# Machine Learning 101

## kNN Speed Run

**Quentin Crain**

# Agenda

- What is ML?
- What is kNN?
- Let's Do It!

# Topics

- Machine Learning (ML): https://en.wikipedia.org/wiki/Machine_learning
    - https://developers.google.com/machine-learning/crash-course
    - https://www.coursera.org/articles/what-is-machine-learning
    - https://www.geeksforgeeks.org/machine-learning/ml-machine-learning/
- kNN: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
    - https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/
    - https://towardsdatascience.com/k-nearest-neighbor-classifier-explained-a-visual-guide-with-code-examples-for-beginners-a3d85cad00e1/
- `plotly`: https://plotly.com/
- Scatter plot: https://en.wikipedia.org/wiki/Scatter_plot

# What is Machine Learning?

"Learn" from data to
*infer/generalize/predict*
labels for new data.

- ***Supervised***: **Label** new data given a labeled **dataset** from a set of **features**.
  - Ex: Regression (linear), Classification (kNN, ...), Decision Trees, SVM
- ***Unsupervised***: **Label** new data based on similarity/difference to a unlabeled **dataset** from a set of **features**.
  - Ex: Clustering (k-means, PCA)
- ***Reinforcement***: Create a **policy** to achieve a **goal** through **rewards**.
  - Ex: NNs (CNN, RNN,), aka "AI"

https://ischool.syracuse.edu/what-is-machine-learning/

# What is kNN?

## It is a classification algorithm.

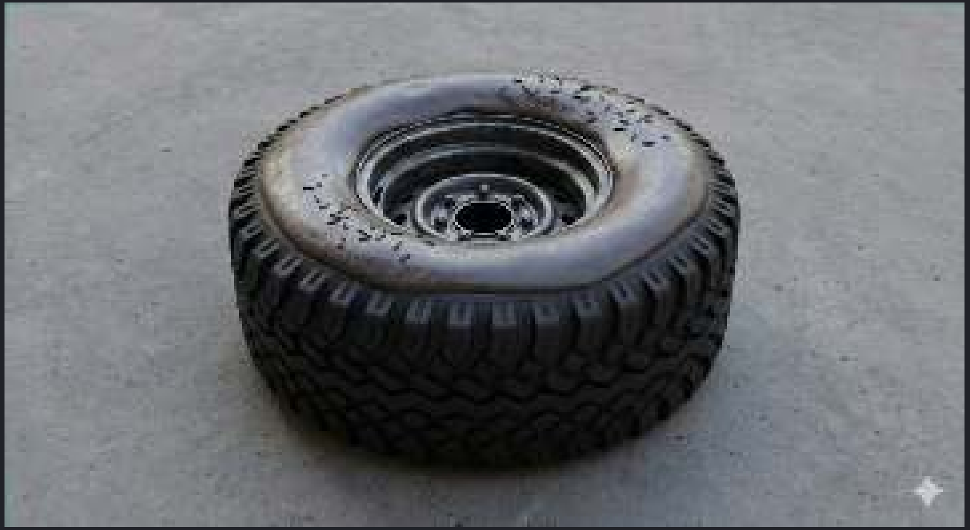
> “ A machine learning (ML) algorithm whereby data is "classified", ie: *labeled*, using existing labeled data. ”

# What is Classification?

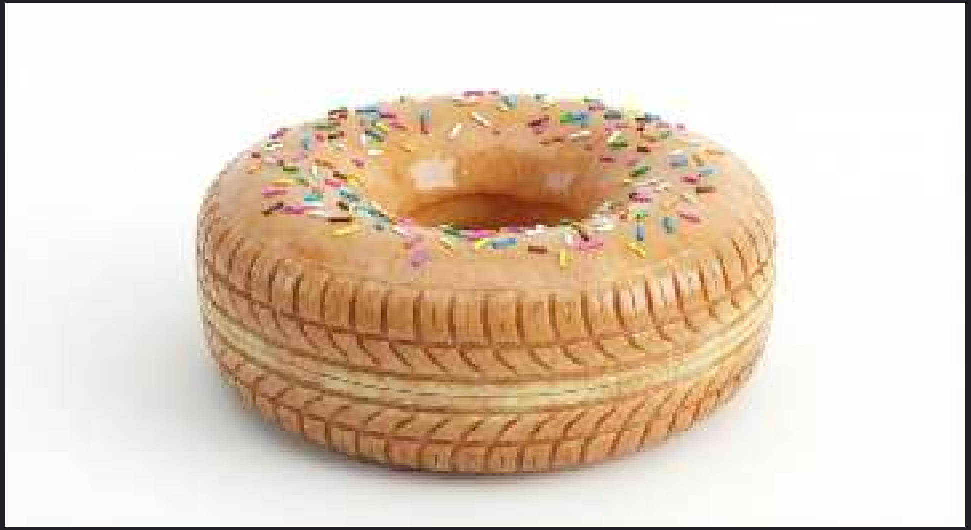
Dataset

# Classify This!

Ready?



tire



donut

# Classification Level: Extreme!

Ready?

# Let's Do It!

We are going to classify penguins!

# Define Your Question

## Lost Penguin!!!

I found this penguin, who are their pals?

00:00

# Penguin Dataset

The penguin dataset is the "Hello World" for kNN learning.

penguins_raw.csv

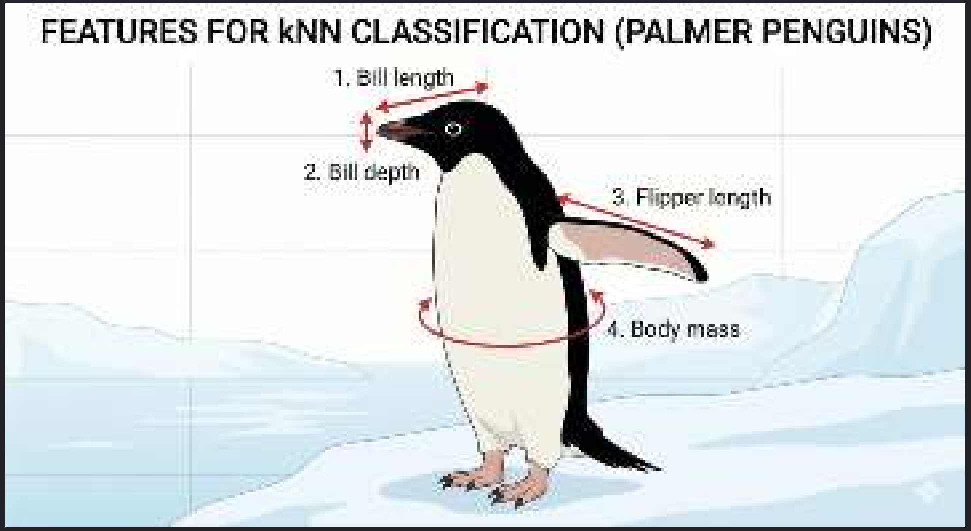
(Good repo for ML datasets: https://archive.ics.uci.edu/ )

00:00

# Extract Features & Clean

Extract 4 features,
2 of your choice:

1. Species
2. Sex
3. *Choice 1*
4. *Choice 2*



Clean file: `penguins_clean.csv`

00:00

# Divide Your Data

You need 3 datasets in ML projects:

- ***Training***: This is your known good data to build your model on/with.
  - 70% of your records
- ***Validation***: This the small subset of known good data to verify your model's goodness.
  - 20% of your records
- ***Test***: This the data for your model to label/predict.
  - 10% of your records

00:00

# Visualize Your Validation Data

Explore your validation data:

- https://plotly.com/python/px-arguments/#passing-dictionaries-or-arraylikes-as-the-dataframe-argument
- https://plotly.com/python/line-and-scatter/#setting-size-and-color-with-column-names

Start here: viz.py viz.csv

# kNN: k Nearest Neighbors

“ You are who you are nearest. ”

# kNN: The Intuition

When a new entity is to be classified via the **features** under consideration "distance" implies similarity.

# kNN: The Math

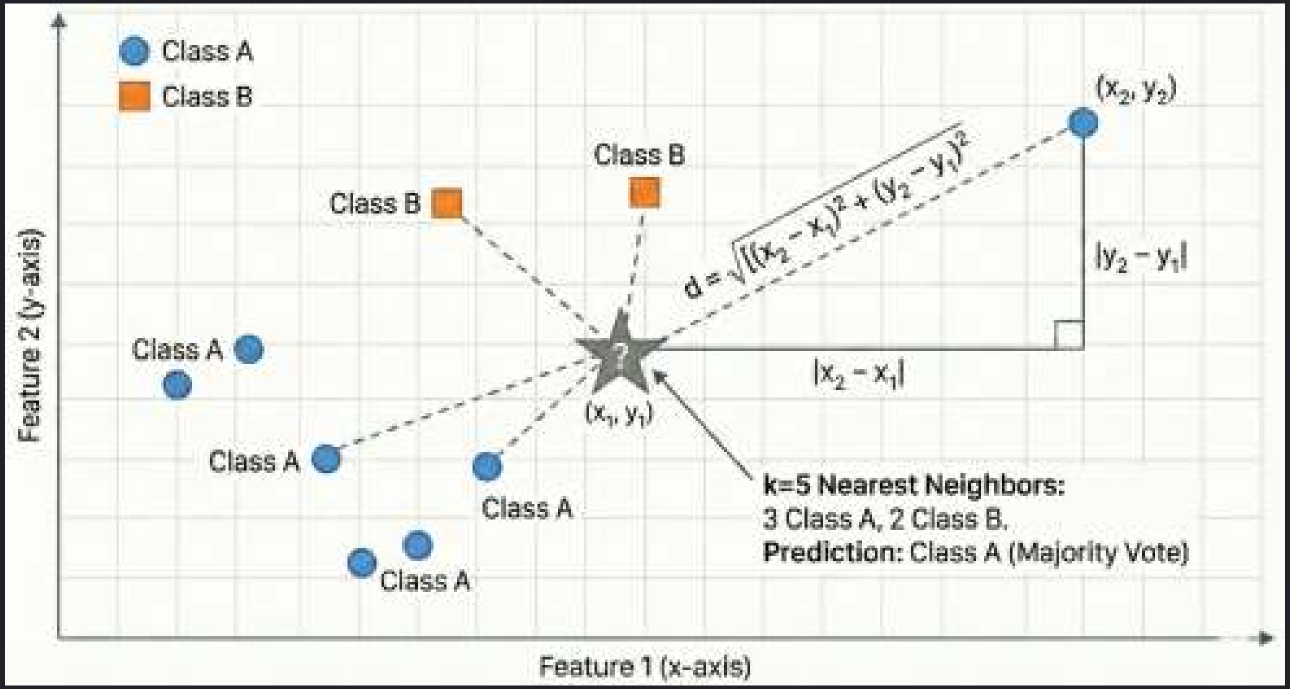
## Euclidean distance

$$\overline{dist_x} = \sqrt{\sum_n \left(\vec{x} - \vec{y_n}\right)^2}$$

## Manhattan distance

$$\overline{dist_x} = \sum_n \left|\vec{x} - \vec{y_n}\right|$$

# kNN: Vizualized



# Code Your Distance Function

# Classify Your Validation Data

Classify your validation against the training data.

Compare (text or visual) against original validation.

# Classify Your Test Data

Now time to classify your test data.

Plot training and classified test data and analyze.

# Readings

## PDFs

- This presentation ml101.pdf
- Early paper hinting towards kNN 1951-NonparametricDiscriminationConsistencyProperties-FixHodges.pdf
- First paper on kNN 1967-NearestNeighborPatternClassification-Cover.pdf
- 2023-KNNClassificationAReview-SyriopoulosKalampalikis.pdf
-

## More Presentations

- DS 101
- kNN 101

# END