

Artificial Intelligence for Medicine II

Spring 2025

Lecture 2-1: Data & Data Types

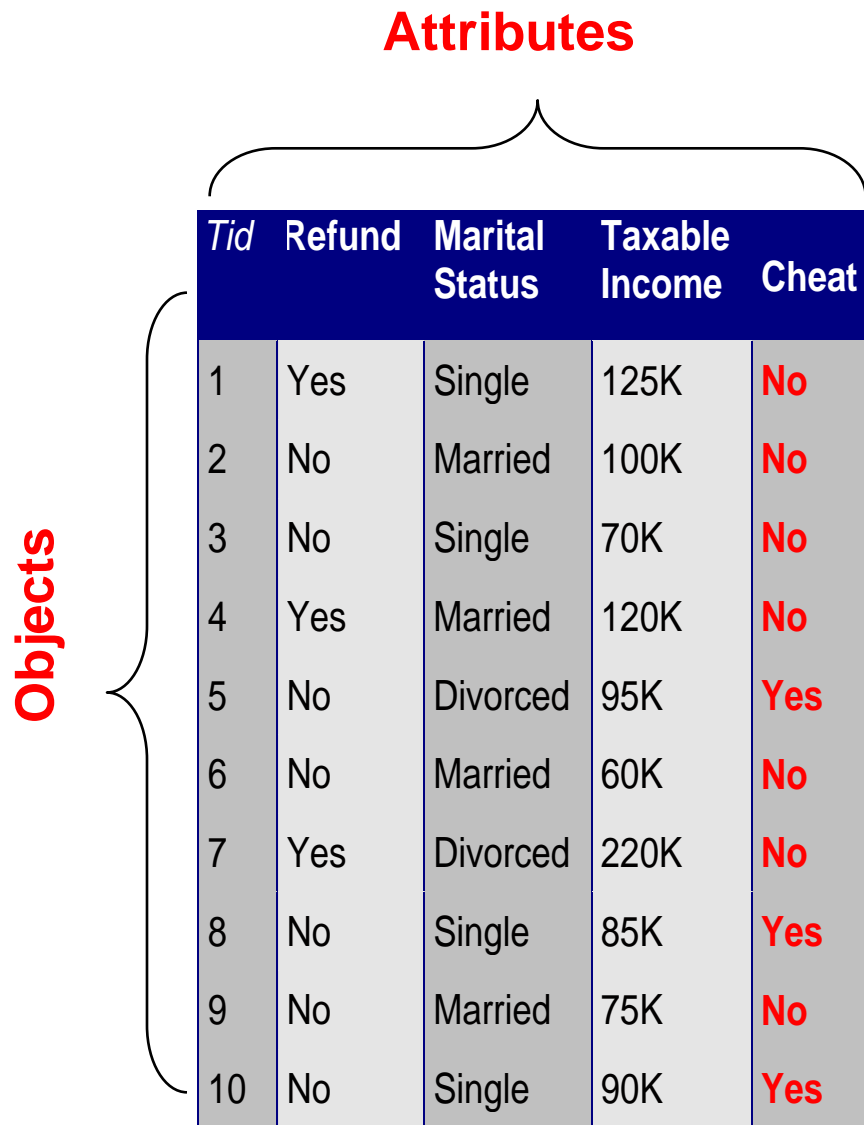
(Many slides adapted from Han, Kamber & Pei; Tan, Steinbach, Kumar
and the web)

Machine learning (ML)

- Machine learning (ML) is a subfield of artificial intelligence (AI) that enables computers to **learn from data** without being explicitly programmed. Essentially, it allows computers to find patterns and make predictions or decisions based on data.
- **Core Concepts:**
- **Data:**
 - ML algorithms rely on data to learn. This data can be in various forms, such as numbers, text, images, or audio.
 - The **quality and quantity of data** significantly impact the performance of ML models.
- **Algorithms:**
 - These are the sets of rules and statistical techniques that ML models use to learn from data.
 - Different algorithms are suited for different types of tasks and data.
- **Models:**
 - A model is the output of an ML algorithm after it has been trained on data.
 - It represents the learned patterns and can be used to make predictions or decisions on new data.
- **Learning:**
 - The process of an ML algorithm adjusting its parameters based on the data it receives.
 - The goal is to minimize errors and improve the model's accuracy.

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as variable, field, characteristic, dimension, or **feature**
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, tuple or instance



The diagram illustrates the relationship between data objects and attributes. A table contains 10 rows of data. A bracket labeled 'Attributes' spans the columns: Tid, Refund, Marital Status, Taxable Income, and Cheat. A bracket labeled 'Objects' spans the rows, indicating each row represents a single data object.

Attributes				
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
 - Same attribute can be mapped to different attribute values
 - Example: height can be measured in feet or meters
 - Different attributes can be mapped to the same set of values
 - Example: Attribute values for ID and age are integers
 - But properties of attribute can be different than the properties of the values used to represent the attribute

Types of Attributes

- There are different types of attributes
 - **Nominal**
 - Examples: ID numbers, eye color, zip codes
 - **Ordinal**
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
 - **Interval**
 - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - **Ratio**
 - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

		Attribute Type	Description	Examples	Operations
Categorical Qualitative		Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
		Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative		Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
		Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

		Attribute Type	Transformation	Comments
Categorical Qualitative		Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
		Ordinal	An order preserving change of values, i.e., $new_value = f(old_value)$ where f is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative		Interval	$new_value = a * old_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
		Ratio	$new_value = a * old_value$	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete and Continuous Attributes

- Discrete Attribute
 - Has only a finite or countably infinite set of values
 - Examples: zip codes, counts, or the set of words in a collection of documents
 - Often represented as integer variables.
 - Note: **binary attributes** are a special case of discrete attributes
- Continuous Attribute
 - Has real numbers as attribute values
 - Examples: temperature, height, or weight.
 - Practically, real values can only be measured and represented using a finite number of digits.
 - Continuous attributes are typically represented as floating-point variables.

Important Characteristics of Data

- Dimensionality (number of attributes)
 - High dimensional data brings a number of challenges
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Size
 - Type of analysis may depend on size of data

The types of Data in ML/AI

- In machine learning, datasets are typically composed of different types of data, which can be categorized based on their structure, format, and the nature of the information they represent. Here are the main types of data in ML/AI :
- **1. Structured Data**
- **Definition:** Data that is organized in a tabular format, such as rows and columns (e.g., databases, spreadsheets).
- **Examples:**
 - Numerical data (e.g., age, salary, temperature)
 - Categorical data (e.g., gender, product categories)
 - Date/time data (e.g., timestamps, dates)
- **Use Cases:** Regression, classification, and other traditional ML tasks.
- **2. Unstructured Data**
- **Definition:** Data that does not have a predefined structure or format.
- **Examples:**
 - Text data (e.g., emails, social media posts, articles)
 - Image data (e.g., photos, medical scans)
 - Audio data (e.g., voice recordings, music)
 - Video data (e.g., surveillance footage, movies)
- **Use Cases:** Natural language processing (NLP), computer vision, speech recognition.

The types of Data in ML/AI

- **3. Semi-Structured Data**
- **Definition:** Data that does not fit neatly into a tabular structure but has some organizational properties.
- **Examples:**
 - JSON, XML files
 - Emails (structured headers with unstructured body text)
 - Log files
- **Use Cases:** Data integration, web scraping, and processing hierarchical data.
- **4. Time Series Data**
- **Definition:** Data points collected or recorded at specific time intervals.
- **Examples:**
 - Stock prices
 - Weather data
 - Sensor data (e.g., IoT devices)
- **Use Cases:** Forecasting, anomaly detection, and trend analysis.

The types of Data in ML/AI

- **5. Geospatial Data**
- **Definition:** Data that includes geographic or location-based information.
- **Examples:**
 - GPS coordinates
 - Maps
 - Satellite imagery
- **Use Cases:** Location-based recommendations, route optimization, and environmental monitoring.
- **6. Graph Data**
- **Definition:** Data represented as nodes and edges, often used to model relationships.
- **Examples:**
 - Social networks (e.g., friends on Facebook)
 - Knowledge graphs
 - Recommendation systems
- **Use Cases:** Social network analysis, fraud detection, and recommendation systems.

The types of Data in ML/AI

- **7. Text Data**
- **Definition:** A subset of unstructured data, specifically consisting of written or spoken language.
- **Examples:**
 - Documents
 - Tweets
 - Customer reviews
- **Use Cases:** Sentiment analysis, text classification, and machine translation.
- **8. Image Data**
- **Definition:** A subset of unstructured data, consisting of visual information.
- **Examples:**
 - Photographs
 - Medical images (e.g., X-rays, MRIs)
 - Satellite images
- **Use Cases:** Object detection, facial recognition, and medical diagnosis.

The types of Data in ML/AI

- **9. Audio Data**
- **Definition:** A subset of unstructured data, consisting of sound waves.
- **Examples:**
 - Speech recordings
 - Music
 - Environmental sounds
- **Use Cases:** Speech-to-text, music recommendation, and sound classification.
- **10. Video Data**
- **Definition:** A subset of unstructured data, consisting of a sequence of images (frames) and often audio.
- **Examples:**
 - Surveillance footage
 - Movies
 - Video calls
- **Use Cases:** Activity recognition, video summarization, and object tracking.

Examples of data sets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Class</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such a data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a 'term' vector
 - Each term is a component (attribute) of the vector
 - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

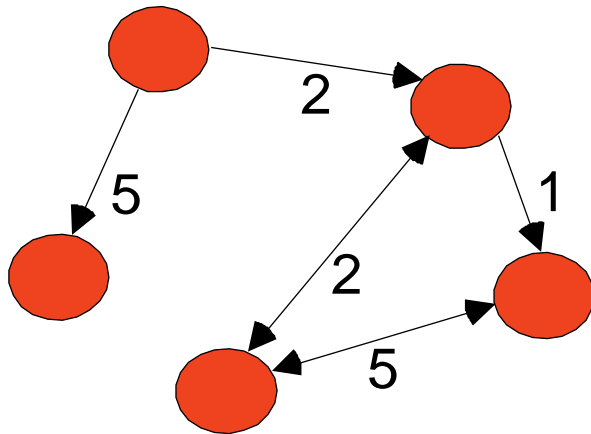
Transaction Data

- A special type of data, where
 - Each transaction involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
 - Can represent transaction data as record data

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Generic graph, a molecule, and webpages



Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

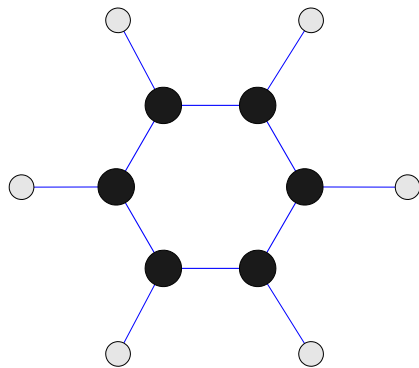
Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.



Benzene Molecule: C₆H₆

Ordered Data

- Sequential transaction data

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Ordered Data

- Genomic sequence data

**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCCGCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Image Data

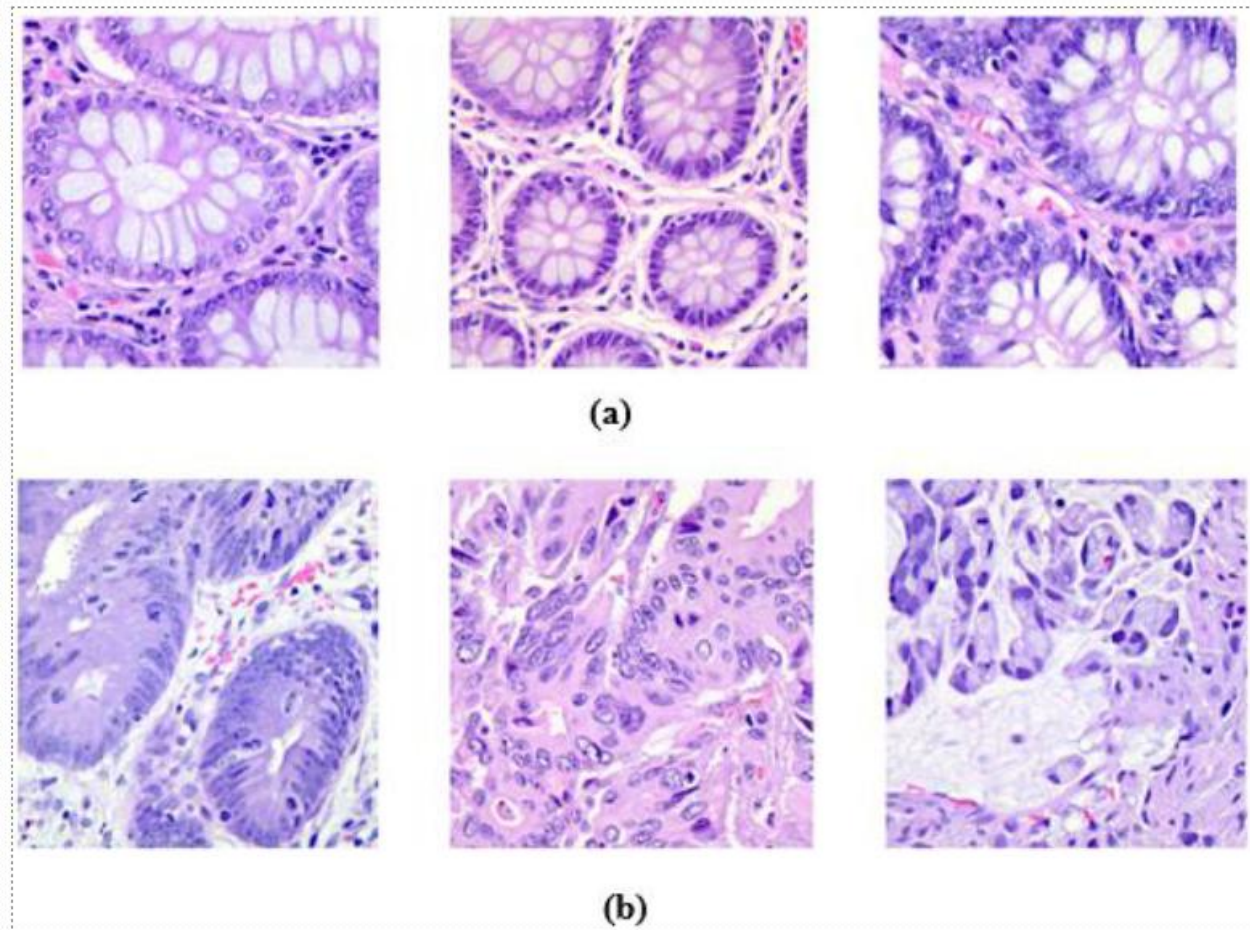


Figure - Examples of histopathological images in LC25000 dataset:

- a) Colon Benign Tissue,
- b) Colon Adenocarcinoma

Common Image Data Types

Data Type	Bit Depth	Range	Applications
Binary	1-bit	0 or 1	Masking, thresholding
8-bit Grayscale	8-bit	0 to 255	Photography, medical imaging
16-bit Grayscale	16-bit	0 to 65,535	High-precision imaging
8-bit RGB	8-bit/channel	0 to 255 per channel	Digital photography
16-bit RGB	16-bit/channel	0 to 65,535 per channel	High-dynamic-range imaging
Floating-Point	32-bit/64-bit	0.0 to 1.0	Scientific data, HDR imaging
Indexed	8-bit	0 to 255 (indices)	GIF images, low-memory storage
Depth	16-bit/float	Varies	3D reconstruction, AR/VR

Spatio-Temporal Data

- Spatio-Temporal Data

**Average
Monthly
Temperature
of land and
ocean**

Jan

