

Recap: What is Data Science?

DS701 Session 2 — Wed Sep 9, 2026

Today's plan

Knowledge-Check Review

KC 1: The taxonomy

List the six question types from the lecture's taxonomy (Leek & Peng), in order from least to most demanding, and for any three of them say in a few words what that type of question asks.

KC 2: The churn model

A team builds a churn model with excellent held-out accuracy and notices that customers who received a discount rarely churn. They recommend launching the discount to everyone. Which question type was asked, which was answered, and why does the model's accuracy not settle the matter?

KC 3: Your own question

Pick one question you would like to answer with data — from your own background, a Spark pitch, or anything you are curious about. State the question in one sentence, classify it using the taxonomy, and name the row of the lecture’s techniques map (“You want to... / Technique”) you would reach for first, with one sentence on why that row fits.

Highlights

Data science in one line

Data science is the discipline of **drawing defensible conclusions and making decisions from data**, combining statistical reasoning, computation, and domain understanding.

First, classify the question

Before choosing a technique, classify what kind of question you are asking (after Leek & Peng, *Science* 2015):

Question type	Asks	Example
Descriptive	What does the data say, summarized?	Average ride time by hour
Exploratory	What patterns are worth a closer look?	Riders cluster into commuter/leisure groups
Inferential	Does the pattern hold beyond this sample?	Do commuters <i>in general</i> ride longer in rain?
Predictive	What will happen for a new case?	Will this user churn next month?
Causal	What happens if we <i>intervene</i> ?	Does the discount <i>cause</i> retention?
Mechanistic	Exactly how does the effect	Through which pathway

Type errors are the common failure

The lifecycle is a loop

Every project cycles through roughly six stages — rarely in a straight line:

The techniques map: this course in one table

You want to...	Technique	Sessions
Group similar things, no labels	k-means, hierarchical, GMM	4–7
Say how alike two things are	distance & similarity measures	3
Predict a category	trees & random forests , k-NN, logistic regression	9–13
Predict a number	linear regression + regularization	12–13
Compress / de-noise / visualize high-dimensional data	SVD, PCA, t-SNE	14–15
Ask “did X cause Y?”	causal graphs, effect	16–19

In-Class Activity

Activity: Question types, then Pandas

Goal: (**core**) classify ten short scenarios by question type and spot a type error; (**stretch**) load a small real dataset (Titanic passengers) in Pandas and walk the clean-and-explore stage — read, inspect, filter, [groupby](#), plot — then say which question type your [groupby](#) result actually answers.

- Work in groups of 2–3. ~45 minutes today.
- Parts marked (**autograded**) are submitted to Gradescope; the rest is participation.
- Staff will circulate — be ready to explain any part of your work.
- Dependencies: [pandas](#), [numpy](#), [matplotlib](#) only.
- Pandas is new to some of you: the [Pandas tutorial](#) is the reference; the notebook links the sections you need.

Section A1: [Open in Colab](#) [Open on GitHub](#)

Section B1: [Open in Colab](#) [Open on GitHub](#)

Going deeper

Full lecture notes: [What is Data Science?](#)

- [A working definition](#) — the three ingredients
- [First, classify the question](#) — the taxonomy, with examples
- [The taxonomy is not academic](#) — type errors and standards of evidence
- [The lifecycle of a data science project](#) — the six-stage loop
- [Two cultures, one practitioner](#) — understand vs. predict
- [Types of data science roles](#) — where each job sits on the taxonomy
- [The techniques map](#) — print it
- [Pandas tutorial](#) — reading, inspecting, filtering, [groupby](#), plotting
- Next up: distances and similarity (02) — the row half the map is built on.