Data Science Basics: Pipeline and Literacy

July 7th , 2025

Before going to science, what is data?

- **Data** is a collection of **facts**, such as numbers, words, measurements, observations, or even just descriptions of things.
- Data can be **qualitative** (descriptive) or **quantitative** (numerical).
- Not useful in general.
- It becomes useful when it's **organized, analyzed, and interpreted**.
- Structural vs Unstructured

Category	Description	Examples	Easy to Analyze?
Structured Data	Organized into rows & columns (like a spreadsheet or database).	Age, height, price, test scores	✅ Yes
Unstructured Data	No fixed format or structure; harder to organize.	Text messages, images, videos, emails, social media posts	❌ No (requires special processing)

Can you think of an unstructured data source from your daily life?

- your photo gallery
- TikTok videos
- voice messages
- the questions you asked
- WIFI
- ……..

So…….

- **Over 80%** of the world's data is unstructured
- **Computers can only understand structured data.** Unstructured data (like images, text, audio) **must be transformed** into structured forms before analysis
- The raw data is **messy and unstructured**.
- Your job as a data scientist is to:
  - **Collect** the raw data.
  - **Preprocess** and **structure** it.
  - **Analyze** and **model** it.

Type	Description	Example
Structured	Neatly organized (e.g., rows & columns)	Spreadsheets, SQL databases
Unstructured	No fixed format	Texts, images, videos
Semi-structured	Some organization, but flexible	JSON, XML, HTML, emails

Example one: Reviews to structured data

- reviews = [

  "I love this ice cream! It's so creamy and delicious.",

  "Terrible service. I waited 20 minutes, and no one helped me.",

  "The flavor was okay, but a bit too sweet for my taste.",

  "Amazing staff and fast service. Will come again!"

]

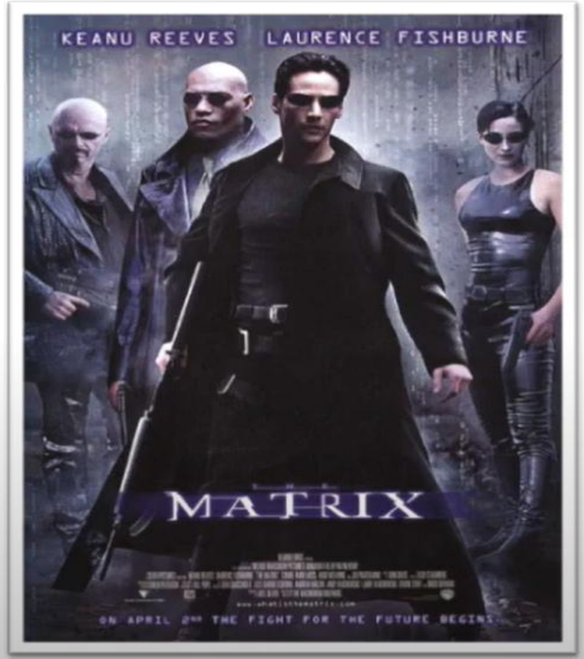
Review	Sentiment
I love this ice cream...	Positive
Terrible service...	Negative
The flavor was okay...	Neutral
Amazing staff...	Positive

Example one: Image to numbers

Each image is made up of **pixels** — tiny squares, each with color values:

- Grayscale: One number per pixel (0 = black, 255 = white)
- Color (RGB): Three numbers per pixel (Red, Green, Blue)

📷 Example: A 100×100 image = 10,000 pixels = 10,000 numbers (or 30,000 for RGB)



$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Example three: Turning Emails into Structured Data

- From: jane@company.com
- Subject: URGENT: Meeting rescheduled
- Body: Hi team, our project meeting is moved to 3 PM today. Please confirm.

Sender	Subject	Urgency	Keyword(s)
jane@company.com	Meeting rescheduled	Yes	meeting, 3 PM

Data Science Pipeline

What is Data Science?

- Data Science is an interdisciplinary field that uses methods from statistics, computer science, and domain knowledge to extract insights from data.
- Core components:
- Statistics: Understand patterns, test hypotheses, conduct inferences.
- Machine Learning: Predict, classify and model complex systems.
- Domain Knowledge: Interpret findings meaningfully.

Example

Student	SAT	GPA	Volunteer Hours	AP Classes	Essay Score (1–10)	Interview Score (1–10)	Admitted?
A	1400	3.8	50	5	9	8	Yes
B	1250	3.5	20	2	6	5	No
C	1500	4.0	60	6	10	9	Yes
D	1100	3.2	15	1	4	3	No
E	1350	3.7	40	4	8	7	Yes

What’s the role of "Essay Score" and "Interview Score"?

A. Help models detect spam
B. Add more features about student qualities
C. Measure how fast someone can respond
D. Predict sports outcomes

B

Take away

🧠 **Data Science Combines:**

- **Statistics** – Making sense of numbers.
- **Computer Science** – Writing code to work with data.
- **Domain Knowledge** – Understanding the topic you're studying (like sports, medicine, business, etc.)

What is Data Science Pipeline?

- 1*. Problem formulation
- 2*. Data acquisition
- 3. Preprocessing/Data cleaning
- 4. Exploratory analysis (EDA)
- 5. Modeling, inference, and/or prediction
- 6. Evaluation, validation and communication
- 7. Deployment or presentation

Research: 1, 2

Industry/ Commercial: 2,1

Step	Purpose
Problem	Define the question
Collection	Get relevant data
Cleaning	Fix and prepare the data
Exploration (EDA)	Discover patterns
Modeling	Predict or infer
Evaluation	Measure model performance
Communication	Share results and insights
Deployment	Apply insights / model in real life

Step 1 - Problem Formulation

- Ask: What is the goal of the project?
- Define a specific, measurable, and relevant question.
- Translate business/real-world problems into analytical terms.
- Examples:
- Can we predict student dropout based on early academic performance?
- How does weather affect monthly energy consumption?
- Involves stakeholder interviews and background research.

🧭 Concrete Example: Problem Formulation

- **Context:** A local school district wants to reduce student dropout rates.
- **Initial Goal (Vague):** We want to prevent students from dropping out.
- **Clarified Question**: Can we identify students at risk of dropping out before the end of the academic year?
- **Refined Data Science Problem**: Using historical student data (attendance, grades, disciplinary actions, etc.), can we build a model to predict whether a student will drop out in the next semester?

Step 2 -Data Acquisition

- **What--Data can be:**
  Structured (tables, spreadsheets)
  Semi-structured (JSON, XML, Emails)
  Unstructured (text, images, audio)
- **Where and How--Sources:**
  Public datasets (UCI, Kaggle, government portals)
  APIs (Twitter, OpenWeatherMap)
  Internal databases (CRM, ERP systems)
  Web scraping, sensors, logs
- **Considerations:**
  Ethics: Informed consent, anonymization
  Coverage: Is the data representative?
  Granularity: Level of detail

Concrete Example: Problem Formulation

- **Target Variable**: Dropout (Y/N)
- **Potential Features:** Attendance rate, GPA over last 3 terms, Number of suspensions, Socioeconomic indicators, Parent-teacher meeting participation,.....
- **Data Sources:**

1. Internal school records (grades, attendance, behavior logs)
2. Free/reduced lunch program (proxy for socioeconomic status)
3. External: State education databases or census ZIP-code data

Class discussion

- 🎧 **Specific Example**: Sentiment Analysis on a Popular Artist or Song
- Find out how people feel about the release of a new song or album by a popular artist among students (e.g., Olivia Rodrigo, Drake, BTS, Taylor Swift, etc.).
- Data Acquisition Plan: What, Where and How?

Data Sources:

- Twitter hashtags (e.g., #NewDrakeAlbum, #Swifties)
- YouTube video comments on the official song
- TikTok reactions or comment threads

How to Collect:

- Search for public comments/posts about the song
- Copy a sample of 10–20 comments into a spreadsheet
- Label each as Positive / Neutral / Negative

Comment	Platform	Sentiment
“This song is 🔥 🔥 🔥 ”	Twitter	Positive
“Not her best work, honestly.”	YouTube	Neutral
“So overrated. Skip.”	TikTok	Negative

Step 3 - Data Cleaning (Wrangling)

- **Raw data is often messy. Common issues:**
  - Unstructured
  - Missing values (NaN, NULL)
  - Duplicates or inconsistencies
  - Incorrect formats (e.g., dates as strings)
  - Outliers or noise
- **Solutions:**
  - Imputation, removal, standardization
  - Parsing and transformation
- **Tools:**
  - Python: pandas, numpy
  - Spreadsheets (for small datasets)

Class discussion

- Students in a class collected survey responses about how many hours they sleep before school and how alert they feel during first period.

Student	Sleep Hours	Alertness (1–5)
A	7.5	4
B	8 hrs	5
C	-3	3
D	6	N/A
E	missing	2

Problems Identified:

•Mixed formats: “8 hrs” should be a number (8.0)

•Invalid values: -3 hours of sleep isn’t possible

•Missing or blank entries: “missing”, “N/A”

Class discussion

Cleaning Steps:

1. Convert all entries to numeric format (e.g., remove text like “hrs”)
2. Flag and remove or correct invalid entries (e.g., -3)
3. Handle missing values:
   1. Drop the row, or
   2. Impute (fill in) with the average or median

Student	Sleep Hours	Alertness (1–5)
A	7.5	4
B	8.0	5
D	6.0	3.5 *(imputed)*
E	7.0 *(imputed)*	2

Step 4 - Exploratory Data Analysis (EDA)

- **Purpose: Learn about your dataset before modeling**
- **Techniques:**

  Descriptive statistics: mean, median, standard deviation

  Visual tools: Histograms (distribution), Boxplots (spread, outliers)

  Scatter plots (relationships), Heatmaps (correlations) and Dimension Reduction (geometric)
- **Helps refine questions and guide modeling choices**
- **Tools**: matplotlib, seaborn, plotly, pandas profiling

A small example

Student	Sleep Hours	Test Score (%)
A	7.0	88
B	5.5	72
C	8.0	93
D	4.0	65
E	6.5	80

Summary Statistics
- Average Sleep: 6.2 hours
- Average Test Score: 79.6%
- Score Range: 65% to 93%

Correlation
- Calculate Pearson correlation coefficient:

Example result: **r = 0.82** → strong positive relationship

Step 5 - Modeling and Inference

- **Use statistical/machine learning/AI models to:**
  1. **Predict**: future or unseen data (e.g., regression, classification)
  2. **Cluster**: group similar items (e.g., k-means)
  3. **Inference**: relationships (e.g., correlation, causality, p-value)
- **Workflow:**
  1. Select features
  2. Split into training/testing data
  3. Train model on training set
  4. Validate on test set
  5. Interpret coefficients or model parameters
- Tools: scikit-learn, statsmodels, TensorFlow, caret, keras.....

Step 6 – Evaluation and validation

- **How well does your model perform?**
- **Metrics:**
- Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC
- Regression: Mean Absolute Error (MAE), Root Mean Square Error (RMSE)
- **Practices:**
- Cross-validation (e.g., k-fold)
- Confusion matrix, residual plots
- **Watch out for:**
- Overfitting (model too complex)
- Underfitting (model too simple)

Step 6 - Communication

- **Share findings with non-technical audiences**
- **Focus on:**
- Clarity of purpose
- Visual summaries (charts, dashboards)
- Key takeaways
- **Good practices:**
- Use storytelling (context + data)
- Minimize jargon
- Address potential limitations
- Tools: PowerPoint, Jupyter Notebooks, Tableau, Datawrapper, Github

Step 7 - Deployment and Decision Making

- **Turn insights into action:**
- Inform decisions (manual or automated)
- Integrate models into apps or systems
- Build dashboards or live reports
- **Monitor performance over time:**
- Retrain models
- Update data pipelines
- Maintain transparency and accountability

- **Real data science often loops back!**
- **Structured workflows lead to better outcomes!**

Discussion Questions

- Where do you think most time is spent in real projects?
- Which step is most prone to human error?

Data Science Vocabulary and Literacy

Data Structure Vocabulary

- Dataset: A structured collection of data, often tabular
- Variable (Feature): A measurable property or characteristic
- Observation (Instance): A single row of data, one example
- Data Type: Type of variable (e.g., numeric, categorical, boolean)
- **Resources:**
- Surveys, experiments, sensors, logs
- Public datasets: data.gov, World Bank
- APIs: Twitter, weather, finance
- Ethical issues: bias, privacy, consent

Statistical Vocabulary

- Mean: Average value
- Median: Middle value
- Standard Deviation: Spread of data around the mean
- Correlation: Strength and direction of relationship between two variables
- Distribution: How values of a variable are spread (e.g., normal distribution)
- Missing Data: Absence of a value
- Outlier: Value significantly different from others
- Noise: Random variation in data that obscures patterns
- Bias: Systematic error leading to incorrect conclusions

Modeling Vocabulary

- Model: Mathematical or computational representation of a system
- Training: Teaching a model using existing data
- Testing: Evaluating the model on new/unseen data
- Validation: Intermediate evaluation to tune models
- Overfitting: Model is too complex and memorizes training data
- Underfitting: Model is too simple and fails to capture patterns

- Generalization: Model's ability to perform well on new data
- Feature Engineering: Creating input variables to improve model performance

Machine Learning Vocabulary

- Supervised Learning: Learning from labeled data
- Unsupervised Learning: Learning from unlabeled data
- Classification: Predicting categories
- Regression: Predicting continuous values
- Clustering: Grouping similar data points

Task	**Metric**	**When to Prefer**
Classification	**Accuracy**	Classes are balanced, and all errors are equally costly
	Precision	False positives are costly (e.g., spam filter)
	Recall	False negatives are costly (e.g., cancer screening)
	F1 Score	Need balance between precision and recall
Regression	**Mean Squared Error**	Penalizes large errors more (sensitive to outliers)
	Mean Absolute Error	More robust to outliers, treats all errors equally
	R^2 Score	Measures proportion of variance explained by the model

Programming & Computation Vocabulary

- Algorithm: Step-by-step procedure for solving a problem
- Function: Reusable block of code that performs a task
- Library/Package: Collection of pre-written code for common tasks (e.g., pandas, NumPy)
- API: Interface that allows software to communicate
- Filtering: Selecting rows based on conditions
- Aggregation: Summarizing data (e.g., sum, mean)
- Pivoting: Reshaping data from long to wide format (or vice versa)
- Merging/Joining: Combining multiple datasets