Data Science

Guide to AI and Data Courses · Part IV: Data

Question and data | Data wrangling | Exploratory analysis | Visualization | Inference and story | Neighboring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides: https://sercankulcu.github.io/files/ai_courses/slides/Chapter_26_Data_Science.html

INTRODUCTION

What Question Does It Ask?

> **How do we ask the right questions of data and turn the answers into decisions?**

Data science is the process that starts with a question and covers finding, cleaning and analyzing data, then communicating the result so that it **leads to a decision**. It combines statistics, programming and domain knowledge; data mining and machine learning are its tools.

At a glance

APPROACH question, data, analysis, story	**INPUT** raw, messy data	**OUTPUT** report, visualization, model
SUCCESS MEASURE impact on decisions, accuracy	**MATH LOAD** medium: statistics	**DATA NEEDS** moderate to large

A Short History

1962
In "The Future of Data Analysis", Tukey argues for data analysis as a science separate from statistics.

1977
Tukey publishes "Exploratory Data Analysis": look first, then model.

2001
Cleveland proposes "data science" as an academic field that extends statistics.

2010
Kaggle is founded; data science competitions and open datasets spread.

2012
Harvard Business Review calls data scientist "the sexiest job of the 21st century".

2014
Project Jupyter splits from IPython; code, plots and text meet in one notebook.

THE COURSE

Core Concepts

Defining the question
What do we want to know, and what decision will follow?

Tidy data
Each row an observation, each column a variable.

Data cleaning
Fixing missing, wrong and inconsistent records.

Exploratory analysis
Getting to know data with summaries and plots.

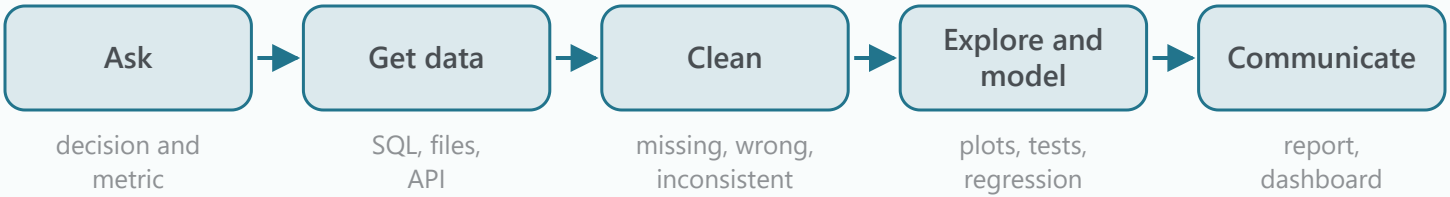
Visualization
Showing the pattern and result with the right chart.

Hypothesis testing
Checking whether a difference is due to chance.

A/B testing
Comparing two options with a randomized experiment.

Reproducibility
Getting the same result again with code, data and versions.

A Typical Workflow



Example: A city council asks, "Should we add more bus trips?" Card tap records are collected, bad timestamps removed, occupancy by line and hour explored with plots, demand estimated with regression, and the result presented to the decision maker in a one-page report.

THE COURSE

What Does the Course Cover?

1. The data science process and asking good questions
2. Reading data in Python or R; tidy data
3. Data cleaning: missing values and outliers
4. Exploratory data analysis
5. Principles of visualization
6. Sampling and descriptive statistics
7. Confidence intervals and hypothesis tests
8. A/B testing and experimental design
9. Regression and predictive models
10. Correlation, causation, confounders
11. Getting data from a database: SQL
12. Reproducible reports and telling the story

Skills you will gain

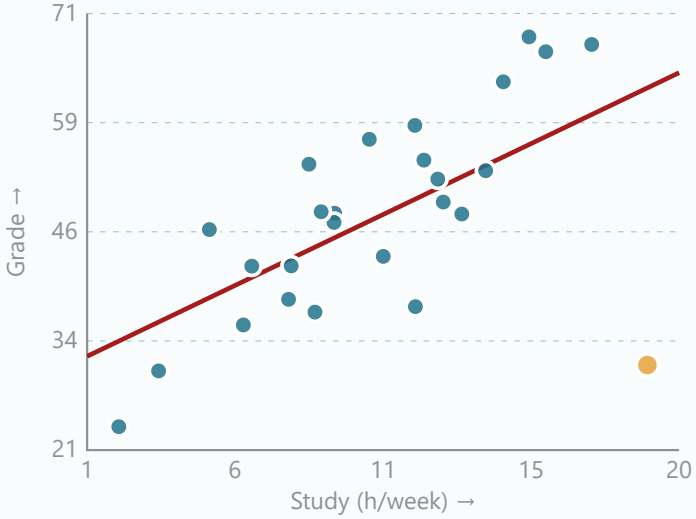
pandas or the tidyverse | SQL | Matplotlib and seaborn | interpreting hypothesis and A/B tests | reproducible reports in Jupyter

A typical semester; the order and coverage vary between universities.

TRY IT

What Drives Grades?

The school asks: “Which habit should we advise students on?” We have survey and grade data for 26 students. Pick variables to see the relationship; click a point to drop it from the analysis.



Horizontal axis (x)

Weekly study (hours)

Vertical axis (y)

Term grade

Outlier student included

n = 26 students
Pearson r = **0.61** (moderate, positive)
Line: grade ≈ 30.3 + 1.71 × study

The orange point is a student who was ill on exam day. With it r = 0.61, without it r = 0.84: one record can change the result, so get to know the data first.

Data science starts with a **question** and ends in a decision: exploratory analysis, cleaning, measuring relationships and communicating the result. Correlation is not causation: students who study more also attend more; separate the two before giving advice.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Probability and Statistics	hypothesis testing, regression, visualization
Programming	reading files, loops and functions in Python or R
Databases and SQL	querying, joining and grouping data

CONTEXT

Are You Ready for This Course?

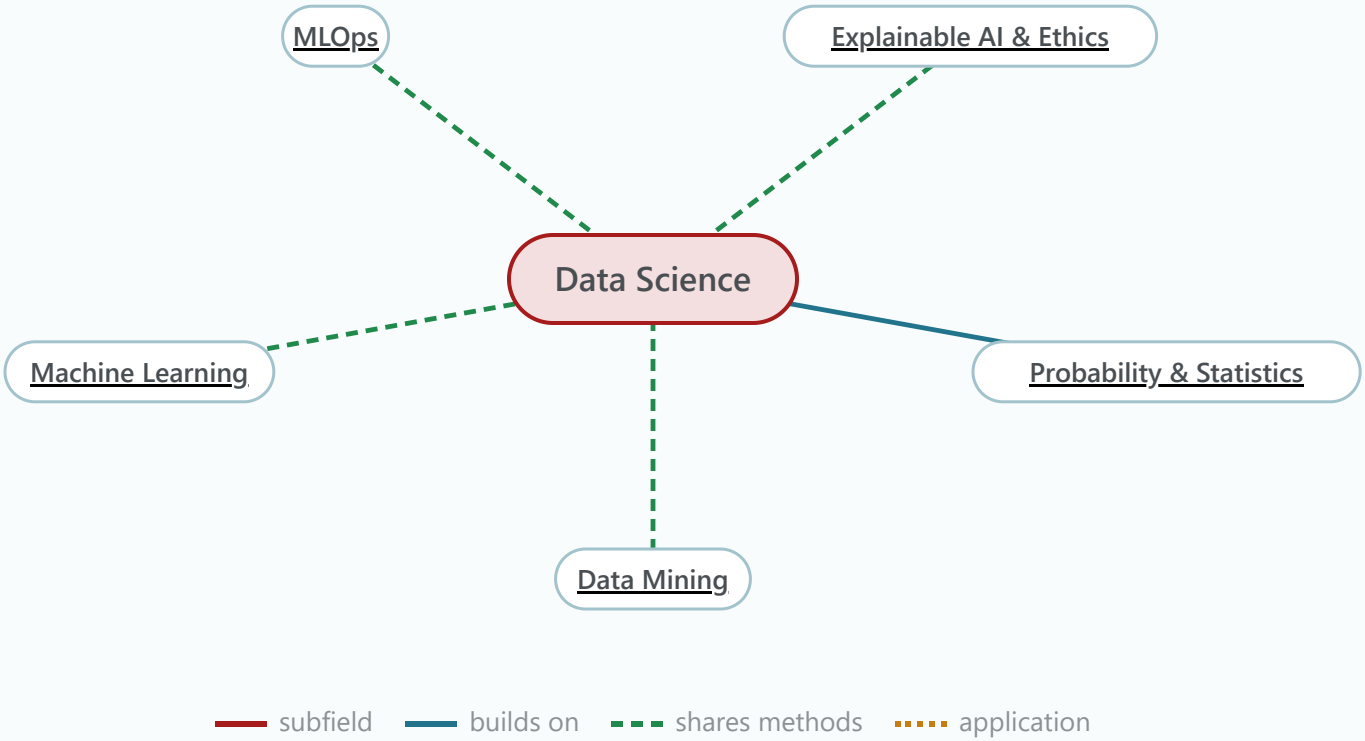
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you explain the mean, median, standard deviation and the normal distribution?	Probability and Statistics
Can you read a table in Python or R and group it by a column?	An intro programming course, then a short pandas or dplyr tutorial
Can you write a simple SQL query with SELECT, WHERE and GROUP BY?	The SQL part of a database course, or online SQL exercises
Can you tell when a chart is misleading?	Not required, the course teaches it; if curious, Tufte's book

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



OVERLAP

Overlapping Courses

Relation	Course	Shared ground
builds on	This course builds on Probability and Statistics	hypothesis testing, regression, visualization
shares methods	Shares methods with Machine Learning	predictive models
shares methods	Shares methods with Explainable AI and AI Ethics	data bias, fairness metrics
shares methods	Shares methods with Data Mining	exploratory data analysis
shares methods	Shares methods with Machine Learning Systems and MLOps	experiment tracking, data versioning

DIFFERENCES

Where It Differs (1/2)

Criterion	Data Science	Data Mining	Machine Learning
Core question	How do we ask the right questions of data and turn the answers into decisions?	How are hidden, useful patterns found in large piles of data?	How does a program generalize from examples?
Approach	question, data, analysis, story	discovery-driven analysis	statistical learning
Input	raw, messy data	large tables or transaction data	labeled or unlabeled examples
Output	report, visualization, model	rules, clusters, outliers	a model that makes predictions
Success measure	impact on decisions, accuracy	interestingness, support, confidence	error on unseen data
Math load	medium: statistics	medium	high: linear algebra, probability, optimization
Data needs	moderate to large	large	moderate to large

- **Data Mining:** Data mining focuses on finding patterns in large data; data science covers the whole path from question to decision.
- **Machine Learning:** ML builds models that predict well on new examples; data science uses a model as one tool for answering a question.

Where It Differs (2/2)

Criterion	Data Science	Probability & Statistics	Big Data
Core question	How do we ask the right questions of data and turn the answers into decisions?	How is uncertainty measured, and how do we draw reliable conclusions from limited data?	How is data that does not fit on one machine stored and processed?
Approach	question, data, analysis, story	probability models and inference	distributed storage and computation
Input	raw, messy data	sample data	huge, fast-flowing data
Output	report, visualization, model	estimate, confidence interval, test result	processed data, query results
Success measure	impact on decisions, accuracy	unbiasedness, power, significance	scalability, latency, cost
Math load	medium: statistics	the theory itself	low
Data needs	moderate to large	a sample	very large

- **Probability & Statistics:** Statistics provides the theory of inference; data science combines it with programming, messy real data and storytelling.
- **Big Data:** Big data systems store and process the data; data science draws meaning and decisions from it.

PRACTICE

Applications

A/B testing
Is the new design really better?

Marketing analytics
Which campaign brings customers?

Public health
Tracking outbreaks and disease trends.

Sports analytics
Measuring player and team performance.

Public policy
Evaluating services with open data.

Product metrics
What keeps users, and what drives them away?

PRACTICE

Risks and Ethics

Confusing correlation with causation

Reading causes into correlations leads to wrong decisions: ice cream sales and drownings rise together because both happen in summer.

p-hacking

Run enough tests and something will come out "significant"; the hypothesis must be fixed before looking at the data.

Re-identification

In data with names removed, ZIP code, birth date and sex are enough to re-identify most people.

Misleading charts

A truncated axis or a cherry-picked time window can tell the opposite story with the same data.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

Python | pandas | Jupyter | R and tidyverse | Matplotlib and seaborn | SQL | statsmodels

Well-known datasets

Titanic (Kaggle) | Gapminder | NYC taxi trips | Palmer Penguins | House Prices (Kaggle) | Anscombe's quartet

PRACTICE

Common Misconceptions

Common belief	In fact
Data science is training models.	Most of the time goes into sharpening the question, finding and cleaning data, and communicating results.
Data speaks for itself.	Without a question, context and domain knowledge, the same data can lead to wrong conclusions.
A significant p-value is an important result.	Statistical significance says nothing about effect size; many tests produce misleading results.
A chart is decoration for the result.	Anscombe's quartet shows it: data with identical summaries look completely different when plotted.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

Data scientist | Data analyst | Product analyst | Business intelligence specialist | Research analyst

Data science is the umbrella of the data courses. From here you can move to machine learning to deepen predictive modeling, to data mining for discovery, to big data systems for scale, or to explainable AI and ethics for questions of fairness and bias.

NEXT STEPS

Further Reading

- Wickham, Çetinkaya-Rundel and Grolemund, *R for Data Science*, 2nd ed., O'Reilly, 2023.
- McKinney, *Python for Data Analysis*, 3rd ed., O'Reilly, 2022.
- James, Witten, Hastie and Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., Springer, 2021.
- Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

The End

Chapter 26: Data Science

← Previous chapter · Map of the field · Next chapter →