Data labeling and metrics

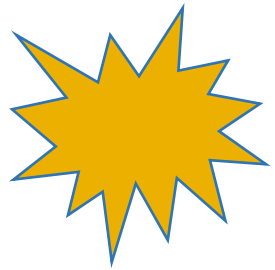
Introduction to Data Science

3rd lecture

Izv. prof. dr. sc. Ana Sović Kržić

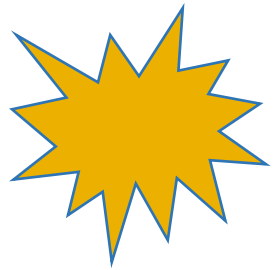
2025/2026

Content



- Data labeling

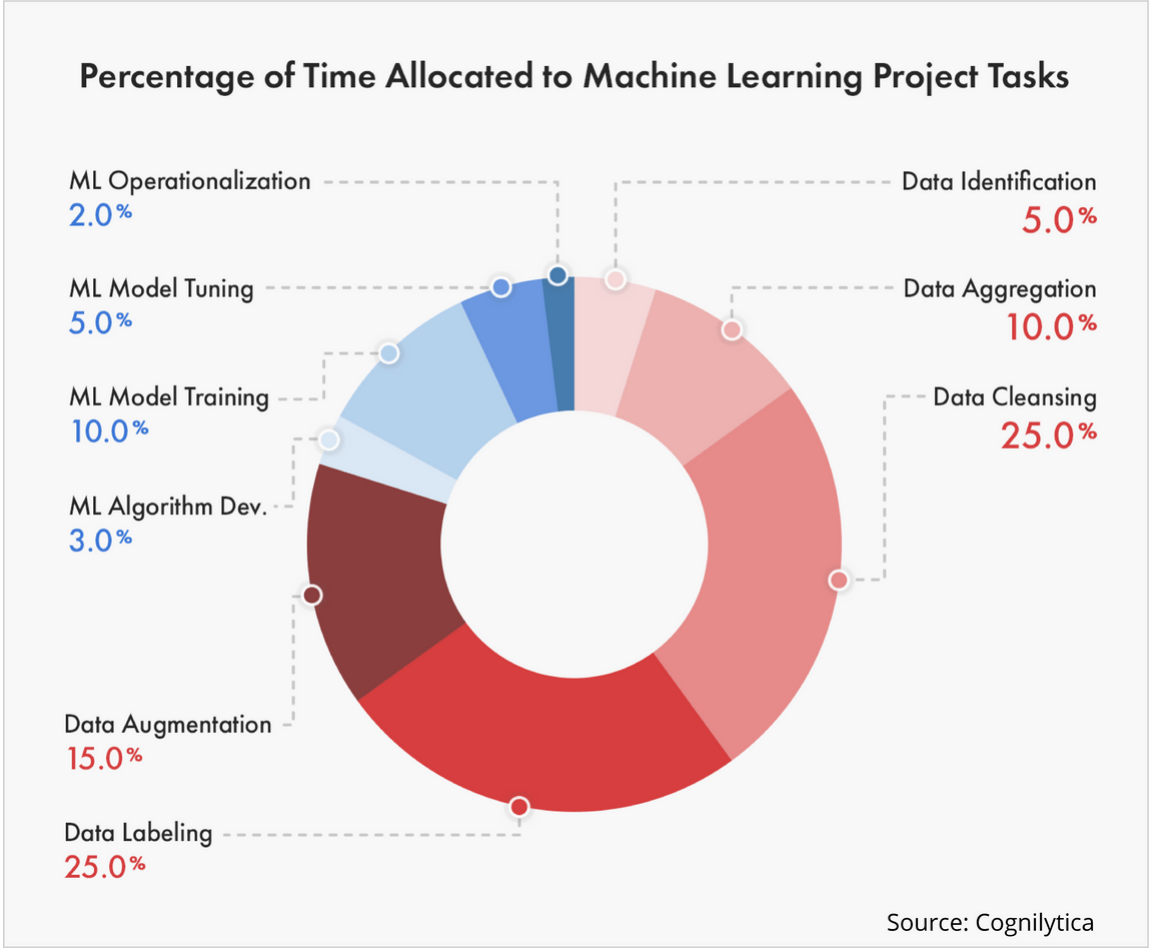
Data labeling



- the process of **adding tags** to raw data (e.g. images, video, text, audio)
- these tags mark which class (group) of objects the observed data belongs to and help machine learning algorithms to identify a certain class (group) of objects when found in unlabeled data

Dana labeling

„Salaries for data scientists can cost up to $190,000/year. It’s expensive to have some of your highest-paid resources wasting time on basic, repetitive work.”



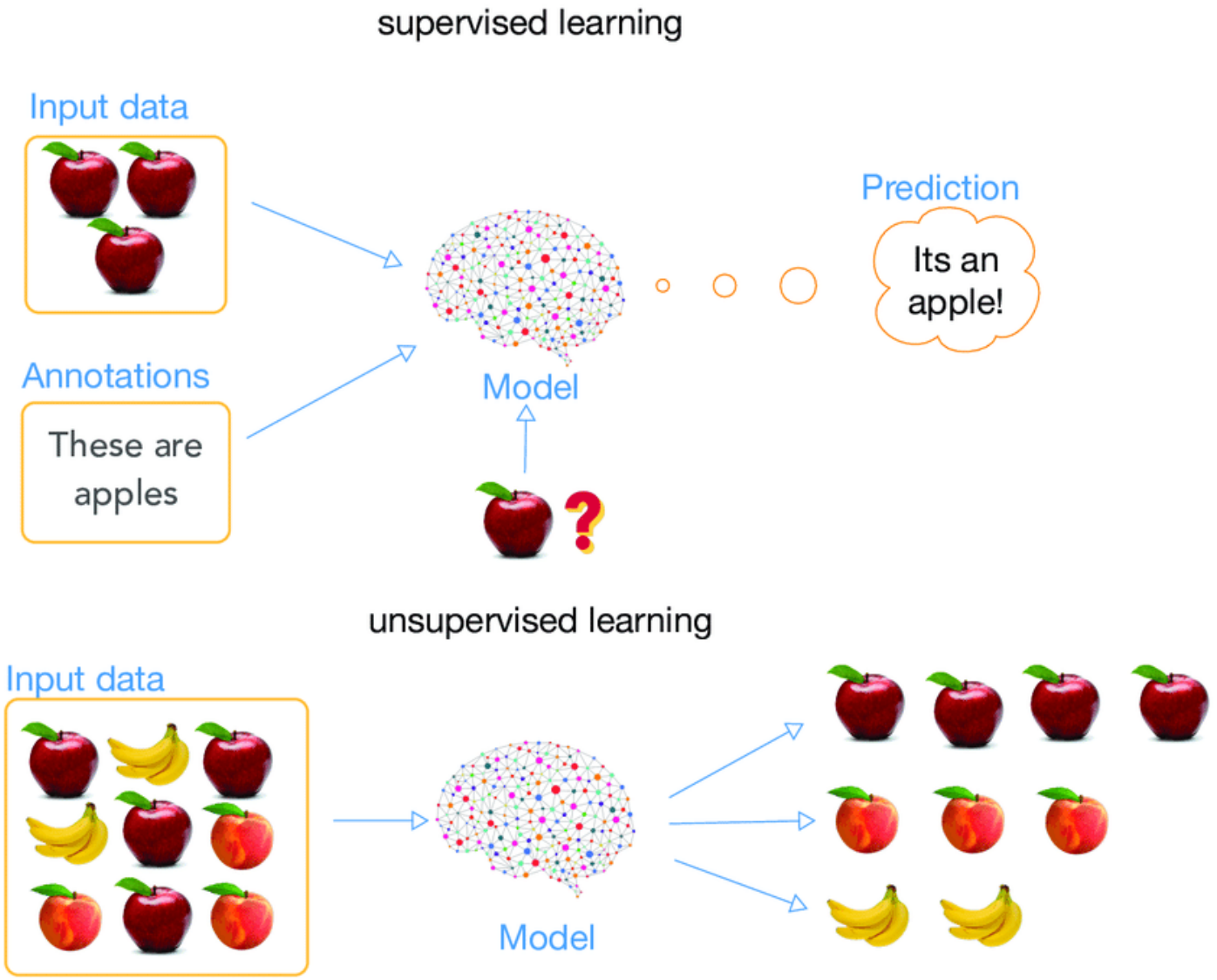
Training data

- the data that are collected and with which the machine learning model learns about the data
- they can take different forms: images, voice, text, features – depending on the used machine learning model and the goal to be achieved
- data can be **annotated or unannotated** - when they are annotated, they are taken as "**ground truth**", i.e. "reference value"

Ground truth

- it is used for information that is known in advance to be **true**
- e.g. image sharpening algorithm - we take a sharp image (ground truth) - we blur it - we sharpen such a blurred image with our algorithm - and compare it with the ground truth to see the success of the algorithm

Types of machine learning (1)

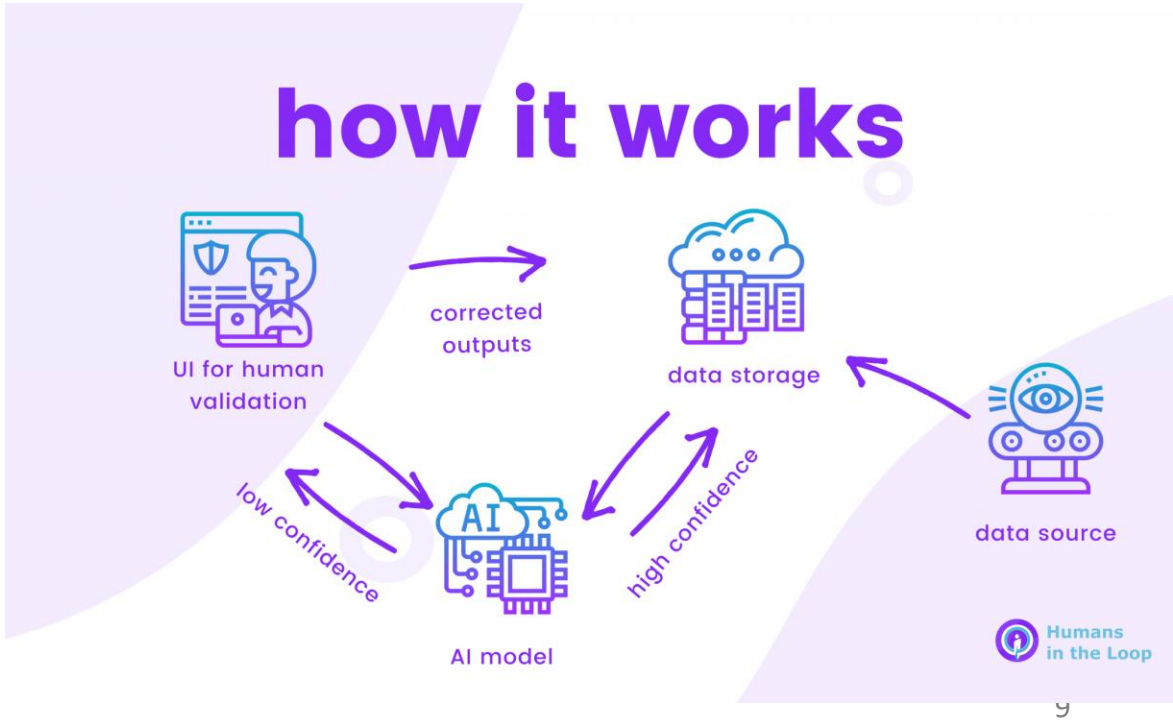


Types of machine learning (2)

Supervised learning	Unsupervised learning	Semi-supervised learning
annotated data	unannotated data	annotated and unannotated data
used for prediction	used for analysis	
has a feedback mechanism	no feedback mechanism	
regression, classification, segmentation	autoencoders that have outputs equal to inputs, clustering	protein sequence classification, Internet content analysis

Human-in-the-loop HITL

- constant human **monitoring and verification of the results of the AI model**: checking whether the prediction is working correctly, identifying gaps in the learning data, providing feedback to the model
- Usage:
  - **annotating learning data**: human annotators annotate learning data fed into (supervised/semi-supervised) machine learning models
  - **learning the model**: people learn the model by constantly monitoring details of the model such as the loss function and prediction. From time to time, people validate the performance of the model and predictions, and the validation results are fed back to the model



Scale

- enable a **flexible tagging process** that allows for scaling as needs and use cases evolve

- E.g. 1h of video – about 800 man-hours for annotation
- 10 min video (30-60 fps) = 18,000-36,000 pictures (frames)

Approaches to data labeling (1)

- it depends on the problem, the time frame of the project, the number of people available

1. In-house data labeling

- ensures the **highest quality**
- they are usually marked by **scientists or people in the organization** (they can also be part-time employees / people for whom data marking is not in the job description)
- correct labeling is essential in, for example, insurance or healthcare
- often requires consulting with experts in the field to correctly label the data
- for higher quality tags – time increases drastically → the whole process is **very slow**

Approaches to data labeling (2)

2. Crowdsourcing

- **with the help of a large number of freelancers registered on the crowdsourcing platform** (they have tens of thousands of registered data annotators)
- annotated datasets mostly consist of trivial data, e.g. images of animals, plants and natural environments, and do not require additional expertise
- https://www.youtube.com/watch?v=6E_IJR22oXk

Approaches to data labeling (3)

3. Outsourcing

- a middle ground between crowdsourcing and in-house labeling
- the task of annotating data is left to **an individual or organization that has trained annotators**
- one of the advantages of outsourcing to individuals is that they can be evaluated on a particular topic before the work is handed over to them
- for projects that do not have a lot of funding, but require significant quality data marking
- https://www.bbc.com/news/technology-46055595

Approaches to data labeling (4)

4. Machine-based annotation

- using **annotation tools and automation** that can drastically increase annotation speed without reducing quality
- automation uses unsupervised (clustering) and semi-supervised machine learning methods

Quality assurance

- **extremely important due to the accuracy of machine learning** that will use this data, e.g. marking passers-by, signs and other vehicles for a self-driving car
- two views:
  - **Accuracy** – measures how well an item is labeled compared to real-world conditions
  - **Quality** – accuracy for the entire data set – does the work of all annotators look the same

Quality measurement methods

1. **Gold standard** – there is a correct answer for an assignment – quality is measured based on correct and incorrect results
2. **Sample review** – a random sample of completed tasks is selected, a more experienced worker (eg team or project manager) reviews the sample
3. **Consensus** – several people perform the same task, and the correct answer is the one that comes from the majority of annotators
4. **Intersection over union (IoU)** – a consensus model often used in object detection within images – combines humans and automation to compare the "bounding boxes" of manually labeled images with predicted bounding boxes from the model

Cronbach alpha

- a measure of the average correlation or consistency of items in a data set
- depending on the characteristics of the study (for example, homogeneity), it can help to quickly access the overall reliability of the labels
- measure of reliability – **tells how internally consistent the scale is**
- N – number of items (questions), $\bar{r}$ – mean correlation between items

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N - 1) \cdot \bar{r}}$$

Cronbach Alpha	internal consistency
0.9 ≤ α	excellent
0.8 ≤ α < 0.9	good
0.7 ≤ α < 0.8	acceptable
0.6 ≤ α < 0.7	questionable
0.5 ≤ α < 0.6	poorly
α < 0.5	unacceptable

Cronbach alpha example

- the restaurant owner wants to measure the overall satisfaction of the visitors
- 10 visitors were asked to answer 3 questions (Q1, Q2, Q3 → N = 3 questions) with grades 1 – 2 – 3

	Q1	Q2	Q3
0	1	1	1
1	2	1	1
2	2	1	2
3	3	2	1
4	2	3	2
5	2	3	3
6	3	2	3
7	3	3	3
8	2	3	2
9	3	3	3

calculate the correlation matrix: $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

	Q1	Q2	Q3
Q1	1.000000	0.429945	0.507630
Q2	0.429945	1.000000	0.662842
Q3	0.507630	0.662842	1.000000

$$\bar{r} = \frac{0.429945 + 0.507630 + 0.662842}{3} = 0.5335$$

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}} = \frac{3 \cdot 0.5335}{1 + (3-1) \cdot 0.5335} = 0.774$$

Types of labels - examples

- it depends on the desired result of machine learning
- most common areas:
  - Computer vision
    - labeled visual data in the form of images
    - https://www.youtube.com/watch?v=UdxRgZJf6dE
  - Natural language processing
    - analysis of human language and their forms during interaction with other humans and machines
  - Speech / sound

Image classification

- **adding a tag to an image**
- the number of unique tags in the entire database is the number of classes that the model can classify
- the classification problem can be divided into:
  - **Binary classification** (consisting of only two labels)
  - **Multi-class classification** (containing several labels)
- classification with **multiple tags** is also possible, e.g. when detecting diseases → each image has more than one tag

Object detection

- **detection of objects and their locations**
- each object is marked with the smallest possible rectangle that surrounds it (**bounding box**)
- usually a **label** is attached to each rectangle
- the **coordinates of the rectangle** and the corresponding **label** are remembered and stored in a JSON file in dictionary format where the key of the dictionary is the number or ID of the image

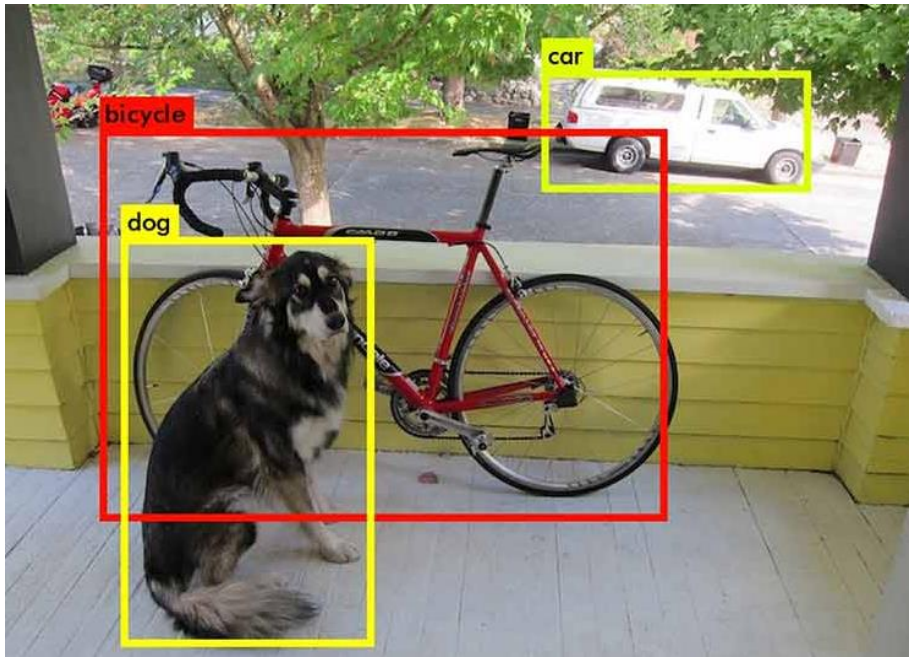
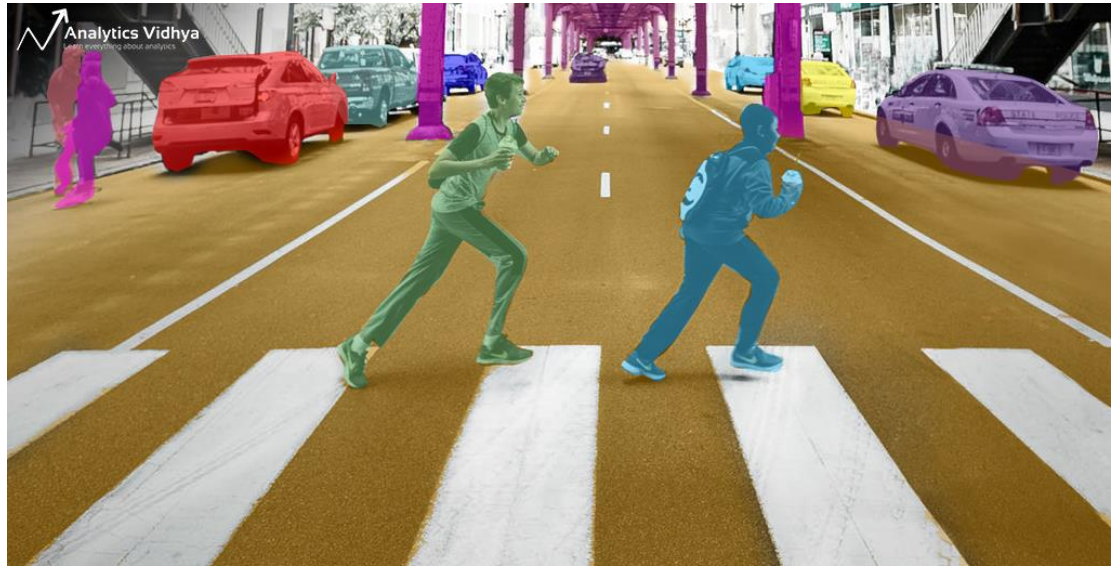


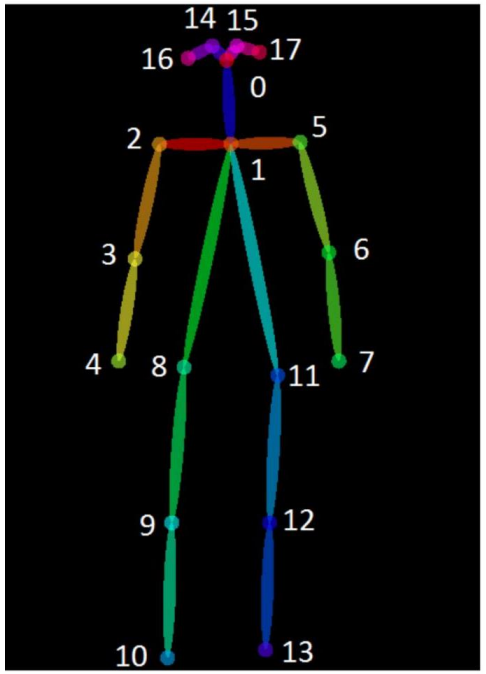
Image segmentation

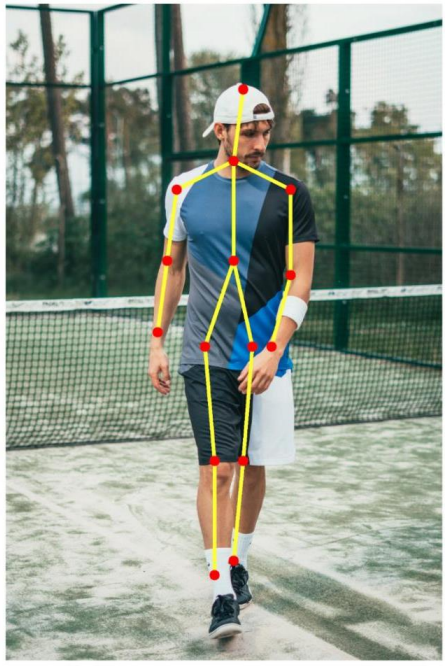
- **separate image objects from their background and other objects** in the image → usually result in images of the same size as the original image, containing 1 where the object is present and 0 otherwise
- if several objects are segmented within one image, each object is marked on a separate channel → the sum of all channels is the reference image

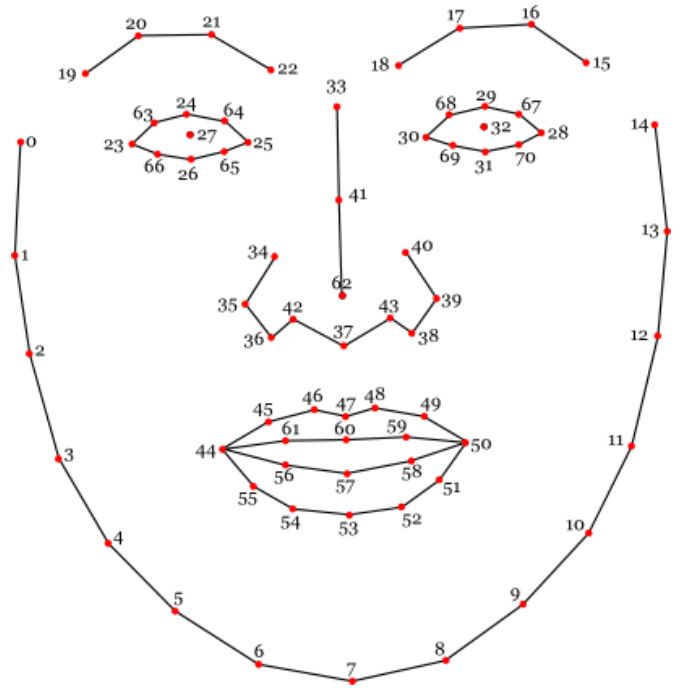


Human pose estimation

- detection of **key points** on the body and linking them to the pose
- reference value: coordinates and associated markings of key points

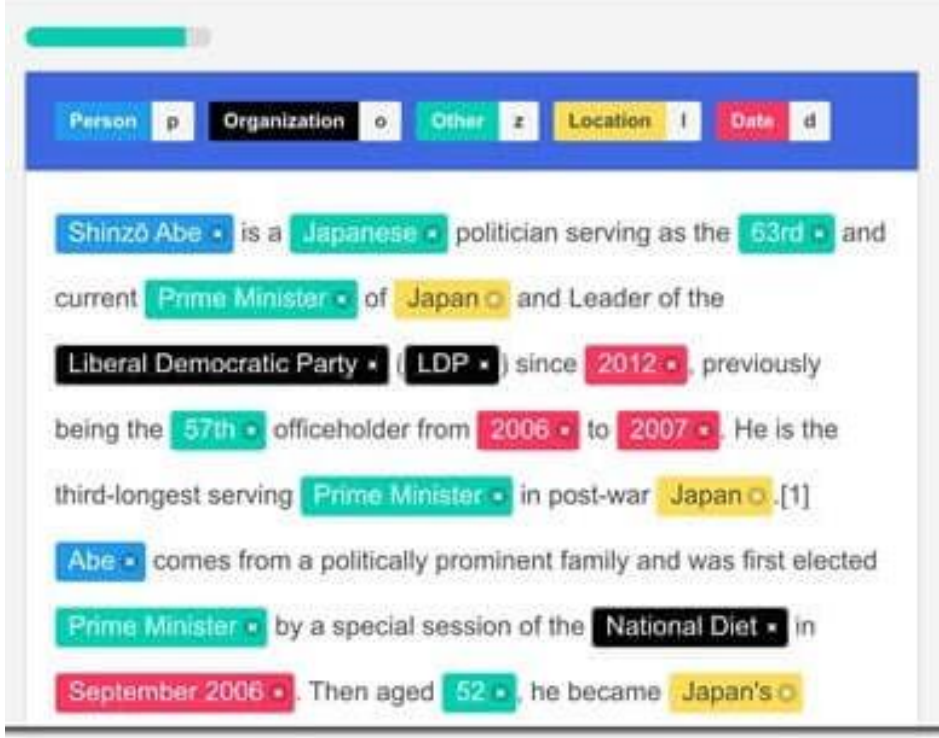






Entities marking

- locating, extracting and marking entities in the text
- annotators read the text, locate the target entities and mark them using predefined tags
- the word "entity" can have different forms depending on the task:
  - **proper nouns** → entity marking refers to the identification and marking of names in the text
  - for **phrase analysis** → key words or key phrases are marked
  - for analyzing and marking **functional elements** of any text such as verbs, nouns, prepositions → marking **parts of speech (POS)** → used for parsing, machine translation and generation of linguistic data

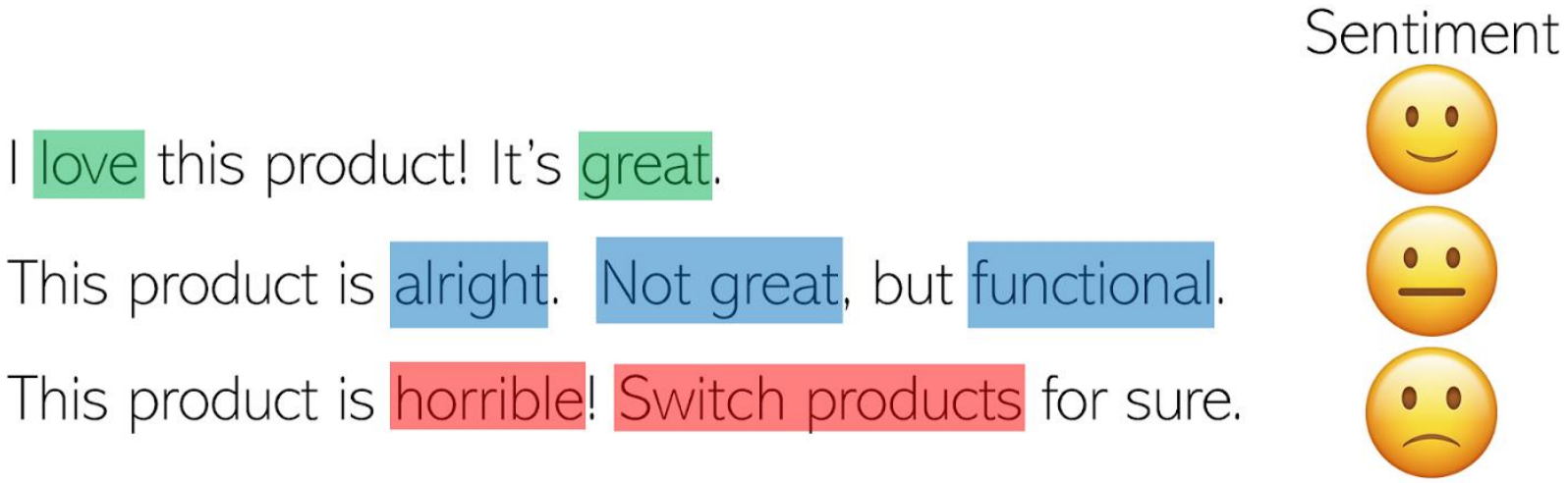


Text classification

- adding one or more **tags to blocks of text**
- the text is viewed as a whole (content, subject, intention, feeling within the text) and a label is assigned to that whole text (based on a known list of labels)
- types of text classification:
  - classification based on **feelings or opinions** (for sentiment analysis)
  - classification based on the **topic** that the text wants to convey (for categorizing topics)
  - **document classification** – for sorting and retrieving documents based on content

Emotion annotation

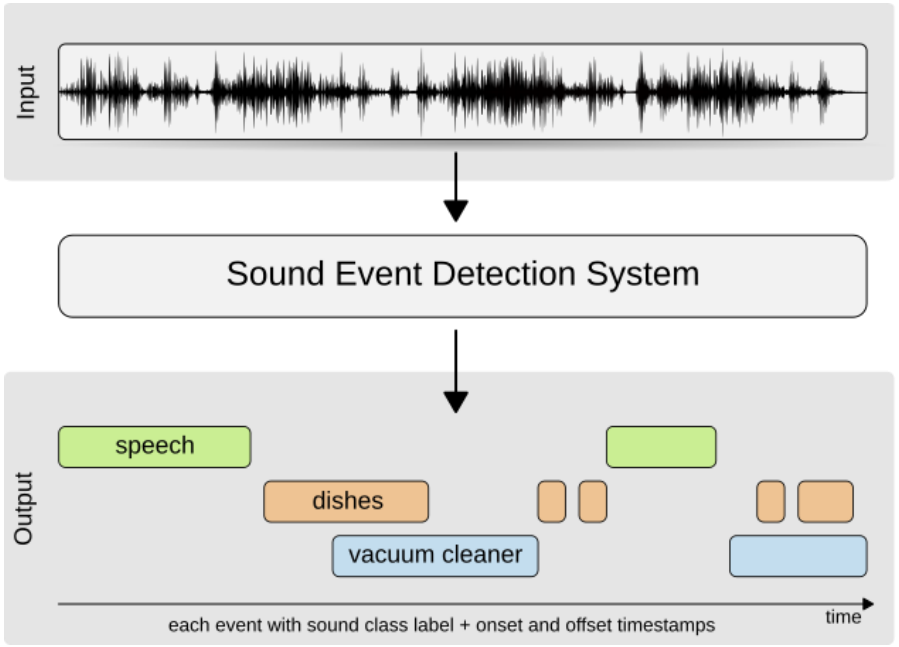
- detection of hidden connotations, sarcasm, wit - real emotions under the text
- marking: **emotions, opinions, sentiments**
- E.g. analysis of guest reviews - based on the given reviews, the annotators should choose a label: positive / neutral / negative

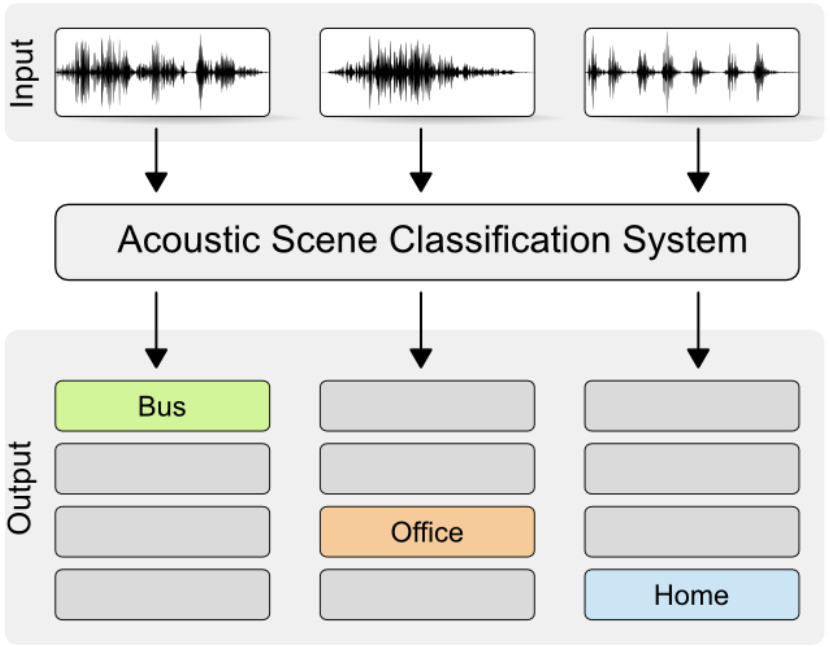


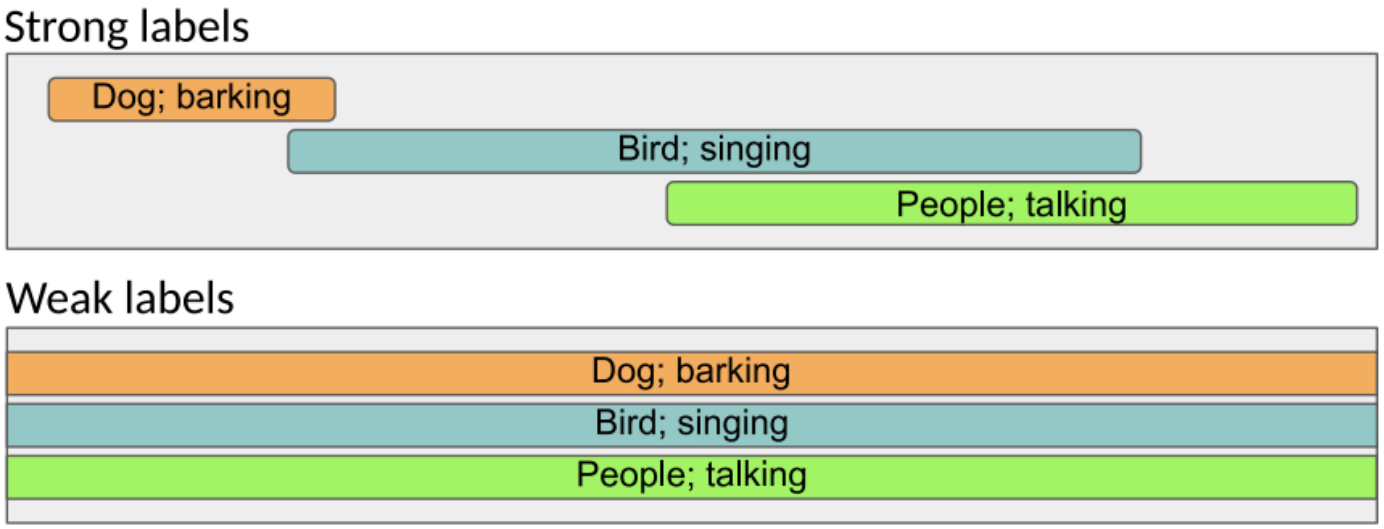
Linguistic annotation

- identification and marking of **grammatical, semantic or phonetic elements** in text or speech
- specially used in chatbots, virtual assistants, search engines, translators
- Types:
  - **Discourse annotation** – connecting anaphora and cataphora with their preceding or following subjects. Example: Ivan broke the chair. He felt really bad about it.
  - **Part-of-speech (POS)** tagging – annotation of different function words within a text
  - **Phonetic** annotation – marking of intonation, emphasis and natural pauses in speech
  - **Semantic** annotation – annotation of word definitions

Sound

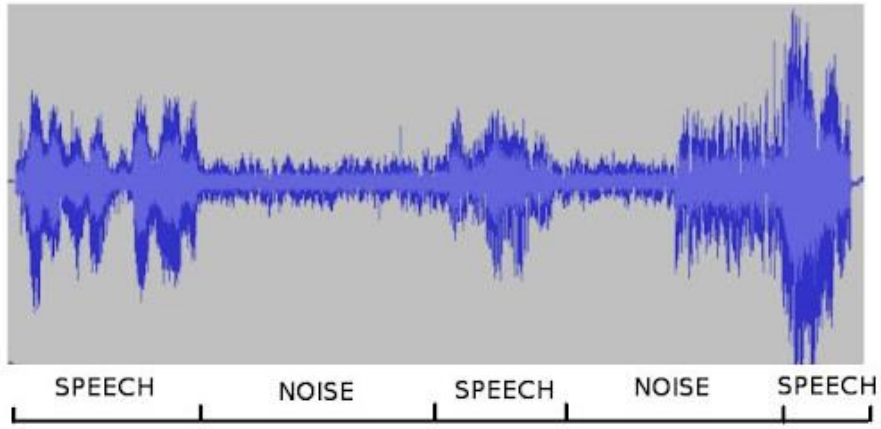

Input
Sound Event Detection System
Output
speech
dishes
vacuum cleaner
each event with sound class label + onset and offset timestamps
time


Input
Acoustic Scene Classification System
Output
Bus
Office
Home

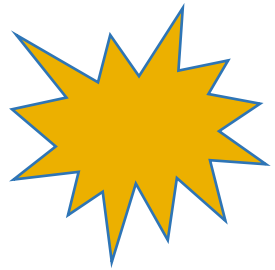

Strong labels
Dog; barking
Bird; singing
People; talking
Weak labels
Dog; barking
Bird; singing
People; talking

Speech

- speaker identification
  - adding a tag to an audio file
- linguistic data
  - language regions are marked first because the audio is not expected to contain 100 percent of speech
  - the surrounding sounds are marked and a speech transcript is created for further processing with the help of NLP algorithms



References



https://www.v7labs.com/blog/data-labeling-guide

https://appen.com/blog/data-labeling/

https://www.cloudfactory.com/data-labeling-guide

https://labelyourdata.com/articles/data-labeling-quality-and-how-to-measure-it

https://towardsdatascience.com/cronbachs-alpha-theory-and-application-in-python-d2915dd63586

https://www.statology.org/cronbachs-alpha-in-python/

https://hackernoon.com/introduction-5-different-types-of-text-annotation-in-nlp-78523wwo