ELEC3544 · Fall 2026

Data Science with Foundation Models

Yang Liu

The University of Hong Kong



Founding Members of ELEC3544!

Instructor

Invented and polished this course from 3 years ago!



Yanchao Yang

Assistant Professor
School of Computing and Data Science
Electrical and Computer Engineering
Institute of Data Science
The University of Hong Kong
Email: yanchaoy at hku dot hk
Office: Room 714, Chow Yei Ching Building, HKU

https://yanchaoyang.github.io/

TAs and Students

From last year:

Zhengyang Hu:
u3010250@connect.hku.hk

Lulu Cai:
u303011209@connect.hku.hk

Chengwei Zhang:
u3013969@connect.hku.hk

Yushi Du:
u3012725@connect.hku.hk

And ALL TAs and students from previous years!!!

> Thanks for starting and contributing to such a unique and visionary course!

Quick survey – Are you using AI agents?



Anthropic Fable 5



OpenAI GPT-5.6 Sol



Kimi K3



GLM 5.2



Google Gemini 3.7



Grok 4.5



Qwen 3.8 Max



DeepSeek V4 Pro



Muse Spark 1.2



ByteDance Doubao

...

Why This Course, Why Now?

1. **A Convergence Moment**

   Powerful AI tools are now accessible to everyone. The barrier to entry has never been lower.

2. **First Generation of AI-Native Data Scientists**

   You are the first a few cohorts learning data science with AI tools from day one — not as an afterthought.

3. **The Rules Are Being Rewritten**

   Traditional workflows are transforming. Those who adapt thoughtfully will thrive.

> This course prepares you to think critically about **when** and **how** to leverage AI tools effectively.

The Evidence Is Clear: AI Is Transforming Data Science

TOP SKILLS IN JOB POSTINGS (2025)

Python	85%
Machine Learning	77%
SQL	59%
NLP Skills	5% → 19%

NLP demand nearly quadrupled due to LLM proliferation

NEW ROLES EMERGING

LLM Engineer
$123K - $224K

AI Engineer
Building apps on foundation models

MLOps Engineer
Model deployment lifecycle

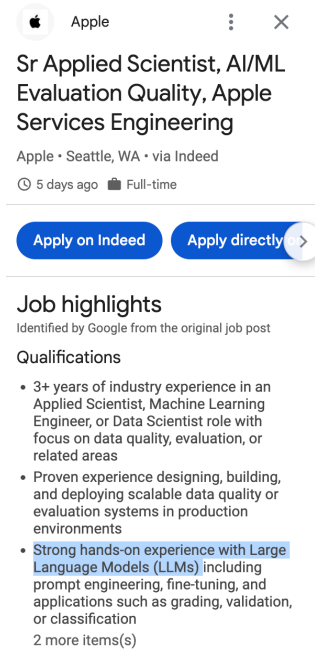
AI Product Manager
Bridging tech and business

MARKET REALITY

$152K
Entry-level salary (up $40K from 2024)

But experience requirements increasing.

Employers want **fewer, more capable** practitioners.



The message: Data science is evolving rapidly. Those who combine strong fundamentals with AI fluency will thrive.

The Course Vision: From Doer to Director

When AI can write code, clean data, and generate analyses, your value shifts from **producing outputs** to **directing, evaluating, and ensuring correctness**.

THE OLD MODEL

Value = ability to produce

Write code from scratch
Memorize syntax and formulas
Execute procedures manually
Speed of production

THE NEW MODEL

Value = ability to direct and evaluate

Know when AI is wrong
Understand why methods work
Ask the right questions
Quality of judgment

This course trains you for the new model: Strong fundamentals + AI fluency + critical judgment = irreplaceable professional value.

Meet Your Teaching Team



Instructor

Yang Liu

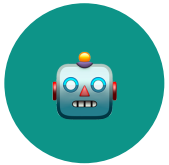
Lectures, course design, office hours, and overall guidance



Teaching Assistants

Human TAs

Grading, project guidance, and office hours support



Chief Teaching Assistant

ChatGPT / DeepSeek

Available 24/7 for learning support, debugging, and exploration

> The CTA is your always-available learning partner — use it to explain concepts, debug code, and explore ideas.

CTA Options



Anthropic Fable 5



OpenAI GPT-5.6 Sol

Best! Needs subscription and VPN



Kimi K3



GLM 5.2



Google Gemini 3.7

Free! Available in HK; free student plan



Grok 4.5



Qwen 3.8 Max



DeepSeek V4 Pro

Free! Available in HK



Muse Spark 1.2



ByteDance Doubao

...

Grading Breakdown

10% **Attendance and Participation**
Be present and engaged in lectures. Speak up when you have questions!

35% **Homework Assignments (3~4 in total)**
Bi-weekly problem sets applying lecture concepts

25% **Midterm Exam (in-class 1.5~2h; around reading week)**
Closed-book — tests foundational understanding

30% **Group Project (3 members per group)**
End-to-end data science project with AI collaboration

Note: Late submissions penalized 10% per day. Extensions available with advance notice.

What Success Looks Like

By the end of this course, you should be able to answer YES to these questions:

1. Can I take a real-world problem and design a data science approach to solve it?

2. Can I use AI tools to accelerate my work while maintaining quality and accuracy?

3. Can I critically evaluate AI outputs and know when to trust or question them?

Today's Journey

Experience the complete data science workflow using your own data

What is Data Science?

Understanding the landscape, the workflow, and where foundation models fit in

Defining Data Science

Data science is the practice of extracting meaningful insights from data using a combination of domain expertise, programming skills, and statistical knowledge.



Domain Expertise

Understanding the problem context and what questions matter



Programming Skills

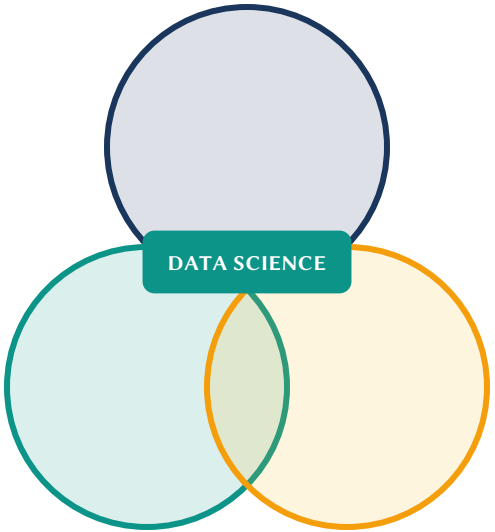
Tools to collect, clean, and process data at scale



Statistical Knowledge

Methods to analyze, model, and draw valid conclusions

The Data Science Venn Diagram



- **Mathematics and Statistics**
  Probability, inference, modeling
- **Computer Science**
  Programming, algorithms, data structures
- **Domain Expertise**
  Industry knowledge, business context

> Data science lives at the intersection. You need all three to be effective.

Data Science vs. Related Fields

FIELD	PRIMARY FOCUS	KEY DISTINCTION
Data Science	Extract insights and inform decisions	End-to-end problem solving
Machine Learning	Build predictive models from data	Subset focused on algorithms
Statistics	Analyze and interpret data rigorously	Mathematical foundations
Data Engineering	Build data infrastructure and pipelines	Infrastructure focus
Business Analytics	Drive business decisions with data	Business strategy focus

These fields overlap significantly. Data scientists often use techniques from all of them.

The Data Science Workflow



Reality check: This process is rarely linear. You'll often loop back as you discover new questions or data issues.

Data Types You'll Encounter



CATEGORICAL

Major, device type, country — distinct groups, no order



ORDINAL

Year in school, ratings (1-5) — ordered categories



NUMERICAL

Sleep hours, age, GPA — measurable quantities



TEXT

Comments, goals, open responses — unstructured text

> **Key insight:** Real datasets contain all these types. Each requires different approaches.

From Theory to Practice

You now have the vocabulary. Next, you'll experience the full workflow using your own data.

WHAT WE'LL DO

Complete a 5-minute survey to generate real data. Then we'll clean it, analyze it, visualize it, and even have AI extract insights from your text responses — all using data you just created.

Why this matters

Experiencing the workflow with your own data makes these concepts concrete.

Hands-On Data Collection

Creating and gathering real data

Why Create Your Own Survey?

CONTROL

You decide exactly what questions to ask and how to phrase them

FRESHNESS

Data is current and collected specifically for your analysis

RELEVANCE

Questions tailored precisely to your research question

OWNERSHIP

Full rights to use, share, and publish your collected data

Key insight: Primary data collection is often the only way to answer novel research questions.

Choosing a Survey Tool

GOOGLE FORMS

Free, easy sharing, integrates with Google Sheets for analysis

MICROSOFT FORMS

Free with Office 365, similar features to Google Forms

QUALTRICS

Advanced logic and branching, academic standard, HKU has access

TYPEFORM

Beautiful conversational UI, higher engagement rates

Key insight: We'll use Google Forms today—free and powerful enough for most projects.

Anatomy of a Good Survey

Clear language

Avoid jargon, technical terms, and ambiguous wording

One concept per question

Don't ask two things at once (double-barreled questions)

Logical flow

Group related questions, order from easy to sensitive

Appropriate length

Respect respondents' time—5-10 minutes is ideal

Test first

Always pilot with a few people before full deployment

Warning: Poorly designed surveys = unreliable data that can't be fixed later

Question Types Matter

MULTIPLE CHOICE → Categorical

Major, device type, yes/no—distinct groups, no order

LINEAR SCALE → Ordinal

Satisfaction 1-5, confidence levels—ordered categories

NUMBER INPUT → Numerical

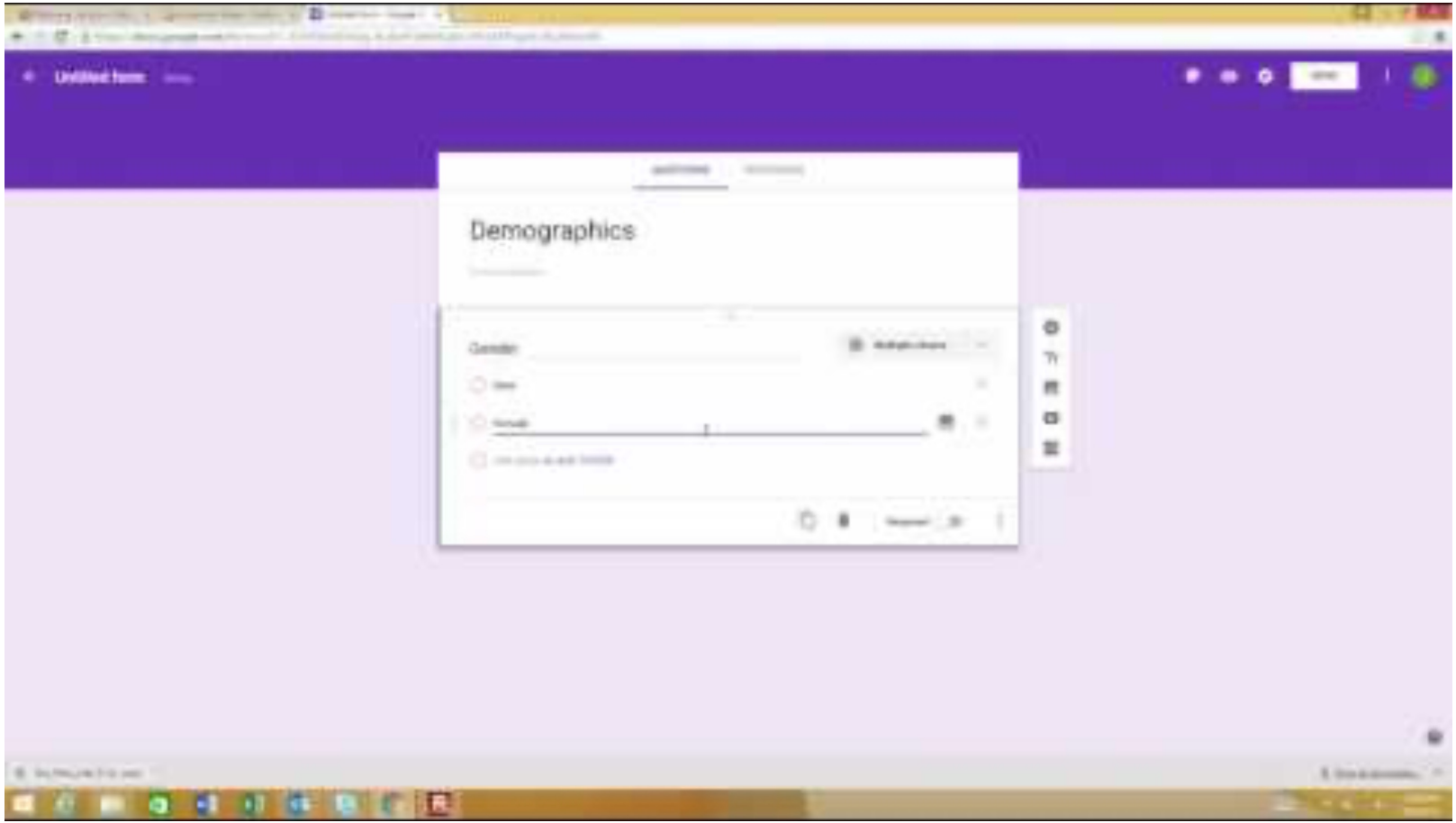
Age, hours of sleep, GPA—measurable quantities

PARAGRAPH → Text

Goals, concerns, feedback—free-form language for NLP

Key insight: Think from a top-down view: choose question types based on how you plan to analyze the data.

What the manual procedure looks like



LIVE DEMO

Building Our Class Survey with CTA

Watch how we construct the survey you're about to take

WHAT WE'LL COVER

- ☑ Create a new Google Form
- ☑ Add different question types
- ☑ Set required vs optional questions
- ☑ Preview and generate share link

All these can be done by talking with Foundation Models

Take the Survey Now!



Scan with your phone camera

DIRECT LINK

https://forms.gle/TKDAaLKkutpxVBWd9

QUICK TIPS

The survey is anonymous. Some questions are optional — skip any you prefer not to answer. Takes about 5 minutes.

You have **5 minutes** — start now!

DATA COLLECTION

Where Does Data Come From?

PRIMARY

Data you collect yourself — surveys, experiments, observations, interviews

Like our class survey!

SECONDARY

Existing datasets — government data, research datasets, public repositories

Kaggle, UCI ML Repository

APIs

Programmatic access — Twitter, weather services, financial data providers

Real-time, structured

WEB SCRAPING

Extracting data from websites — news articles, product info, reviews

Requires ethical care

The best data scientists know multiple collection methods and choose based on the question at hand.

Finding Existing Datasets

Kaggle
Competitions, datasets, notebooks from ML community

UCI ML Repository
Classic datasets for machine learning research

Google Dataset Search
Search engine specifically for datasets

Government Open Data
data.gov, census data, health statistics

Academic Repositories
Harvard Dataverse, Zenodo, figshare

Hugging Face
Datasets for NLP, computer vision, and more

Pro tip: Always check data licenses and documentation. Understanding how data was collected is crucial for valid analysis.

Data Cleaning

Turning messy reality into analysis-ready data

What Data Did We Just Collect?



CATEGORICAL

Major, device type, social media platform — distinct categories



ORDINAL

Year in school, confidence levels — ordered categories



NUMERICAL

Sleep hours, commute time, courses — measurable quantities



TEXT

Learning goals, concerns, mood — unstructured language

> **Key insight:** Real datasets often contain all these types. Each requires different cleaning and analysis approaches.

Common Data Formats

CSV

Comma-Separated Values

Simple tabular data. Easy to read and write. Universal support.

```
name,age,city
Alice,25,HK
```

JSON

JavaScript Object Notation

Nested, hierarchical data. API standard. Human-readable.

```
{"name": "Alice",
"age": 25}
```

Excel

.xlsx, .xls files

Multiple sheets, formatting, formulas. Business standard.

```
Use pandas:
pd.read_excel()
```

Parquet

Columnar format

Efficient for big data. Compressed. Fast queries.

```
Use pandas:
pd.read_parquet()
```

In this course: We'll primarily work with CSV and JSON. Pandas makes loading any format easy with read_csv(), read_json(), etc.

Why Data Cleaning Matters

Garbage In, Garbage Out

Even the most sophisticated analysis or model will produce misleading results if the input data is flawed.

WITHOUT CLEANING

✗ Incorrect conclusions from outliers

✗ Models that fail on edge cases

✗ Visualizations that mislead

✗ Biased or unfair predictions

WITH PROPER CLEANING

✓ Reliable statistical summaries

✓ Robust, generalizable models

✓ Clear, honest visualizations

✓ Trustworthy insights

Real talk: Data cleaning isn't glamorous, but it's where good data science is won or lost.

Common Data Quality Issues

Missing Values
Empty cells, NaN, null, "N/A", blanks — need strategy to handle

Wrong Data Types
"eight" instead of 8, "20 mins" instead of 20

Duplicates
Same person submitted twice, copy-paste errors

Inconsistent Formats
"Year 1" vs "yr 1" vs "1st year" — same meaning, different strings

Outliers
999 minute commute, -2 hours sleep — errors or genuine extremes?

Invalid Values
Age = 500, email without @, dates in wrong format

Key insight: Each issue type requires a different cleaning strategy. There's no one-size-fits-all solution.

Your Raw Survey Data

survey_responses.csv (sample rows showing common issues)

major, year, sleep_hours, commute_min, python_conf

CE, Year 2, 7, 30, 4

engineering, yr 3, eight, 20 mins, 3

Math, Year 2, -2, 999, 5

Issues We Can Spot

Inconsistent formats, text in number fields, impossible values, missing responses

What We'll Fix

Standardize categories, convert types, handle outliers, deal with missing values

This is real! Even with careful survey design, humans enter data inconsistently.

Handling Missing Values

1. **Delete rows with missing data**
   Simple but loses information. Only if missing data is small percentage.

2. **Fill with a constant value**
   Use 0, "Unknown", or a placeholder. Quick but may introduce bias.

3. **Impute with statistics (mean, median, mode)**
   Preserves distribution shape. Most common approach for numerical data.

4. **Use predictive models to impute**
   Most sophisticated. Predict missing values from other features.

Critical question: Why is the data missing? Random vs systematic missingness affects your strategy choice.

Dealing with Outliers

HOW TO DETECT

Visual inspection
Box plots, scatter plots, histograms

Statistical methods
Z-score (>3), IQR method (1.5×IQR)

Domain knowledge
Is 999 min commute possible? Is -2 hrs sleep?

WHAT TO DO

If error → Fix or remove
Typos, data entry mistakes

If genuine → Keep or cap
Real extreme values — document decision

Consider robust methods
Median instead of mean, robust regression

Key principle: Never blindly remove outliers. Investigate first — sometimes outliers are the most interesting data points!

AI-Assisted Data Cleaning

WHAT AI DOES WELL

- ✓ Generates boilerplate cleaning code
- ✓ Suggests multiple approaches to try
- ✓ Explains what each step does
- ✓ Handles common patterns automatically

WHAT STILL NEEDS YOU

- ⚠ Deciding what counts as "valid" data
- ⚠ Domain-specific validation rules
- ⚠ Verifying whether the code works
- ⚠ Ethical decisions about data handling

Key takeaway: AI accelerates the mechanical work; you provide judgment and context.

Data Normalization and Standardization

Min-Max Normalization

Scales values to a fixed range, typically 0 to 1

$$x_norm = (x - \min) / (\max - \min)$$

Use when: Bounded range needed, no extreme outliers

Z-Score Standardization

Centers data with mean=0 and std=1

$$z = (x - \text{mean}) / \text{std}$$

Use when: Comparing across features, algorithms need it

Why Normalize?

Many ML algorithms (like neural networks, KNN, SVM) are sensitive to feature scales. Without normalization, features with larger ranges dominate the model unfairly.

Tip: Know when to use each method. This concept appears frequently in ML courses and interviews.

Data Analysis

Extracting insights from your cleaned data

Exploratory Data Analysis (EDA)

EDA is the process of examining data to understand its structure, patterns, anomalies, and relationships before formal modeling. It's detective work for data.



Summarize

Mean, median, mode, standard deviation, quartiles. Get the big picture numbers.



Visualize

Histograms, box plots, scatter plots. See distributions and relationships.



Question

What patterns exist? What's surprising? What needs more investigation?

EDA philosophy: Let the data speak first. Resist the urge to jump to conclusions before thoroughly exploring.

Descriptive Statistics: Summarizing Your Data

MEASURES OF CENTER

Mean (Average)
Sum of all values ÷ count. Sensitive to outliers.

Median (Middle)
Middle value when sorted. Robust to outliers.

Mode (Most Common)
Most frequent value. Works for categorical data.

MEASURES OF SPREAD

Range
Max – Min. Simple but affected by outliers.

Standard Deviation
Average distance from mean. Most common measure.

IQR (Interquartile Range)
Q3 – Q1. Middle 50% of data. Robust to outliers.

> **Pro tip:** Always report both center and spread. A mean without standard deviation tells only half the story.

Your Survey Statistics

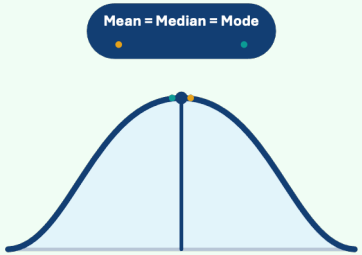
You can compute descriptive statistics on your actual survey responses using Python and pandas.

```
# What we'll run:
df.describe()  # Summary stats
df['sleep_hours'].mean()  # Average sleep
df['major'].value_counts()  # Major distribution
```

WATCH FOR

How do YOUR stats compare to what you expected? Any surprises in the class data?

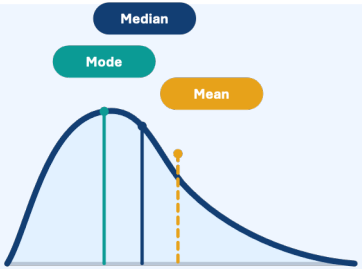
Understanding Distributions



Normal (Bell Curve)

Symmetric around mean.

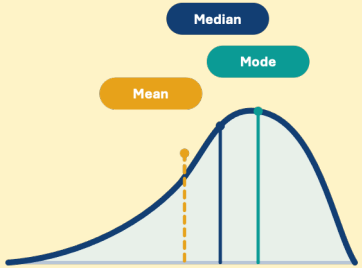
Mean ≈ median.
Example: Heights, test scores



Right-Skewed

Long tail to the right.

Mean > median.
Example: Income, home prices



Left-Skewed

Long tail to the left.

Mean < median.
Example: Age at retirement

> **Why it matters:** The shape of your distribution determines which statistics and methods are appropriate. Skewed data often needs median over mean.

Correlation: Measuring Relationships

Correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, where -1 indicates perfect negative correlation, 0 indicates no correlation, and +1 indicates perfect positive correlation. Ranges from -1 (perfect negative) to +1 (perfect positive).

+1

Positive Correlation

As X increases, Y increases. More study hours ⟶ higher grades

0

No Correlation

No linear relationship. Shoe size ⟶ GPA (hopefully!)

-1

Negative Correlation

As X increases, Y decreases. More partying ⟶ lower grades

⚠️ **CRITICAL:** Correlation ≠ Causation. Ice cream sales and drowning deaths are correlated, but ice cream doesn't cause drowning!

Data Visualization: Seeing Your Data

> *"The greatest value of a picture is when it forces us to notice what we never expected to see."*
> — John Tukey, pioneer of exploratory data analysis

Reveal Patterns

Trends, clusters, and relationships that numbers alone can't show

Spot Problems

Outliers, errors, and data quality issues become obvious visually

Communicate

A good chart can convey insights instantly to any audience

Coming up: We'll create visualizations of your survey data using matplotlib and seaborn.

Choosing the Right Chart

Histogram
Distribution of one numerical variable
e.g., sleep hours distribution

Box Plot
Distribution + outliers + comparison
e.g., sleep by year level

Pie Chart
Parts of a whole (use sparingly!)
e.g., device usage breakdown

Bar Chart
Comparing categories
e.g., students per major

Scatter Plot
Relationship between two variables
e.g., commute vs study hours

Line Chart
Trends over time
e.g., survey responses over time

> **Rule of thumb:** Choose based on what question you're answering, not what looks prettiest.

Visualizing Your Survey Data

Creating charts from your actual responses using matplotlib and seaborn

CHARTS WE'LL CREATE

- Histogram of sleep hours distribution
- Bar chart of majors represented
- Box plot comparing sleep by year level
- Scatter plot exploring relationships

DISCUSSION QUESTIONS

- What patterns do you see?
- Anything surprising in the data?
- What additional charts would help?

[CTA demonstrates in Google Colab notebook using matplotlib or seaborn]

AI for Analyzing Text Responses

The challenge: How do you analyze 100+ free-text responses to "What do you hope to learn?"

TRADITIONAL APPROACH

- Manual coding of themes
- Word frequency counts
- Keyword matching

Time-consuming, subjective, misses nuance

AI-ASSISTED APPROACH

- Automatic theme extraction
- Sentiment analysis
- Semantic understanding

Fast, consistent, but requires validation

Critical thinking: How do we know the AI's categorization is accurate? Always validate AI outputs.

Next: We'll analyze YOUR free-text survey responses using AI and critically evaluate the results.

Foundation Models (AI Agents)

Understanding AI as your learning partner

What is a Foundation Model?

> A **foundation model** is an AI system trained on massive amounts of data that can be adapted to many different tasks. Examples include ChatGPT, Claude, and DeepSeek.

🔄

Pre-trained

Learned from billions of text examples before you use it

🎯

General Purpose

Can answer questions, write code, analyze data, and more

💬

Conversational

Responds to natural language prompts in dialogue format

In this course: We can use DeepSeek as your Chief Teaching Assistant (CTA) to help you learn and work with data.

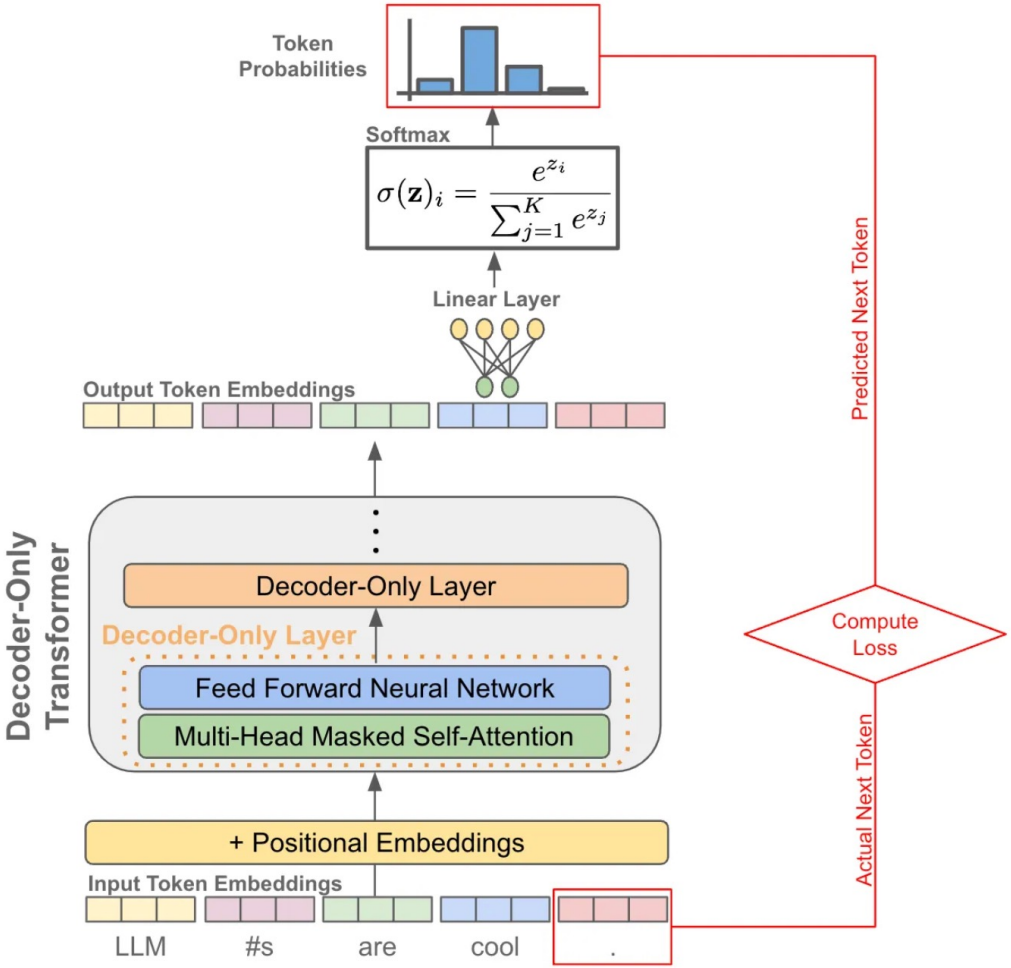
Watch an LLM Think

Let's observe how a foundation model generates responses token by token, revealing how it "thinks."

WHAT WE'LL OBSERVE

- → Token-by-token generation
- → How context shapes output
- → The role of probability
- → Why responses vary

Key insight: LLMs predict the next word based on everything that came before.


Token Probabilities
Softmax
$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
Linear Layer
Output Token Embeddings
Decoder-Only Transformer
Decoder-Only Layer
Decoder-Only Layer
Feed Forward Neural Network
Multi-Head Masked Self-Attention
+ Positional Embeddings
Input Token Embeddings
LLM
#s
are
cool
.
Predicted Next Token
Compute Loss
Actual Next Token

What Can Foundation Models Do?

EXPLAIN CONCEPTS

Break down complex ideas into understandable explanations at your level

WRITE CODE

Generate, debug, and explain Python, SQL, and other programming languages

ANALYZE DATA

Help interpret results, suggest visualizations, and identify patterns

DERIVE MATH

Walk through statistical formulas and mathematical proofs step by step

DEBUG ERRORS

Help identify and fix bugs in your code with clear explanations

BRAINSTORM

Generate ideas for projects, approaches to problems, and creative solutions

Important: These capabilities make AI a powerful learning tool, but you must verify outputs and understand the underlying concepts.

AI Limitations: What Can Go Wrong?

Critical: Understanding these limitations is essential for using AI responsibly and effectively.

HALLUCINATIONS

AI can generate confident-sounding but completely false information, including fake citations and invented facts.

MATH ERRORS

LLMs predict words, not compute math. They can make arithmetic and calculation mistakes.

OUTDATED KNOWLEDGE

Training data has a cutoff date. AI may not know recent events, updated APIs, or new research.

BIAS & STEREOTYPES

AI can reflect biases from training data, including gender, racial, and cultural stereotypes.

The rule: Always verify AI outputs. Trust but verify — especially for facts, calculations, and code.

Prompt Engineering Principles

The quality of AI output depends entirely on the quality of your input. Master these four principles.

1. Be Specific and Clear

Vague prompts produce vague answers. State exactly what you need, including format, length, and scope.

2. Provide Context

Share relevant background, constraints, and goals. The AI cannot read your mind — tell it what it needs to know.

3. Iterate and Refine

First responses are starting points. Follow up, ask for clarification, request changes. Prompting is a conversation.

4. Verify and Validate

Check outputs against known facts. Test code before using. You are responsible for the final product.

Remember: The quality of your output is directly proportional to the quality of your input.

Effective Prompting Techniques

CHAIN OF THOUGHT

Ask the model to reason step by step before giving an answer.

"Think through this step by step before answering..."

When: Math, logic, complex analysis

ROLE PROMPTING

Assign the model a specific persona or expertise level.

"Act as a senior data scientist reviewing this code..."

When: Code review, explanations, feedback

STRUCTURED OUTPUT

Request a specific format for the response.

"Respond in JSON format with fields for..."

When: Data processing, APIs, automation

Pro tip: These techniques dramatically improve response quality for complex tasks. Combine them as needed.

Data Science-Specific Prompts

Different tasks require different prompting patterns. Here are templates you'll use throughout this course.

FOR CODE

Specify language, libraries, input/output formats, and error handling

"Write Python code using pandas to... Input is CSV with columns... Return a DataFrame with..."

FOR ANALYSIS

Describe data structure, analysis goal, and what insights you're looking

"I have a dataset with [columns]. I want to understand [goal]. Suggest approaches to..."

FOR DEBUGGING

Include error message, what you've tried, and expected vs. actual behavior.

"I get this error: [error]. I expected [X] but got [Y]. I've already tried [Z]..."

FOR EXPLANATION

State your current understanding level and what specifically confuses you.

"I understand [A] but I'm confused about [B]. Can you explain with a concrete example?"

These patterns will be your daily tools. Save them and adapt them to your specific needs.

Prompting Techniques

Prompt Engineering helps to effectively design and improve prompts to get better results on different tasks with LLMs.

While the previous basic examples were fun, in this section we cover more advanced prompting engineering techniques that allow us to achieve more complex tasks and improve reliability and performance of LLMs.

- Zero-shot Prompting
- Few-shot Prompting
- Chain-of-Thought Prompting
- Meta Prompting
- Self-Consistency
- Generate Knowledge Prompting
- Prompt Chaining
- Tree of Thoughts
- Retrieval Augmented Generation

Check it out: https://www.promptingguide.ai/techniques

Iterative Prompt Refinement

Watch how a prompt evolves through iteration — from basic to highly effective.

THE REFINEMENT PROCESS

STEP 1
Basic prompt

→

STEP 2
Add context

→

STEP 3
Request format

→

STEP 4
Handle edge cases

Key insight: Prompting is a conversation, not a single query. Each iteration improves the result.

AI Ethics and Responsible Use

ACADEMIC INTEGRITY

AI helps you learn, but you must understand your work.

✓ Use AI to understand concepts

✗ Submit AI output as your own work

BIAS AWARENESS

Question data sources and check for fair representation.

✓ Consider who might be harmed

✗ Accept AI outputs uncritically

TRANSPARENCY

Be honest about how AI contributed to your work.

✓ Acknowledge AI assistance

✗ Hide your use of AI tools

Simple rule: Can you explain and defend every part of your work? If not, you haven't learned it yet.

Working with AI: Best Practices

DO ✓	DON'T ✗
Use AI to **understand** concepts you're learning	Submit AI-generated work as your own
Ask for **explanations** of code or formulas	Trust AI outputs without verification
Verify all outputs against trusted sources	Use AI as a shortcut to avoid learning
Use AI to **debug** and learn from your mistakes	Hide or misrepresent your AI usage

Remember: AI is a learning tool, not a replacement for learning. You are responsible for your own understanding.

Critical Thinking with AI: The VERIFY Framework

Use this framework every time you work with AI outputs to ensure quality and accuracy.

V **Validate** — Check facts against reliable sources. Don't trust claims without evidence.

E **Execute** — Run the code. Test the solution. See if it actually works as claimed.

R **Reason** — Does the logic make sense? Can you follow and explain the reasoning?

I **Inspect** — Look for edge cases, potential errors, and things that might break.

F **Fix** — Iterate and improve. Use AI to help fix issues you identify.

Y **Yours** — Make it your own. Understand it well enough to explain and defend it.

We'll practice this framework throughout the course. It becomes second nature with repetition.

Summary: Foundation Models

Foundation models are powerful learning partners

Pre-trained on massive data, they can explain, code, analyze, and debug — but you must verify their outputs.

Prompt engineering is a learnable skill

Be specific, provide context, iterate, and verify. Use techniques like chain-of-thought, role prompting, and structured output.

AI has real limitations you must understand

Hallucinations, outdated knowledge, math errors, and bias. Trust but always verify using the VERIFY framework.

Responsible use is non-negotiable

Use AI to learn and understand — never as a shortcut. Be transparent about AI assistance. You own your work.

Closing & Next Steps

Wrapping up and looking ahead

Today's Key Takeaways

1. **Data science is a workflow** — from question to collection to cleaning to analysis to insight.

2. **Domain knowledge is important** — expertise can help with more efficient survey design.

3. **Data cleaning is 80% of the work** — garbage in, garbage out. Quality data enables quality insights.

4. **AI is your learning partner** — foundation models help you learn, but you must understand the work.

5. **Always VERIFY** — Validate, Execute, Reason, Inspect, Fix, make it Yours.

These principles will guide everything we do this semester.

What You'll Build This Semester



Homework Assignments

Weekly practice applying concepts to real datasets. Build skills progressively throughout the semester.

35% of grade



Course Project

Team project on a real data science problem. End-to-end experience from question to presentation.

30% of grade



Exams

Midterm tests foundations. No access to AI.

25% of grade

+ **10% Attendance:** Active participation matters. By semester's end, you'll have a portfolio demonstrating your data science capabilities.

Your Learning Resources

AI-powered learning with traditional support systems

PRIMARY RESOURCE

Your CTA (Chief Teaching Assistant)

Claude and ChatGPT are your AI learning partners. Use them to understand concepts, debug code, explore ideas, and prepare for assignments.

HOW TO USE YOUR CTA

- → Explain concepts step-by-step
- → Debug your code with guidance
- → Generate practice problems
- → Review before submitting work

Remember: Always VERIFY — Validate, Execute, Reason, Inspect, Fix, make it Yours.

Course Materials

Syllabus, slides, assignments, and all materials.

Discussion Forum

Ask questions, share insights. TAs monitor daily.

Office Hours

Weekly drop-in with instructor and TAs.

Start with your CTA for immediate help. Escalate to human support for deeper discussion.

Assignment 0: Getting Started

Due: Before next class — This ensures you're ready to work from Day 1.

1

Set Up Your Environment

Install Python and Jupyter, or set up Google Colab. Run the test notebook to confirm everything works.

2

Have a CTA Conversation

Ask the AI to explain a data science concept. Practice the prompting techniques we discussed today.

3

Submit Reflection

Answer brief reflection questions about your setup experience and your AI conversation.

Purpose: This ensures everyone starts Lecture 2 with working tools and initial AI experience.

Questions?

Thank you for an engaging first lecture!

CONTACT INFORMATION



yangliu.mit.edu

Instructor: Prof. Yang Liu

Email: yliu12@hku.hk

Office Hours: 5:00 - 5:50 PM, Thursdays

Location: CB-714, Chow Yei Ching Building

Let me know if you want remote!

Remember: You're entering data science at an extraordinary moment.

The skills you build this semester will serve you throughout your career. Let's make it count.

TA contacts

Zhengyang Hu:
u3010250@connect.hku.hk

Lulu Cai:
u303011209@connect.hku.hk

Yushi Du:
u3012725@connect.hku.hk

Jun Dai:
jundai332@gmail.com

TA office hours to be released later this week! Two slots (1 hour each) on Tuesday and Thursday / Friday

Watching materials



https://youtu.be/7xTGNNLPyMI



Andrej Karpathy @karpathy · Dec 27, 2025

I've never felt this much behind as a programmer. The profession is being dramatically refactored as the bits contributed by the programmer are increasingly sparse and between. I have a sense that I could be 10X more powerful if I just properly string together what has become available over the last ~year and a failure to claim the boost feels decidedly like skill issue. There's a new programmable layer of abstraction to master (in addition to the usual layers below) involving agents, subagents, their prompts, contexts, memory, modes, permissions, tools, plugins, skills, hooks, MCP, LSP, slash commands, workflows, IDE integrations, and a need to build an all-encompassing mental model for strengths and pitfalls of fundamentally stochastic, fallible, unintelligible and changing entities suddenly intermingled with what used to be good old fashioned engineering. Clearly some powerful alien tool was handed around except it comes with no manual and everyone has to figure out how to hold it and operate it, while the resulting magnitude 9 earthquake is rocking the profession. Roll up your sleeves to not fall behind.

2.6K 10K 55K 16M

One year later after the first launch of our course

