Deep Learning from Scratch

Session #2: Working with Data

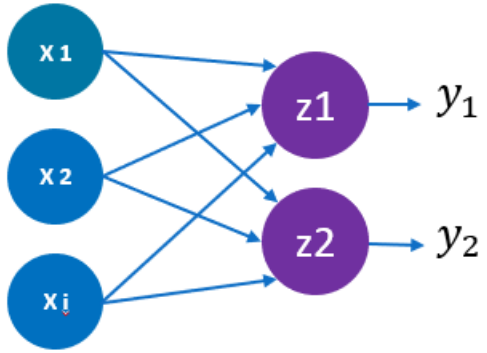
by: Ali Tourani – Summer 2021

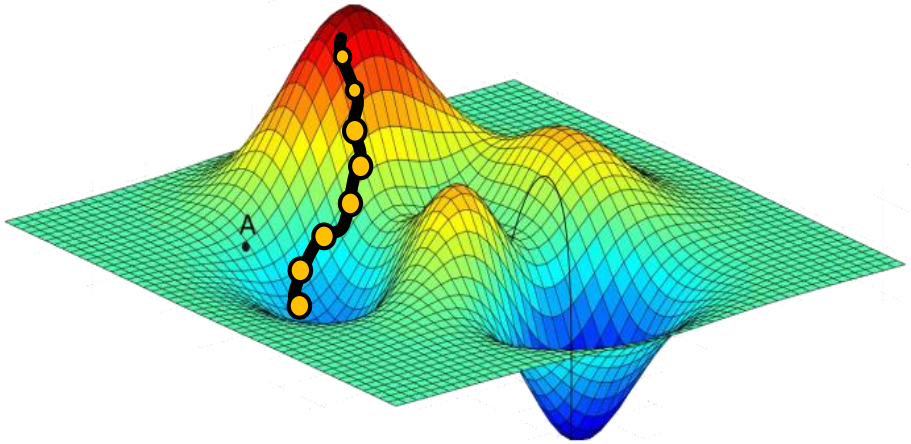
Agenda

- Warm-up and Review
- Importance of Data
- Datasets
- Where to Find Data
- Deep Learning and Data
- Assignment and Homework
- References

Warm-up and Review

- A brief introduction to Deep Learning
- Artificial Neural Networks (ANNs)
  - Bias
  - Perceptron
  - Activation Functions
  - Forward & Backward pass
- Loss Functions
  - Loss Optimization
  - Gradient Descent Algorithm





Importance of Data

- Main requirements of Deep Learning
  - **Big Data**
  - Powerful Hardware
  - Efficient Software

We are here!

Why is data so important?

- ✓ Data helps to make informed decisions based on **reality**, not guess
- ✓ With data, we can tune the DNNs parameters and create **generalizable models**
- ✓ Without proper data, we cannot **expect** the model to work correctly

Image source: https://www.informatica.com

Importance of Data

The analogy to Deep Learning is that the rocket engine is the deep learning model, and the fuel is the huge amounts of data we can feed to these algorithms.

Andrew Ng
Computer Scientist and ML pioneer

Image source: https://www.seamgen.com

Importance of Data

Types of Data

- Deep Learning can be applied to solve various problems
  - Thus, the type of data we use depends on the problem!
- General classification of Data
  - **Image** (Computer Vision)
  - **Video** (Motion Detection/Tracking)
  - **Sound** (Voice Recognition)
  - **Text** (Reviews Classification)
  - **Time Series** (Sensor Data)

Image source: https://www.newbanking.com

Datasets

- Collections of data, generally of the same type
- Generally contain hundreds or thousands of samples
  - They can be enormous (e.g., 20GB)
- Generated by companies, organizations, universities, etc.
- Characteristics of the dataset we choose:
  - Must be relevant to the problem
  - Must have adequately labeled data and accurate classification
  - Must be easy to access
  - Should have proper formatting of data (no need to pre-process)

In many cases, datasets are not released to the general public

Image source: https://www.pngkey.com

Datasets

Sample: MS-COCO Dataset

Type of data: Image
Size: 25GB
# of Instances: 1,500,000
# of classes: 80
More info:
- Perfect for object detection
- Contains captions

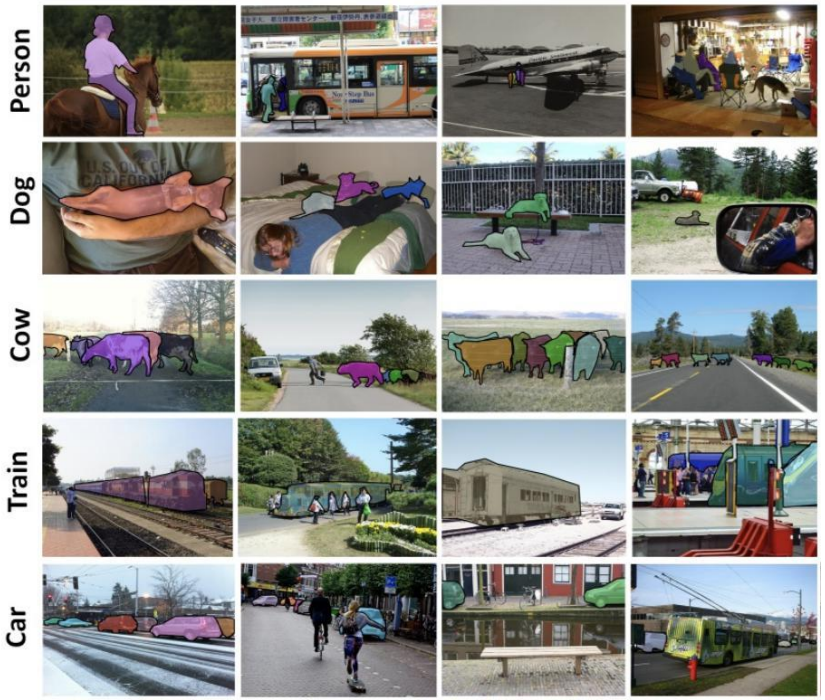


Image source: https://www.cocodataset.org

Datasets

Sample: ImageNet Dataset

Type of data: Image
Size: 150GB
# of Instances: 1,500,000
# of classes: 1000
More info:
- Perfect for object detection
- Contains bounding boxes

Image source: https://www.paperswithcode.com

Datasets

Sample: Open Images Dataset

Type of data: Image
Size: 500GB
# of Instances: +9,000,000
# of classes: -
More info:
- Perfect for object detection
- Object detection, image classification, and visual relationship detection

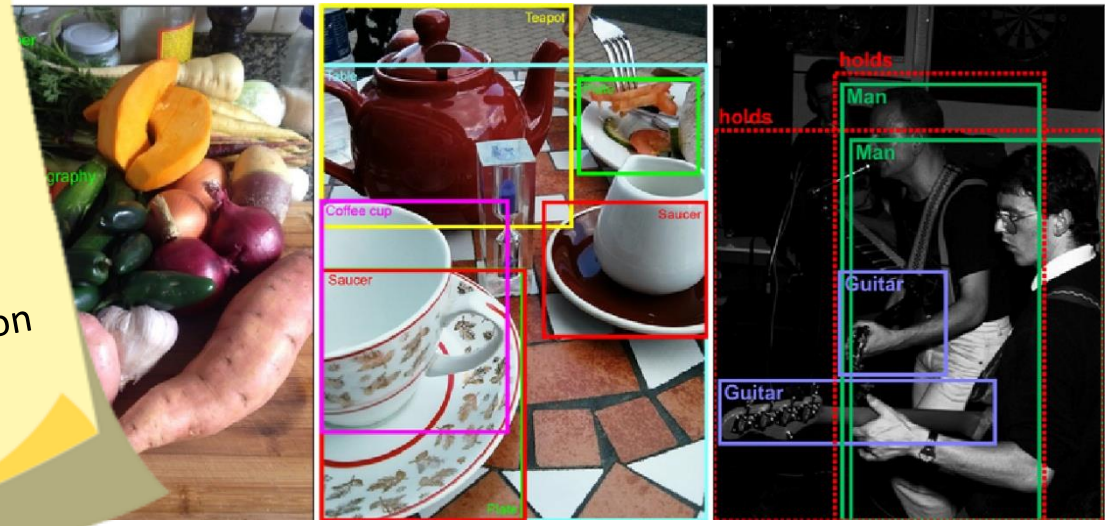


Image source: https://www.semanticscholar.org

Datasets

Sample: CIFAR-10 Dataset

Type of data: Image
Size: 170MB
# of Instances: 160,000
# of classes: 10
More info:
Contains various objects

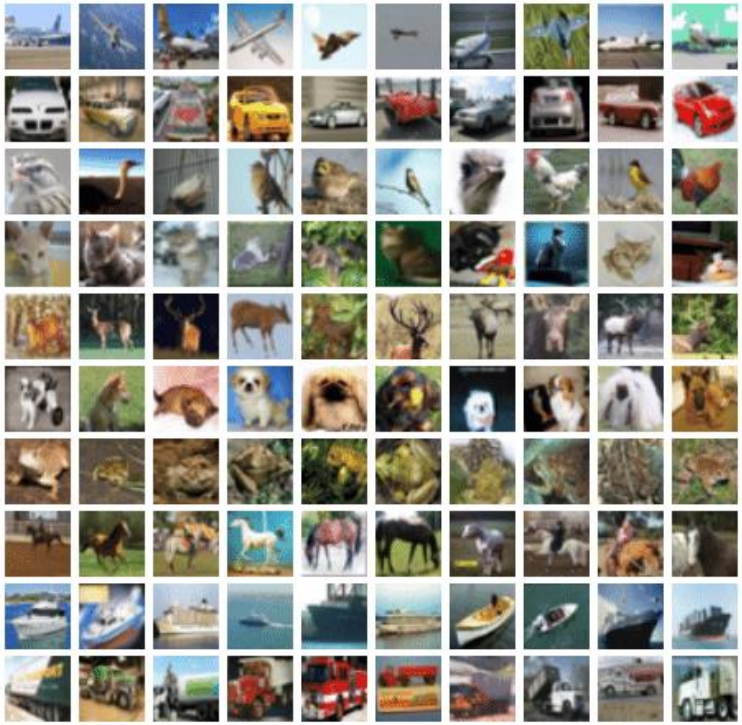


Image source: https://www.researchgate.net

Datasets

Sample: MNIST Dataset

Type of data: Image
Size: 50MB
# of Instances: 70,000
# of classes: 10
More info:
Contains handwritten digits

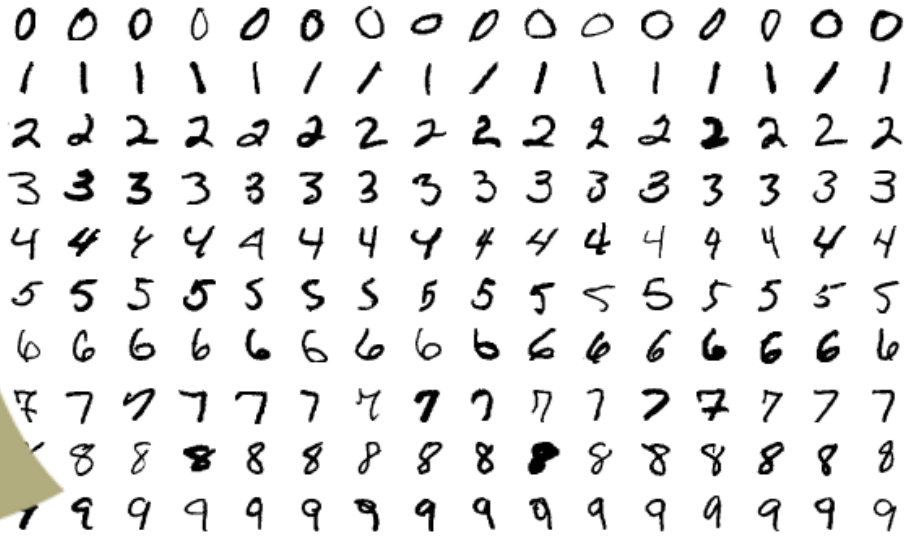


Image source: https://www.wikipedia.org

Datasets

Sample: Youtube-8M Dataset

Type of data: Video
Size: 50GB
# of Instances: 6,100,000
# of classes: 3862
Total length: 350,000 h
More info:
Contains short video clips with high quality annotations

Image source: https://www.research.google.com

Datasets

Sample: Free Music Archive (FMA) Dataset

Type of data: Sound
Size: 1TB
# of Instances: 100,000
More info:
- Proper for music analysis
- Contains tracks metadata and genres

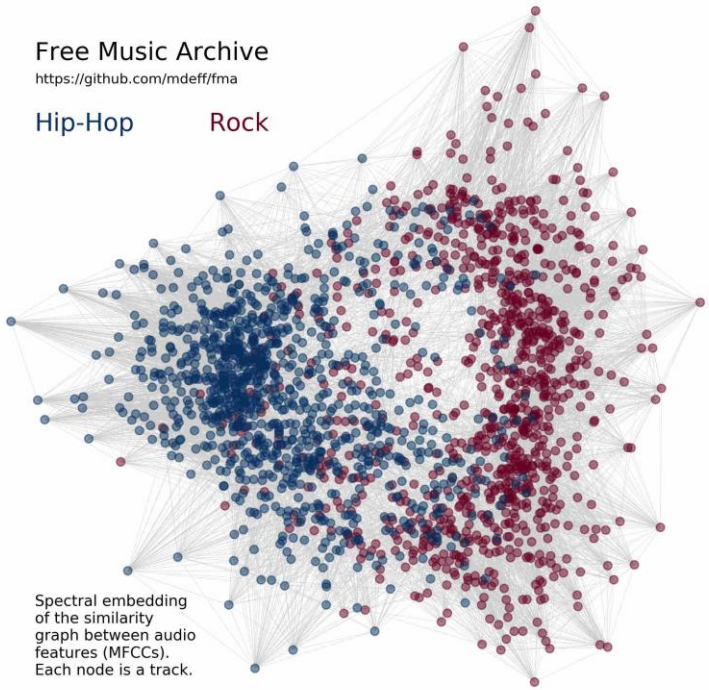


Image source: https://www.paperswithcode.come

Datasets

Sample: Ballroom Dance Dataset

Type of data: Sound
Size: 14GB
# of Instances: 698
Duration of each file: ~30 s
More info:
- Contains ballroom dancing audio

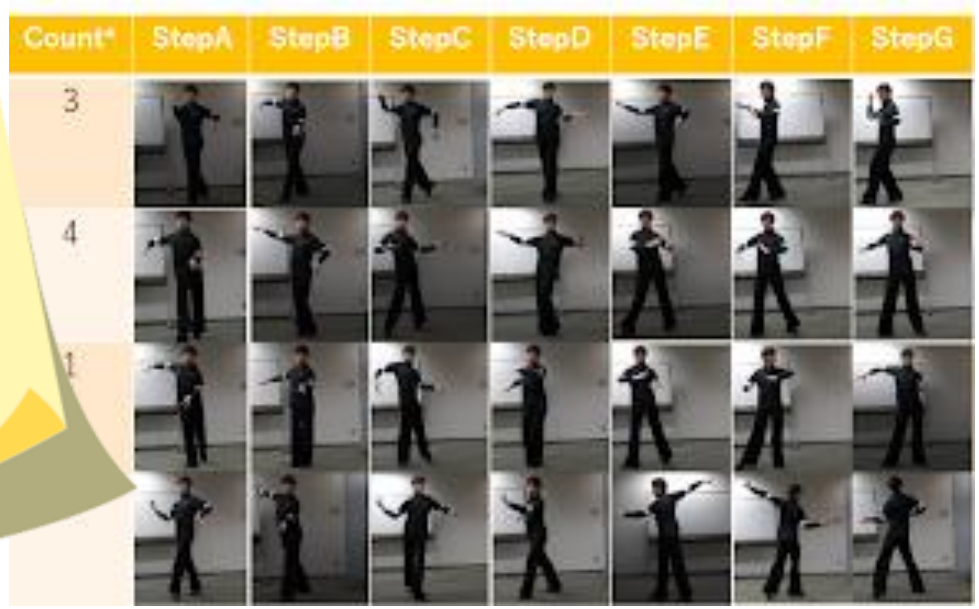


Image source: https://www.hasc.jp

Datasets

Sample: Sentiment140 Dataset

Type of data: Text
Size: 80MB
# of Instances: 1,600,000
More info:
- Text classification for sentiment analysis
- Twitter data

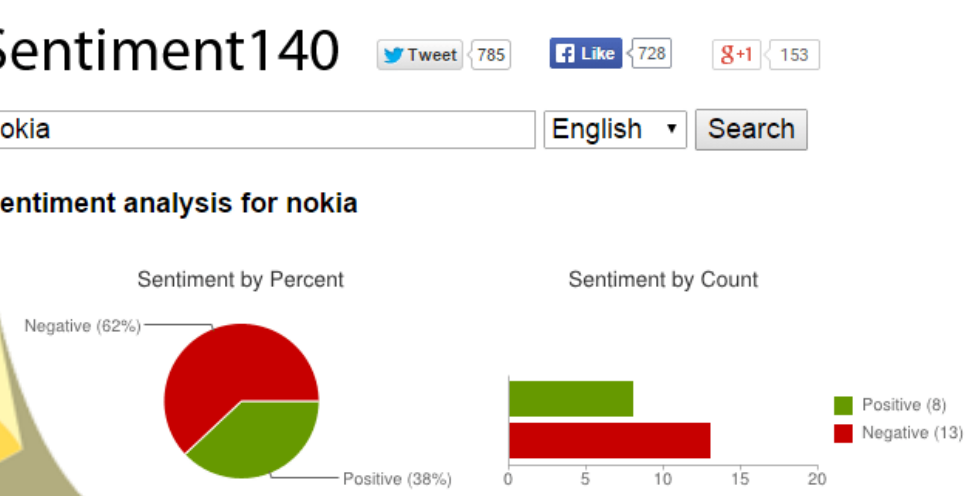


Image source: https://www.researchgate.net

Datasets

Sample: Large Movie Review Dataset

Type of data: Text
Size: 80MB
# of Instances: 50,000
More info:
- Binary Sentiment Classification
- Polar movie reviews from IMDB

Image source: https://www.imdb.com

Datasets

Sample: Minimum Daily Temperatures Dataset

Type of data: Time Series
Size: ~70KB
# of Instances: 3650
More info:
- Contains temperature data over 10 years in Melbourne

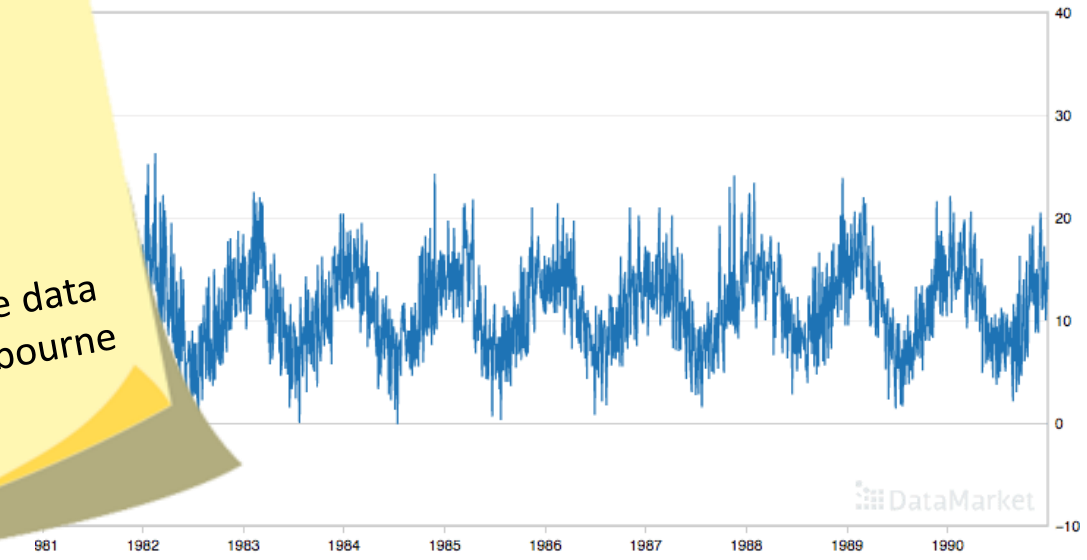


Image source: https://www.machinelearningmastery.com

Datasets

Standard Data

Great for academic works

Clean data classified into balanced categories

Collected in standard and normal conditions

Generally can help us in the pre-training (*) stage

May trick the DNN in some scenarios!

Real-world Data

Messy, varied, and evolving

Collected from various sources in standard and challenging conditions

Perfect for practical usage and DL solutions

Needs pre-processing (*)

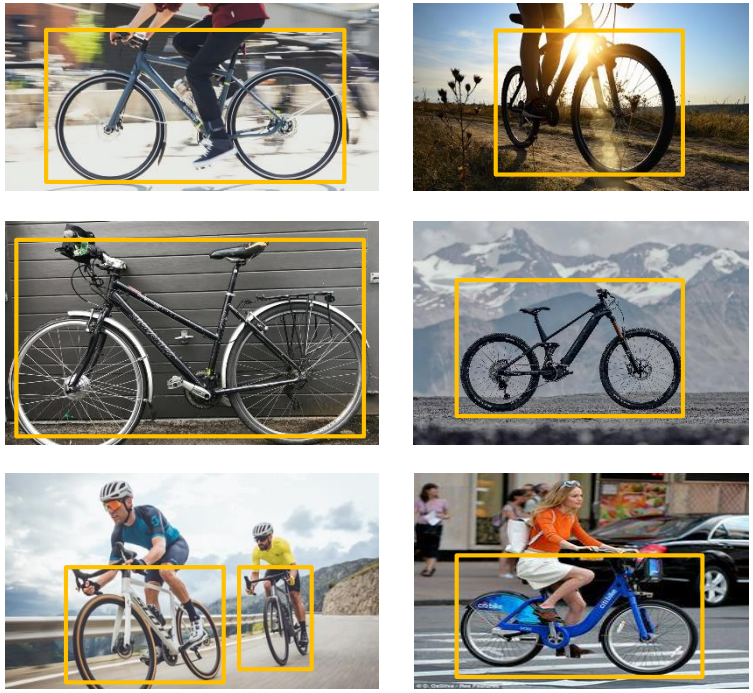
Needs labeling/annotation (in supervised approaches)

*** We'll talk about them soon!**

Datasets

Standard Data

Real-world Data



Datasets

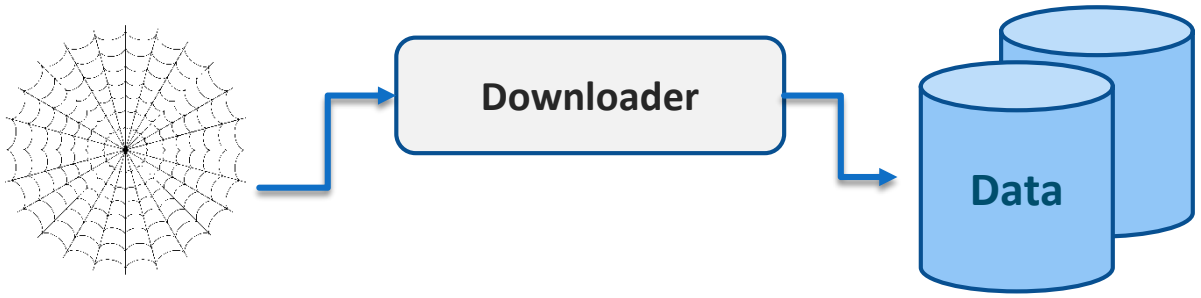
Standard Data

Real-world Data

Datasets

Generate Data

- Collecting data might be needed in some cases
  - Using various sensors like cameras, sound recorders, thermometers, etc.
- Data crawling
  - Helpful when there are no public/accessible datasets
  - Automatic collection of data from websites using web crawlers



Datasets

Generate Data

- *Q1: But, why do we need real-world data?!*
  - To make **practical applications**
  - To find out how the model can **perform better**
  - To make sure that the model cannot be **tricked**!
  - To understand how the model can make **wrong predictions**

- *Q2: How much data do we need?*
  - The minimum requirement depends on the problem

Image source: https://www.pngimg.com

Datasets

Generate Datasets

- *Q3: What about rear/scarce instances?*
  - We need to balance the classes
  - **Approach#1: Collect more data**
    - Sometimes not possible due to cost, accessibility, time, etc.
  - **Approach #2: Data augmentation**
    - Increase the amount of data by adding modified copies of existing samples
    - Sample techniques: flip, rotation, crop, scale
    - Check this repository for a simple try

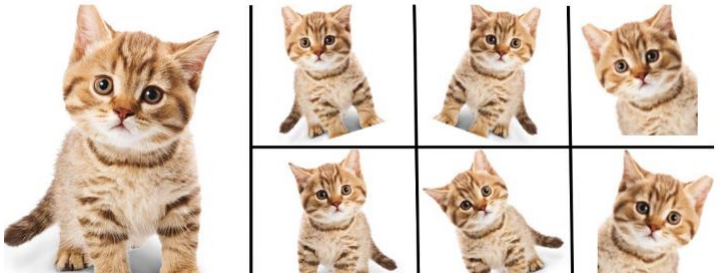


Image source: https://www.nanonets.com

Datasets

Generate Datasets

- *Q4: How to prepare a generated dataset for training?*
  - Clean the data as much as possible
    - Remove outliers, missing values, low-quality samples, duplications, etc.
  - Collected data should be cleansed, standardized and de-identified
    - Prevent any personal identity from being revealed
    - Always consider privacy issues!
  - Separate data into proper classes
  - Add labels (manually/automatically)

Datasets

Data Annotation/Labeling

- Categorizing and labeling of data for AI applications
- Providing naming conventions that describe data
- It can be done manually or automatically
- Essential factors: specificity + accuracy
- The output of this stage:
  - JSON files, CSV files, files containing marked images or videos, sound
- List of tools for annotating data:
  - https://github.com/taivop/awesome-data-annotation

Datasets

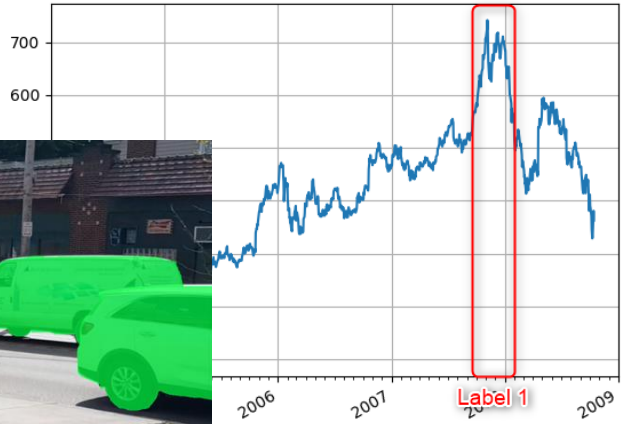
Data Annotation/Labeling

Type of Data	Annotation Sample
Image	Drawing bounding boxes, polygons, 2D & 3D points, etc.
Video	Adding tracking data, etc.
Sound	Transcriptions of languages, intonations, non-speech sounds, etc.
Text	Tag important keywords, titles, headlines, etc.
Time Series	Time labeling, timestamps, etc.

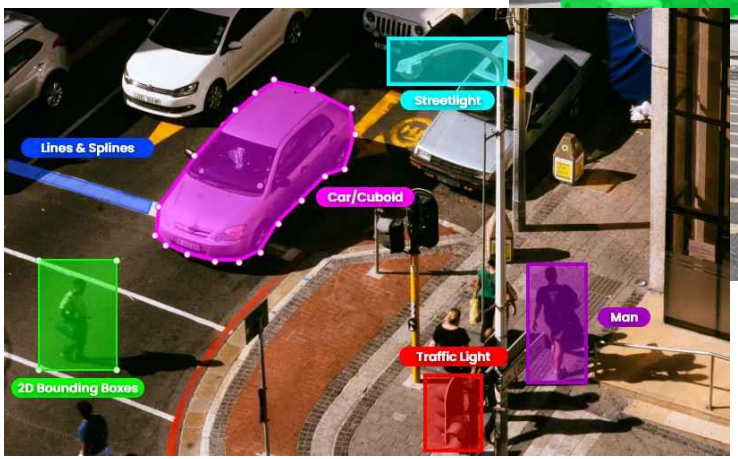
Datasets

Data Annotation/Labeling

Time Series



Image



Video

Image source: https://www.amnetdigital.com, www.innotescus.io, and www.stackexchange.com

Datasets

Data Annotation Tools

I. LabelBox link

- ✓ A simple and powerful tool for images
- ✓ It can be used for classification, object detection, and image segmentation
- ✓ Simply draw bounding boxes using Pen Tool

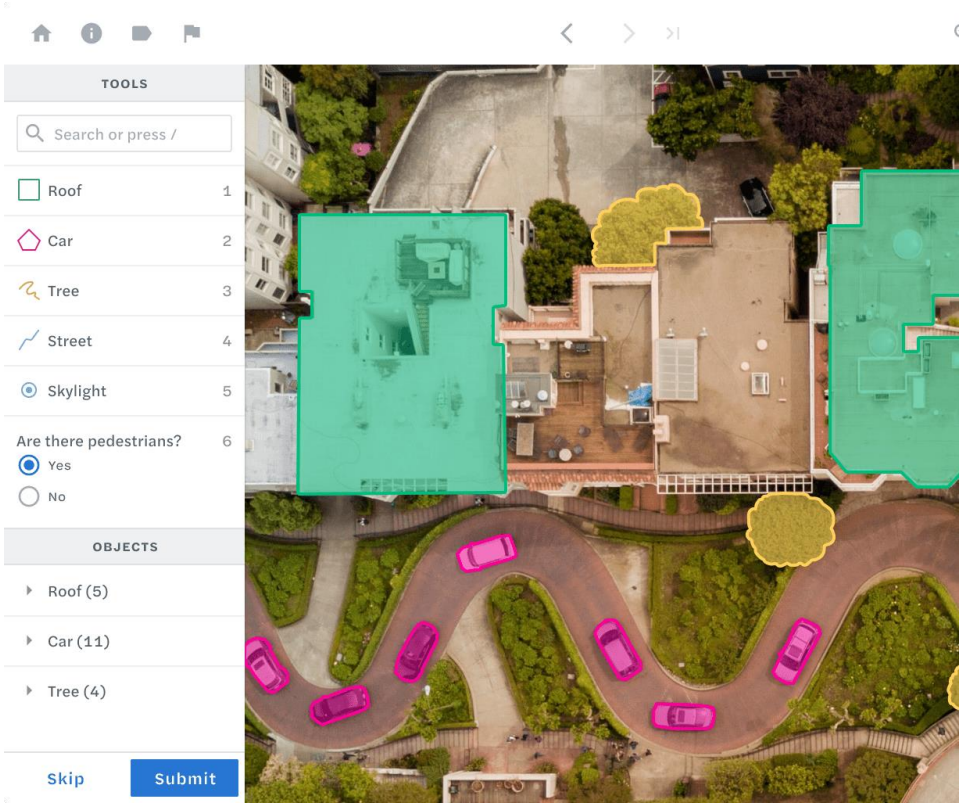


Image source: https://www. labelbox.com

Datasets

Data Annotation Tools

II. MakeSense link

- ✓ Free, open-source, and web-based
- ✓ Supports multiple image annotations, including points, bounding boxes, and polygons
- ✓ Different formats including CSV, JSON, XML, YOLO, etc.

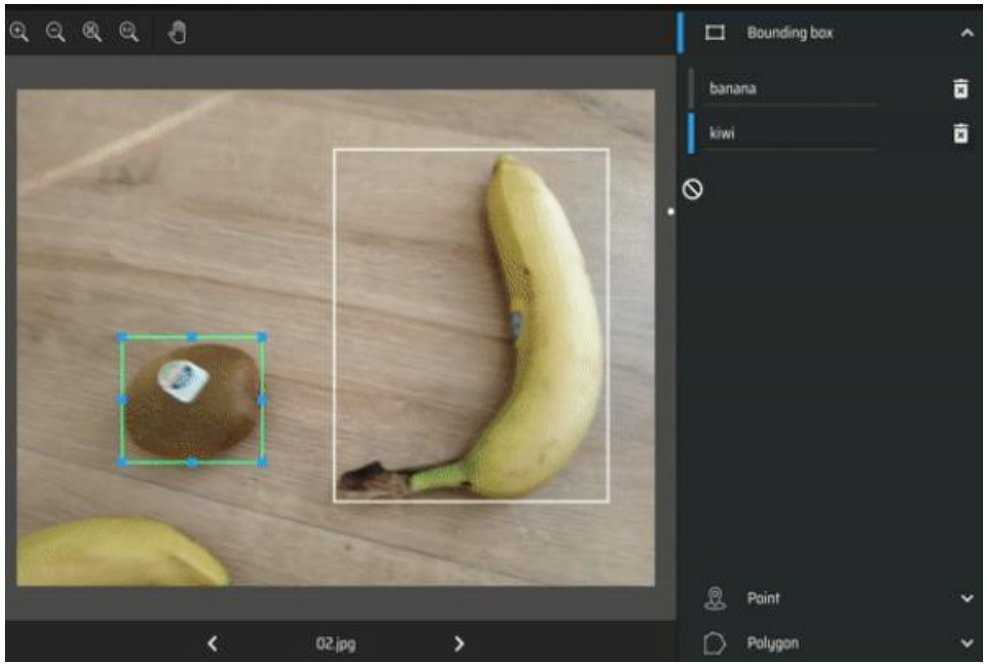


Image source: https://github.com/SkalskiP/make-sense

Datasets

Data Annotation Tools

III. DLabel link

- ✓ A free image and video labeling tool
- ✓ Provides a web-based annotation interface
- ✓ Great tool for adding track data labels (videos)
- ✓ High quality annotated data

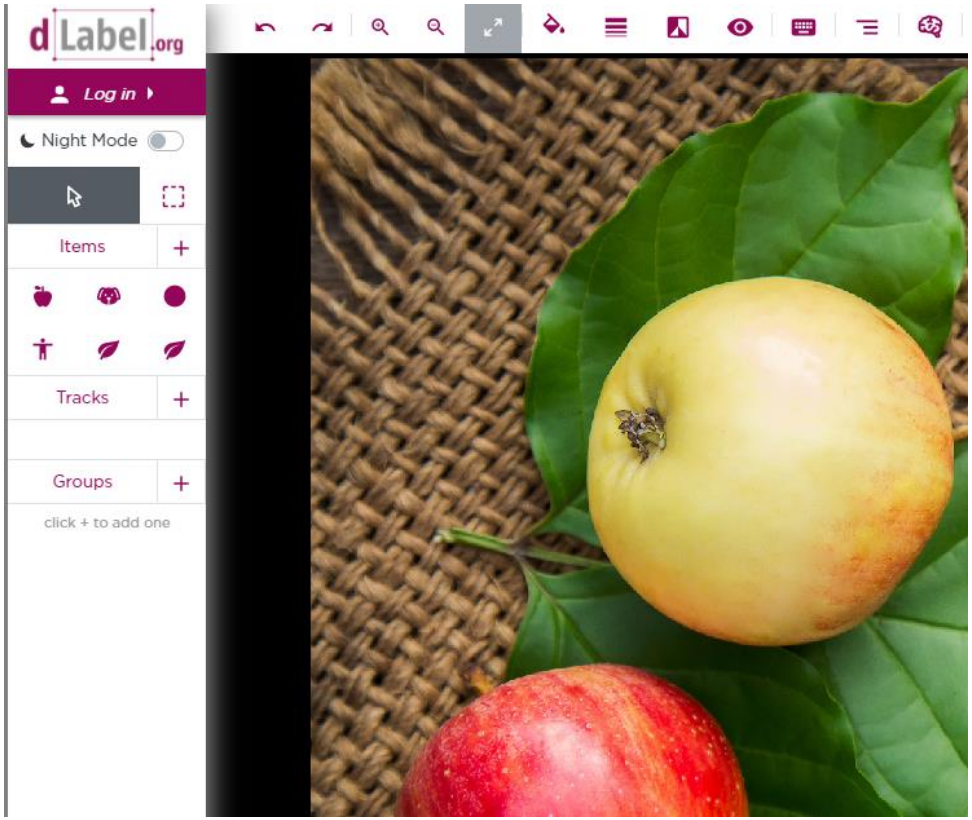


Image source: https://www.dlabel.org/app

Datasets

Data Annotation Tools

III. Doccano link

- ✓ An open-source text annotation tool
- ✓ Practical for sentiment analysis, named entity recognition, and also text summarization

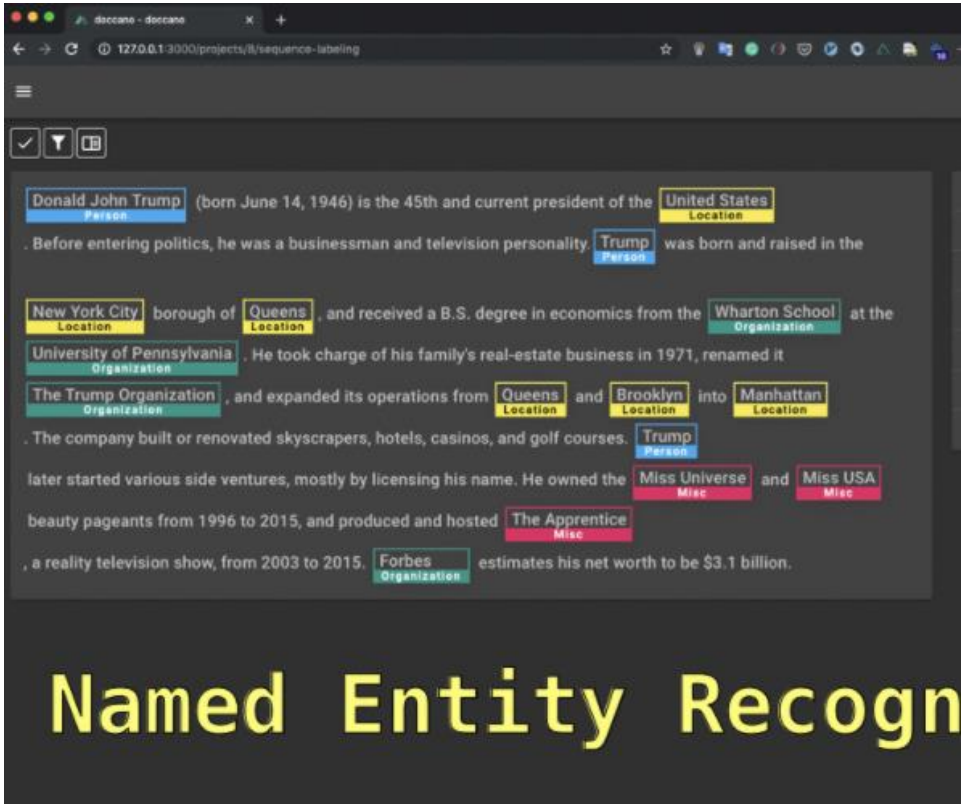


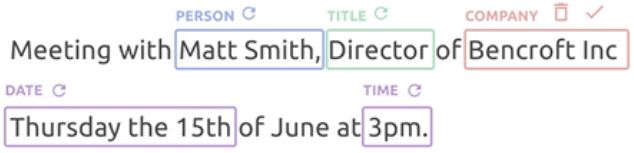
Image source: https://github.com/doccano/doccano

Datasets

Data Annotation Tools

III. LightTag link

- ✓ A text labeling tool for Natural Language Processing (NLP) applications
- ✓ Easy to use with lots of features



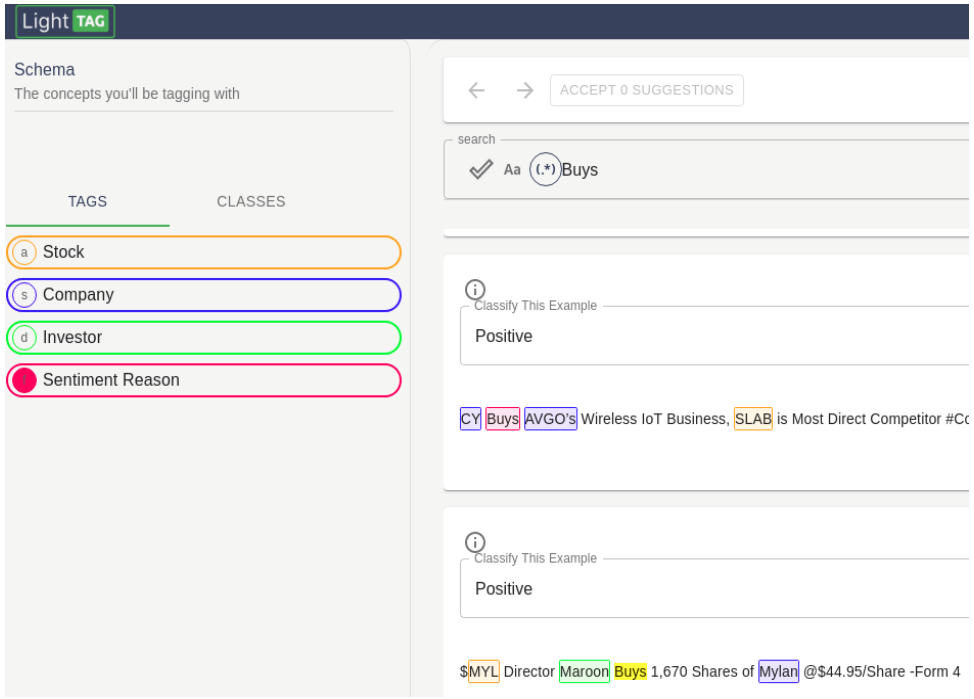
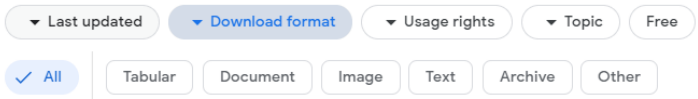


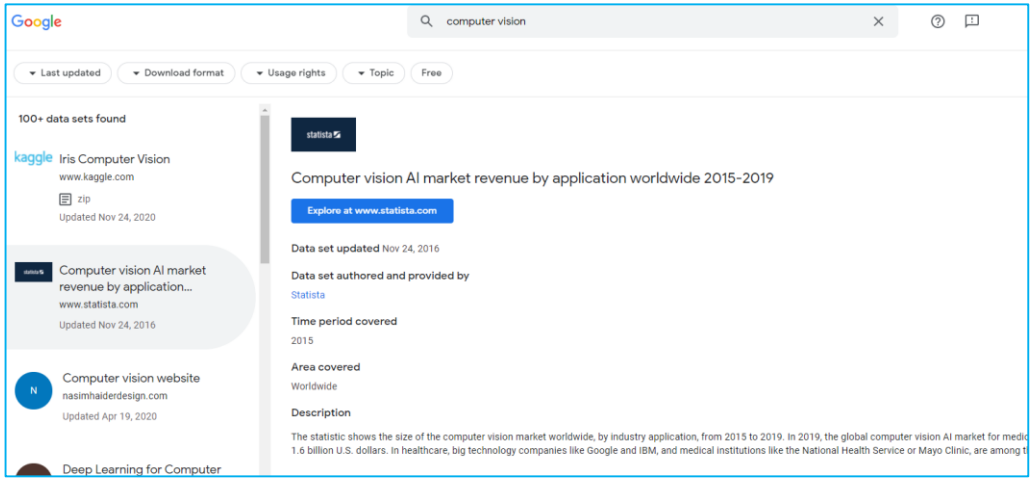
Image source: https://www.lighttag.io/

Where to Find Data

I. Google's Datasets Search Engine



- Easily find datasets by typing keywords
  - Filter results and add constraints to find appropriate data
- Easily publish your generated dataset

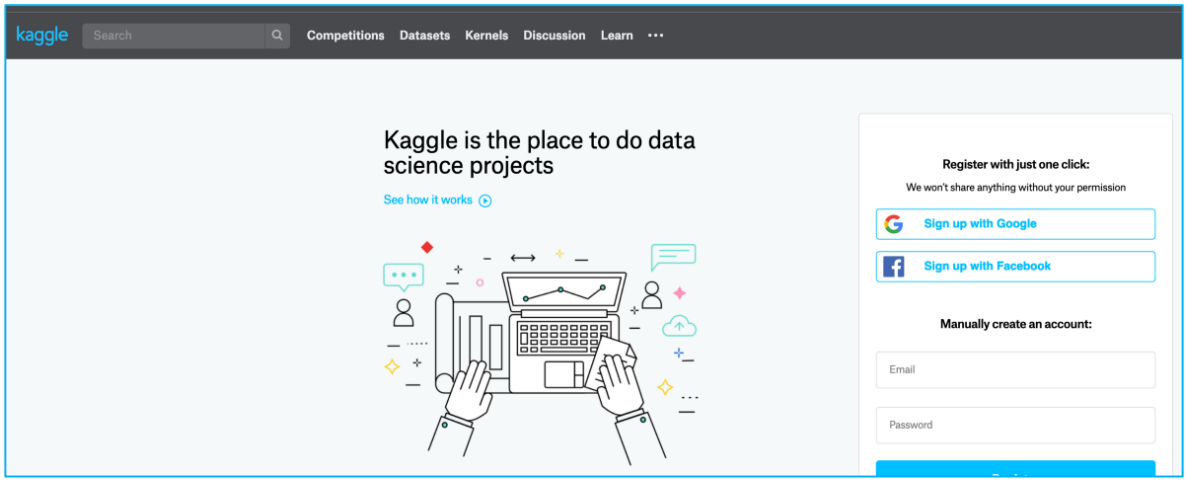


https://datasetsearch.research.google.com/

Where to Find Data

II. Kaggle

- One of the best places to find data
  - The leading platform for Data Science
- Lots of great features for data scientists

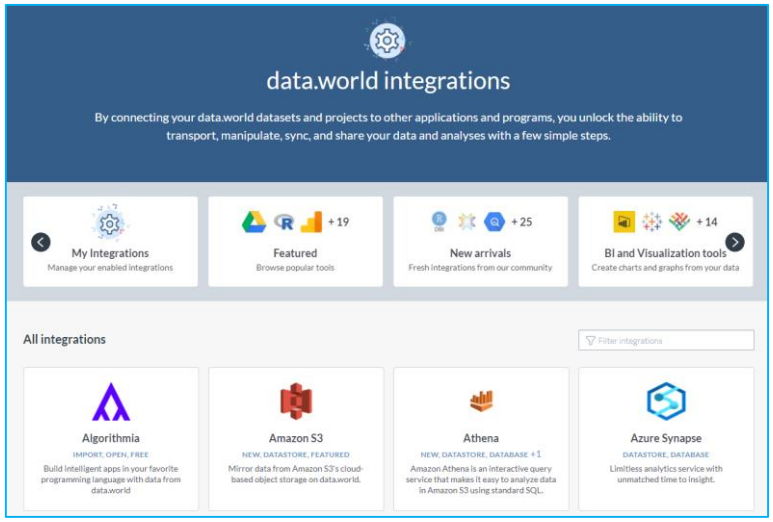


https://www.kaggle.com/datasets

Where to Find Data

III. Data World

- Another great service to get access to public datasets
- Excellent user interface, integrations, and tools



data.world

https://data.world/

Where to Find Data

Others

- UCI Machine Learning Repository
- Lionbridge AI Datasets
- CERN Open Data Portal
- Amazon Open Data
- Microsoft Azure Public Datasets
- Awesome Public Datasets

Deep Learning and Data

Why do we need data in Deep Learning?

1. To train the Deep Neural Network (DNN) and fit the model
2. To validate the model and provide an unbiased evaluation of it
3. To test and evaluate the final model

```
# Split data
train, validation, test = dataset()

# Tune model
for all parameters:
    model = fit(train, parameters)
    skill = eval(model, validation)

# Evaluate
model = fit(train)
skill = eval(model, test)

```

Deep Learning and Data

Training-set

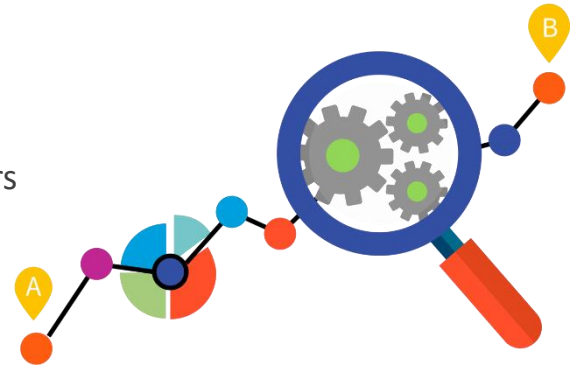
- A (subset of) dataset used to train the DNN (learning stage), AKA **training data**
- Contains all the observations that the model needs to learn
- The DNN adjusts its weights and biases using training data
- A pre-training stage might be needed in DNNs
- Training data should:
  - Follow a general probability distribution
  - Cover all variations of required data

Image source: https://www.theamlshop.ca

Deep Learning and Data

Validation-set

- AKA *Development-set* or *dev set*
- A sample of data **held back from training** for frequent evaluation
- Used to cross-check whether the model is correctly trained or not
  - We can re-check the training process and the algorithm in case of low accuracy
- Should follow the same probability distribution as the training set
- Commonly used to:
  - Decrease the probability of **Overfitting (*)**
  - Evaluate the error function and check the DNN parameters



*** We'll talk about them soon!**

Image source: https://www.pinclipart.com

Deep Learning and Data

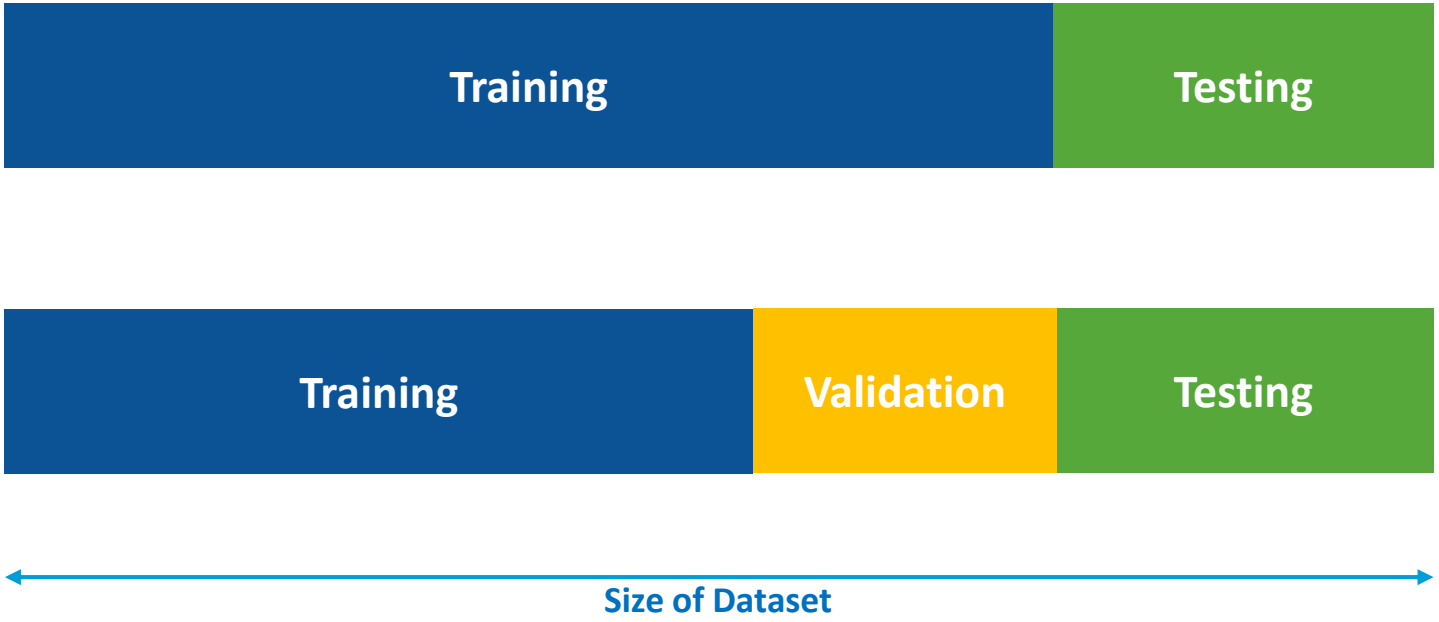
Test-set

- It contains **unseen data**
  - Data that has not been used in the training process
- Used to assess the performance of the DNN
- Independent of the training dataset
- Should follow the same probability distribution as the training-set
- **Holdout:** isolating a part of the training-set for test (evaluation) as the test-set and do not use them for training the network

Image source: https://www.flaticon.com

Deep Learning and Data

Let's put them all together!



Deep Learning and Data

Let's put them all together!

- Dataset split ratio
  - It considerably depends on the amount of data and our model
  - **Cross Validation?**
    1. Split the dataset into two (**train-set**, **test-set**)
    2. Randomly choose *X%* of the train-set as the **validation-set**
    - A popular approach: K-Fold Cross Validation

Deep Learning and Data

Let's put them all together!

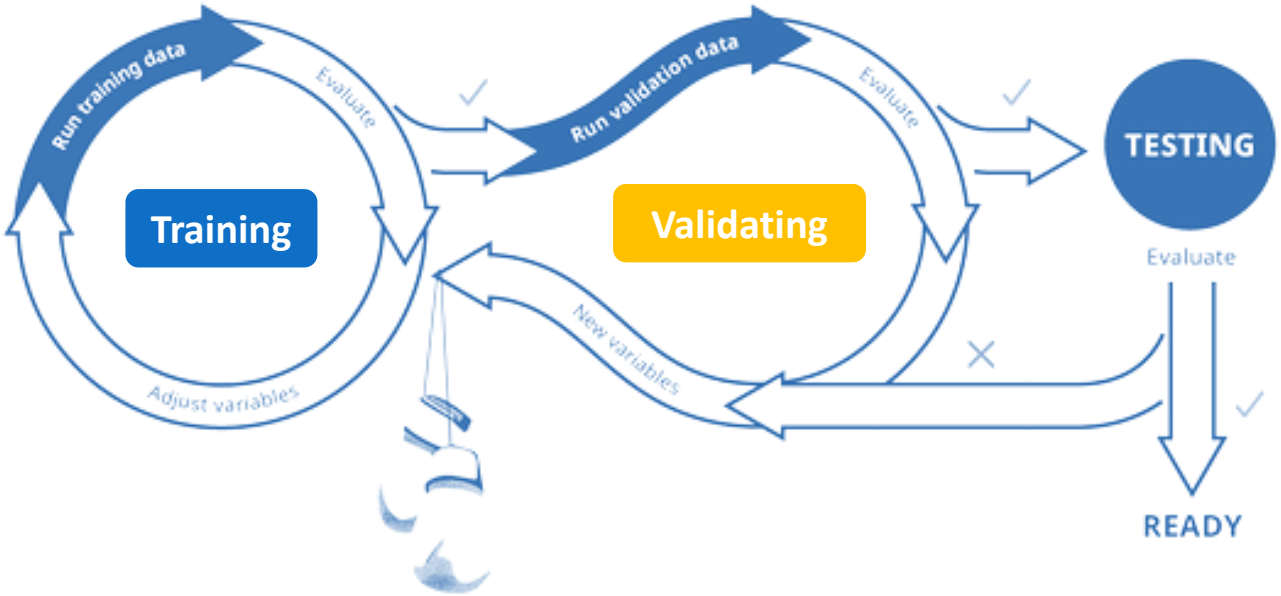


Image source: https://www.insidebigdata.com

Assignment & Homework

Find your Field of Interest

- Find an appropriate dataset for the subject that you have chosen
- Let me know about your choice
- Discuss your ideas for using the dataset

References

Web pages and Articles

- https://scrape.works/blog/go-large-data-deep-learning/
- https://wiki.pathmind.com/data-for-deep-learning
- https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/
- https://medium.com/merantix/applying-deep-learning-to-real-world-problems-ba2d86ac5837
- https://www.bmc.com/blogs/data-annotation/
- https://towardsdatascience.com/9-best-places-to-find-machine-learning-datasets-dfdba8af5220

Questions?