DATASTAX
DEVELOPERS

CASSANDRA SUMMIT
MARCH 13-14, 2023 • SAN JOSE, CA

TRAINING DAY

CASSANDRA SUMMIT
MARCH 13-14, 2023 • SAN JOSE, CA

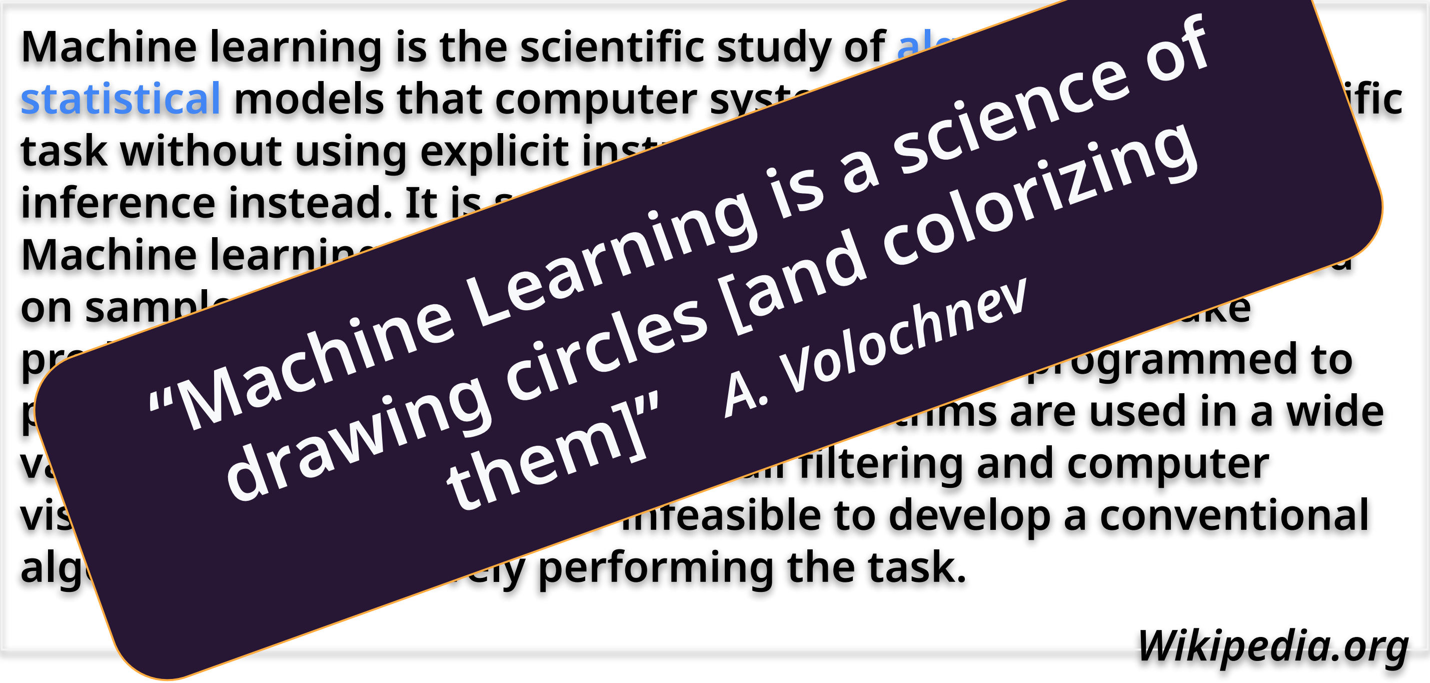
APACHE CASSANDRA® FOR ARCHITECTS AND DATA ENGINEERS:

# 4 - Machine Learning with keras and Tensor Flow

#1

Introduction to Machine Learning

What is machine leanrning



Machine learning is the scientific study of al
statistical models that computer syst ific
task without using explicit inst
inference instead. It is s
Machine learnin d
on sampl ke
pr programmed to
p ithms are used in a wide
va l filtering and computer
vis infeasible to develop a conventional
alg ely performing the task.

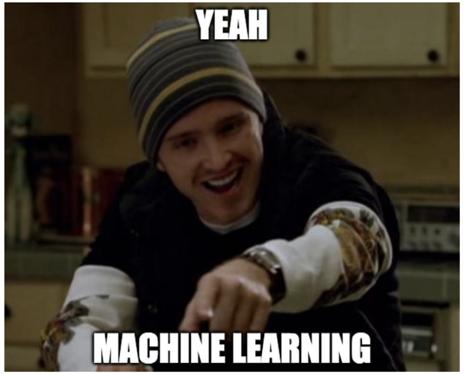
Wikipedia.org

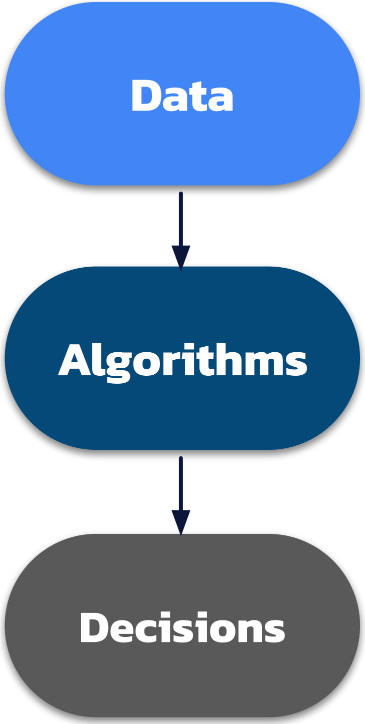
“Machine Learning is a science of drawing circles [and colorizing them]” *A. Volochnev*

❯ How it works ?

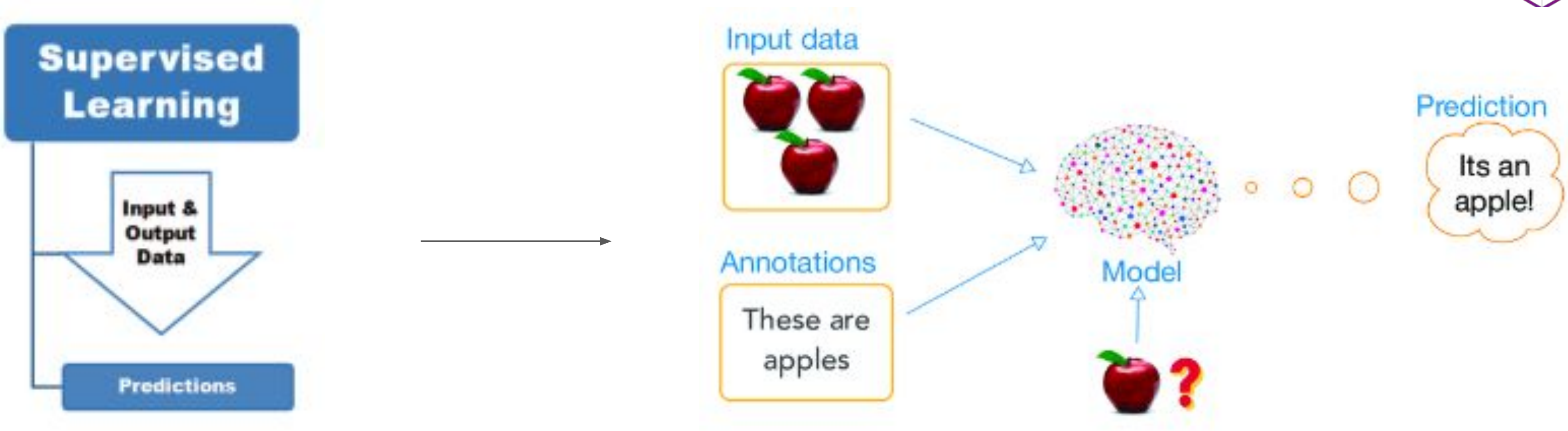
Machine Learning is a scientific way to process raw data using algorithms to make better decisions.

No magic, just billions rows of data and two buckets of mathematics. Voilà!



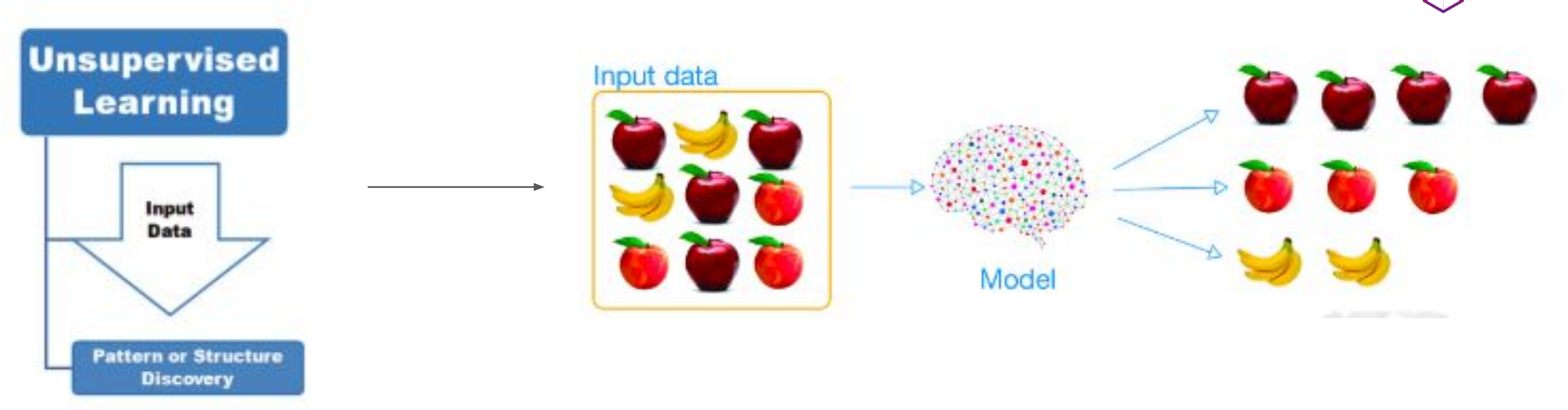


Supervised learning



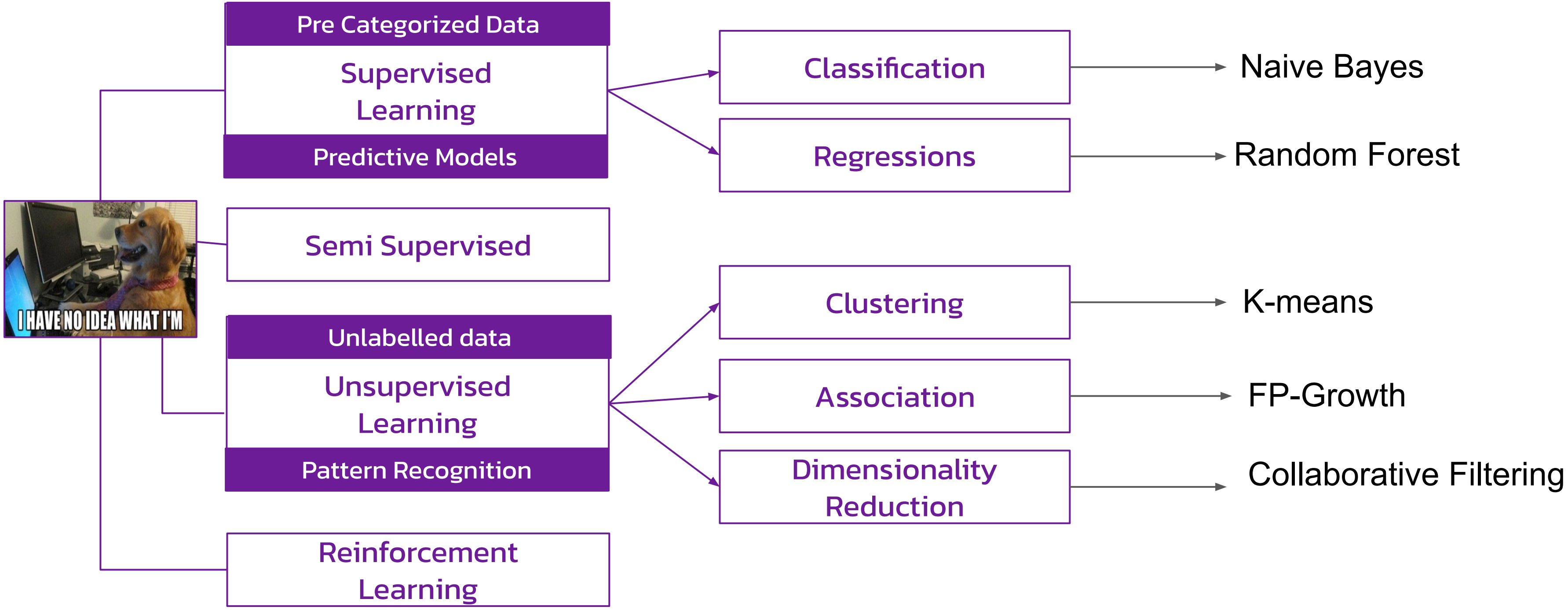
- **Knowledge of the output: learn with expert**
  - **Data are labelled with class or value**
  - **Goal: predict the class**

❯ Unsupervised learning



- **No Knowledge of the output: self-guided**
  - **Data are not labelled with class or value**
  - **Goal: Determine Patterns of Grouping**

Machine Learning Algorithms



Pre Categorized Data

Supervised Learning

Predictive Models

Classification

Naive Bayes

Regressions

Random Forest

Semi Supervised

Unlabelled data

Unsupervised Learning

Pattern Recognition

Clustering

K-means

Association

FP-Growth

Dimensionality Reduction

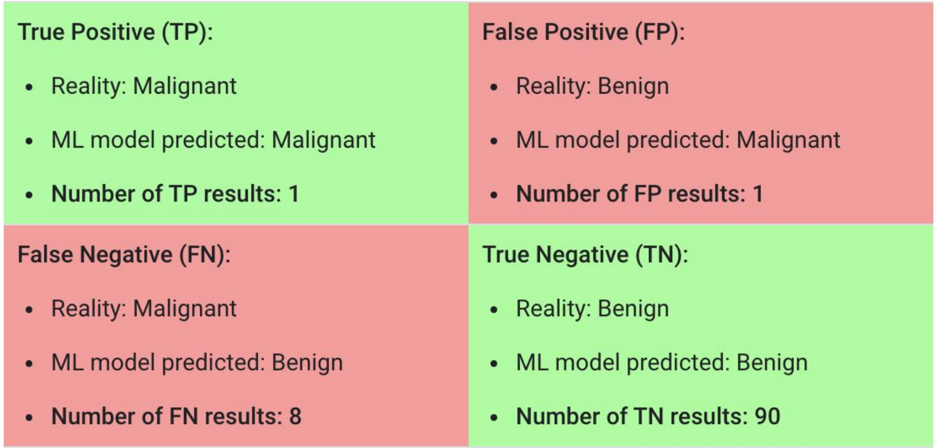
Collaborative Filtering

Reinforcement Learning

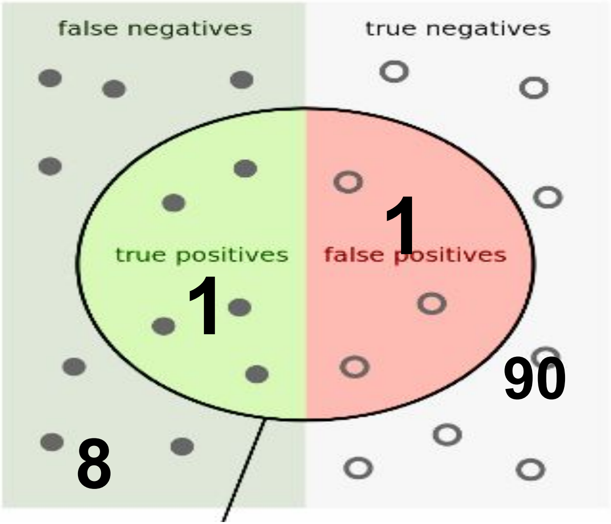
#2
Evaluating AI Models

❯ Evaluating data model

100 people, 9 have malignant tumor (very bad), 91 have benign tumor (bad)



True Positive (TP): • Reality: Malignant • ML model predicted: Malignant • **Number of TP results: 1**	**False Positive (FP):** • Reality: Benign • ML model predicted: Malignant • **Number of FP results: 1**
False Negative (FN): • Reality: Malignant • ML model predicted: Benign • **Number of FN results: 8**	**True Negative (TN):** • Reality: Benign • ML model predicted: Benign • **Number of TN results: 90**

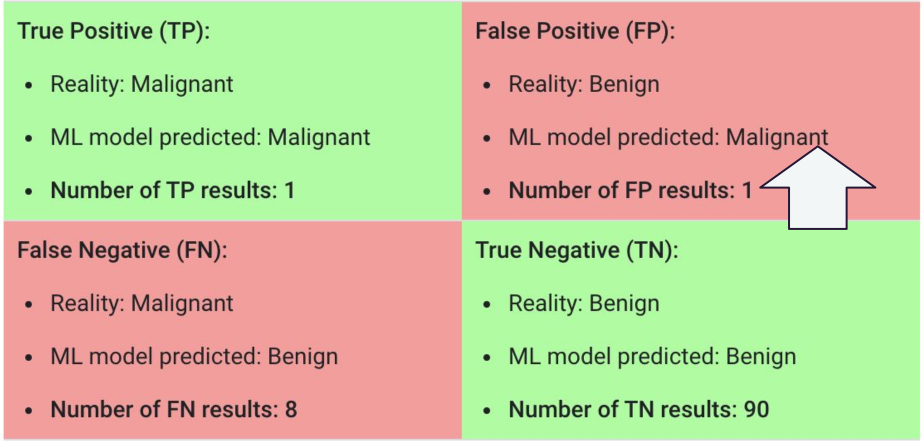


Example and images from:
https://shiffdag.medium.com/what-is-accuracy-precision-and-recall-and-why-are-they-important-ebfcb5a10df2

❯ Accuracy

Accuracy is an evaluating classification models metric, it is the fraction of predictions model identified correctly.

$$\textbf{Accuracy} = \frac{\textbf{Correct Prediction (TP + TN)}}{\sum \textbf{Predictions}}$$

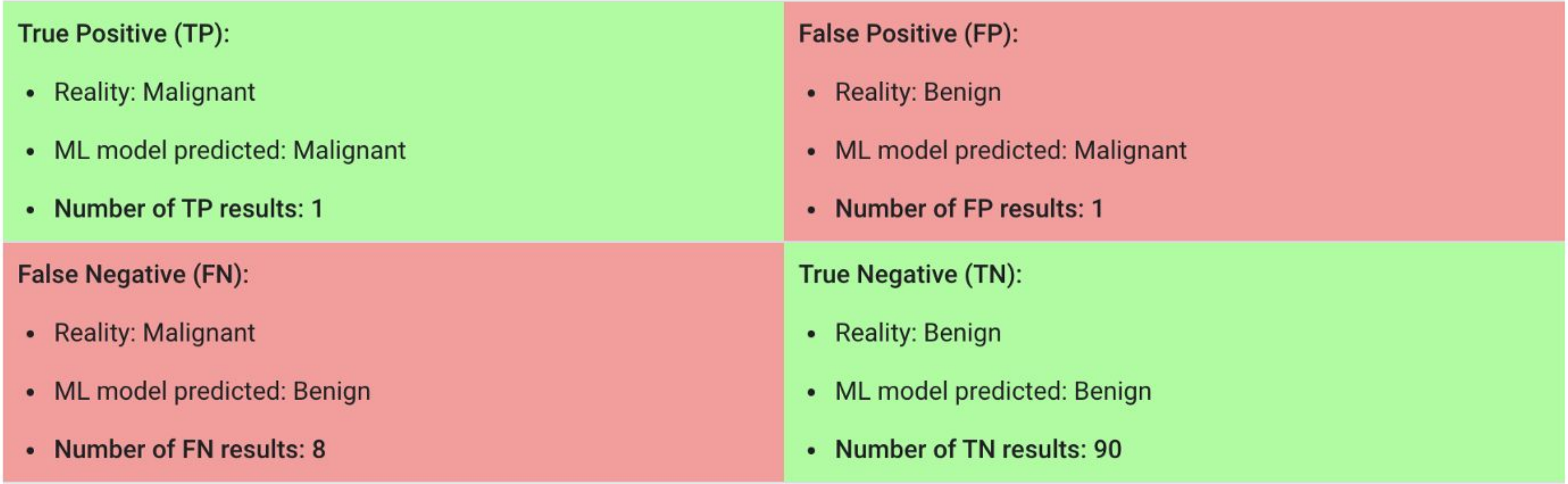


True Positive (TP): • Reality: Malignant • ML model predicted: Malignant • **Number of TP results: 1**	False Positive (FP): • Reality: Benign • ML model predicted: Malignant • **Number of FP results: 1**
False Negative (FN): • Reality: Malignant • ML model predicted: Benign • **Number of FN results: 8**	True Negative (TN): • Reality: Benign • ML model predicted: Benign • **Number of TN results: 90**

What is the accuracy here ?

How many go home without proper treatment ?

› Accuracy



True Positive (TP): • Reality: Malignant • ML model predicted: Malignant • **Number of TP results: 1**	**False Positive (FP):** • Reality: Benign • ML model predicted: Malignant • **Number of FP results: 1**
False Negative (FN): • Reality: Malignant • ML model predicted: Benign • **Number of FN results: 8**	**True Negative (TN):** • Reality: Benign • ML model predicted: Benign • **Number of TN results: 90**

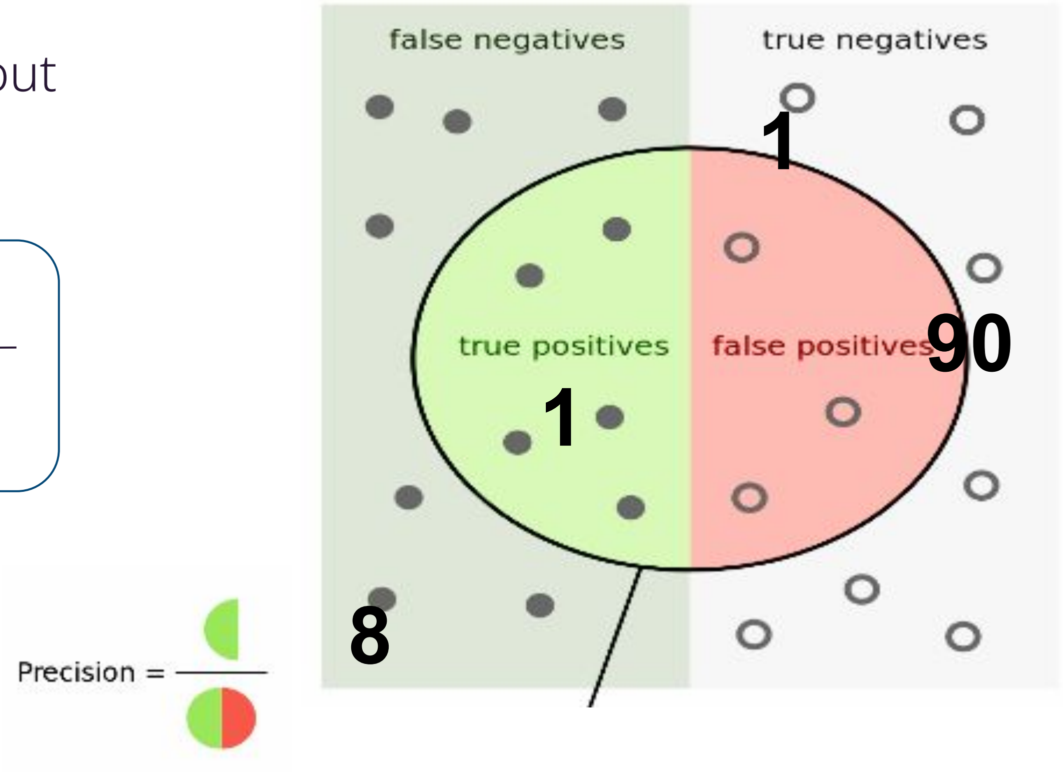
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1+90}{1+90+1+8} = 0.91$$

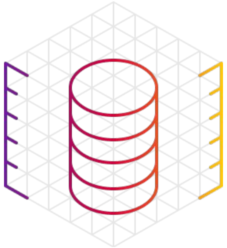
› Precision

Precision counts true positives out of all true and false positives.

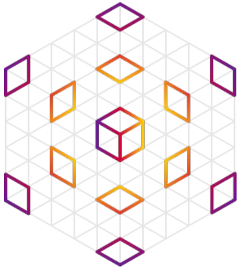
$$\text{Precision} = \frac{\text{True Positives (TP)}}{\sum \text{Positives (TP + FP)}}$$

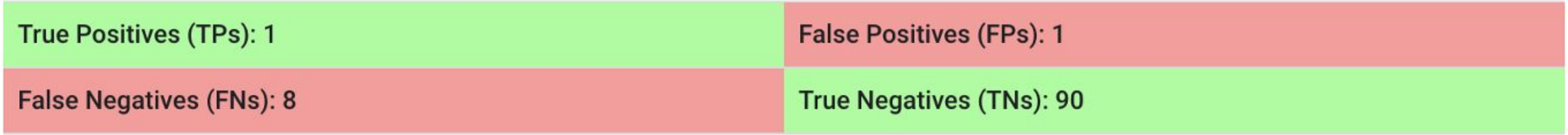
What is the precision here?





❯ Precision





True Positives (TPs): 1	False Positives (FPs): 1
False Negatives (FNs): 8	True Negatives (TNs): 90

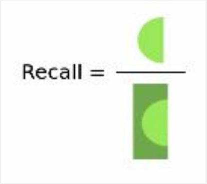
$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+1} = 0.5$$

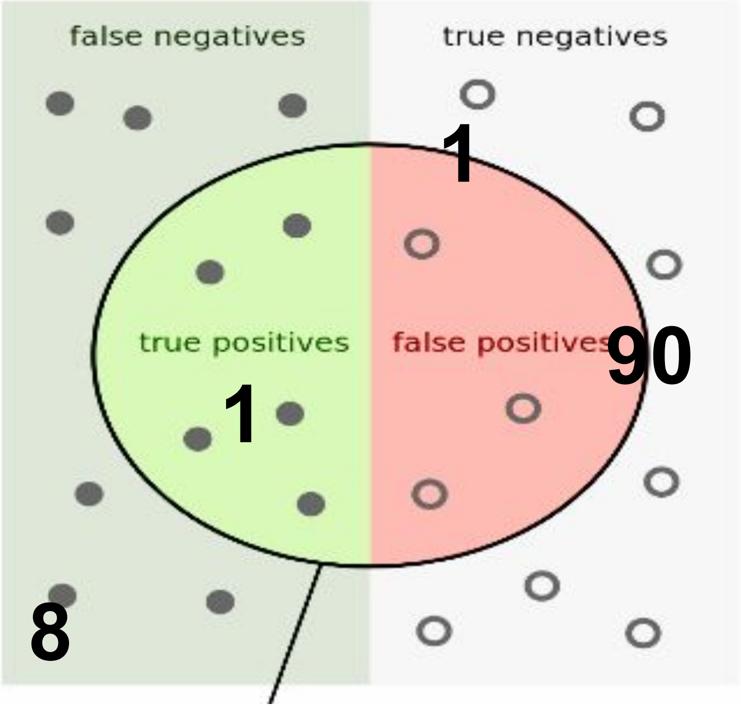
❯ Recall

Recall correctly identified positives out of all real positives.

$$\textbf{Prediction} = \frac{\textbf{True Positives (TP)}}{\sum \textbf{Correct (TP + FN)}}$$

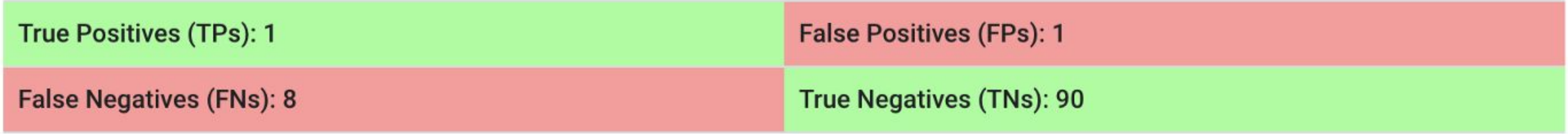
What is the recall here?





Recall

Let's calculate recall for our tumor classifier:

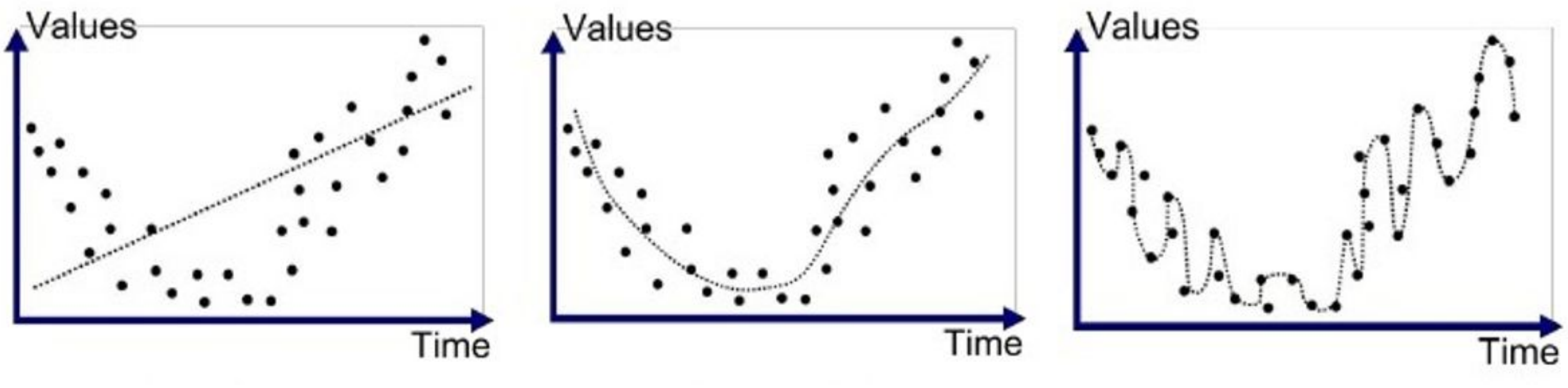


True Positives (TPs): 1	**False Positives (FPs): 1**
False Negatives (FNs): 8	**True Negatives (TNs): 90**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1}{1+8} = 0.11$$

Anub Bhande @ Medium.com

❯ Under fitted vs over-fitted model



Underfitted

Good Fit/Robust

Overfitted

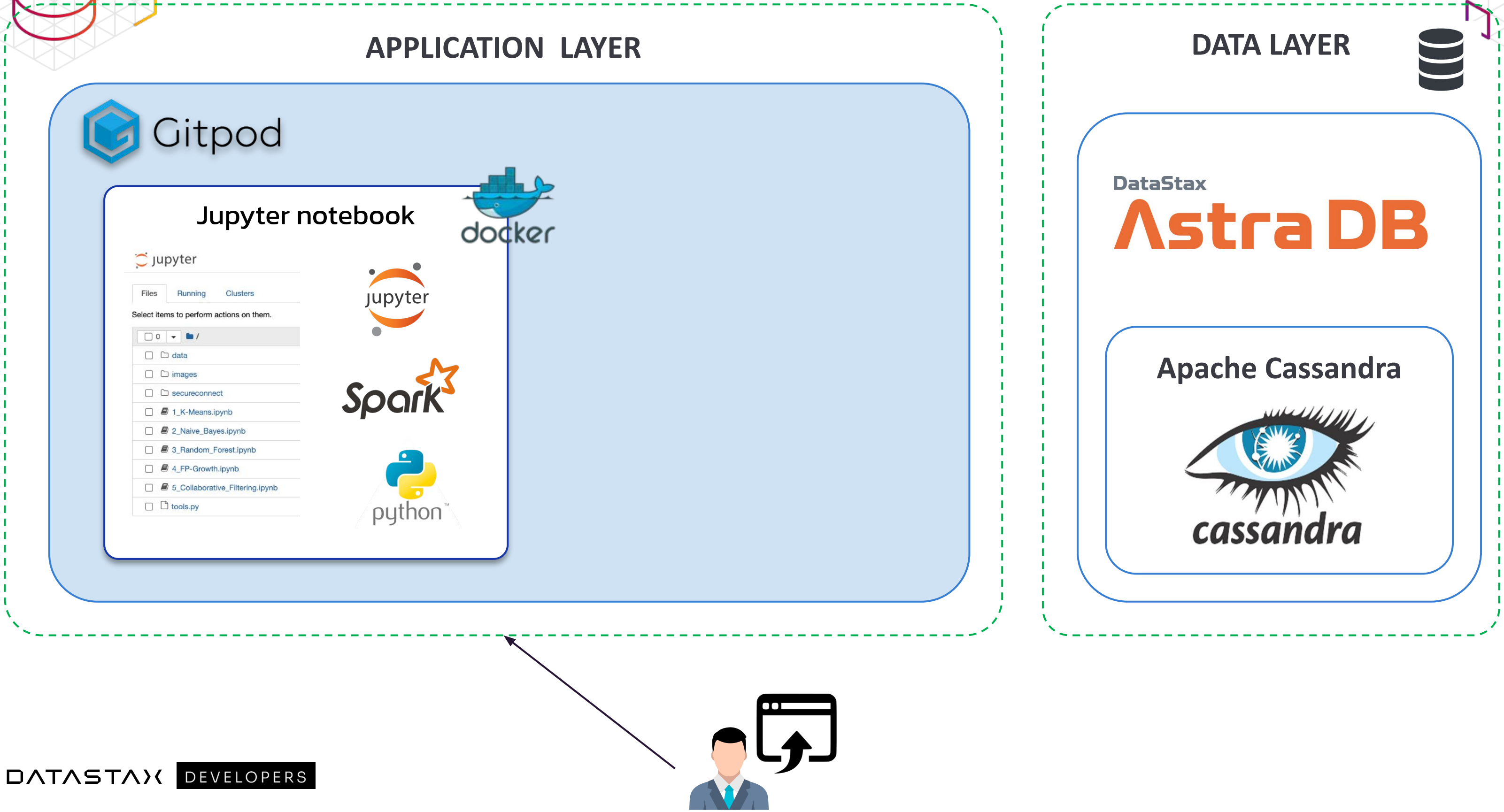
Not accurate, too simple

Good, well generalised

Over-trained, perfect on train data, fails on test data

#3
Tooling

❯ Jupyter Notebook



APPLICATION LAYER

Gitpod

Jupyter notebook

jupyter

Files Running Clusters

Select items to perform actions on them.

□ 0 ▾ /

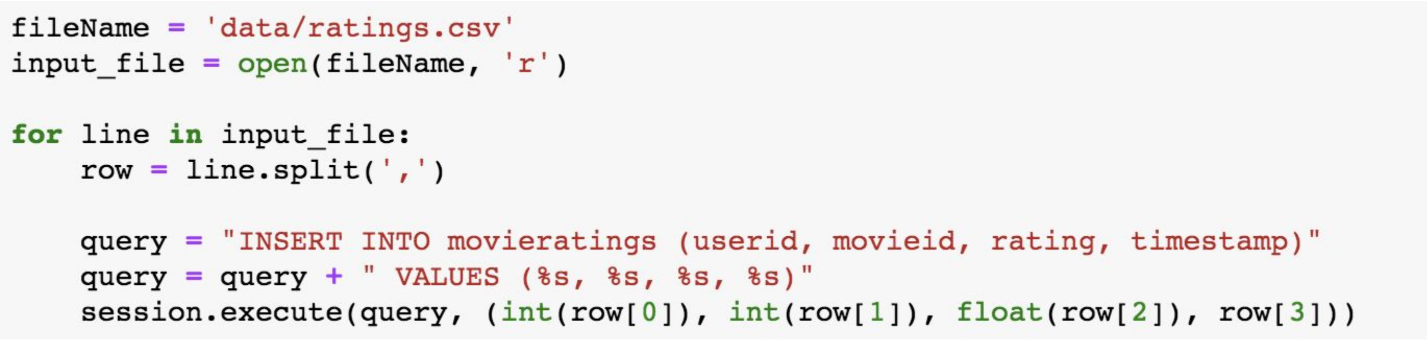
- □ data
- □ images
- □ secureconnect
- □ 1_K-Means.ipynb
- □ 2_Naive_Bayes.ipynb
- □ 3_Random_Forest.ipynb
- □ 4_FP-Growth.ipynb
- □ 5_Collaborative_Filtering.ipynb
- □ tools.py

DATA LAYER

Apache Cassandra

Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.



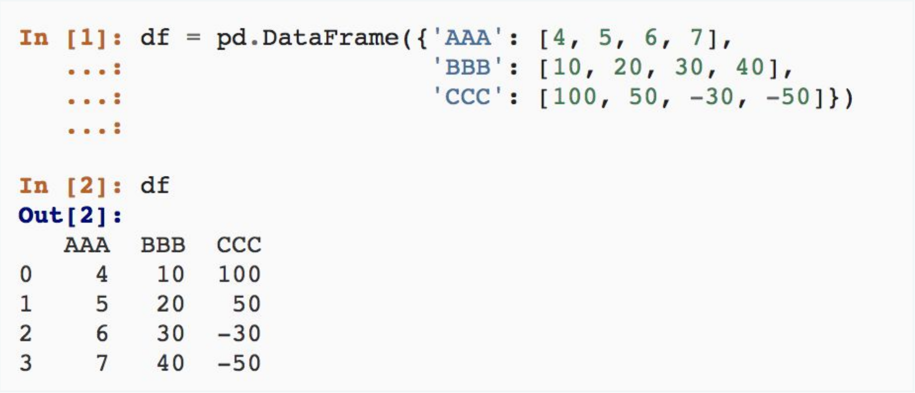
```python
fileName = 'data/ratings.csv'
input_file = open(fileName, 'r')

for line in input_file:
    row = line.split(',')

    query = "INSERT INTO movieratings (userid, movieid, rating, timestamp)"
    query = query + " VALUES (%s, %s, %s, %s)"
    session.execute(query, (int(row[0]), int(row[1]), float(row[2]), row[3]))
```

❯ Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.



```
In [1]: df = pd.DataFrame({'AAA': [4, 5, 6, 7],
   ...:                    'BBB': [10, 20, 30, 40],
   ...:                    'CCC': [100, 50, -30, -50]})
   ...:

In [2]: df
Out[2]:
   AAA  BBB  CCC
0    4   10  100
1    5   20   50
2    6   30  -30
3    7   40  -50
```

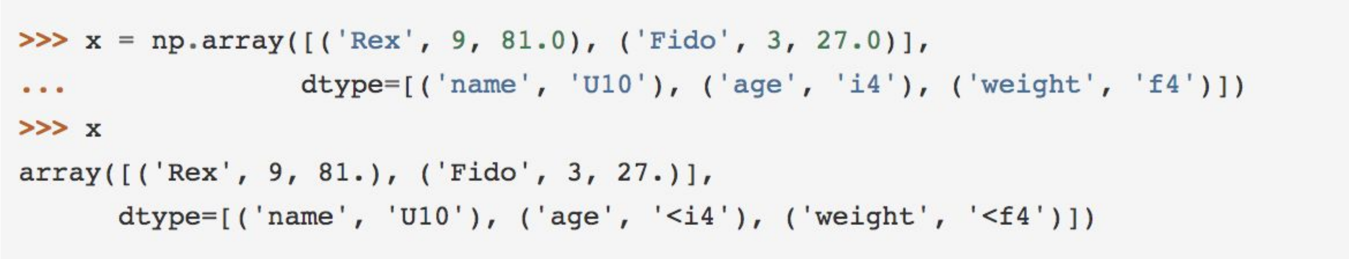
› Py Spark

Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In addition, PySpark, helps you interface with Resilient Distributed Datasets (RDDs) in Apache Spark and Python programming language.

❯ Num Py

NumPy is the fundamental package for scientific computing with Python.

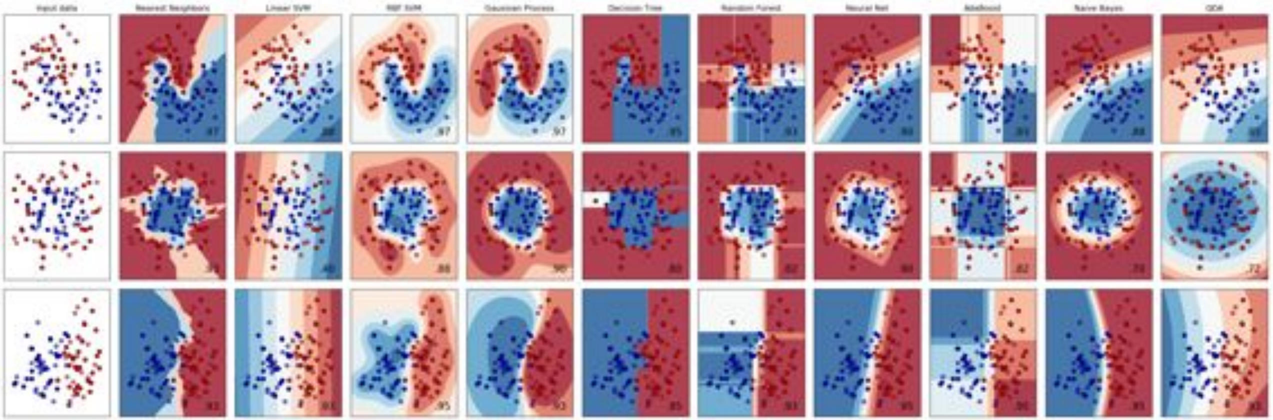
It contains among other things: a powerful N-dimensional array object, sophisticated functions, useful linear algebra, Fourier transform, and random number capabilities.



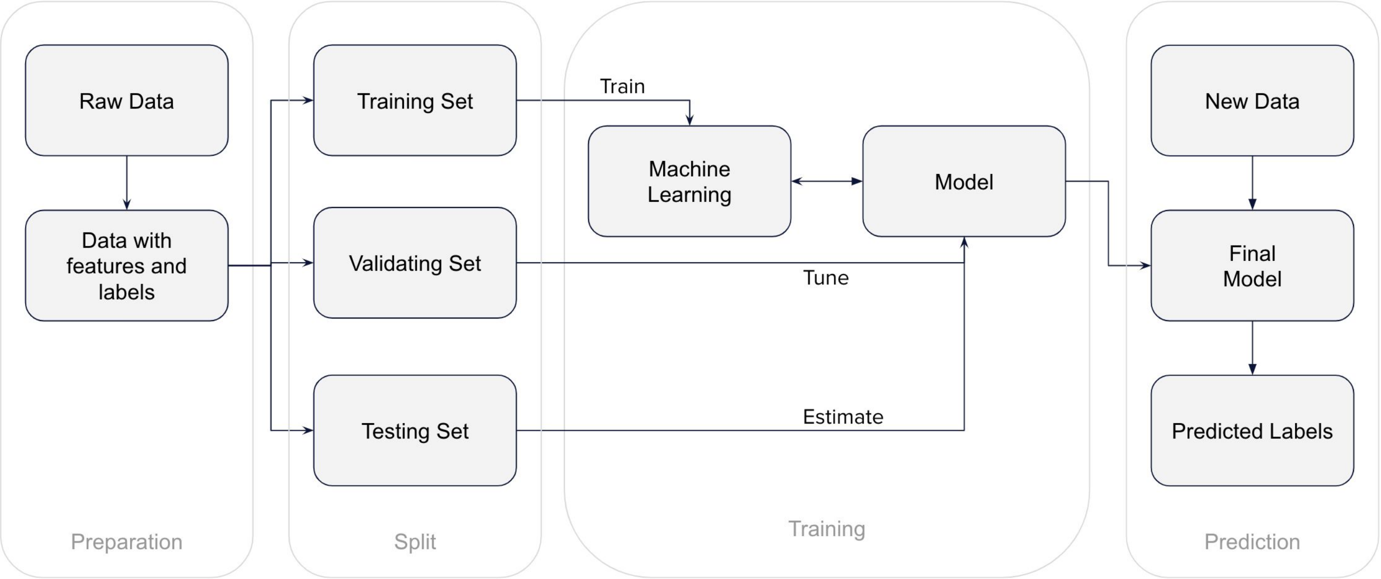
```
>>> x = np.array([('Rex', 9, 81.0), ('Fido', 3, 27.0)],
...              dtype=[('name', 'U10'), ('age', 'i4'), ('weight', 'f4')])
>>> x
array([('Rex', 9, 81.), ('Fido', 3, 27.)],
      dtype=[('name', 'U10'), ('age', '<i4'), ('weight', '<f4')])
```

❯ Scikit Learn ("sklearn")

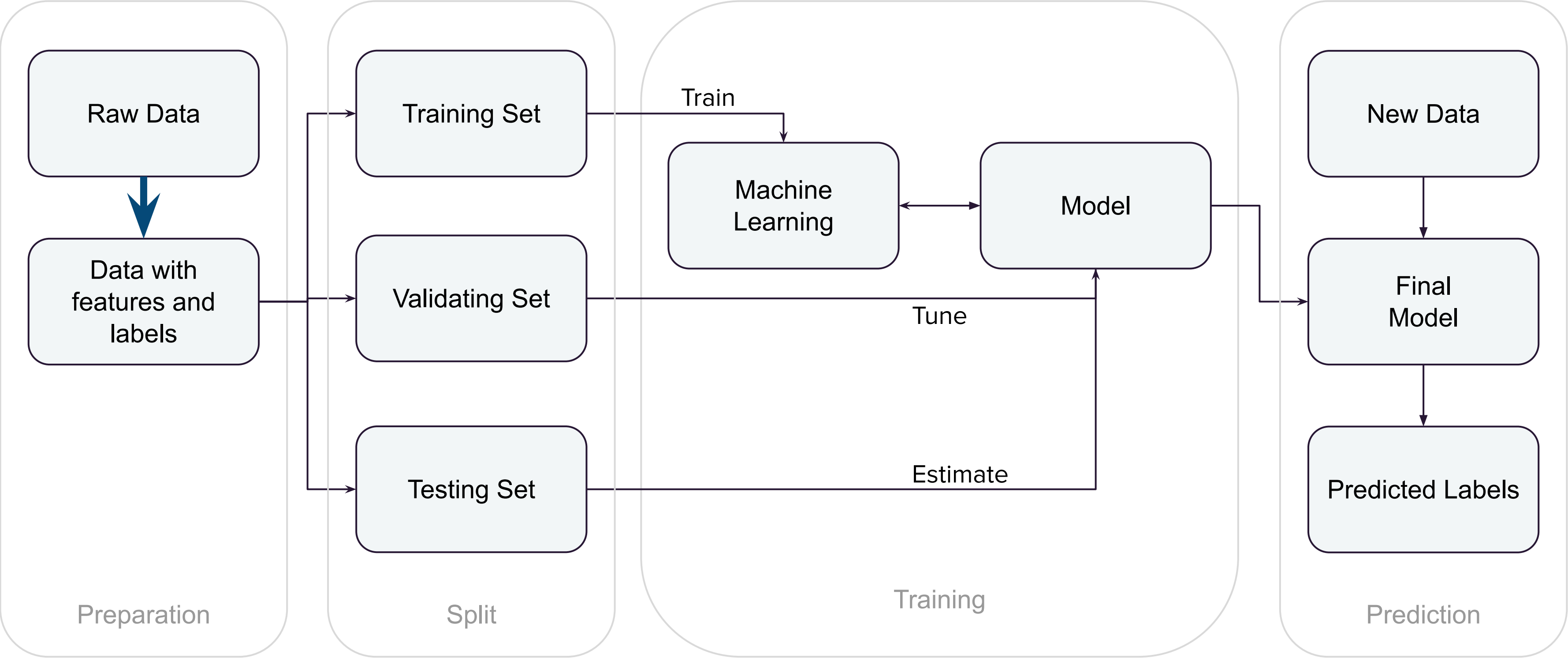
An open source, simple and efficient tool for predictive data analysis, accessible to everybody, and reusable in various contexts. Built on NumPy, SciPy, and matplotlib.



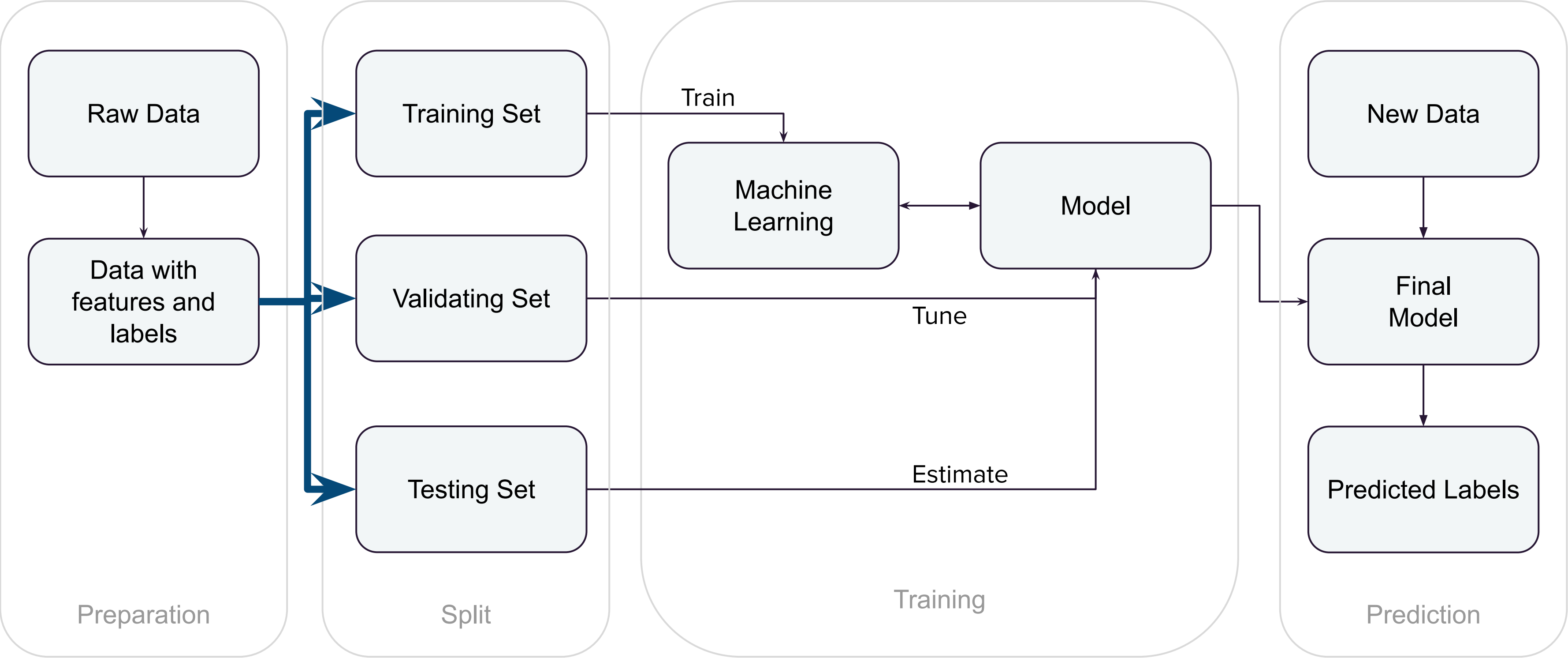
- **Question / Hypothesis**
- **Algorithm Selection**
- **Data Preparation**
- **Data Split**
- **Training**
- **Tuning**
- **Testing**
- **Analysis**
- **Repeat**



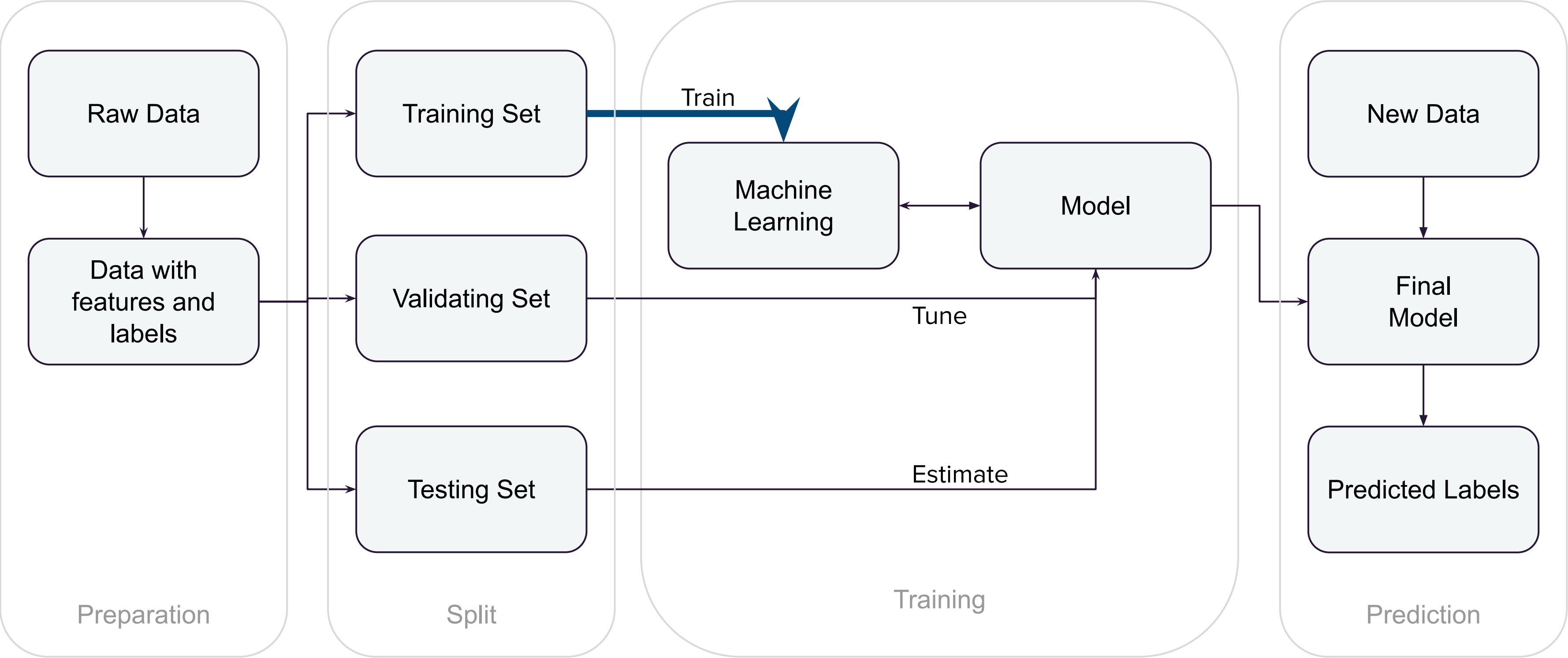
Learning Workflow



Data Preparation


Raw Data
Data with features and labels
Preparation
Training Set
Validating Set
Testing Set
Split
Train
Machine Learning
Model
Tune
Estimate
Training
New Data
Final Model
Predicted Labels
Prediction

Data Split


Raw Data
Data with features and labels
Preparation
Training Set
Validating Set
Testing Set
Split
Train
Machine Learning
Model
Tune
Estimate
Training
New Data
Final Model
Predicted Labels
Prediction

Training

FUN FACT: this image was created by … an algorithm, starting from the textual prompt: **"a metallic cyborg in a gym"**

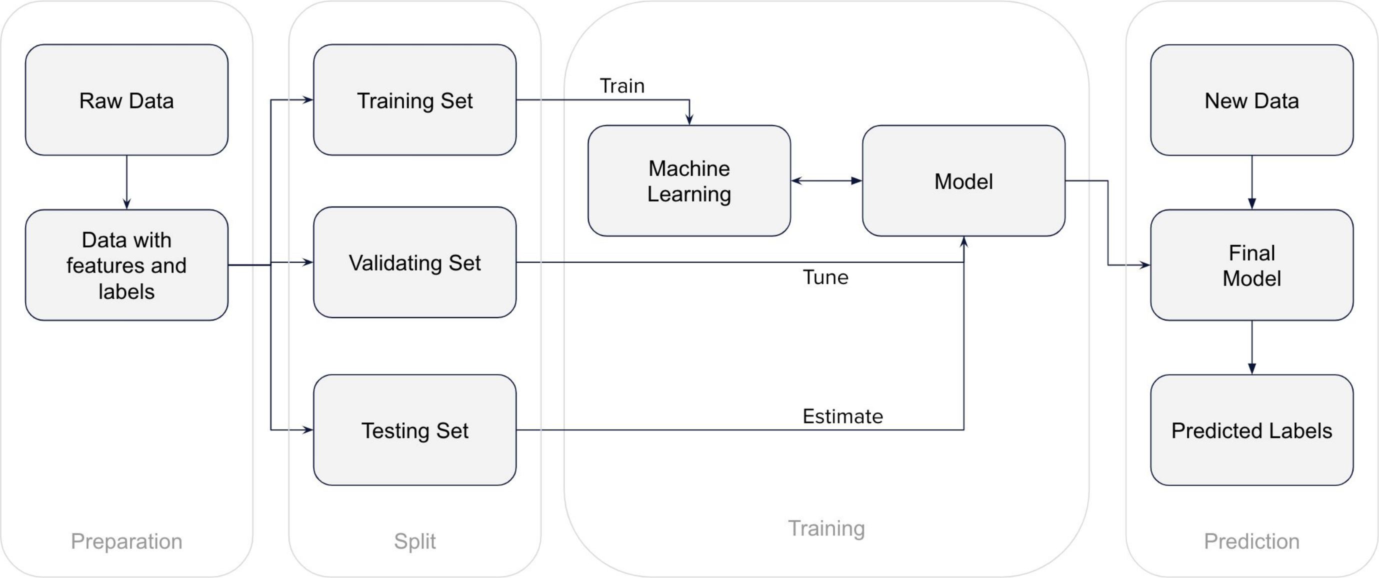
`https://huggingface.co/spaces/stabilityai/stable-diffusion`

Intermezzo: "training the machine"

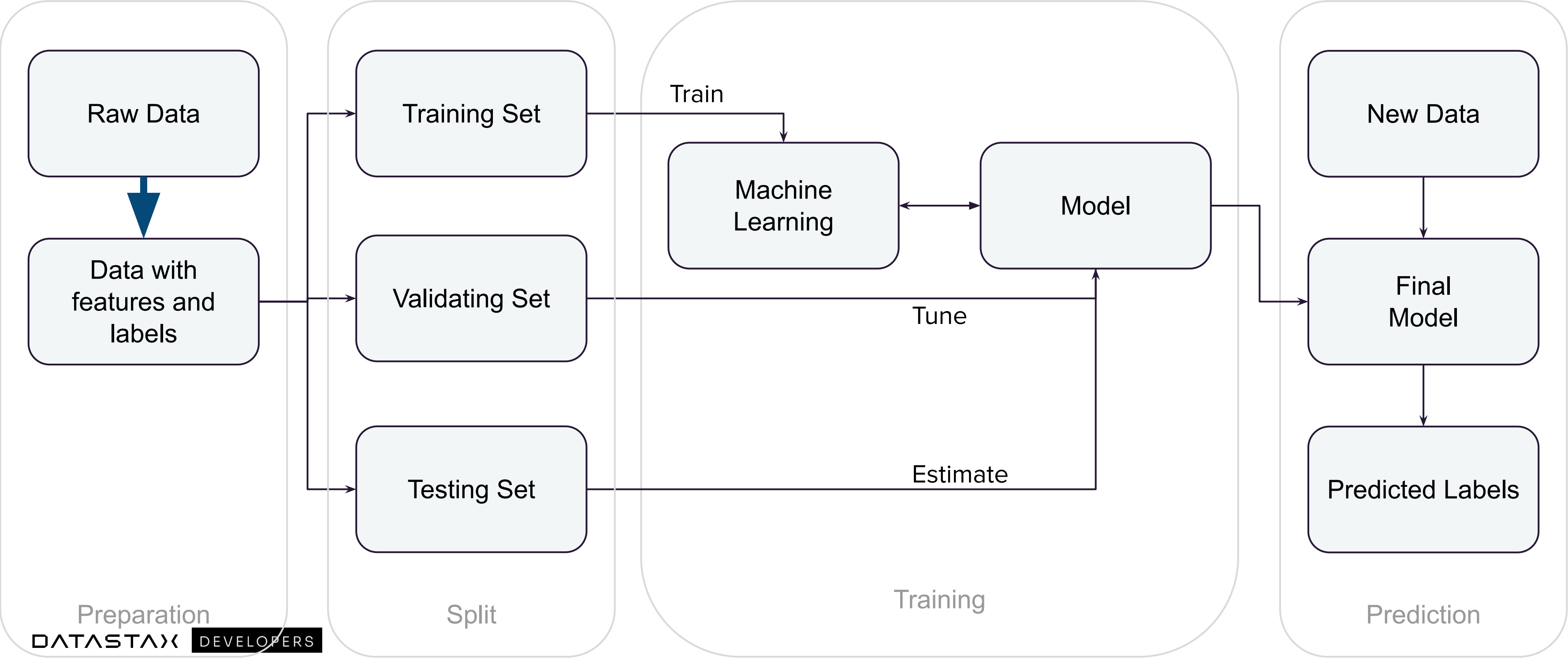
#3
Methodology

❯ Learning Workflow

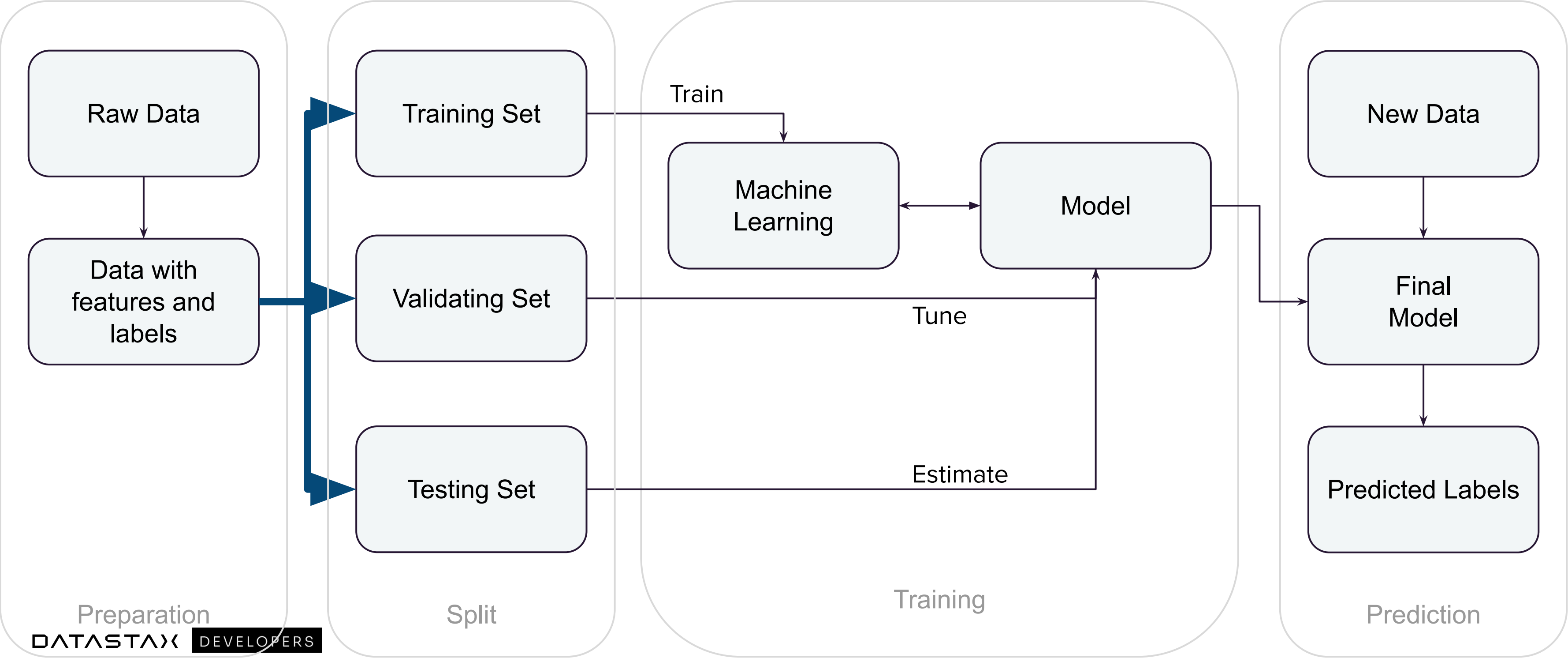
- Question / Hypothesis
- Algorithm Selection
- Data Preparation
- Data Split
- Training
- Tuning
- Testing
- Analysis
- Repeat



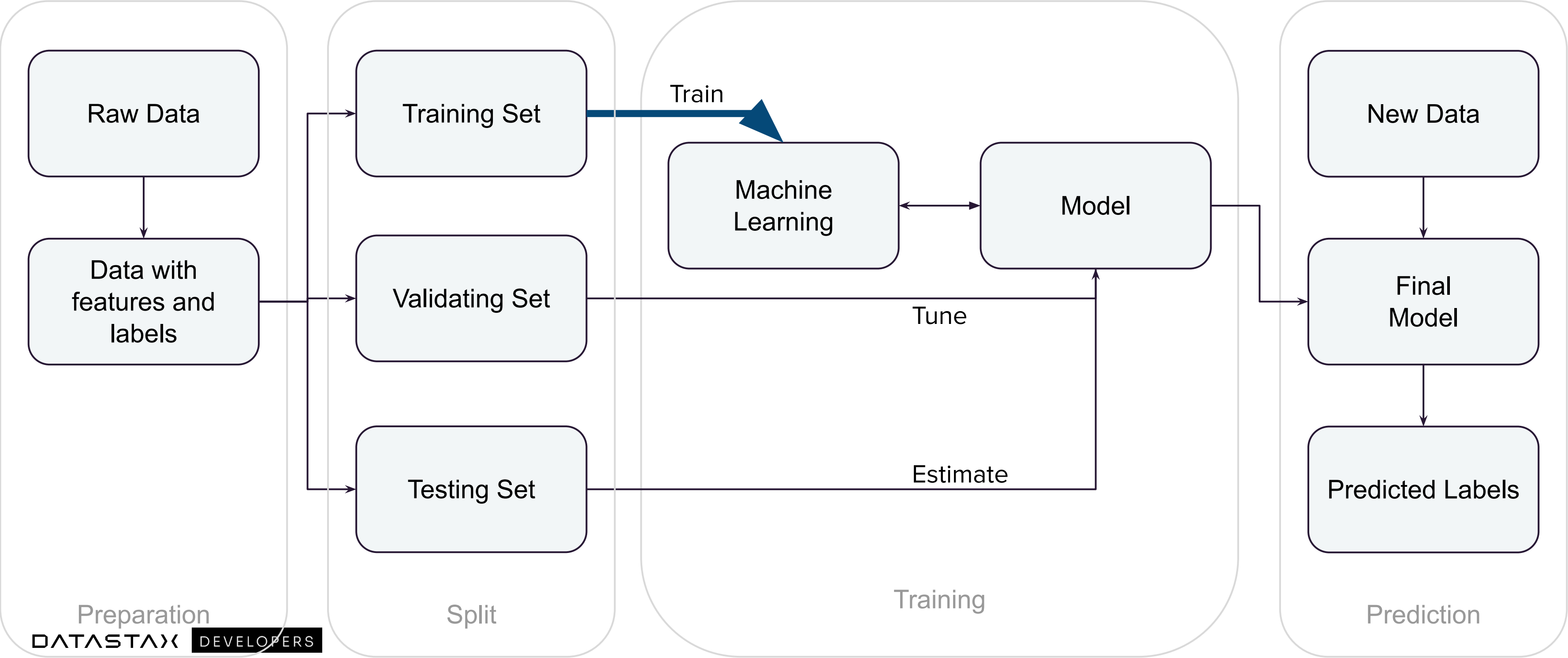
❯ Data preparation



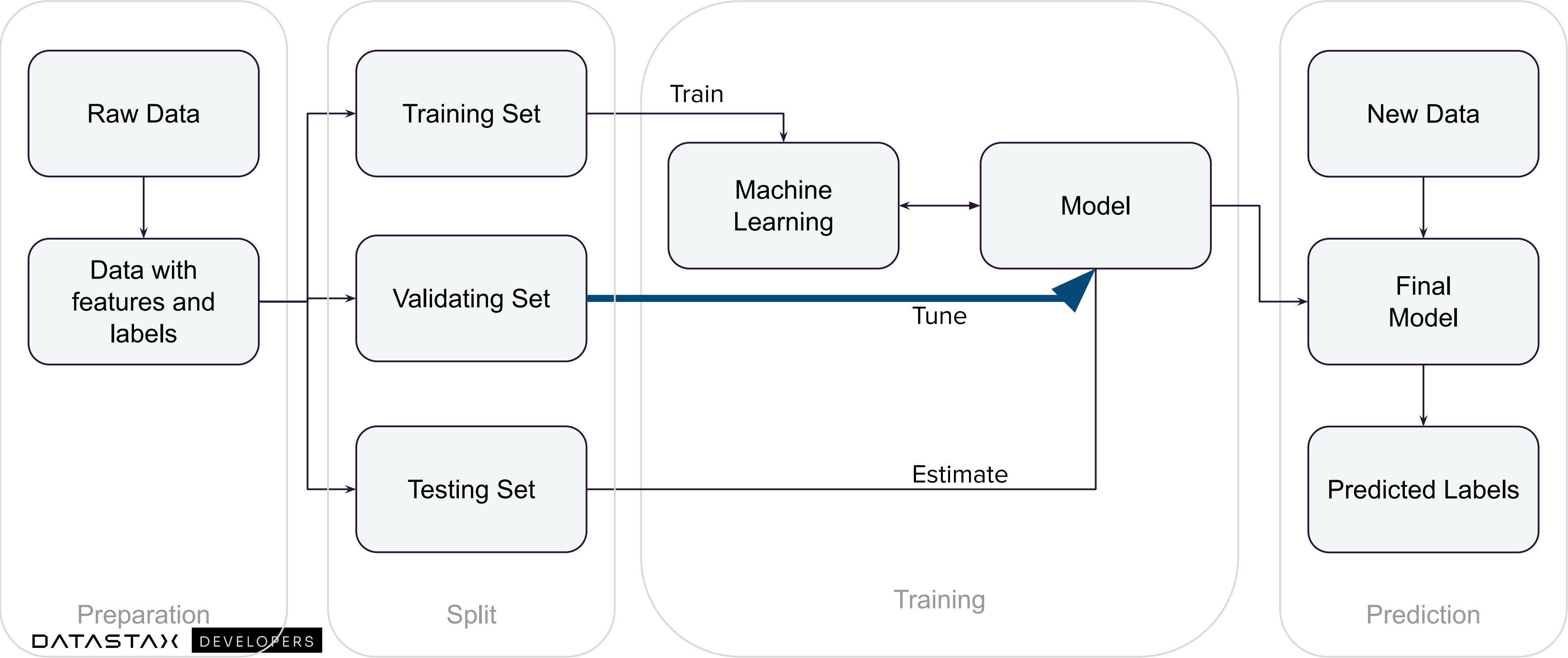
❯ Data Split


Raw Data
Data with features and labels
Training Set
Validating Set
Testing Set
Train
Machine Learning
Model
Tune
Estimate
New Data
Final Model
Predicted Labels
Preparation
Split
Training
Prediction

❯ Training


Raw Data
Data with features and labels
Preparation
Training Set
Validating Set
Testing Set
Split
Train
Machine Learning
Model
Tune
Estimate
Training
New Data
Final Model
Predicted Labels
Prediction

❯ Tuning



Raw Data

Data with features and labels

Training Set

Validating Set

Testing Set

Train

Machine Learning

Model

Tune

Estimate

New Data

Final Model

Predicted Labels

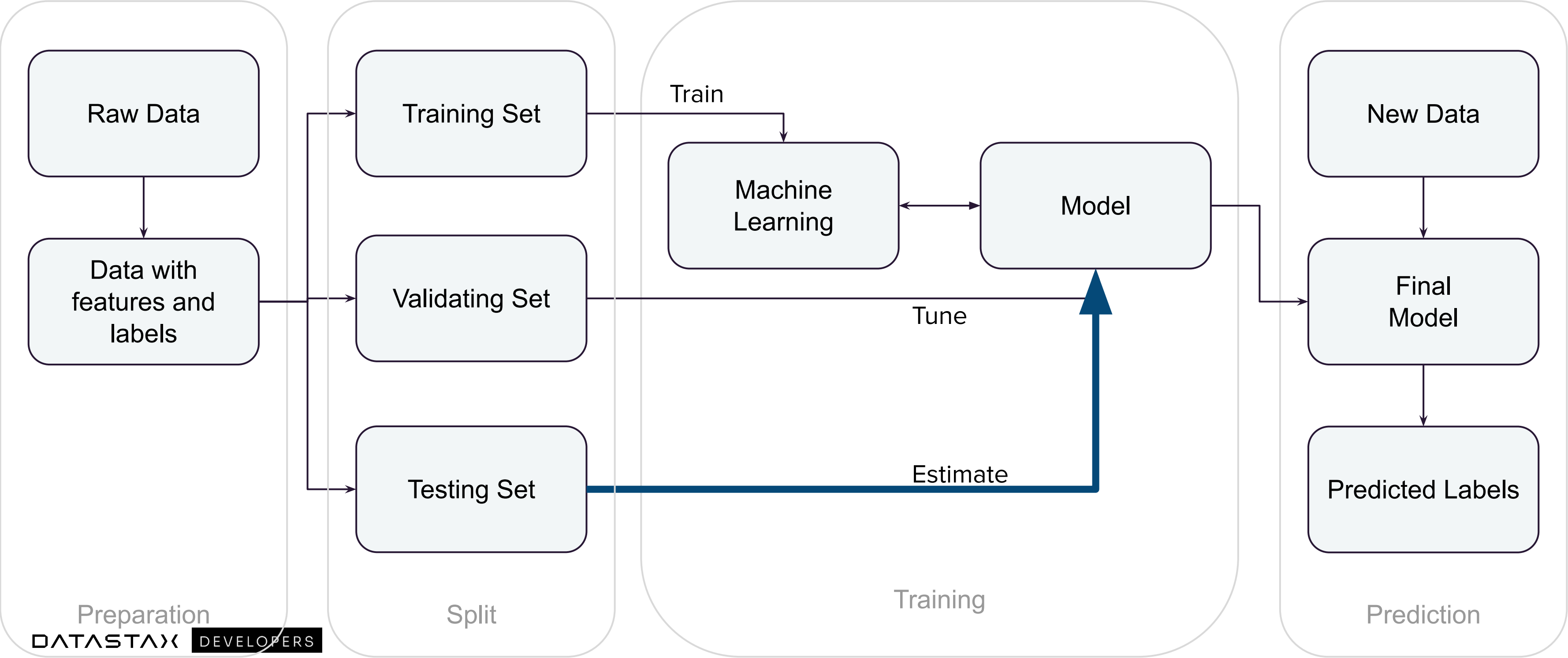
Preparation

Split

Training

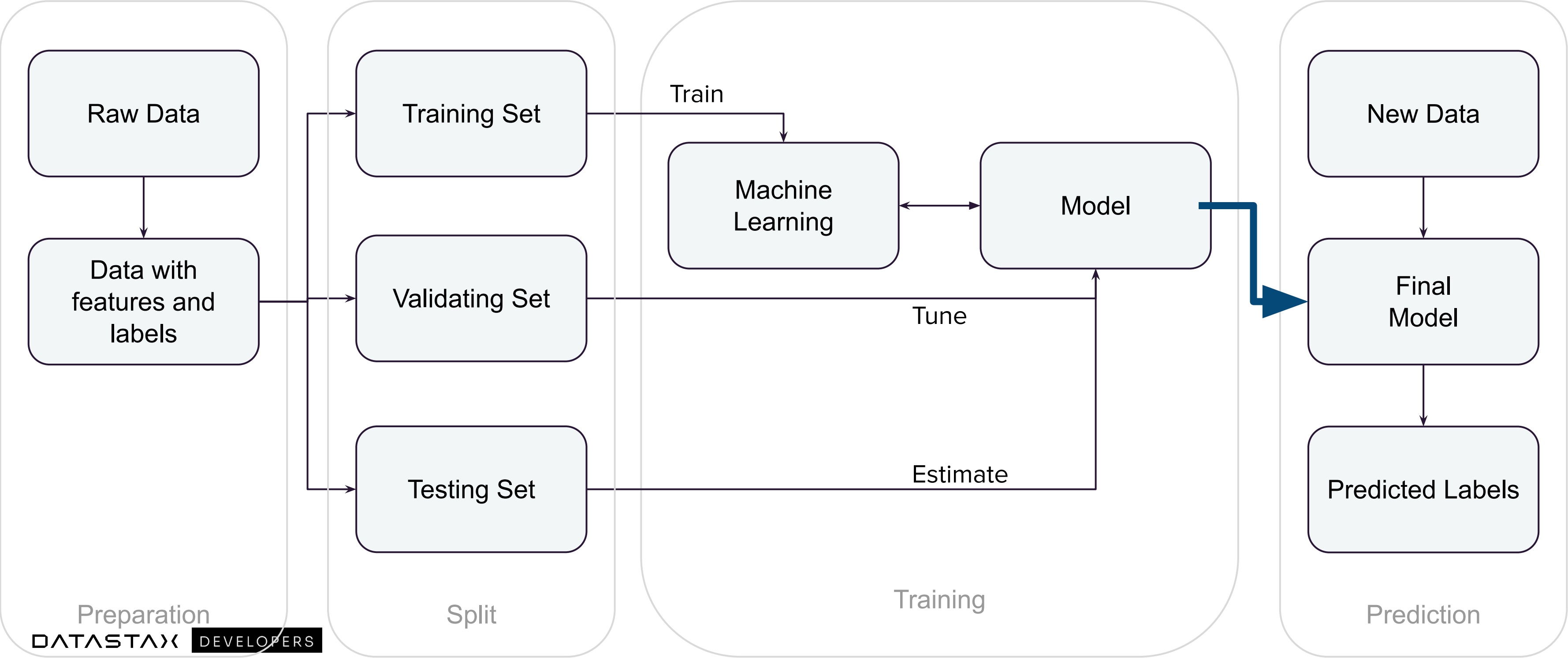
Prediction

❯ Testing



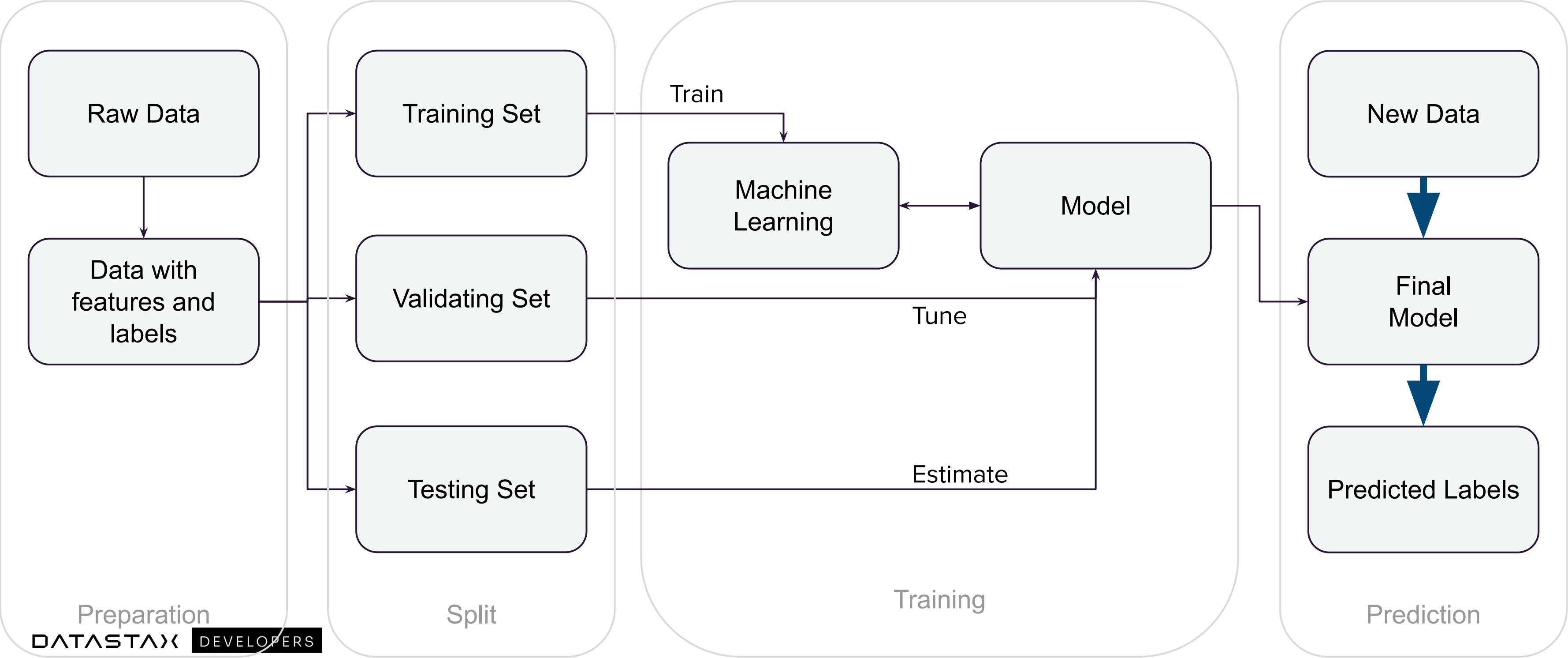
DATASTAX DEVELOPERS

❯ Testing



DATASTAX DEVELOPERS

❯ Testing


Raw Data
Data with features and labels
Training Set
Validating Set
Testing Set
Train
Machine Learning
Model
Tune
Estimate
New Data
Final Model
Predicted Labels
Preparation
Split
Training
Prediction

Thank You

DATASTAX DEVELOPERS