C A graph that represents the composition ratio of categories by the area of sectors.

D A graph that represents the changes in time-series data by lines.

(2) From the following options (A–D), choose the one that is NOT an appropriate basic element of data storytelling.

A Make it possible to express in one sentence what you want to convey through the graph (clarification of the message).

B Remove elements unrelated to the message and emphasize the parts you want to draw attention to (selection of elements).

C Properly arrange the title, axis labels, units, and notes (provision of supporting information).

D Include as many elements unrelated to the message as possible to increase the amount of information.

(3) For each of analysis purposes (a–c), choose the most appropriate graph from the choices below (A–C).

a You want to grasp the monthly average temperatures of multiple cities together at a glance.

b You want to show differences in purchase ratios for age group × product category using areas.

c You want to see the relationship between two quantitative variables in a case where the number of points is small enough that the density can also be sufficiently read.

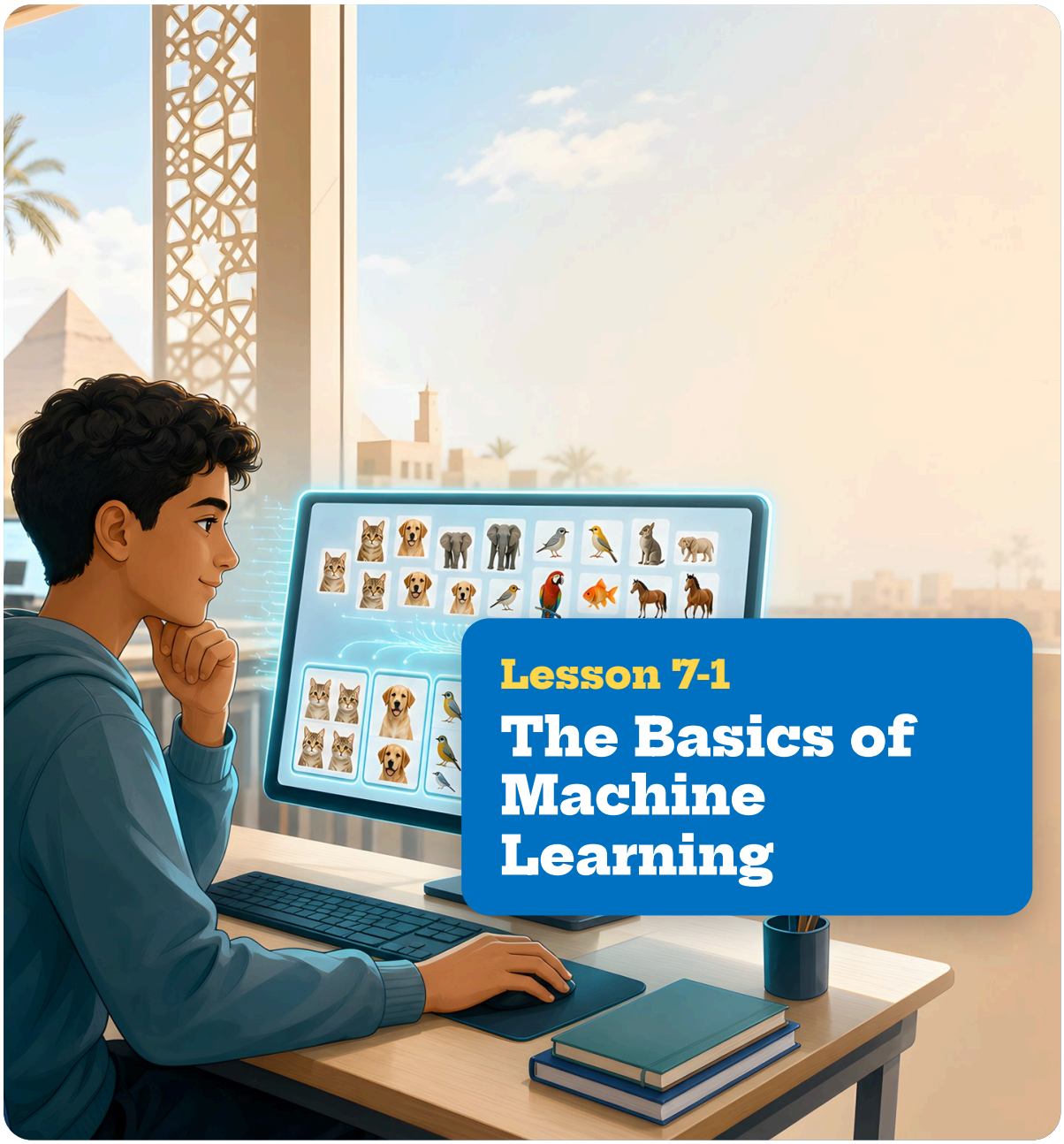
[Choices] **A** Scatter plot **B** Heatmap **C** Mosaic plot

REFLECT & CHALLENGE

Reflect: Recall a chart that confused you. Was it the wrong graph, too many elements, or missing labels? **Challenge:** You are given a heatmap whose color scale does not start at zero, so small differences look dramatic. Describe the false impression this creates, and the one change that would make it honest without changing the data.

Key takeaway

Remember:
Match the graph to the data (heatmap for a value across two axes; mosaic plot for two categories by area; scatter/bar/line for one or two variables), then tell a data story: one message, only the elements that serve it, honest labels — and check it lands with the audience.



Lesson 7-1

The Basics of Machine Learning

Learning Objectives

- Explain the idea of machine learning and distinguish supervised, unsupervised, and reinforcement learning.
- Describe classification, clustering, and reinforcement learning with examples, and explain the limits caused by data bias and the black-box problem.

How do you teach a computer to tell a cat from a dog when you cannot write down a rule for it? You do not write the rule — you show it thousands of examples and let it find the rule itself. That is machine learning, the engine behind most of today's AI. It learns patterns from data to make judgments on new data. Yet it is not magic: if the examples are biased, the judgments are biased, and its reasoning can be a black box even to its makers. Machine learning is therefore best used to support human judgment, not to replace it.

? Guiding question: How does machine learning solve problems differently from ordinary programming, and what are its main types and limits?

Point!

1. The Idea of Machine Learning

(1) **(Machine learning)**…a technology that learns patterns and regularities from large amounts of data so that judgments and predictions can be made on new data. It is the most common method used to make AI work.

(2) Conventional programming and machine learning differ greatly in how they solve problems.

Approach	What humans provide	What the computer does
Conventional programming	Decision rules (procedures) and input data	Produces answers by following the rules
Machine learning	Large amounts of data (with correct answers, where available)	Finds the decision rules from the data on its own

(3) Machine learning is used for problems where it is difficult for humans to express the rules in words. *e.g.,* judging whether a photo shows a dog or a cat; distinguishing whether an email is spam; transcribing audio into text

(4) For such problems, the rules are too complex for humans to write out completely, so the approach of letting the computer learn from examples (data) becomes effective.

(5) However, machine learning is not a panacea. Because the quality and bias of the training data affect the results, biased data leads to biased judgments (bias). There is also the black-box problem, in which the basis for the model's judgments becomes hard for humans to see. Given these limitations, it is important to use machine learning not as a replacement for human judgment, but as a tool that supports human judgment.

2. Three Types of Learning Methods

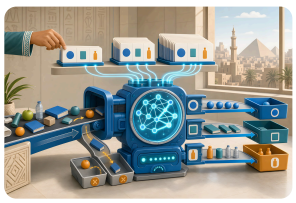
(1) Machine learning is broadly divided into the following three types, depending on the form of data given and the method of learning.

Key Fact:
Machine learning finds the rules from data instead of being given them. It comes in three forms — supervised (learn from labeled examples), unsupervised (find hidden groups), and reinforcement (learn from rewards) — but its results depend on the data, so bias and the black-box problem make it a support for human judgment, not a replacement.

Key Concepts:

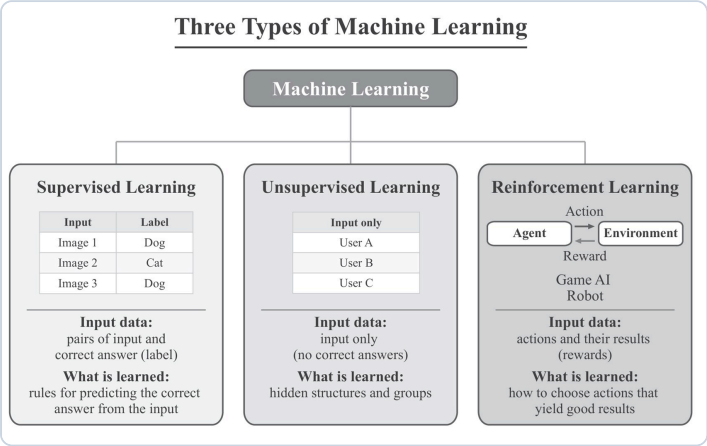
- machine learning
- supervised / unsupervised / reinforcement
- classification, feature
- clustering
- bias, black-box problem

KEY TERMS
machine learning — learning rules from data to judge new data.
bias — skewed data leading to skewed judgments.
black-box problem — the basis of a judgment is hard to see.



Machine learning improves from more examples.

Learning method	Data given	What is learned
(Supervised learning)	Pairs of input and the correct answer (label)	Rules for predicting the correct answer from the input
(Unsupervised learning)	Input only (no correct answers)	Hidden structure or groups in the data
(Reinforcement learning)	Actions and their results (rewards)	How to choose actions to obtain better results



The three types of machine learning and the data each uses

(2) Which learning method to use is determined by the type of data at hand and the nature of the problem to be solved.

Situation	Suitable learning method
Past data has correct answers, and you want to predict the correct answer for new data	Supervised learning
Data is available but without correct answers, and you want to find tendencies or groups in the data	Unsupervised learning
You want the system to learn how to choose actions for better results through trial and error	Reinforcement learning

3. Classification (A Representative Example of Supervised Learning)

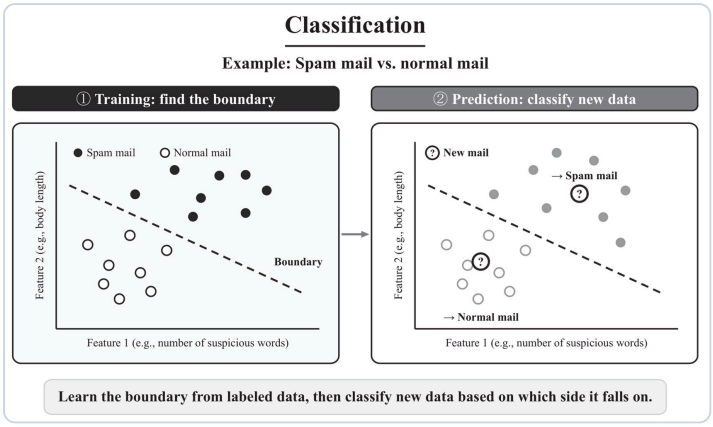
(1) **(Classification)**…a method of predicting which type (class) the input data belongs to. It is a representative example of supervised learning. *e.g.,* sorting emails into "spam" or "normal"; sorting medical images into "abnormal" or "normal"; classifying animals in photos as "dog," "cat," or "bird"

(2) In classification, large amounts of data with correct answers (labels) attached are given in advance for learning. After training is complete, the model can predict which class new data without known labels belongs to.

KEY TERMS
unsupervised learning — finds groups with no labels.
reinforcement learning — learns actions from rewards.
label — the known correct answer on training data.

KEY TERMS
supervised learning — learns from labeled examples.
classification — predict which class an input belongs to.
feature — a clue in the data useful for prediction.

Stage	What is done
(Training)	Find rules for classification from pairs of input data and correct answers (labels)
(Prediction)	Use the learned rules to determine which class new data belongs to



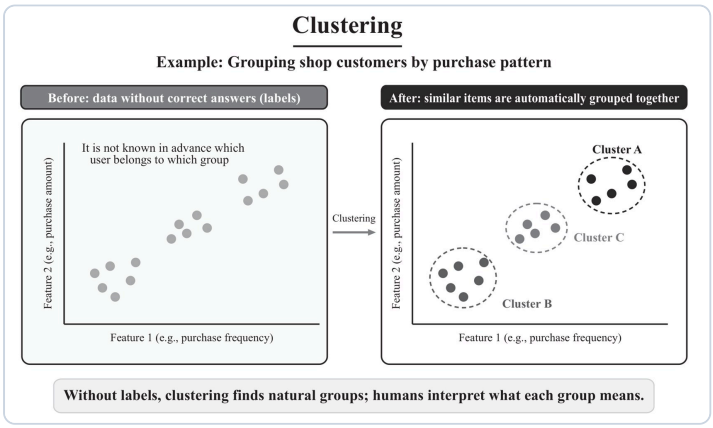
Classification: learn a boundary from labeled data, then classify new data

(3) Information in the training data that serves as a clue useful for prediction is called a **(feature)**. *e.g.,* In spam classification, the types of words contained in the email, the sender's address, and the length of the body are features.

(4) The choice of features greatly affects the accuracy of classification. Choosing features that match the purpose is important for building a good classification model.

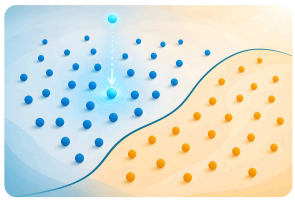
4. Clustering (A Representative Example of Unsupervised Learning)

(1) **(Clustering)**…a method that groups similar items together into groups (clusters) within data. It is a representative example of unsupervised learning. *e.g.,* grouping users of an online shop by similar purchase tendencies; grouping users with similar listening patterns on a music streaming service for recommendations; automatically grouping news articles with similar content



Clustering: without labels, similar items are automatically grouped

(2) The difference between clustering and classification lies in whether the correct answers (labels) are given.



Classification draws a boundary between labeled groups.

PAUSE & THINK

Labels given or not given — which one word separates classification from clustering?

Method	Correct answers (labels)	Purpose
Classification	Given in advance	Predict which class new data belongs to
Clustering	Not given	Find natural groups within the data

(3) In clustering, the computer itself finds "what is similar." Even when the names or number of groups are not known in advance, grouping can be done automatically based on the features of the data.

(4) Clustering is used when one wants to grasp the overall picture of the data or when one wants to look at the data from a new perspective. Humans interpret the results of grouping and use them in marketing, product development, and so on.

5. Reinforcement Learning

(1) **(Reinforcement learning)**…a method in which actions are chosen in a certain situation, and how to choose better actions is learned based on the reward obtained as a result. *e.g.,* learning how to obtain a high score in a game through trial and error; a robot learning to walk without falling by trying movements; a self-driving car learning safe driving operations

(2) In reinforcement learning, learning is accomplished by repeatedly trying combinations of actions and results.

Element	Description
(State)	The situation in which the agent (the learning side) is placed
(Action)	The choices that can be selected in that situation
(Reward)	The evaluation given as a result of the action (a numerical value indicating goodness or badness)

(3) What distinguishes reinforcement learning from the other two learning types is that it learns from the consequences of its own actions. Unlike supervised learning, where there is a correspondence of "this input has this correct answer," reinforcement learning gradually learns better choices from the rewards obtained for the actions it tries.

(4) Reinforcement learning is suited to problems where the correct answer is not known in advance, but the goodness of the result can be judged. It is often used in combination with environments where trial and error can be repeated (simulators or games).

KEY TERMS

clustering — group similar items with no labels.
reinforcement learning — learn actions from rewards.
reward — a value scoring how good an action was.

Reinforcement learning learns by reward and penalty.

PAUSE & THINK

Learning from rewards through trial and error — which of the three types is that?

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A Machine learning is a technology that learns patterns and regularities from large amounts of data so that judgments and predictions can be made on new data.

B In supervised learning, the model learns from data without correct answers (labels).

C Clustering is a method that groups data according to labels given in advance.

D Reinforcement learning is a method in which how to choose better actions is learned based on the reward obtained as a result of an action.

(2) Match each description (a–c) with the most appropriate term from the choices below (A–C).

a A learning method that learns rules for predicting correct answers for inputs from pairs of input and correct answer

b A method that finds groups of similar items from data without correct answers

c Information in the input data that serves as a clue for prediction in classification

[Choices] **A** Feature **B** Supervised learning **C** Clustering

✓ Solution

(1)

A Correct definition of machine learning. Therefore, ○.

B Data without correct answers (labels) is unsupervised learning. In supervised learning, correct answers are given. Therefore, ×.

C Clustering is a method that finds natural groups from data without given labels. Therefore, ×.

D Correct description of reinforcement learning. Therefore, ○.

(2)

a: Learning from pairs of input and correct answer is supervised learning. **B**

b: Finding groups from unlabeled data is clustering. **C**

c: Information that serves as a clue for prediction is a feature. **A**

Try

1 Answer the following questions.

1. What is the term for the technology that learns patterns and regularities from large amounts of data so that judgments and predictions can be made on new data?
2. In machine learning, what is the term for the information in the input data that serves as a clue useful for prediction?
3. What is the term for the representative method of unsupervised learning that groups similar items together?
4. In reinforcement learning, what is the term for the numerical value that represents the evaluation given as a result of an action?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that is NOT an appropriate situation in which machine learning is used.

A Distinguishing the type of animal in a photograph.

B Transcribing audio into text.

C Calculating the sum of two integers.

D Distinguishing whether an email is spam.

(2) For each of purposes (a–c), choose the most appropriate learning method from the choices below (A–C).

a Predict, from past purchase history data and whether or not a purchase was made, whether a new customer will buy the product.

b Find groups of customers with similar purchase tendencies from customer purchase data.

c Have a robot walk and learn movements that allow it to advance without falling through trial and error.

[Choices] **A** Unsupervised learning **B** Supervised learning **C** Reinforcement learning

EXAM-STYLE QUESTION

Explain why machine learning should support human judgment rather than replace it. **[6] Refer to:** bias in data and the black-box problem.

PAUSE & THINK

Which task has a simple rule a human can already write — so needs no learning?

(3) From the following options (A–D), choose the one that most appropriately describes the difference between classification and clustering.

A Classification divides data into a small number of groups, and clustering divides data into many groups.

B Classification handles only numerical data, and clustering handles only text data.

C Classification learns from data with correct answers (labels), and clustering learns from data without correct answers.

D Classification is unsupervised learning, and clustering is supervised learning.

(4) From the following options (A–D), choose the one that is NOT an appropriate point of caution when using machine learning.

A If the training data is biased, the model's judgments may also be biased.

B The basis for a machine learning model's judgments may be hard to see, so for important judgments human checking is needed.

C Machine learning is a panacea and can make more accurate judgments than humans on any problem.

D It is important to use machine learning as a tool that supports human judgment.

⚙ Think as an Engineer — Investigate & Decide

Your school has years of data and wants help with two tasks: (i) flagging students who may need extra support early, and (ii) discovering natural groups of study habits it did not know about. You advise on how to use machine learning responsibly.

☐ Let the model decide alone **☐ Choose the right method, and keep a human in charge**

1 · Identify the method. For task (i) and task (ii), name which learning type fits (supervised, unsupervised, or reinforcement). For each, name one piece of data the school would need to collect, and one kind of student that data might miss.

2 · Analyze the risks. Name one way the training data for task (i) could be biased (for example, if it only includes students who already asked for help), and say what harm the black-box problem could cause if a teacher acts on a flag without understanding it — a judgment the lesson's tables do not make for you.

3 · Decide the role of the human. State which decisions you would keep a human in charge of, and which you would let the model assist with, and why. **In pairs,** compare plans and agree on the one safeguard you would never remove.

Give evidence for your answer.

KEY TERMS

label — the known correct answer attached to data.

human-in-the-loop — a person checks or approves the model's output.

Stuck? Start from the data you have: are there correct answers to learn from (supervised) or none (unsupervised)? Then ask, for each decision, what it would cost to be wrong — the higher that cost, the more a human must stay in charge.

🌍 In a New Context

A bank wants software that reads each loan application and flags the ones that look risky for a human officer to review. Name which learning type this is, name one feature the model might use, and explain why the officer must stay in the loop rather than letting the model decide alone.

❓ Lesson Question — Answered

The lesson asked how machine learning differs from ordinary programming and what its main types and limits are. In conventional programming, humans provide the rules and the computer follows them; in machine learning, humans provide data and examples, and the computer finds the rules itself — which is why it suits problems whose rules are too complex to write, like recognizing images or speech. It comes in three types: supervised learning (labeled examples, whose representative task is classification, using features to predict a class), unsupervised learning (no labels, whose representative task is clustering, finding natural groups), and reinforcement learning (learning from rewards through trial and error, described by state, action, and reward). Its limits matter as much as its power: results depend on the quality and bias of the data, so biased data gives biased judgments, and the black-box problem hides the basis of a decision. For these reasons machine learning is best used to support human judgment, not to replace it.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

Machine learning is a technology that learns patterns and regularities from large amounts of data so that judgments and predictions can be made on new data. Machine learning is broadly divided into three types depending on the form of data given and the method of learning. There is (a) learning, which learns from pairs of input and correct answer; (b) learning, which finds structure or groups from data without correct answers; and (c) learning, which learns how to choose better actions based on the (d) obtained as a result of an action. In (e), a representative example of supervised learning, the choice of (f), which serves as a clue for prediction, greatly affects the result. In (g), a representative example of unsupervised learning, the computer itself finds “what is similar” and discovers natural groups.

1. Fill in the blanks (a)–(g).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the difference between conventional programming and machine learning.

A In conventional programming, humans provide the decision rules; in machine learning, the computer finds the decision rules from data on its own.

B Conventional programming requires large amounts of data, and machine learning requires only rules.

C Conventional programming is included in AI, but machine learning is not included in AI.

D Conventional programming can handle new problems, but machine learning can handle only fixed problems.

(2) For each of situations (a–c), choose the most appropriate term from the choices below (A–C).

a There are large amounts of news article texts and their categories (politics, economics, sports, etc.), and you want to automatically determine the category of new articles.

b You have records of products purchased by users of an online shop and want to find groups of users with similar purchasing tendencies.

c You want a computer to play a game and learn the operations that lead to high scores through trial and error.

[Choices] **A** Reinforcement learning **B** Clustering **C** Classification

REFLECT & CHALLENGE

Reflect: Where have you likely met machine learning today (recommendations, photo tagging, voice input)? Which type do you think it was?
Challenge: A face-recognition model was trained mostly on one group of people and works poorly on others. Explain how biased training data caused this, and state one change to the data that would reduce the bias.

Key takeaway

Remember:
Machine learning finds rules from data: supervised (labeled → classification with features), unsupervised (unlabeled → clustering), reinforcement (rewards). Mind its limits — bias and the black-box problem — and keep it supporting human judgment.

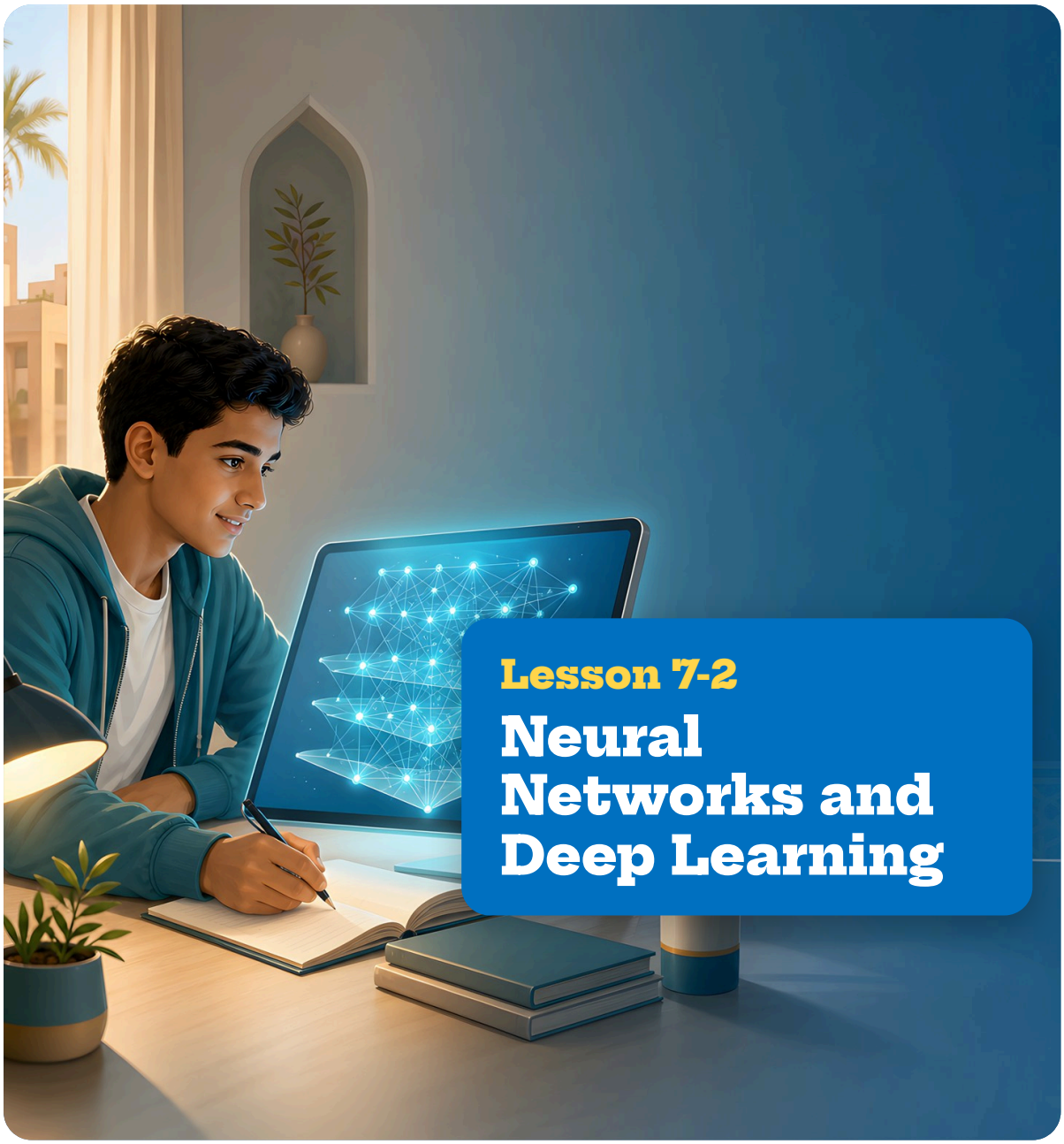
(3) From the following options (A–D), choose the one that is NOT an appropriate description of the characteristics and limitations of machine learning.

A Machine learning is used for problems where it is difficult for humans to express the rules in words.

B Machine learning models are affected in their results by the quality and bias of the training data.

C A machine learning model's basis for judgment is always shown to humans in an easy-to-understand way.

D It is important to use machine learning not as a replacement for human judgment, but as a tool that supports human judgment.



Lesson 7-2

Neural Networks and Deep Learning

Learning Objectives

- Describe the structure of a neural network and explain how a single neuron processes its inputs.
- Explain what deep learning is, and why stacking many hidden layers brought a leap in AI performance.

How can a machine tell a dog from a cat, or turn speech into text? The answer, loosely inspired by the brain, is a neural network: layers of simple units called neurons, each passing signals to the next. A neuron just multiplies its inputs by weights and adds them up — nothing complex on its own — but stack enough of them across enough layers and the network learns, by adjusting those weights on data, to recognize patterns no human could write rules for. Making the stack deep — deep learning — lets the network build up features from simple edges up to whole objects, and it is what pushed AI past humans on many tasks.

? Guiding question: How is a neural network built and trained, and what does making it "deep" add?

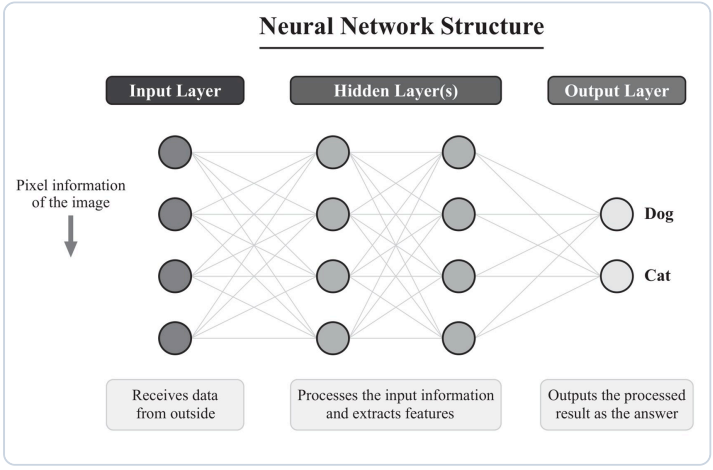
Point!

1. The Structure of Neural Networks

(1) A neural network has a structure in which many **(nodes)**, also called neurons, are connected. Nodes are arranged together in **(layers)**, and the nodes of adjacent layers are connected to each other.

(2) Layers are divided by role into the following three types.

Type of layer	Role
Input layer	Receives data from outside
Hidden layer (middle layer)	Processes the input information and extracts features
Output layer	Outputs the result of processing as an answer



A neural network: input layer, hidden layers, and output layer

(3) Data enters at the input layer, is processed as it passes through the hidden layers, and reaches the output layer in this flow. *e.g.,* In a network that judges whether the animal in a photograph is a dog or a cat, the input layer receives the

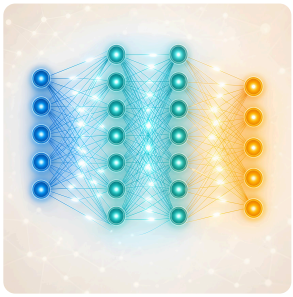
Key Fact:
A neural network is layers of connected neurons: an input layer, hidden layers that extract features, and an output layer. Each neuron multiplies its inputs by weights and sums them; training adjusts those weights from data. Deep learning stacks many hidden layers to build features step by step — the leap behind modern AI.

Key Concepts:

- node (neuron), layer
- input / hidden / output layer
- weight, training
- deep learning
- hierarchical features

KEY TERMS

node (neuron) — a single processing unit.
layer — a group of neurons at one stage.
input / hidden / output layer — receive, process, answer.



A network stacks neurons into connected layers.

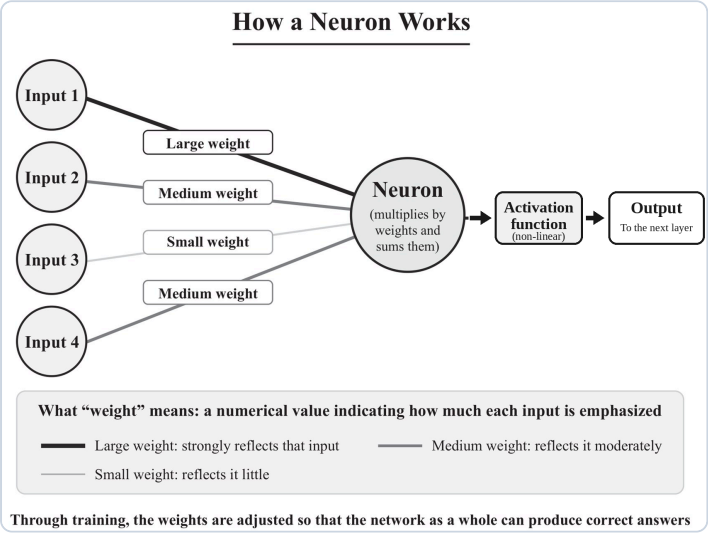
pixel information of the photograph, the hidden layers extract features such as shape and pattern, and the output layer outputs the judgment of “dog” or “cat.”

(4) The number of hidden layers and the number of nodes in each layer are designed according to the difficulty of the problem to be handled. The more complex the problem, the more layers and nodes are needed.

2. How a Neuron Works

(1) Each neuron receives multiple inputs, multiplies them by weights and sums them, then passes the sum through an **(activation function)**, which lets the network learn relationships that are not simply proportional, before passing the result to the next layer.

(2) **(Weight)**…a numerical value that represents how much each input is emphasized. Weights can be negative as well as positive, and the larger a weight is in size, the more strongly that input influences the neuron's result.



A neuron weights its inputs, sums them, applies an activation function, and passes the result on

(3) **(Training)** is the process of gradually adjusting the weights between neurons using a large amount of data.

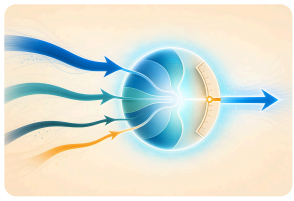
Stage	What happens
Before training	The weights are set to arbitrary values, and correct predictions cannot be made
During training	Data is given, and the weights are gradually adjusted based on the difference between the prediction and the correct answer
After training	With the adjusted weights, correct predictions can be made even on new data

(4) As training progresses, a combination of weights that allows the network as a whole to produce correct answers is found. Even though the function of each individual neuron is simple, complex judgments become possible by combining many neurons.

KEY TERMS

weight — how much an input is emphasized.

training — adjusting weights from data.



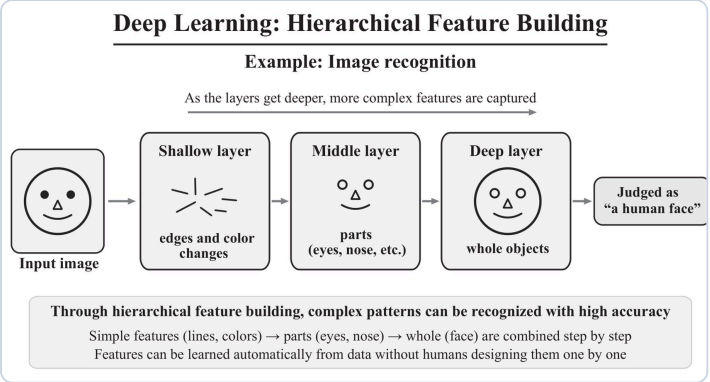
A neuron weights its inputs, sums them, applies an activation function, and passes the result on.

3. Deep Learning

(1) **(Deep learning)** is a technology that uses neural networks with many hidden layers stacked deeply. “Deep” refers to having many hidden layers.

(2) When the hidden layers are made deeper, features of data can be built up step by step and captured. *e.g.,* In an image-recognition network, features are built up from shallow to deep layers as follows.

Layer position	Examples of features extracted
Shallow layers	Simple features such as lines and changes in color
Middle layers	Shapes of parts such as eyes, nose, and ears
Deep layers	Structure of the whole face or object



Deep learning builds features step by step, from edges to whole objects

(3) Through this hierarchical building of features, deep learning can recognize complex patterns with high accuracy.

(4) Deep learning became practical as a result of large amounts of data and high computational power (in particular, specialized processors well suited to AI computation, such as GPUs and TPUs) becoming available. Before that, it was technically difficult to advance learning sufficiently even when layers were made deeper.

4. Why Deep Learning Brought a Leap in AI

(1) With the availability of deep learning, accuracy close to or surpassing humans has become achievable in many tasks that were previously considered difficult for computers.

(2) The fields in which there has been particularly great progress include the following.

PAUSE & THINK

Why does stacking more hidden layers let a network capture more complex features?

Deep layers build from edges to whole objects.

PAUSE & THINK

What two things had to become available before deep learning could work in practice?

Field	Examples
(Image recognition)	Face recognition systems, diagnostic support for medical images, object detection for self-driving cars
(Speech recognition)	Voice operation of smart speakers, transcription from speech, speech translation
(Natural language processing)	Machine translation, summarization of text, answering questions

(3) Before deep learning, features had to be designed by humans one by one and given to the computer. With deep learning, features can be learned automatically from data, allowing the limitations of human design to be exceeded.

(4) As a result of accumulating these capabilities, the foundation has been laid for AI that newly creates text and images.

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A A neural network has three types of layers: input layer, hidden layer, and output layer.

B The input layer is the layer that outputs the results of processing as an answer to the outside.

C Deep learning is a technology that uses a neural network with only one hidden layer.

D Training a neural network is the process of adjusting the weights between neurons using a large amount of data.

(2) Match each description (a–c) with the most appropriate term from the choices below (A–C).

a In a neural network, the layer that receives data from outside

b In a neural network, a numerical value that represents how much each input is emphasized

c In a neural network, the layer that processes input information and extracts features

[Choices] **A** Hidden layer **B** Weight **C** Input layer

✓ Solution

(1)

A There are three types of layers: input, hidden, and output. Therefore, ○.

B The output layer is what outputs results to the outside. The input layer is the layer that receives data from outside. Therefore, ×.

C Deep learning is a technology that uses neural networks with many hidden layers stacked deeply. Therefore, ×.

PAUSE & THINK

Which layer receives the data, and which one gives the final answer?

D Correct description of training. Therefore, ○.

(2)

a: The one that receives data is the input layer. **C**

b: The value representing how much an input is emphasized is the weight. **B**

c: The one that processes information and extracts features is the hidden layer. **A**

Try

1 Answer the following questions.

1. What is the term for the individual processing unit that makes up a neural network?
2. In a neural network, what is the term for the layer that processes input information and extracts features?
3. In a neural network, what is the term for the numerical value that represents how much each input is emphasized?
4. What is the term for the technology that uses neural networks with many hidden layers stacked deeply?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the layers of a neural network.

A The input layer has the role of outputting processing results to the outside.

B The hidden layer has the role of receiving data from outside.

C The output layer has the role of processing input information and extracting features.

D Data enters at the input layer, is processed as it passes through the hidden layers, and reaches the output layer in this flow.

(2) From the following options (A–D), choose the one that most appropriately describes the training of a neural network.

A From before training begins, the weights between neurons are set to correct values.

B As training progresses, a combination of weights that allows the network as a whole to produce correct answers is found.

C Training is performed by increasing the number of neurons.

D Once training is complete, the network becomes unable to handle new data.

EXAM-STYLE QUESTION

Explain how a neural network with simple neurons can make complex judgments, referring to weights and training. **[6]**
Refer to: weighting inputs and adjusting weights on data.

PAUSE & THINK

Before training, are the weights already correct? What changes them?

(3) From the following options (A–D), choose the one that is NOT an appropriate characteristic of deep learning.

A It has a structure with many hidden layers stacked deeply.

B It can build up features of data step by step.

C The deeper the hidden layers, the more complex features can be captured.

D Even if the number of hidden layers is increased, the difficulty of problems that can be handled does not change.

(4) For a deep learning network that performs image recognition, reorder the following (A–C) in the order in which features are extracted from shallow layers to deep layers.

A Structure of the whole face or object

B Simple features such as lines and changes in color

C Shapes of parts such as eyes, nose, and ears

PAUSE & THINK

More hidden layers let a network capture more complex features — but what does that extra depth demand in data and computing power?

Think as an Engineer — Investigate & Decide

A team wants an app that reads handwritten digits 0–9 from scanned school registers and turns them into typed numbers. You advise on building it with a deep learning network.

☐ **Write the rules for each digit by hand**

☐ **Train a deep network on many examples**

1 · Design the layers. Say what the input layer would receive, what the output layer would produce, and in one line why hidden layers (rather than a single step) are needed here.

2 · Explain the depth. Using the idea of hierarchical features, describe what shallow, middle, and deep layers might each capture for a handwritten digit — a chain the lesson shows for faces but does not spell out for digits, so you must map it yourself.

3 · Judge the cost and risk. Name the two things this network needs to become practical, and name one way the training data could be too narrow, making the app fail on some students' handwriting. **In pairs,** compare designs and agree on the one dataset change that would most improve fairness.

Give evidence for your answer.

Stuck? Use the lesson's face example as a model, then translate it to a digit: ask what would count as an “edge,” a “part,” and a “whole” for a written number. For the risk, list whose handwriting the network saw during training — then ask whose it never saw.

KEY TERMS

training data — the labeled examples a network learns from.

generalize — work correctly on new, unseen data.

🌍 In a New Context

A voice assistant turns spoken commands into text and often works well even for words it was not explicitly programmed with. Name which field of AI progress this belongs to, and explain, using layers and training, how it can handle speech no one wrote rules for.

❓ Lesson Question — Answered

The lesson asked how a neural network is built and trained and what making it deep adds. A neural network is built from layers of connected neurons: an input layer that receives data, hidden layers that process it and extract features, and an output layer that gives the answer, with data flowing input → hidden → output. Each neuron is simple — it multiplies its inputs by weights, sums them, and passes the sum through an activation function — and the network is trained by gradually adjusting those weights on large amounts of data. Before training, arbitrary weights give wrong answers; after training, tuned weights give correct ones even on new data. Making the network deep — deep learning, many hidden layers — lets features be built hierarchically, from simple edges through parts to whole objects, so complex patterns are recognized with high accuracy. Once large data and powerful processors like GPUs made it practical, deep learning let features be learned automatically instead of designed by hand, driving leaps in image recognition, speech recognition, and natural language processing.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

A neural network has a structure in which many processing units called (a) are connected. These are arranged together in (b), and are divided into three types: (c), which receives data from outside; (d), which processes information and extracts features; and (e), which outputs the result as an answer. Each (a) multiplies its inputs by (f), sums them, passes the sum through an activation function, and passes the result to the next layer. Training gradually adjusts (f) using large amounts of data. The technology that uses neural networks with many hidden layers stacked deeply is called (g), and its characteristic is that it can build up features of data (h) and capture them.

⭐ REFLECT & CHALLENGE

Reflect: Where have you used a tool built on deep learning (voice typing, photo search, translation)? Which field was it?
Challenge: A team makes a network deeper and deeper but accuracy stops improving and training becomes very slow. Using what made deep learning practical, explain what their deeper network now demands more of, and name the one thing — besides adding layers — they should check first.

1. Fill in the blanks (a)–(h).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes how a neuron in a neural network works.

A Each neuron passes one input as is to the next layer.

B Each neuron multiplies its inputs by weights, sums them, passes the sum through an activation function, and passes the result on.

C Each neuron averages the input values and passes the result to the next layer.

D Each neuron passes only the largest input to the next layer.

(2) From the following options (A–D), choose the one that most appropriately describes the background to deep learning becoming practical.

A Computation became simpler by reducing the number of hidden layers to one.

B Large amounts of data and high computational power became available.

C Connections between neurons were eliminated, simplifying processing.

D A method in which humans set weights one by one by hand was established.

(3) From the following options (A–D), choose the one that most appropriately describes why deep learning brought a leap in AI.

A Features could be learned automatically from data without humans designing them one by one.

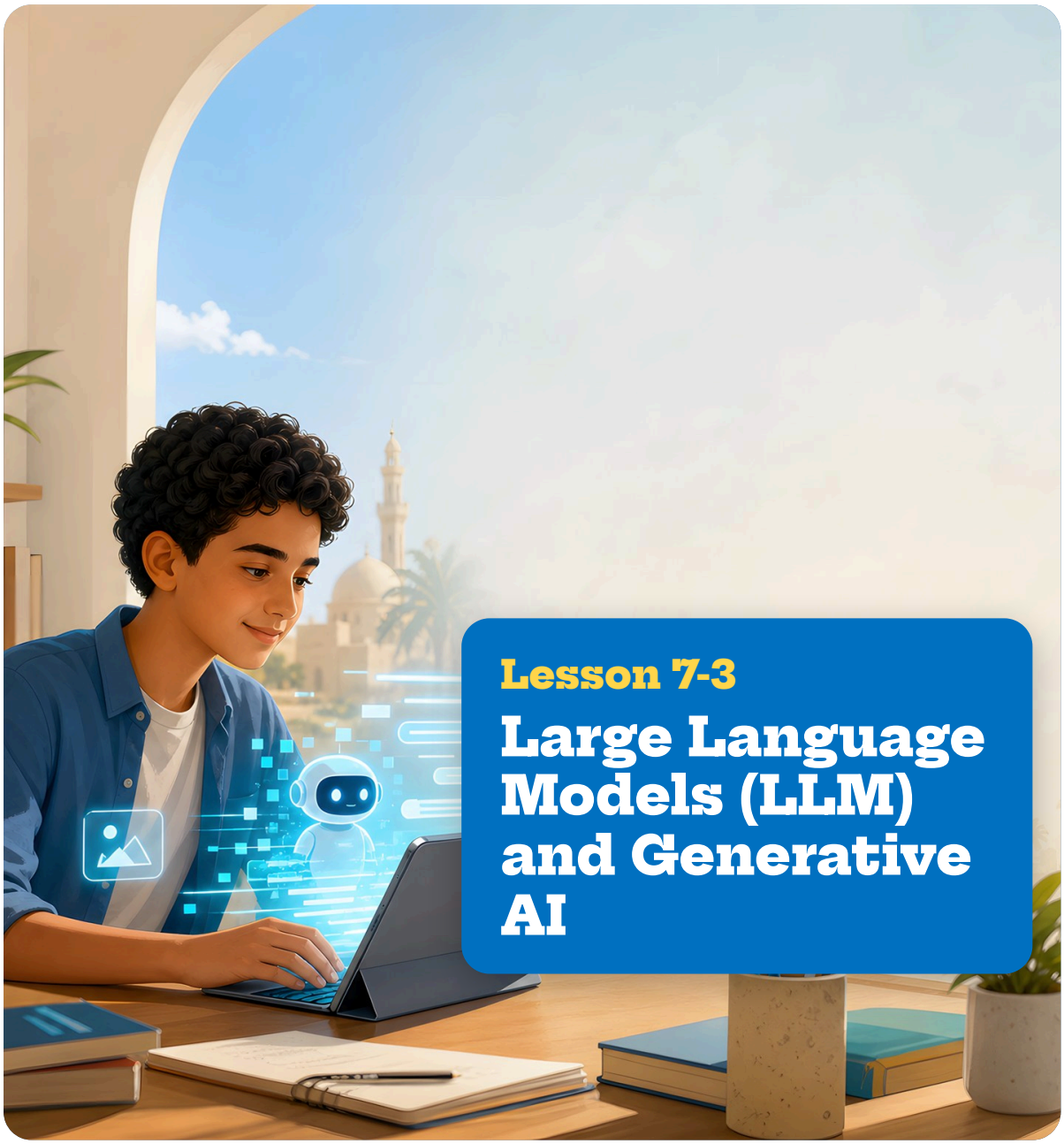
B Computers became able to think with the same mechanism as the human brain.

C Computers became able to fully understand the meaning of words.

D Computers became able to make completely accurate judgments on any problem.

Key takeaway

Remember:
A neural network = layers of neurons (input → hidden → output); each neuron weights and sums its inputs, and training tunes the weights on data. Deep learning stacks many hidden layers to build features (edges → parts → wholes), enabled by big data and GPUs.



Learning Objectives

- Explain how a language model generates text by predicting the next word, and why this causes hallucination.
- Describe what makes a language model large, and explain the roles of RLHF tuning and multimodality.

When a chatbot writes a fluent paragraph, it is not looking up an answer — it is guessing the next word, then the next, using patterns learned from enormous amounts of text. That simple trick, scaled up to billions of parameters, gives large language models their power to translate, summarize, and converse. But because each word is chosen for being statistically plausible rather than certainly true, these models can state falsehoods convincingly — a failure called hallucination. This lesson shows how such models work, how human feedback (RLHF) makes them helpful, and how multimodality lets them handle images and audio too.

? Guiding question: How does a language model generate text, what makes a model "large," and why can it still get facts wrong?

Point!

1. Language Models and "Predicting the Next Word"

(1) **(Language model)**…a model created to learn from large amounts of text and predict which word is likely to come next in a sentence. Through training, it acquires patterns of word usage and phrasing according to context.

(2) Text generation is achieved by repeating the prediction of the next word.

Step	Input sentence	Word predicted next
1st	"Today's weather is"	"sunny"
2nd	"Today's weather is sunny"	"and"
3rd	"Today's weather is sunny and"	"warm"

In this way, a complete sentence is built up by connecting the results of predicting one word at a time.

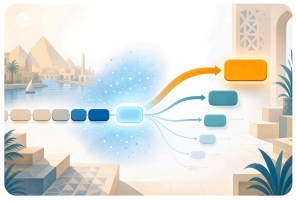
Key Fact:
A language model generates text by predicting the next word again and again — choosing what is statistically plausible, not guaranteed true (the cause of hallucination). Large language models scale this up with vast data and billions to trillions of parameters, are tuned by RLHF to be helpful, and, through multimodality, handle text, images, and audio together.

Key Concepts:

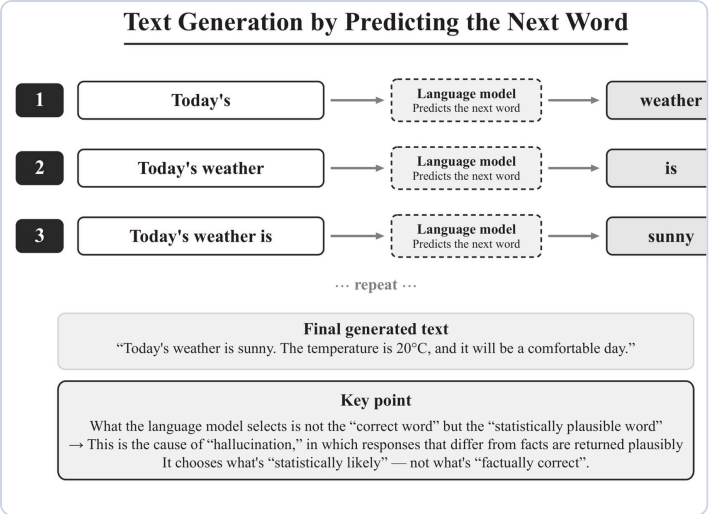
- language model, next-word prediction
- hallucination
- large language model (LLM), parameters
- RLHF
- multimodality

KEY TERMS

language model — predicts the likely next word.
hallucination — a plausible answer that is factually wrong.



A language model predicts the next word.



Text is generated by predicting the next word, one at a time

(3) For each candidate word, the language model calculates a "probability of coming next" and selects words based on that probability. These probabilities come from patterns the model learned during training, so it can also handle wordings that never appeared in the training data.

(4) What a language model selects is not the correct word, but a statistically plausible word. Because of this property, **(hallucination)**—the phenomenon of returning responses that differ from the facts but are stated in a plausible way—occurs.

2. Large Language Models (LLM)

(1) **(Large language model (LLM))**…a language model based on a very large neural network trained using vast amounts of text from the Internet.

(2) "Large" mainly refers to the following two kinds of size.

Aspect	Description
Amount of training data	The model is trained using vast amounts of text containing billions to trillions of words
Number of parameters	The numerical values corresponding to the weights of the neural network number in the billions to trillions

(3) Researchers have observed that when the number of parameters or the amount of training data exceeds a certain scale, qualitatively new abilities begin to appear. *e.g.,* handling complex reasoning; summarizing text; translating into foreign languages; generating program code

(4) Major services built on large language models include ChatGPT, Claude, and Gemini, which provide natural conversation and a wide range of text generation.

3. Tuning by RLHF

(1) Just by learning from vast amounts of text, large language models do not necessarily return responses that are desirable for humans. They may answer

KEY TERMS

large language model (LLM) — a huge network trained on vast text.

parameter — a weight; LLMs have billions to trillions.

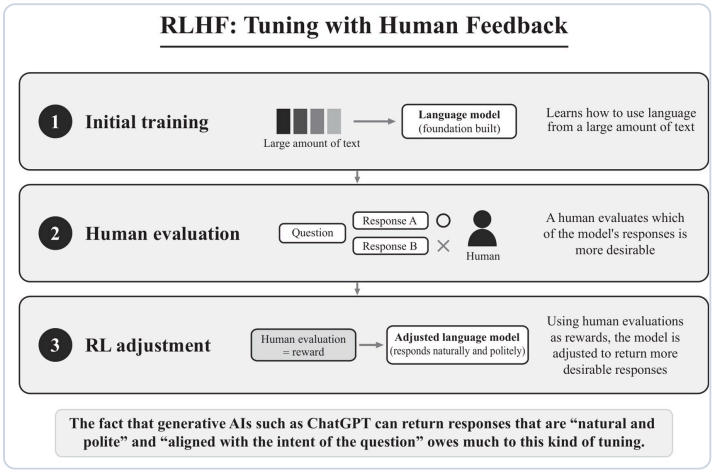
RLHF — tuning by human feedback as reward.

without grasping the intent of the question or return responses containing inappropriate content.

(2) A representative method for solving such problems is **(RLHF (Reinforcement Learning from Human Feedback))**. RLHF is a method that applies reinforcement learning to the tuning of language models.

(3) RLHF proceeds in the following flow.

Stage	Description
① **Initial training**	Build a language model that has learned word usage from large amounts of text
② **Human evaluation**	Humans give evaluations on the model's responses, such as "which is more desirable"
③ **Tuning by reinforcement learning**	Using human evaluations as rewards, tune the model to return more desirable responses



RLHF: human evaluations become rewards that tune the model

(4) Through RLHF, language models begin to return natural and polite responses, responses aligned with the intent of the question, and responses that avoid inappropriate content. The fact that generative AIs such as ChatGPT feel "easy to use" to users owes much to such tuning.

4. Multimodality

(1) **(Multimodality)**…the property of being able to handle multiple kinds of data with a single model, such as text, images, audio, and video. Each kind of data is called a **(modality)**.

(2) Early language models handled only text, but recent generative AIs can handle different kinds of data in combination. *e.g.,* looking at an image and describing its content in text; listening to audio and transcribing it into text; giving instructions in text and generating an image; summarizing the content of a video

(3) The background to multimodality being achieved is that neural networks have become able to handle different kinds of data by converting them into a

PAUSE & THINK

Left alone, a huge model may answer rudely or off-point. What step fixes that?

Human feedback ranks outputs to tune the model.

KEY TERMS

multimodality — one model handling many data kinds.
modality — one kind of data (text, image, audio…).

common numerical representation. Inside the network, text, images, and audio are all processed in the same way as sets of numerical values.

(4) Through multimodality, generative AI is evolving from a technology that handles only language into a technology that integrally handles the various kinds of information that humans deal with daily.

One model can output text, image, and sound.

PAUSE & THINK

If a model sounds confident and fluent, does that guarantee it is correct?

✎ Worked Example

(1) Mark each statement (A–D) with ○ if true or × if false.

A A language model is a model that predicts which word is likely to come next in a sentence.

B A language model finishes processing after predicting just one word and does not generate text.

C In a large language model, when the number of parameters or the amount of training data exceeds a certain scale, qualitatively new abilities sometimes appear.

D Multimodality is the property of handling only text.

(2) Match each description (a–c) with the most appropriate term from the choices below (A–C).

a A language model based on a very large neural network trained using vast amounts of text from the Internet

b A method that uses human feedback as a reward to tune a language model to return more desirable responses

c The phenomenon in which a language model returns responses that differ from the facts in a plausible way

[Choices] **A** Hallucination **B** Large language model **C** RLHF

✓ Solution

(1)

A Correct definition of a language model. Therefore, ○.

B Text generation is achieved by repeating the prediction of the next word. Therefore, ×.

C Correct description of large language models. Therefore, ○.

D Multimodality is the property of being able to handle multiple kinds of data—text, images, audio, video, etc.—with a single model. Therefore, ×.

(2)

a: A huge language model trained on vast text is a large language model. **B**

b: Tuning by reinforcement learning using human feedback as a reward is RLHF. **C**

c: The phenomenon of returning fact-discrepant responses in a plausible way is hallucination. **A**

Try

1 Answer the following questions.

1. What is the term for the model that learns from large amounts of text and predicts which word is likely to come next in a sentence?
2. What is the term for a language model based on a very large neural network trained using vast amounts of text from the Internet?
3. What is the term for the method that uses human feedback as a reward to tune a language model to return more desirable responses?
4. What is the term for the property of being able to handle multiple kinds of data—text, images, audio, video, and so on—with a single model?

2 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes how a language model works.

A It memorizes the input sentence as is and returns the same sentence.

B It generates text by repeatedly predicting the next word in the sentence.

C It selects and returns the closest match from prepared template sentences.

D It fully understands the meaning of the sentence and returns the single correct response.

(2) From the following options (A–D), choose the one that most appropriately describes the cause of hallucination.

A Because the language model always retrieves and returns sentences from the training data as is.

B Because the language model selects and returns words that are statistically plausible rather than correct.

C Because the language model avoids answering questions and returns unrelated text.

D Because the language model is built to return only fact-based responses.

EXAM-STYLE QUESTION

Explain why a language model can produce a fluent answer that is factually wrong (hallucination). **[6] Refer to:** choosing statistically plausible words.

PAUSE & THINK

"Plausible" or "correct" — which one does the model actually optimize for?

(3) From the following options (A–D), choose the one that most appropriately describes RLHF.

A A method of building a language model from scratch using a large amount of text.

B A method of tuning a language model to return more desirable responses, using human evaluations as rewards.

C A method that removes the neural network structure from a language model and simplifies the processing.

D A method that has a language model solve math problems to improve its computational accuracy.

(4) From the following options (A–D), choose the one that is NOT an appropriate application of multimodal AI.

A Showing an image and having it described in text.

B Listening to audio and transcribing it into text.

C Giving instructions in text and generating an image.

D Computing the result of input numerical values using only the rules of the four arithmetic operations and returning it.

⚙ Think as an Engineer — Investigate & Decide

A school club wants to use a generative-AI chatbot to help write short articles about local history for the school website. You advise on using it responsibly.

☐ **Publish whatever the AI writes** ☐ **Use the AI, but verify before publishing**

1 · Decide where it helps. Name one part of the task where the chatbot genuinely saves effort, and one part where its output must not be trusted as-is. Explain each in one line.

2 · Analyze the risk. Using how a language model chooses words, explain why the chatbot might state a wrong date or invented name for a local landmark very confidently, and describe exactly how the club would catch such a hallucination before publishing — a check the lesson does not spell out for you.

3 · Decide the safeguards. State the rule the club should follow for every AI-written article before it goes live, and say what role RLHF-style tuning does and does not solve here. **In pairs,** compare rules and agree on the one check you would never skip.

Give evidence for your answer.

KEY TERMS

generative AI — AI that creates new text, images, or audio.

fact-check — verifying a claim against a reliable source.

Stuck? Remember the model is rewarded for *sounding* right, not for *being* right. Which items in a history article — dates, names, places — are exactly the kind of thing it could get confidently wrong? Think about who or what could confirm each one, and at what point a person should step in.

🌎 In a New Context

A student uploads a photo of a handwritten note and asks an AI assistant to read it aloud and then summarize it in text. Name the property that lets one model move between image, text, and audio, and explain in one line how the network can handle all three.

❓ Lesson Question — Answered

The lesson asked how a language model generates text, what makes a model "large," and why it can still get facts wrong. A language model learns from large amounts of text to predict the likely next word, and it generates text by repeating that prediction, connecting one word at a time into a whole sentence. It chooses each word by a probability of coming next — the statistically plausible word, not the guaranteed-correct one — which is exactly why hallucination (fluent but false responses) can occur. A large language model scales this up: trained on vast Internet text with billions to trillions of parameters, and beyond a certain scale it gains qualitatively new abilities such as reasoning, summarizing, translating, and generating code (ChatGPT, Claude, and Gemini are examples). Because raw training does not guarantee helpful answers, RLHF tunes the model with human evaluations as rewards, making responses natural, on-intent, and safer. And multimodality lets a single model handle text, images, and audio together by converting them all into a common numerical representation. The lasting caution: fluency is not truth, so outputs must still be checked.

Exercise

1 Cover the Point! section with a red plastic sheet and quiz yourself in your notebook in numerical order.

2 Read the following passage and answer each question.

A model that predicts which word is likely to come next in a sentence is called a (a). Text generation is achieved by repeating the prediction of the (b). A language model based on a very large neural network trained using vast

⭐ REFLECT & CHALLENGE

Reflect: When you last used a generative-AI tool, did you check its facts? What could a confident wrong answer have cost? **Challenge:** A classmate says "RLHF makes the AI always tell the truth." Explain why that is not correct, and state what RLHF actually improves.

amounts of text from the Internet is called a (c); when the number of (d) or the amount of training data exceeds a certain scale, qualitatively new abilities sometimes appear. Since just learning from vast text does not necessarily lead to responses desirable for humans, (e), a method of tuning by reinforcement learning using human evaluations as rewards, is used. Also, the property of being able to handle multiple kinds of data—text, images, audio, video, and so on—with a single model is called (f).

1. Fill in the blanks (a)–(f).

3 Answer the following questions.

(1) From the following options (A–D), choose the one that most appropriately describes the mechanism of text generation.

A Text is generated by connecting the results of repeatedly predicting the next word.

B Text is generated by selecting from thousands of prepared templates.

C Text is generated by first deciding the length of the sentence and then deciding all words simultaneously.

D Text is generated by machine-translating the input sentence.

(2) From the following options (A–D), choose the one that most appropriately describes the background to large language models acquiring new abilities.

A Because reducing the number of parameters below a certain level simplified the processing.

B Because limiting training data to a small amount of high-quality text raised quality.

C Because qualitatively new abilities appear when the number of parameters or the amount of training data exceeds a certain scale.

D Because humans wrote down the rules of language one by one and taught them to the model.

(3) From the following options (A–D), choose the one that most appropriately describes the purpose for which RLHF is performed.

A To increase the number of words a language model can handle.

B To make a language model's responses natural and aligned with the intent of the question.

C To enable a language model to handle data other than text.

D To reduce the number of parameters used in a language model's calculations.

⭐ Key takeaway

Remember:
A language model generates text by predicting the next word (plausible, not always true → hallucination). LLMs use vast data and billions to trillions of parameters; RLHF tunes them by human feedback; multimodality unifies text, image, and audio. Fluency ≠ truth — always verify.

Editorial Team

Project Lead & Editorial Director

Masaki Iisaka

Editors & Writers

Biichi Nakajima
Hiromi Fujii

Reviewers

Prof. Mohamed Sayed Farag.

Prof. Eman Serag El-Din Bakr.

Dr. Abeer Hamed Ahmed.

Dr. Taher Abdel Hamid El-Adly

Dr. Manal Ziada Abdel Fadil.

The Central Administration for Curriculum Development

General Supervision

Dr. Gebriel Anwer Hemieda

Head of the Central Administration for Curriculum Development

EDUCATION SPRINGS NEW LIFE

class	name