VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Fundamentals of Machine Learning

Data and their Properties

Jan Platos

November 15, 2023

Data and their Properties

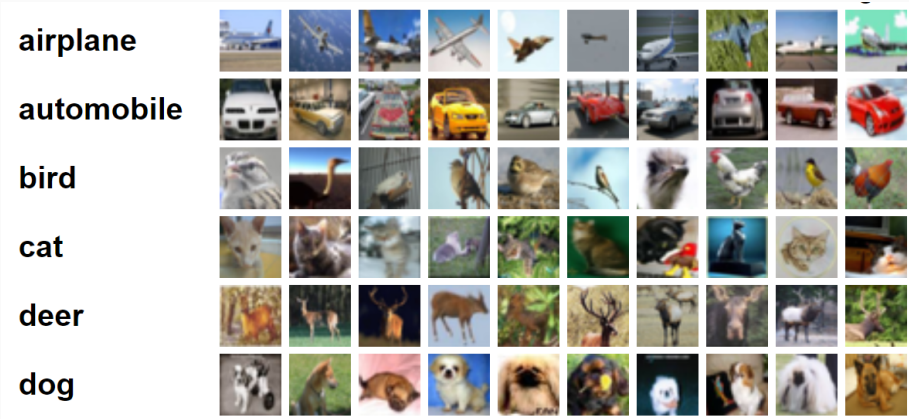
Data and their Properties

- What are data?
- Where are data?
- Why we have data?

Data and their Properties - Data types

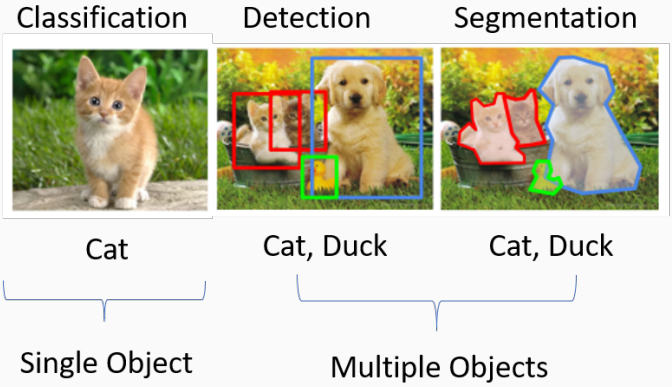
João Batista Neto, CC BY 3.0, via Wikimedia Commons

Data and their Properties- Image Classification



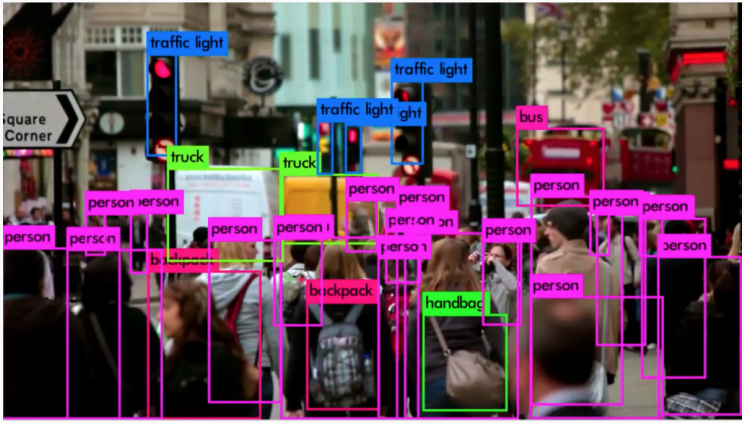
https://becominghuman.ai/cifar-10-image-classification-fd2ace47c5e8

Data and their Properties- Image Classification



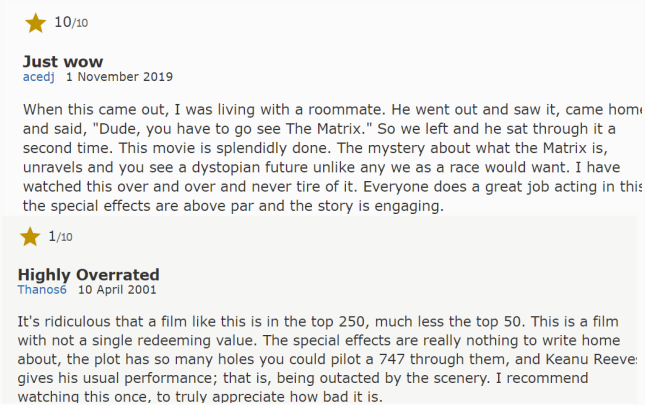
https://medium.com/@kolungade.s/object-detection-image-classification-and-semantic-segmentation-using-aws-sagemaker-e1f768c8f57d

Data and their Properties- Image Classification



https://maxprog.net.pl/artifficial-intelligence-in-practice/ai-in-city-monitoring-and-object-detection/

Data and their Properties- Text Classification



★ 10/10

Just wow
acedj 1 November 2019

When this came out, I was living with a roommate. He went out and saw it, came hom
and said, "Dude, you have to go see The Matrix." So we left and he sat through it a
second time. This movie is splendidly done. The mystery about what the Matrix is,
unravels and you see a dystopian future unlike any we as a race would want. I have
watched this over and over and never tire of it. Everyone does a great job acting in thi
the special effects are above par and the story is engaging.

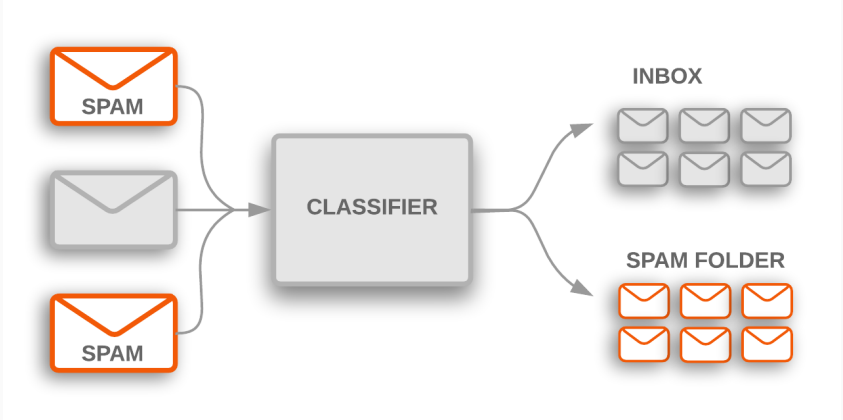
★ 1/10

Highly Overrated
Thanos6 10 April 2001

It's ridiculous that a film like this is in the top 250, much less the top 50. This is a film
with not a single redeeming value. The special effects are really nothing to write home
about, the plot has so many holes you could pilot a 747 through them, and Keanu Reeve
gives his usual performance; that is, being outacted by the scenery. I recommend
watching this once, to truly appreciate how bad it is.

https://www.imdb.com/title/tt0133093/reviews

Data and their Properties- Text Classification



https://developers.google.com/machine-learning/guides/text-classification

Data and their Properties- Time Series

The 12-Lead ECG: Anatomic Locations and Supplying Coronary Arteries

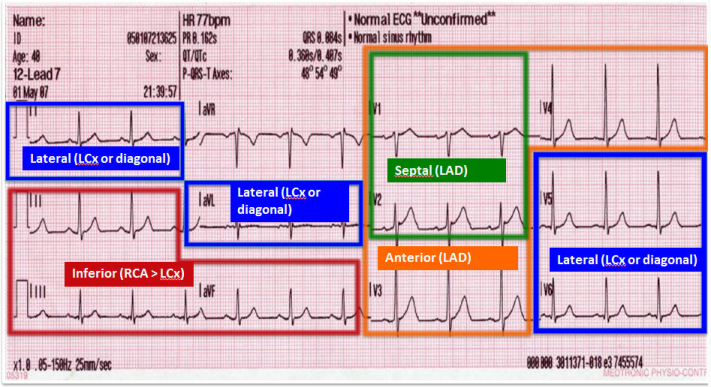
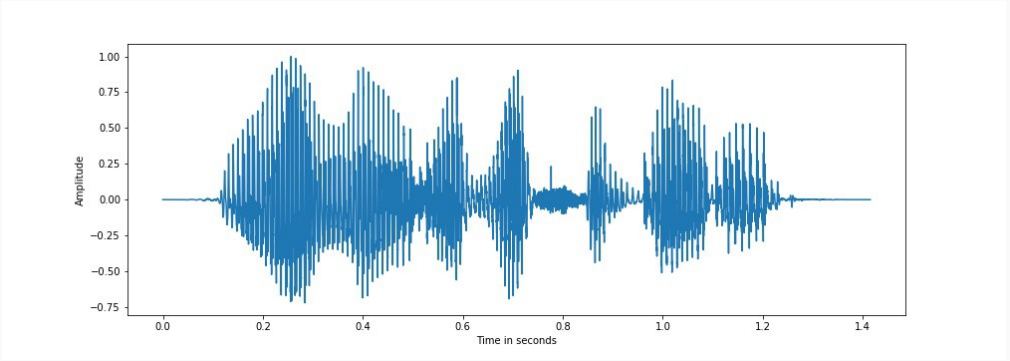


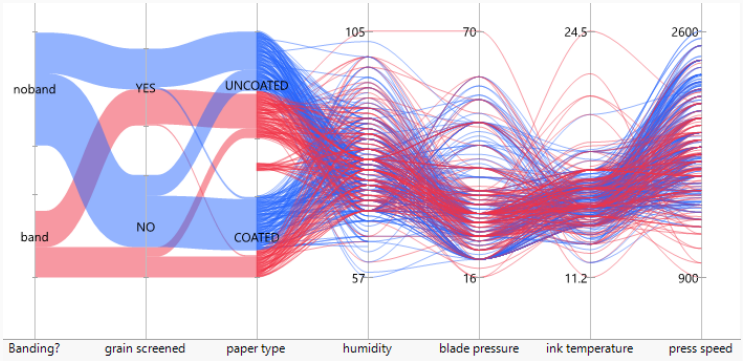
Image adapted by Sumaya Mekkaoui.
Original Image by Glen Larson. Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=2599842

Data and their Properties- Time Series



https://towardsdatascience.com/beginners-guide-to-speech-analysis-4690ca7a7c05

Data and their Properties - Data analysis



https://community.jmp.com/t5/JMP-Blog/Parallel-coordinates-in-JMP/ba-p/31024

Data and their Properties - Data matrix

- A matrix where rows are records and columns are features.
- The matrix usually contains many data types together.
- Combination of strings, numeric, categorical and other data.
- Analysis of such complex dataset is hard.
- Each feature has its own properties.

Data and their Properties - Data matrix

Numeric values

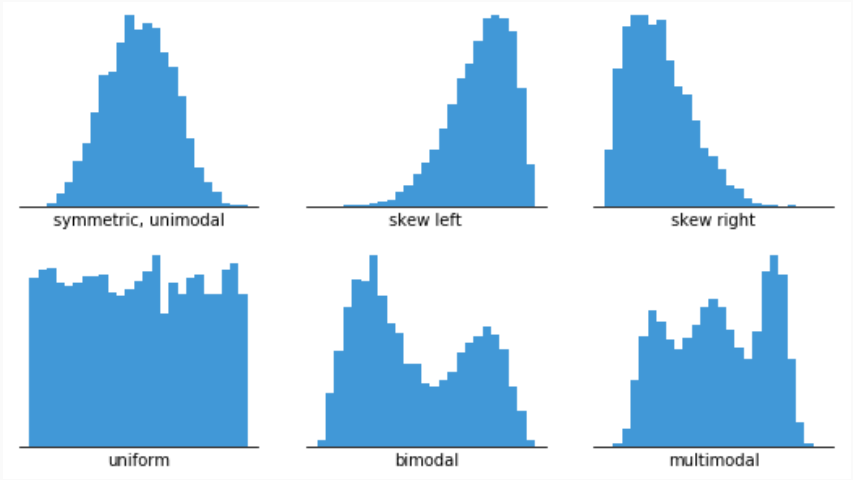
- Numeric values may be integer, real or complex number.
- Integer values belong to a specific range, e.g. byte, short, int, long, etc.
- Real number are usually float, double or decimal.
- Complex numbers are rare case.
- Integer values are processed as real or as categorical type.

Data and their Properties - Data matrix

Real numbers

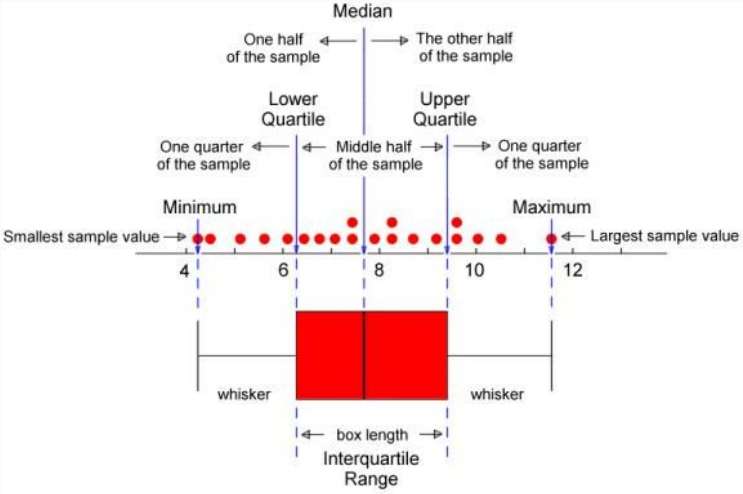
- Real number columns has basic limits - minimum and maximum.
- Other important terms are mean, median, and quartile.
- Distribution of real values is the most important feature.
- Distribution depicts which values are present a how frequently.
- The overview of the feature may be taken from the histogram or a Box plot.

Data and their Properties - Data matrix



Source: https://chartio.com/learn/charts/histogram-complete-guide/

Data and their Properties - Data matrix



Source: https://web.pdx.edu/~stipakb/download/PA551/boxplot.html

Data and their Properties - Data matrix

Categorical data

- Categorical data represents a set of possibilities a feature may take.
- The data may be numeric or textual.
- Very frequent data type unusually together with numeric.
- The overview of the feature may be taken from the (discreet) histogram.
- Very frequent as a class/label for a data.

Data and their Properties - Data matrix

Categorical data

- Processing is done using one of the following process:
  - binarization,
  - ordinal encoding (problem with sorting and distance),
  - one-hot encoding (dummy encoding),
  - embedding,
  - algorithmic encoding (cyclic feature).

Data and their Properties - Data matrix

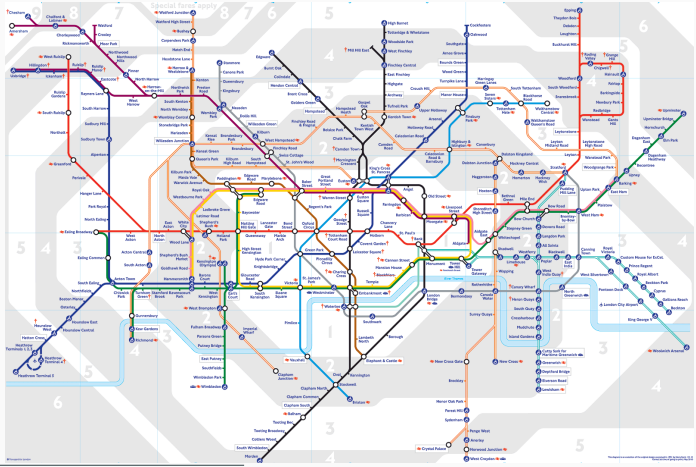
Textual data

- Textual data are in the form of a single word or an open text.
- Single word and short text may represent a categorical value (e.g. METAR).
- Open text columns are hard to process.
- Usually processed separately as a text data:
  - normalized, tokenized,encoded (embedding), ...

Data and their Properties - Graphs

https://creativecartography.sites.grinnell.edu/beck_map_1933-1/

Data and their Properties - Graphs

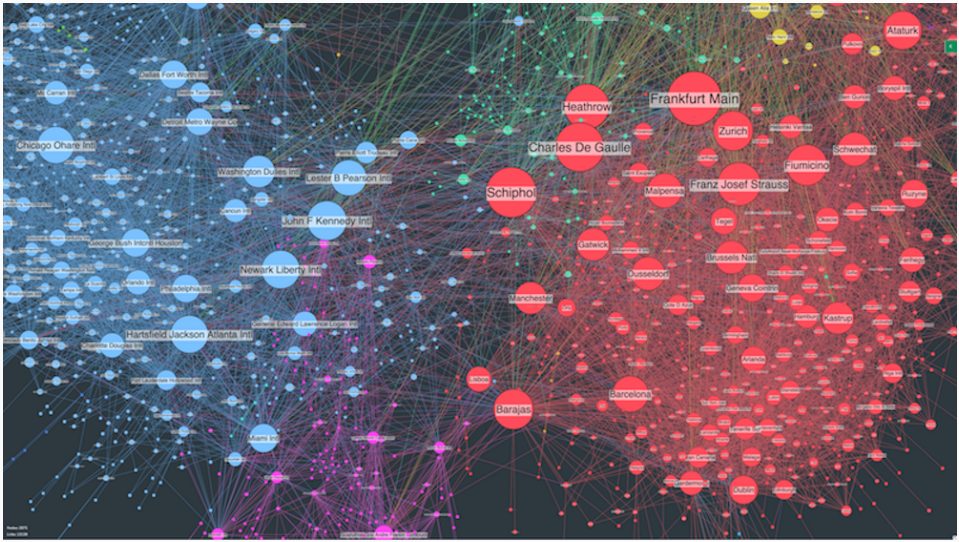


https://www.bbc.co.uk/london/travel/downloads/tube_map.html

Data and their Properties - Graphs

- A completely different data than data matrix.
- Consists of Nodes and Edges (and Values).
- Depicts the structure, or topology of some information or relations.
- May depict many features of data,

Data and their Properties - Graphs



Questions