What is Data Science?

Why this lecture exists

You are one week into a Master’s degree in data science. It is fair to ask:

Defining the field

A working definition

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

Data science, ML, AI, statistics — untangling the terms

The boundaries are blurry and employers use the words loosely. What matters is not the label but the **question you are answering** — which is the next topic, and the most useful idea in this lecture.

How practitioners approach a problem

First, classify the question

Before choosing a technique, classify what kind of question you are asking. A useful taxonomy (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 *in general* ride longer in rain?
Predictive	What will happen for a new case?	Will this user churn next month?
Causal	What happens if we *intervene*?	Does the discount *cause* retention?
Mechanistic	Exactly how does the effect	Through which pathway

The taxonomy is not academic

The lifecycle of a data science project

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

Two cultures, one practitioner

Modeling to understand

- Parameters mean something
- Small, interpretable models
- Inference, intervals, effects
- “Why does Y happen?”

Linear/logistic regression, causal graphs

Modeling to predict

- Accuracy on unseen data is all
- Complexity fine if validated
- Feature importance at best
- “What will Y be?”

Random forests, neural networks

The jobs

Types of data science roles

The field has differentiated. Titles vary by company, but these clusters are stable:

Role	Center of gravity	Typical day
Data analyst	descriptive/exploratory	SQL, dashboards, “what happened and why” for a business team
Data scientist (product/inference)	inferential/causal	A/B tests, metrics design, “did the feature work?”
Machine learning engineer	predictive, in production	training pipelines, serving, monitoring; strong software engineering
Data engineer	the data itself	pipelines, warehouses, quality; makes everyone else possible
Applied / research scientist	novel methods	pushing a model class further; papers or patents

What this means for you

The techniques map

What this course covers, and when to reach for it

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	16–18

The arc of the semester

How to succeed in this course’s format

Recap