Data acquisition

- **Duration:** 1 hr
- **Outline:**

1. Importance of data
2. Dataset building

Data collection

- **Duration:** 1 hr
- **Outline:**

1. Importance of data

2. Dataset building

Importance of data

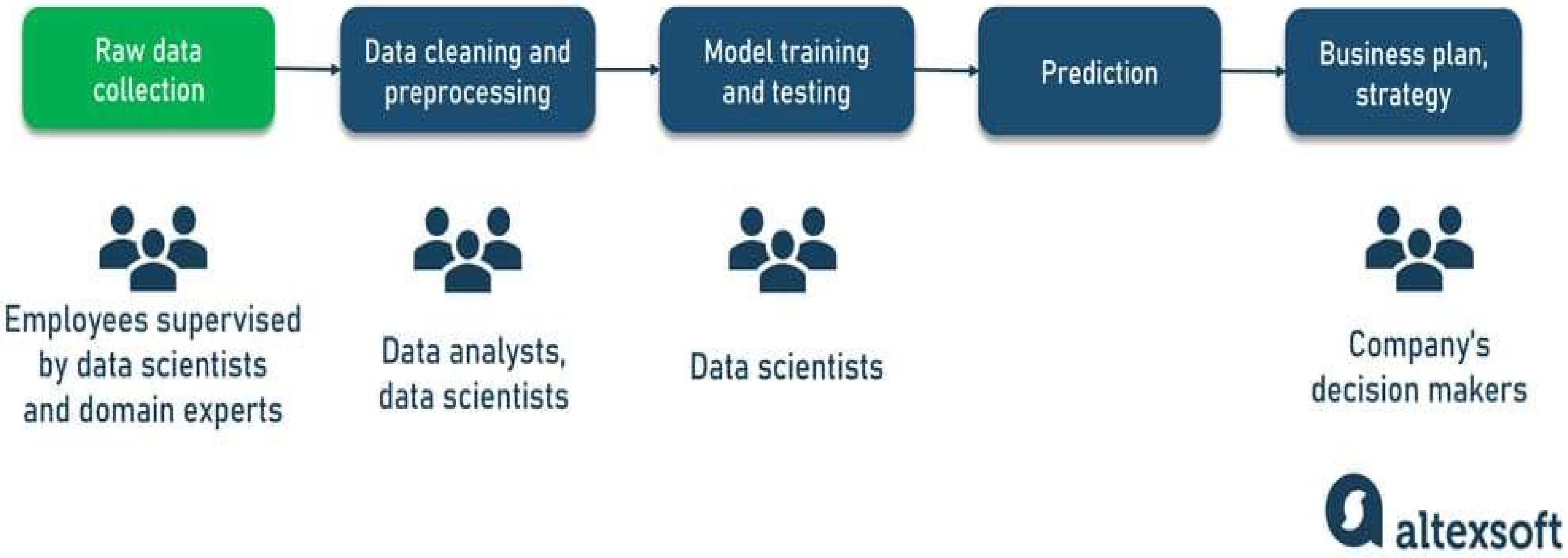
- **Machine Learning: How?**
  - Data Collection
    - **Goals**
      - First requirement: **having good data**
        - » Get meaningful, **representatives examples** of each concept to capture, **balanced across classes**, etc.
        - » Get **accurate annotations**
          - E.g., songs with accurate emotion tags might be hard to get, as emotion is naturally ambiguous…

Importance of data

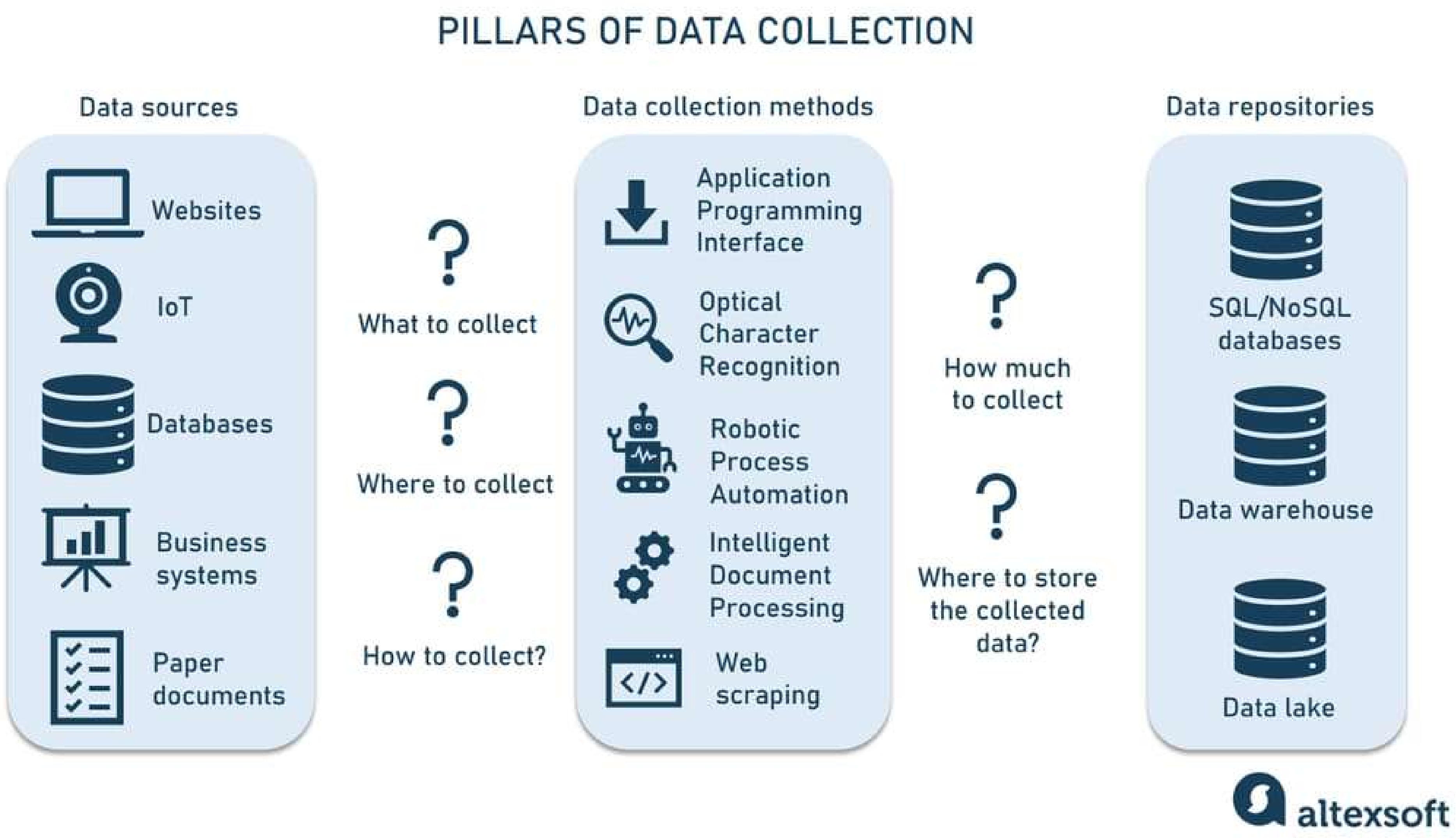
- **Machine Learning: How?**
  - Data Collection
    - **Goals**
      - First requirement: **having good data**
        - » Get meaningful, **representatives examples** of each concept to capture, **balanced across classes**, etc.
        - » Get **accurate annotations**
          - E.g., songs with accurate emotion tags might be hard to get, as emotion is naturally ambiguous…

Importance of data

DATA COLLECTION IN DECISION-MAKING PROCESS


Raw data collection
Data cleaning and preprocessing
Model training and testing
Prediction
Business plan, strategy
Employees supervised by data scientists and domain experts
Data analysts, data scientists
Data scientists
Company's decision makers
altexsoft

Importance of data



PILLARS OF DATA COLLECTION

Data sources
- Websites
- IoT
- Databases
- Business systems
- Paper documents

What to collect

Where to collect

How to collect?

Data collection methods
- Application Programming Interface
- Optical Character Recognition
- Robotic Process Automation
- Intelligent Document Processing
- Web scraping

How much to collect

Where to store the collected data?

Data repositories
- SQL/NoSQL databases
- Data warehouse
- Data lake

altexsoft

Importance of data

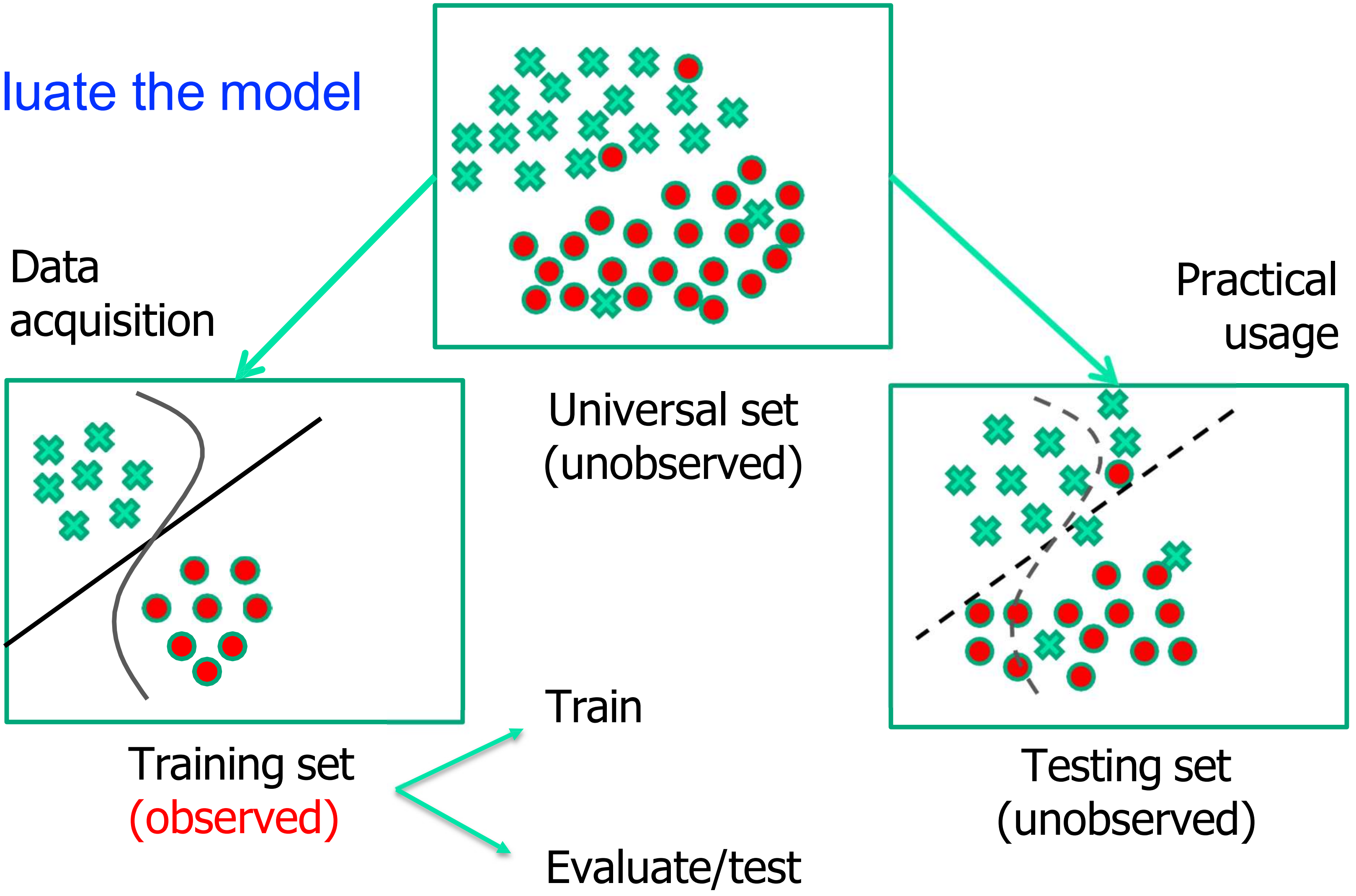
- ML depends heavily on data.
- Data in unorganized format is not useful for machines to ingest the useful information.
- Flawed data can make a ML system harmful.

Ex: The absence of asthmatic death cases in the data used for a healthcare project which aims to cut costs in the treatment of patients with pneumonia

- In every ML/AI projects, data preparation takes most of time

Data is used for…

- Train the model
- Evaluate the model



Data acquisition

Universal set (unobserved)

Practical usage

Training set (observed)

Train

Evaluate/test

Testing set (unobserved)

What factors make a good dataset?

- The right quantity
- The approach to split data
- The past history
- Domain expertise (Two key qualities: independence and identical distribution)
- The right kind of data transformation

https://www.promptcloud.com/blog/what-to-look-for-in-training-dataset/

Data collection

- **Duration:** 1 hr
- **Outline:**

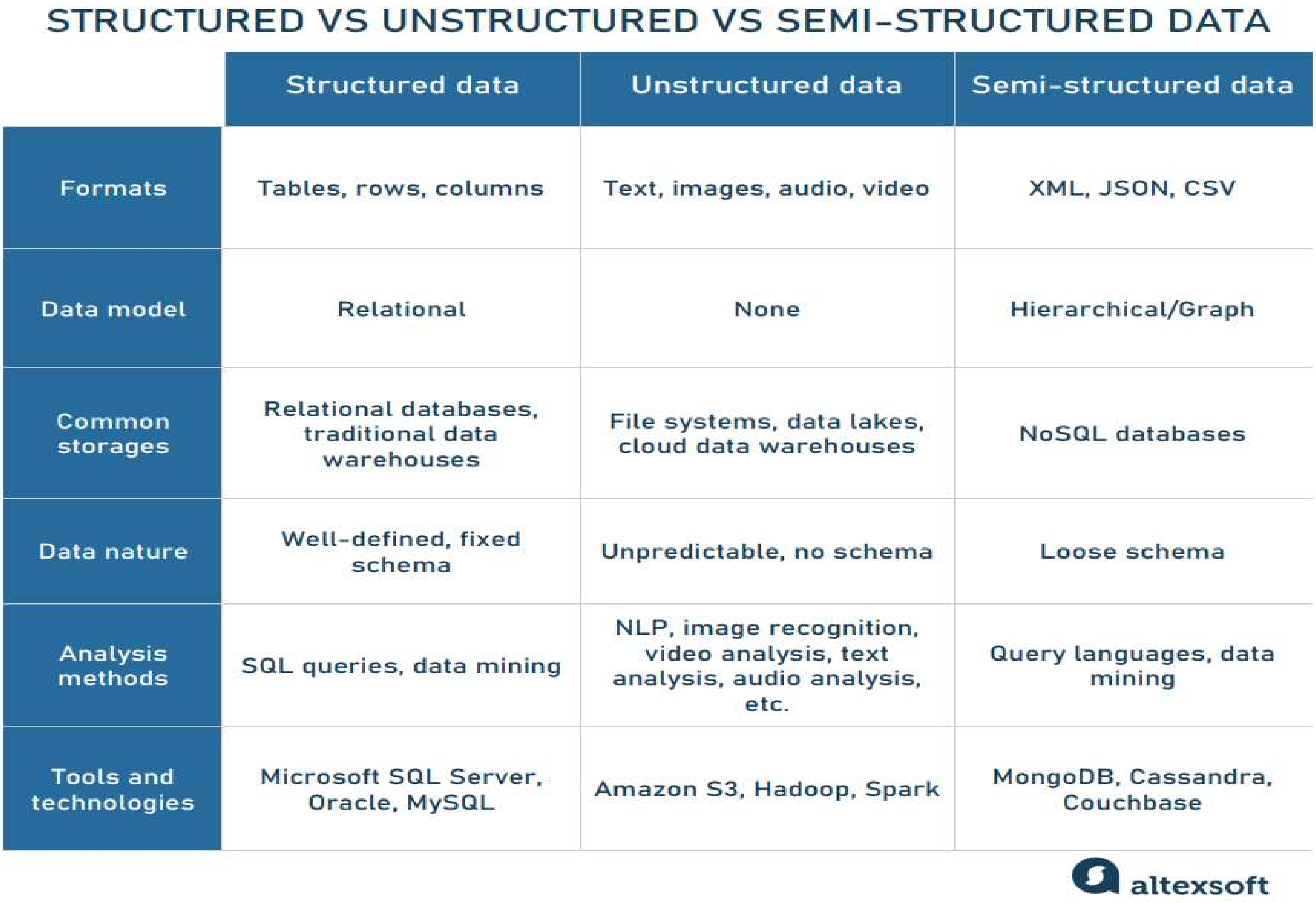
1. Importance of data

2. Dataset building

Dataset structure

- Dataset comprises data and labels:
  - Data: array [m, k] stores the k-D feature vectors of m objects
  - Labels: contain the m object labels
- Label types:
  - Integer numbers
  - String (class name)
  - Soft: real numbers in interval [0,1]
  - Target: numeric values in interval $(-\infty, +\infty)$

Dataset structure



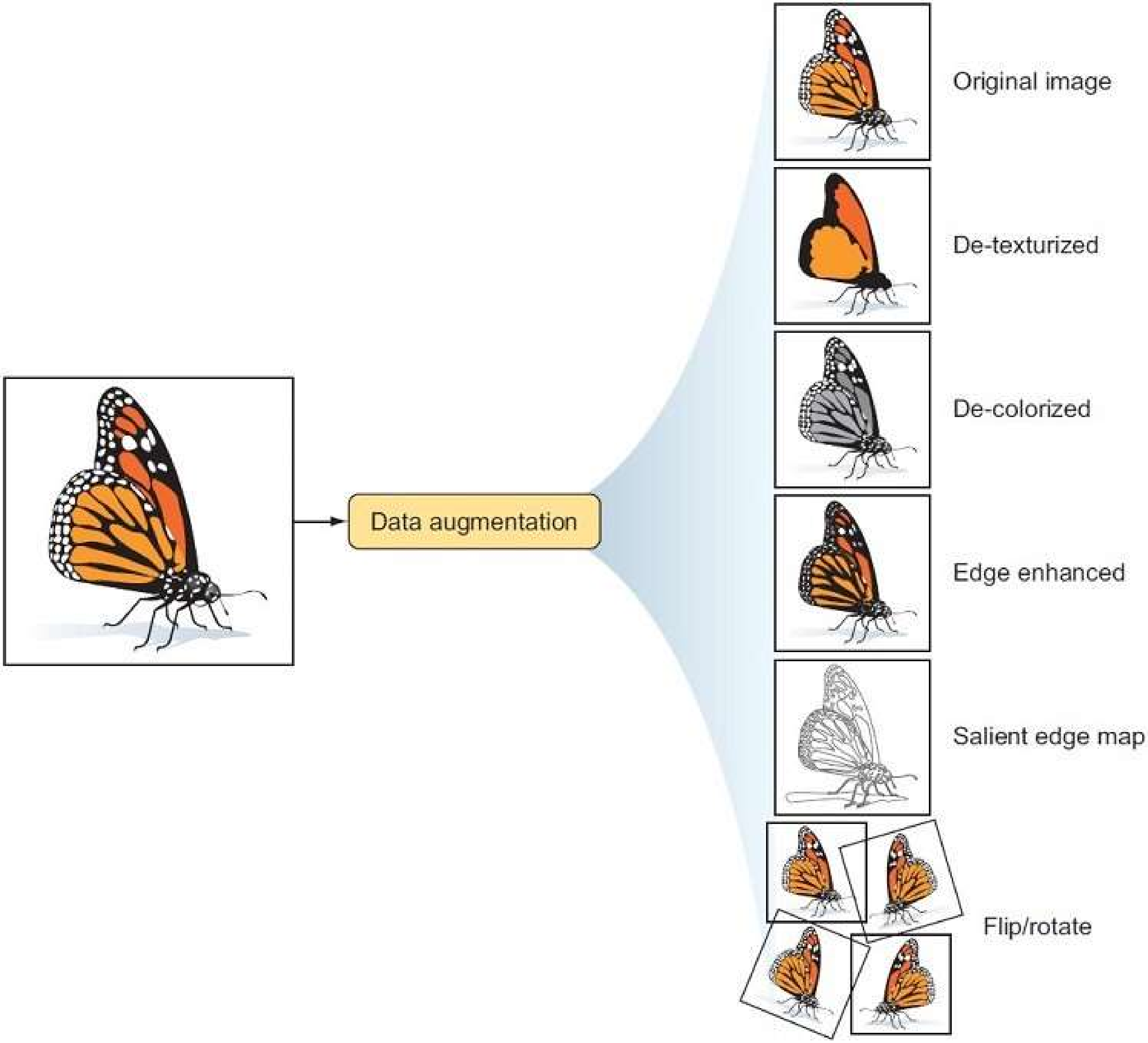
STRUCTURED VS UNSTRUCTURED VS SEMI-STRUCTURED DATA

	Structured data	Unstructured data	Semi-structured data
Formats	Tables, rows, columns	Text, images, audio, video	XML, JSON, CSV
Data model	Relational	None	Hierarchical/Graph
Common storages	Relational databases, traditional data warehouses	File systems, data lakes, cloud data warehouses	NoSQL databases
Data nature	Well-defined, fixed schema	Unpredictable, no schema	Loose schema
Analysis methods	SQL queries, data mining	NLP, image recognition, video analysis, text analysis, audio analysis, etc.	Query languages, data mining
Tools and technologies	Microsoft SQL Server, Oracle, MySQL	Amazon S3, Hadoop, Spark	MongoDB, Cassandra, Couchbase

altexsoft

How to build dataset?

- Start small and reduce the complexity of the data.
- Articulate the problem early (i.e., classification, detection, ranking,…)
- Establish data collection mechanisms
- Check the data quality (human errors, technical problems, missing features, adequate?, imbalanced?)
- Format data
- Clean data
- Segmentation
- Complete **feature engineering**



Data augmentation

Original image

De-texturized

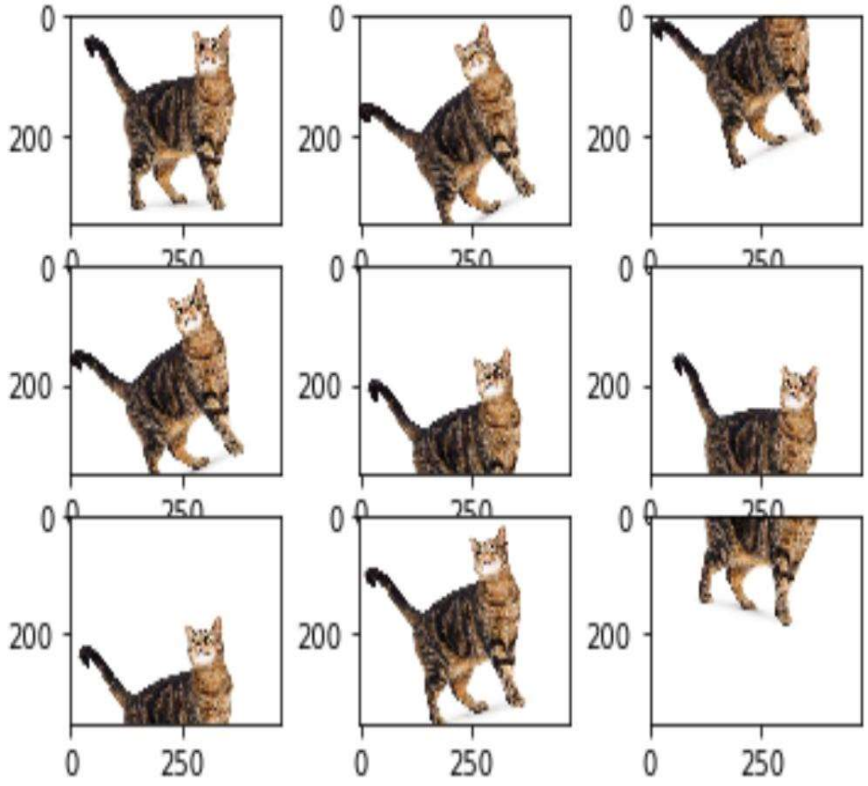
De-colorized

Edge enhanced

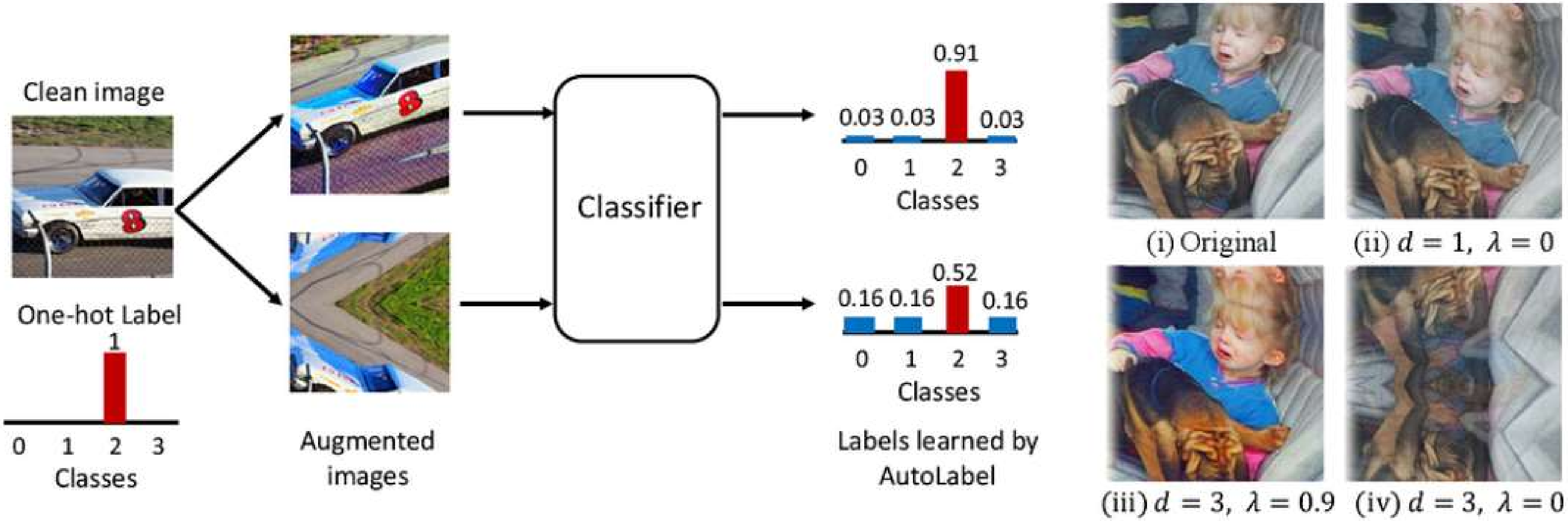
Salient edge map

Flip/rotate

How to build dataset?



How to build dataset?



(i) Original

(ii) $d = 1,\ \lambda = 0$

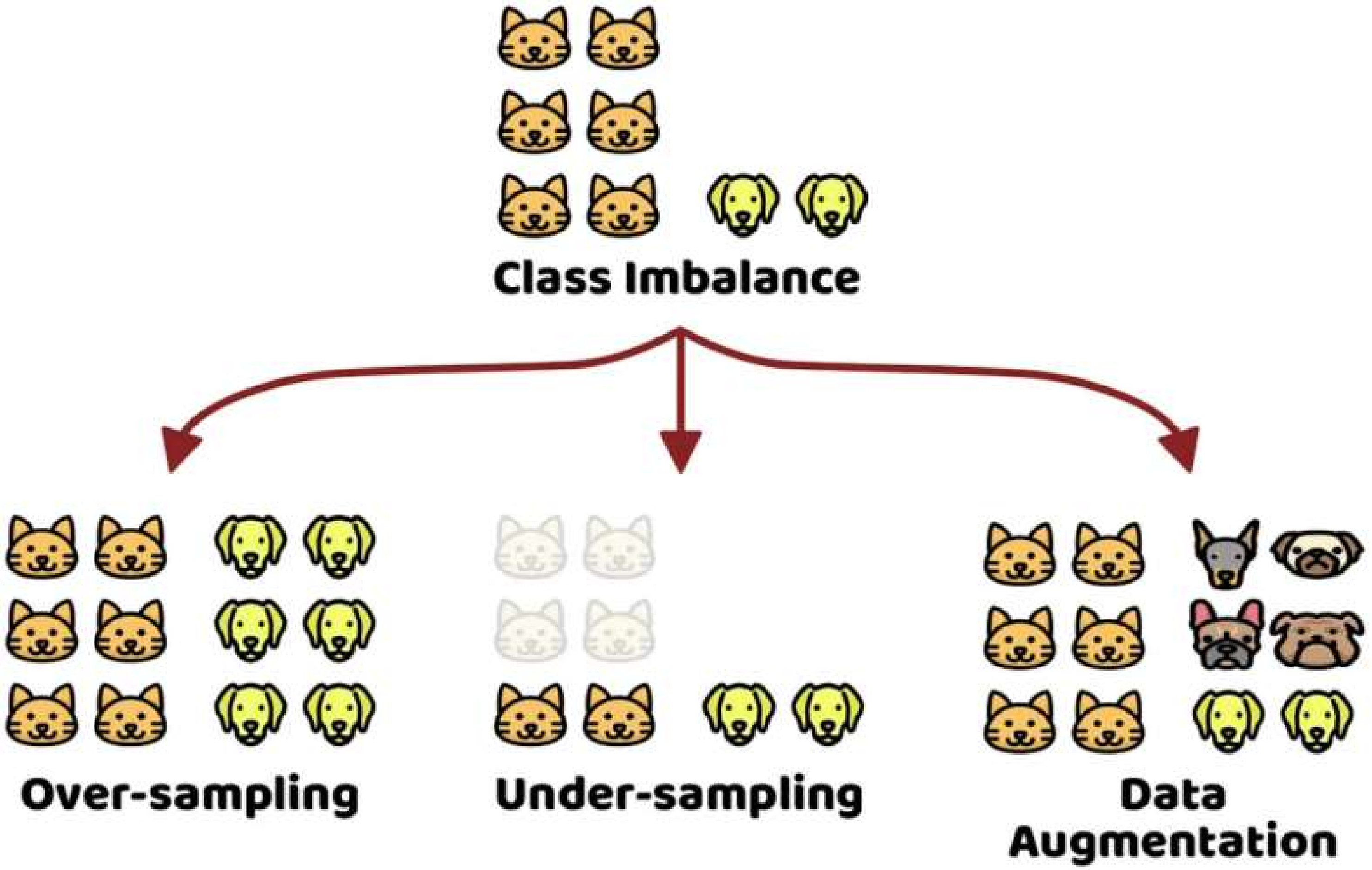
(iii) $d = 3,\ \lambda = 0.9$

(iv) $d = 3,\ \lambda = 0$

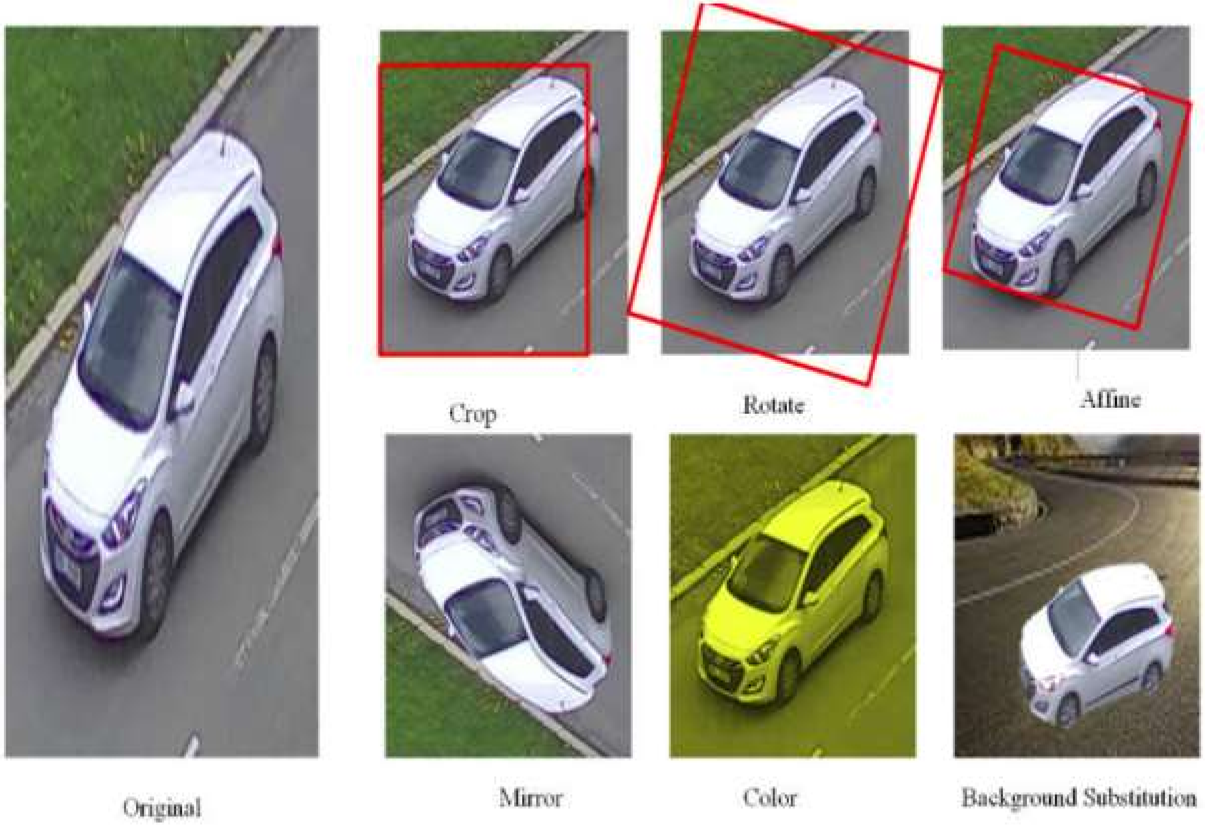
(a) Pipeline

(b) Augmented images of varying distances

How to build dataset?
Balance Act

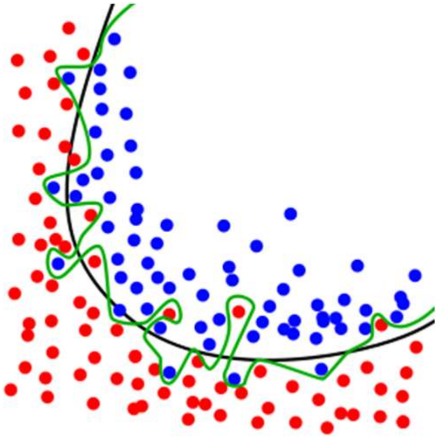


How to build dataset?
Balance Act

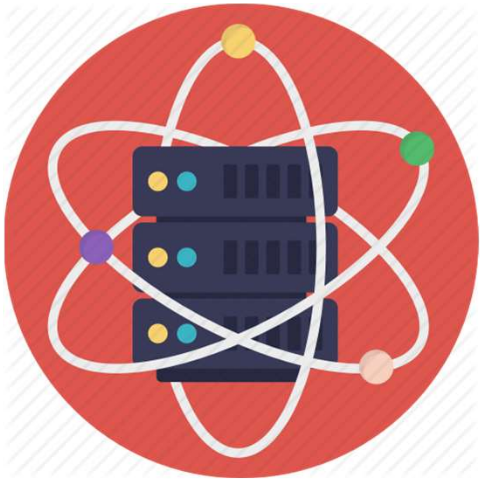


Purpose of Data Augmentation

Enlarge
dataset



Prevent
overfitting



Improve
performance model

What to Augment!

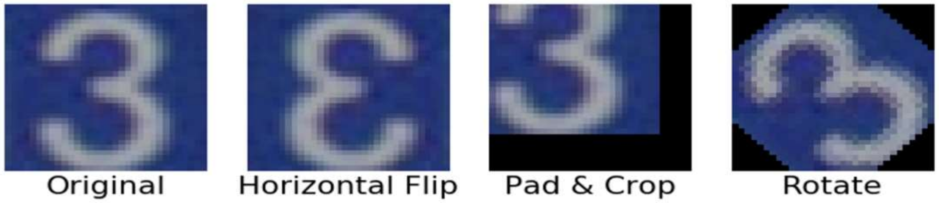
Audio

Texts

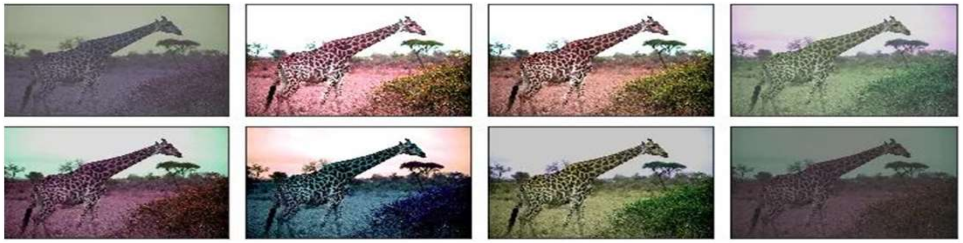
Images

Any other types

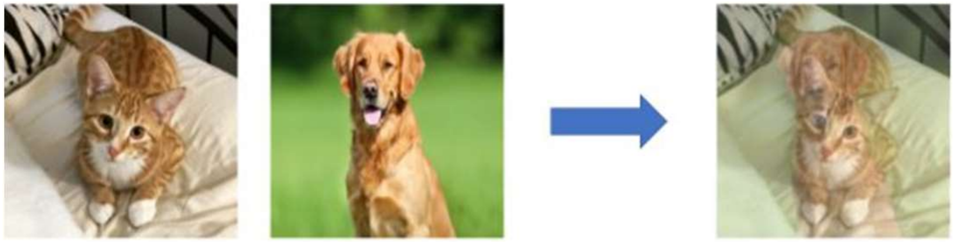
Data Augmentation techniques : For Images



Geometric transformations



Color space transformations



Mixing images

Data Augmentation techniques : For Text

In my presentation focus topic is Data Augmentation.

In my presentation **theme** topic is Data Augmentation -- Synonym Replacement

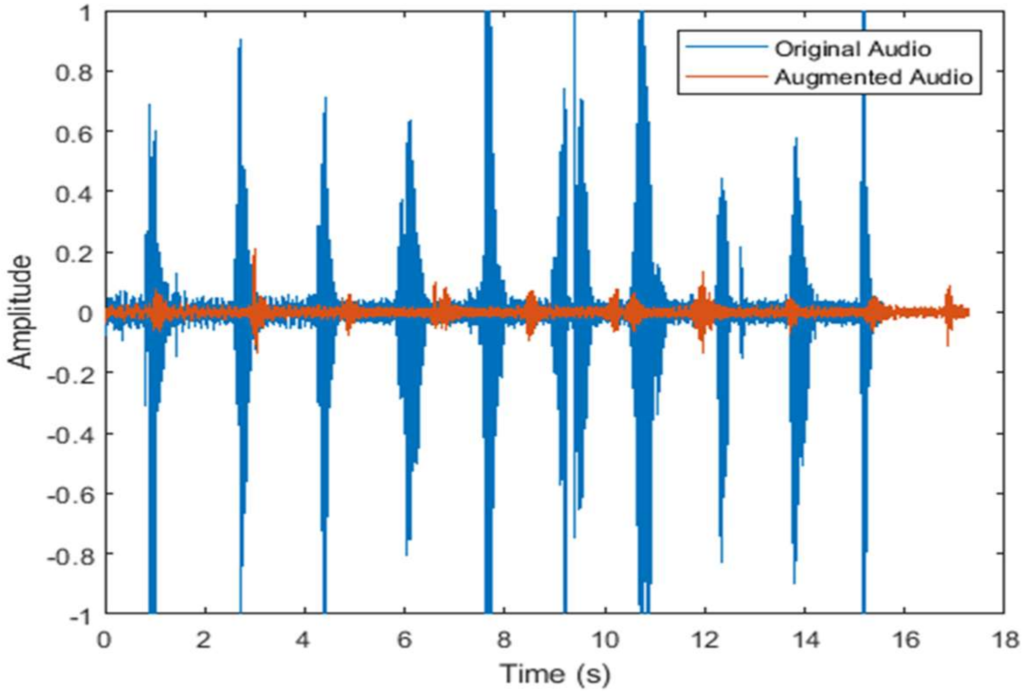
In my presentation focus topic is Data Augmentation **techniques** -- Random Insertion

In my presentation **topic focus** is Data Augmentation -- Random Swap

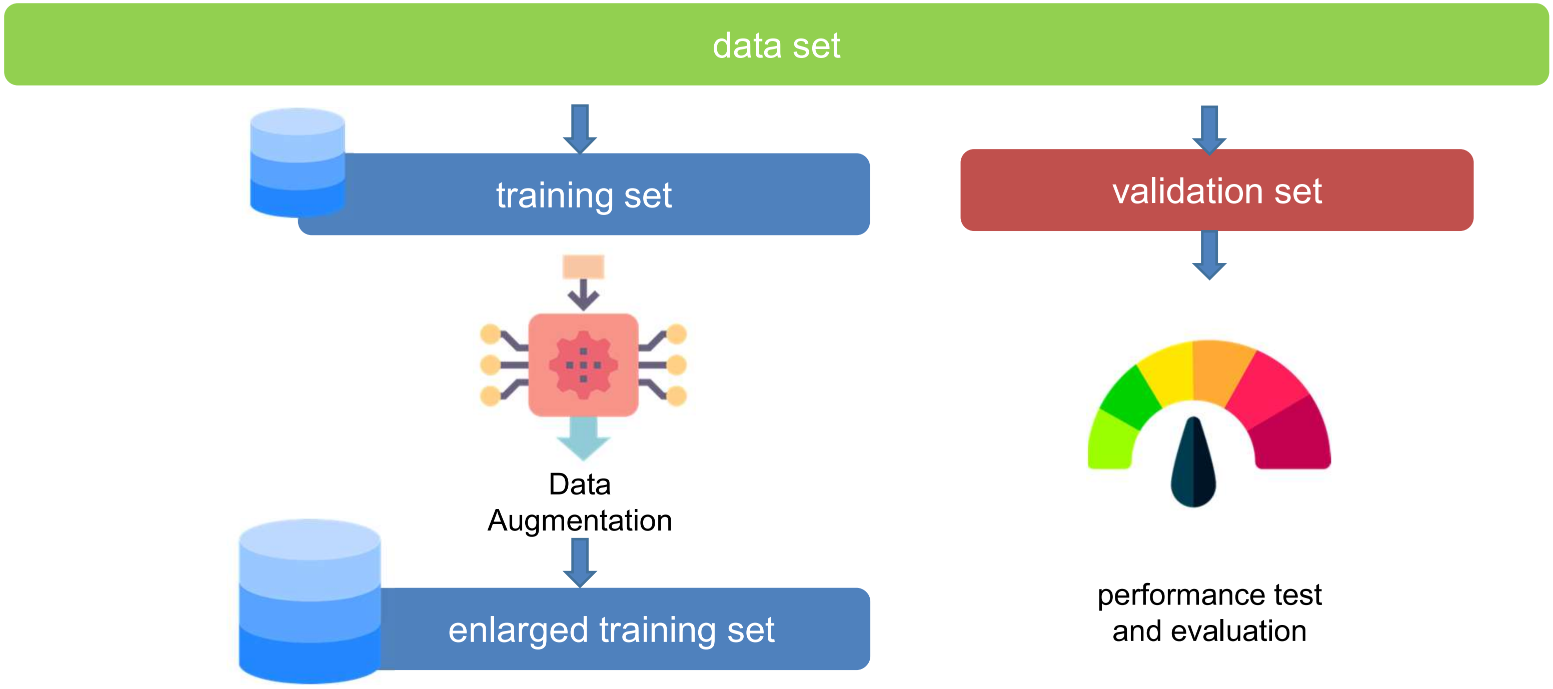
In my presentation focus topic Data Augmentation --Random Deletion

Data Augmentation techniques : For Audio

- Noise Injection
- Shifting
- Changing the speed of the Tape



Workflow of Data Augmentation



data set

training set

validation set

Data Augmentation

enlarged training set

performance test and evaluation

When do I use data augmentation?

small data set

complex problem

transformation of data would be effective and easy

You know how to find the correct method for your data

You have a strategy for dealing with overfitting

Data Augmentation Applications

Natural Language Processing

Image classification

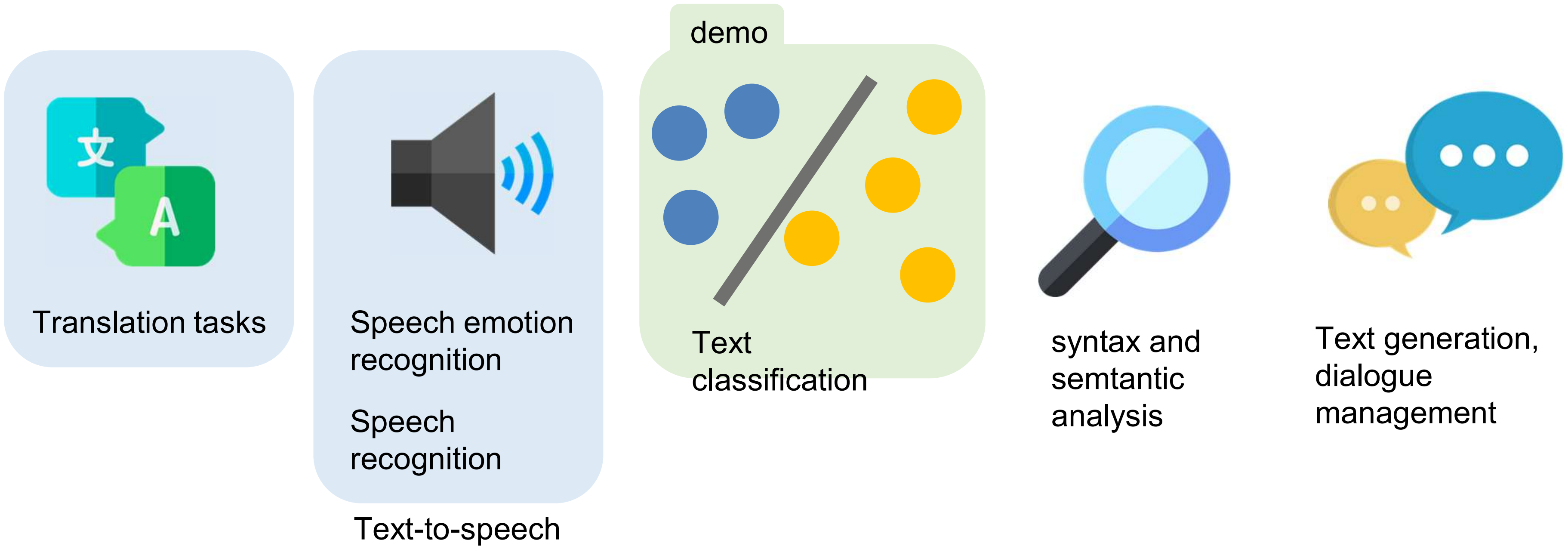
Speech recognition, sound classification

text classification, text generation

Scientific research, medical fields

Human-computer interaction, Understanding human emotion

Natural Language Processing: some tasks



Translation tasks

Speech emotion recognition

Speech recognition

Text-to-speech

demo

Text classification

syntax and semtantic analysis

Text generation, dialogue management

NLP and Data augmentation in speech emotion recognition

C. Etienne and B. Schmauch, **"Speech Emotion Recognition with Data Augmentation and Layer-wise Learning Rate Adjustment"**

Problem:

class imbalance and small dataset

Solution:

Data augmentation by rescaling of spectograms

Results:

Improvement in accuracy of predictions

	Baseline			Best model
Augmentation during training	-	-	+	+
Oversampling (×2) of happiness and anger	-	+	+	+
Frequency range (kHz)	4	4	4	8
Weighted accuracy	66.4	63.5	64.2	64.5
Unweighted accuracy	57.7	59.8	60.9	61.7

10-cross validation scores depending on the techniques applied (for each experiment we present the results corresponding to its best run).

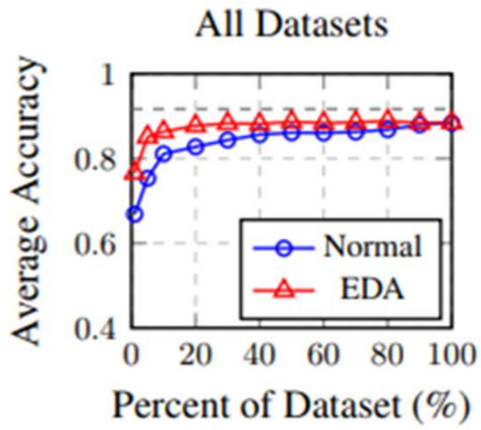
NLP and Data augmentation in classification tasks

J. Wei and K. Zou, **"EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks"**

Problem: Performing text classification depends on quality and quantity of data

Solution: Data augmentation by application of multiple transformations on text

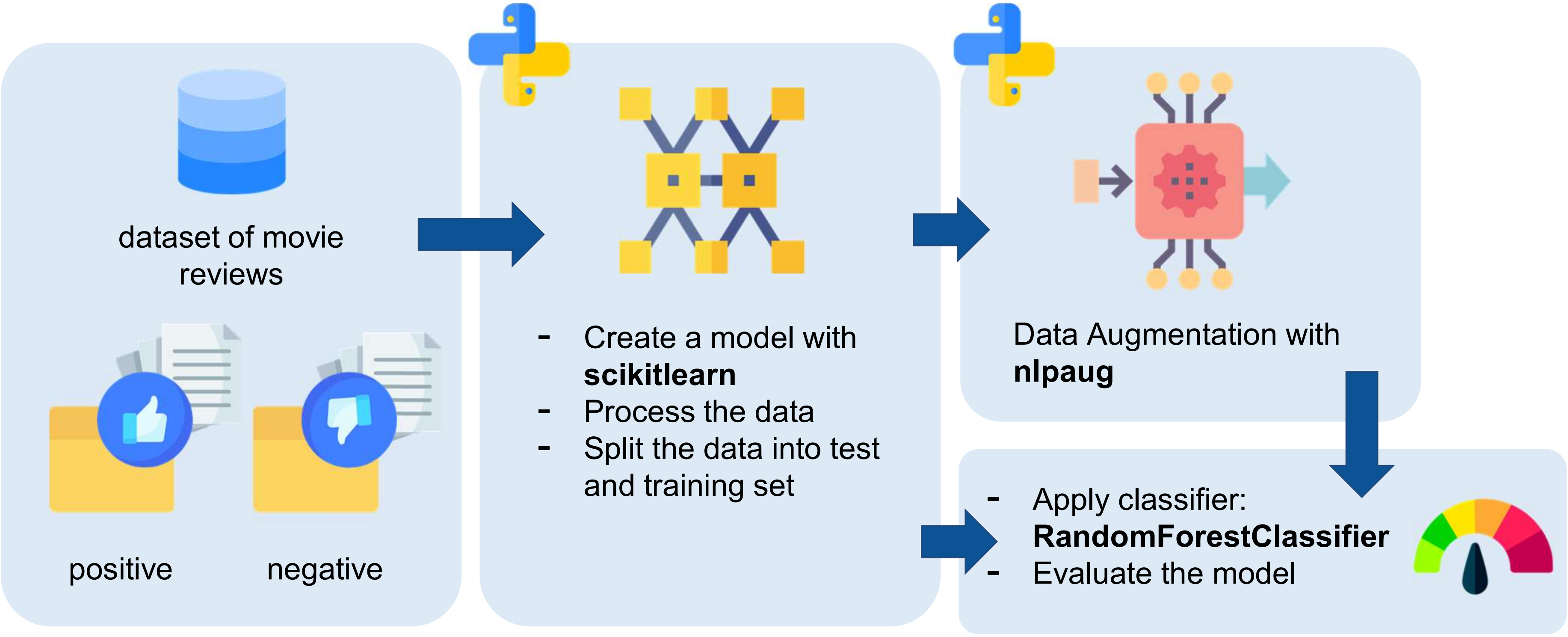
Results: Performance gain of model if right amount of data augmentation chosen



Performance on benchmark text classification tasks with and without EDA, for various dataset sizes used for training. [1]

[1] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 2019, p. 6384.

Demo: NLP text classification and Data Augmentation



dataset of movie reviews

positive

negative

- Create a model with **scikitlearn**
- Process the data
- Split the data into test and training set

Data Augmentation with **nlpaug**

- Apply classifier: **RandomForestClassifier**
- Evaluate the model

https://github.com/mimmimkr/nlp_dataaug

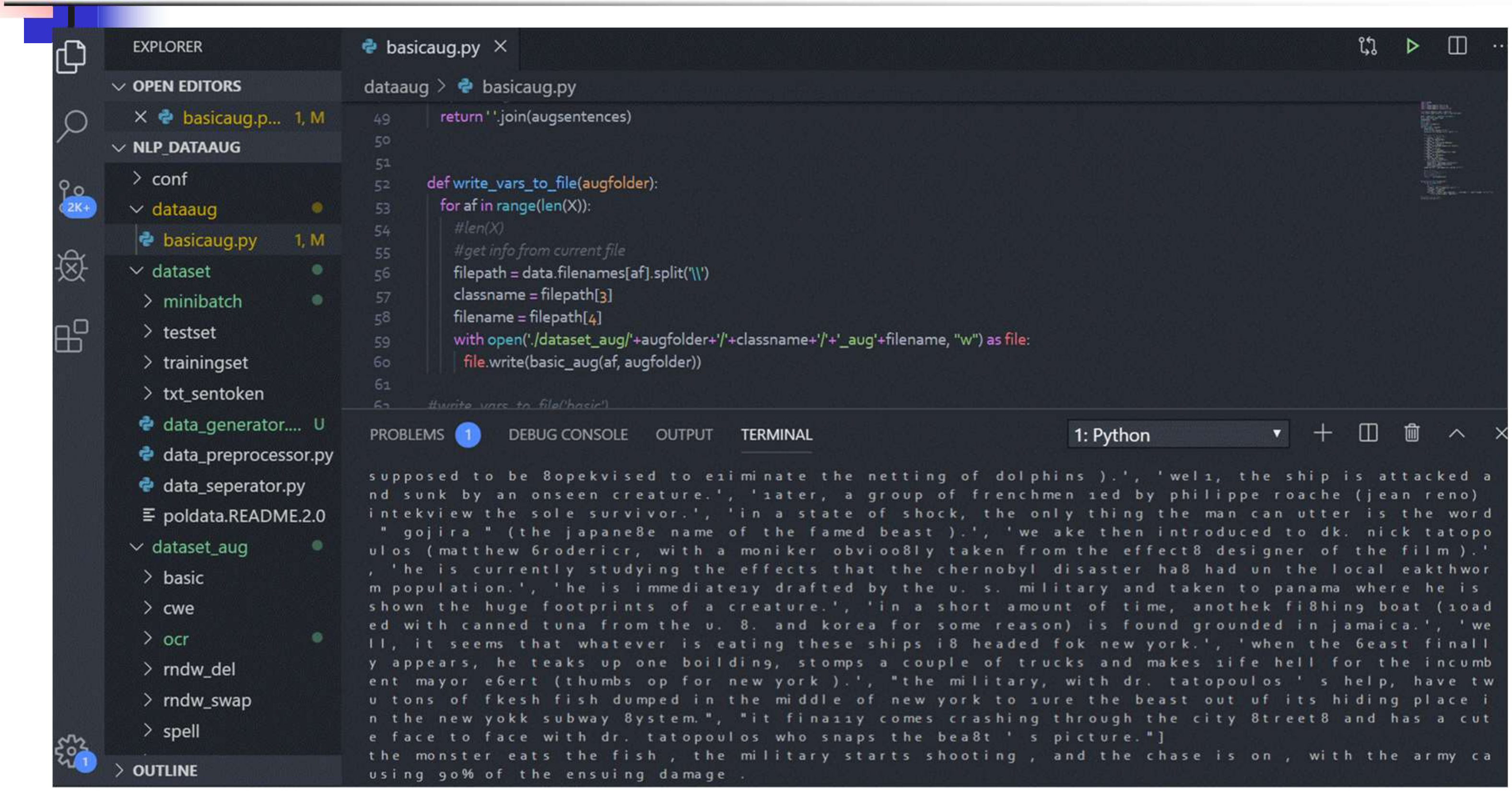
Demo: Data augmentation in nlpaug

```
import nlpaug
import nlpaug.augmenter.word as naw
```

```
def write_vars_to_file(type):
            for i in range(len(textdata)):
                        …
                        with open(path, “w”) as file

            file.write(doaugment(file, type))
```

```
def doaugment(file, augtype):
...
if(augtype==’spell’):
            aug = naw.SpellingAug()

for r in range(len(sentences)):
            augsentence =
      aug.augment(sentences[x])
              augsentences.append(augsentence
…

return ' '.join(augsentences)
```

Demo: performing Data Augmentation



```python
    return ' '.join(augsentences)


def write_vars_to_file(augfolder):
    for af in range(len(X)):
        #len(X)
        #get info from current file
        filepath = data.filenames[af].split('\\')
        classname = filepath[3]
        filename = filepath[4]
        with open('./dataset_aug/'+augfolder+'/'+classname+'/'+'_aug'+filename, "w") as file:
            file.write(basic_aug(af, augfolder))

#write_vars_to_file('basic')
```

```
supposed to be 8opekvised to e1iminate the netting of dolphins ).', 'wel1, the ship is attacked a
nd sunk by an onseen creature.', '1ater, a group of frenchmen 1ed by philippe roache (jean reno)
intekview the sole survivor.', 'in a state of shock, the only thing the man can utter is the word
 " gojira " (the japane8e name of the famed beast ).', 'we ake then introduced to dk. nick tatopo
ulos (matthew 6rodericr, with a moniker obvioo8ly taken from the effect8 designer of the film ).'
, 'he is currently studying the effects that the chernobyl disaster ha8 had un the local eakthwor
m population.', 'he is immediate1y drafted by the u. s. military and taken to panama where he is
shown the huge footprints of a creature.', 'in a short amount of time, anothek fi8hing boat (1oad
ed with canned tuna from the u. 8. and korea for some reason) is found grounded in jamaica.', 'we
ll, it seems that whatever is eating these ships i8 headed fok new york.', 'when the 6east finall
y appears, he teaks up one boilding, stomps a couple of trucks and makes 1ife hell for the incumb
ent mayor e6ert (thumbs op for new york ).', "the military, with dr. tatopoulos ' s help, have tw
u tons of fkesh fish dumped in the middle of new york to 1ure the beast out uf its hiding place i
n the new yokk subway 8ystem.", "it fina11y comes crashing through the city 8treet8 and has a cut
e face to face with dr. tatopoulos who snaps the bea8t ' s picture."]
the monster eats the fish , the military starts shooting , and the chase is on , with the army ca
using 90% of the ensuing damage .
```

Demo: Result of Augmentation with synonym replacement

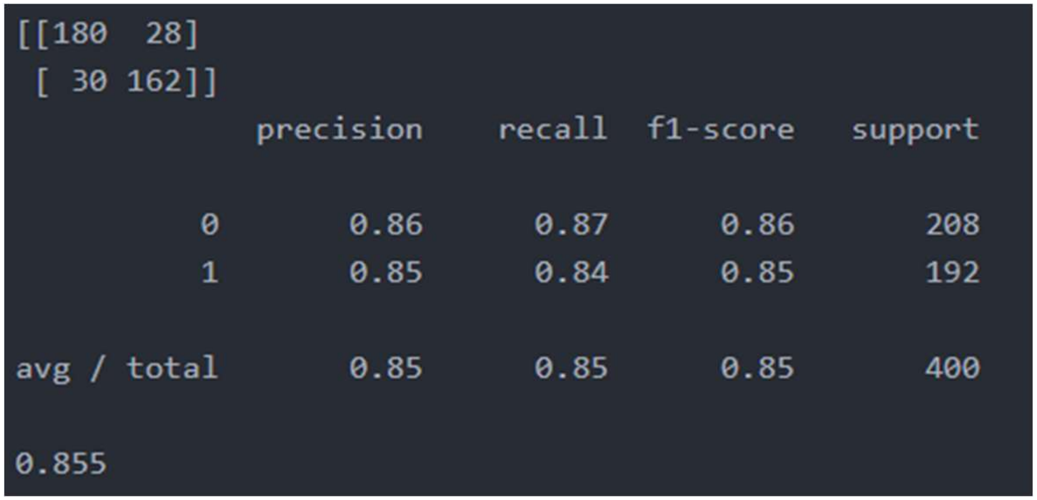
original review

a sci fi/comedy starring jack nicholson , pierce brosnan , annette benning , glenn close , martin short and other stars.
a warner bros picture
the martians have landed in this hillarous tim burton movie. before entering the cinema , i was initially a little bit nervous about what this film would be like .
many people were saying that this film was silly rubbish , and there was no point to it all .
how wrong they were .
i left this film feeling much happier than i was before i entered the cinema .
the story is about martians attacking earth .
using ray guns (hooray !)
they generally cause havoc around the u . s and other countries.

augmented review

a sci fi / funniness star knave nicholson, president pierce brosnan, annette benning, john herschel glenn jr. close, dino paul crocetti short and other stars.
a charles dudley warner bros picture
the martians hold landed in this hillarous tim burton motion picture. before entering the movie theater, i was ab initio a little bit nervous about what this plastic film would comprise similar.
many people were saying that this film be silly rubbish, and at that place was no point to it all. how wrong they were.
i left this motion picture show feeling much happier than i was before single entered the cinema.
the chronicle is about martians attacking worldly concern.
using shaft of light gun (hooray!)
they generally cause havoc around the u. due south and other state.

Demo



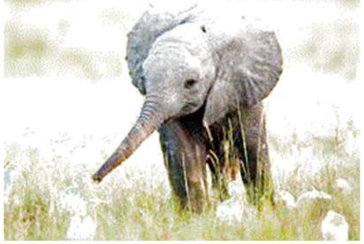
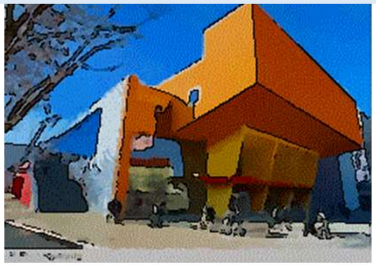
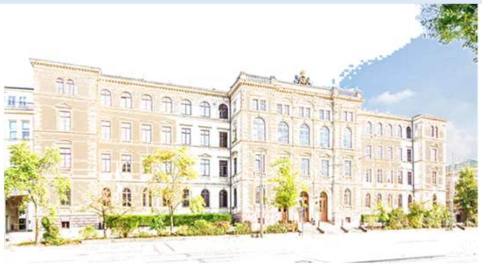
```
[[180  28]
 [ 30 162]]
             precision    recall  f1-score   support

          0       0.86      0.87      0.86       208
          1       0.85      0.84      0.85       192

avg / total       0.85      0.85      0.85       400

0.855
```

accuracy of the predictions of the text classifier without augmentation

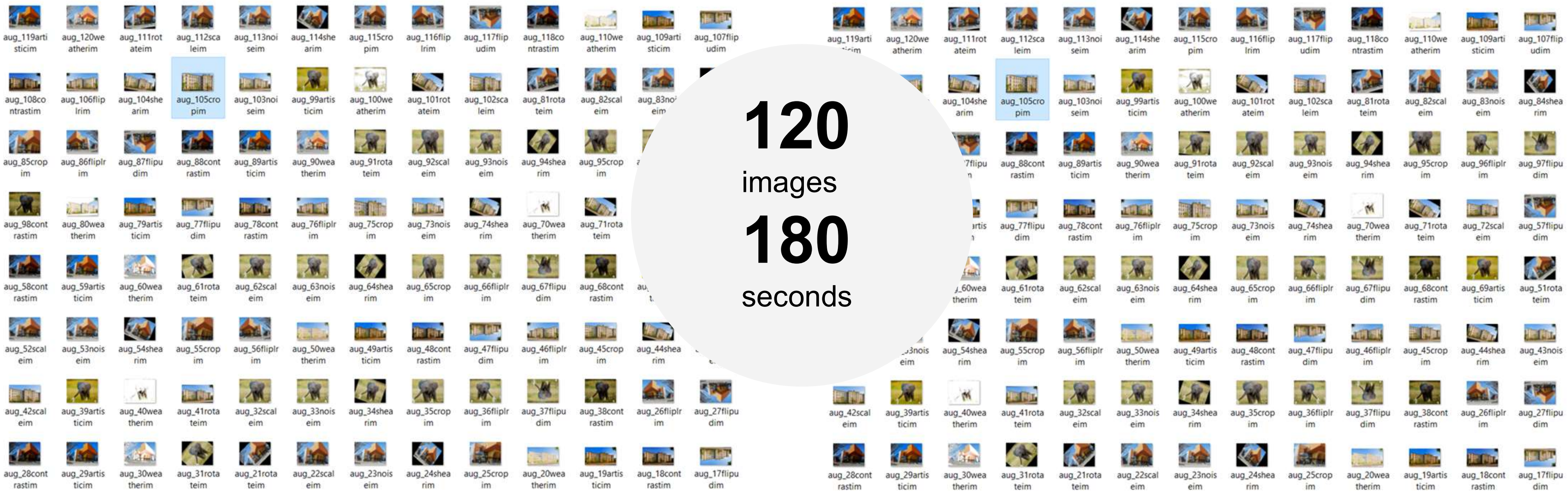
- ✓ performed text classification with an 85% accuracy
- ✓ augmented over 2000 text files
- ✗ improved the model with augmentation

Demo: BONUS - image augmentation with imgaug

Augmentation by: cropping, scaling, artistic filters, weather, blur, rotation, flip...

Demo: BONUS - image augmentation with imgaug

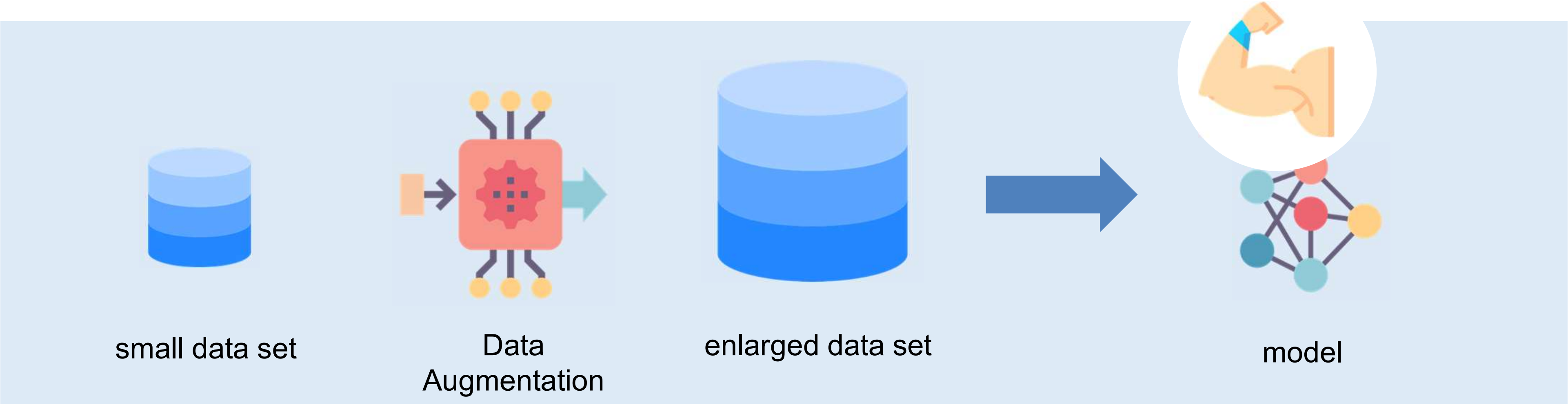


120
images
180
seconds

Data augmentation: pros and cons

+	-
easy implementation, low effort	time consuming (sometimes)
tailored approach, no model changes	trial and error
performance boost: easy to measure	relational gaps between data and augmented data
can help with overfitting	can lead to overfitting if not done right

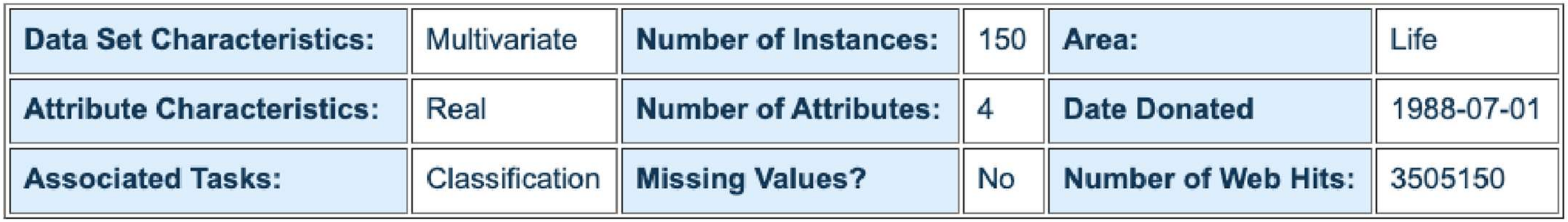
Data augmentation: key takeaway



An example

Iris dataset

Iris dataset



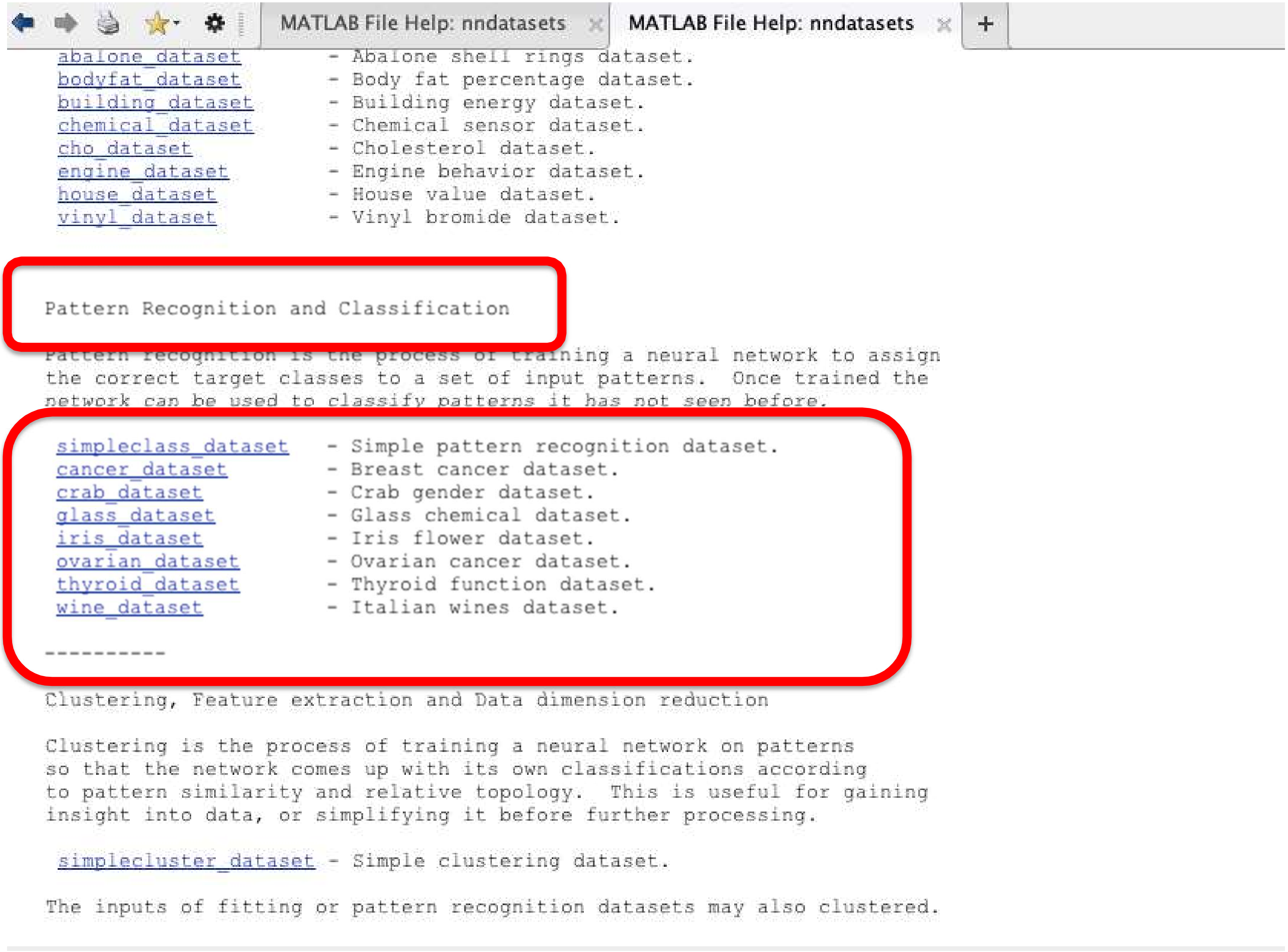
Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	3505150

- Perhaps the best known database
- The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
- Inputs: 1- sepal length in cm, 2- sepal width in cm, 3- petal length in cm, 4- petal width in cm
- Outputs: Iris Setosa, Iris Versicolour, Iris Virginica

Bài tập

- Bài tập theo nhóm đã phân công

- **1:** cơ sở dữ liệu hoa diên vĩ
- **2:** cơ sở dữ liệu cua
- **3:** cơ sở dữ liệu kính
- **4:** cơ sở dữ liệu rượu vang Ý
- **5:** cơ sở dữ liệu giá nhà đất
- **6:** cơ sở dữ liệu cholesterol



```
  abalone_dataset      - Abalone shell rings dataset.
  bodyfat_dataset      - Body fat percentage dataset.
  building_dataset     - Building energy dataset.
  chemical_dataset     - Chemical sensor dataset.
  cho_dataset          - Cholesterol dataset.
  engine_dataset       - Engine behavior dataset.
  house_dataset        - House value dataset.
  vinyl_dataset        - Vinyl bromide dataset.


 Pattern Recognition and Classification

 Pattern recognition is the process of training a neural network to assign
 the correct target classes to a set of input patterns.  Once trained the
 network can be used to classify patterns it has not seen before.

  simpleclass_dataset  - Simple pattern recognition dataset.
  cancer_dataset       - Breast cancer dataset.
  crab_dataset         - Crab gender dataset.
  glass_dataset        - Glass chemical dataset.
  iris_dataset         - Iris flower dataset.
  ovarian_dataset      - Ovarian cancer dataset.
  thyroid_dataset      - Thyroid function dataset.
  wine_dataset         - Italian wines dataset.

 ----------

 Clustering, Feature extraction and Data dimension reduction

 Clustering is the process of training a neural network on patterns
 so that the network comes up with its own classifications according
 to pattern similarity and relative topology.  This is useful for gaining
 insight into data, or simplifying it before further processing.

  simplecluster_dataset - Simple clustering dataset.

 The inputs of fitting or pattern recognition datasets may also clustered.
```