Machine Learning for Beginners

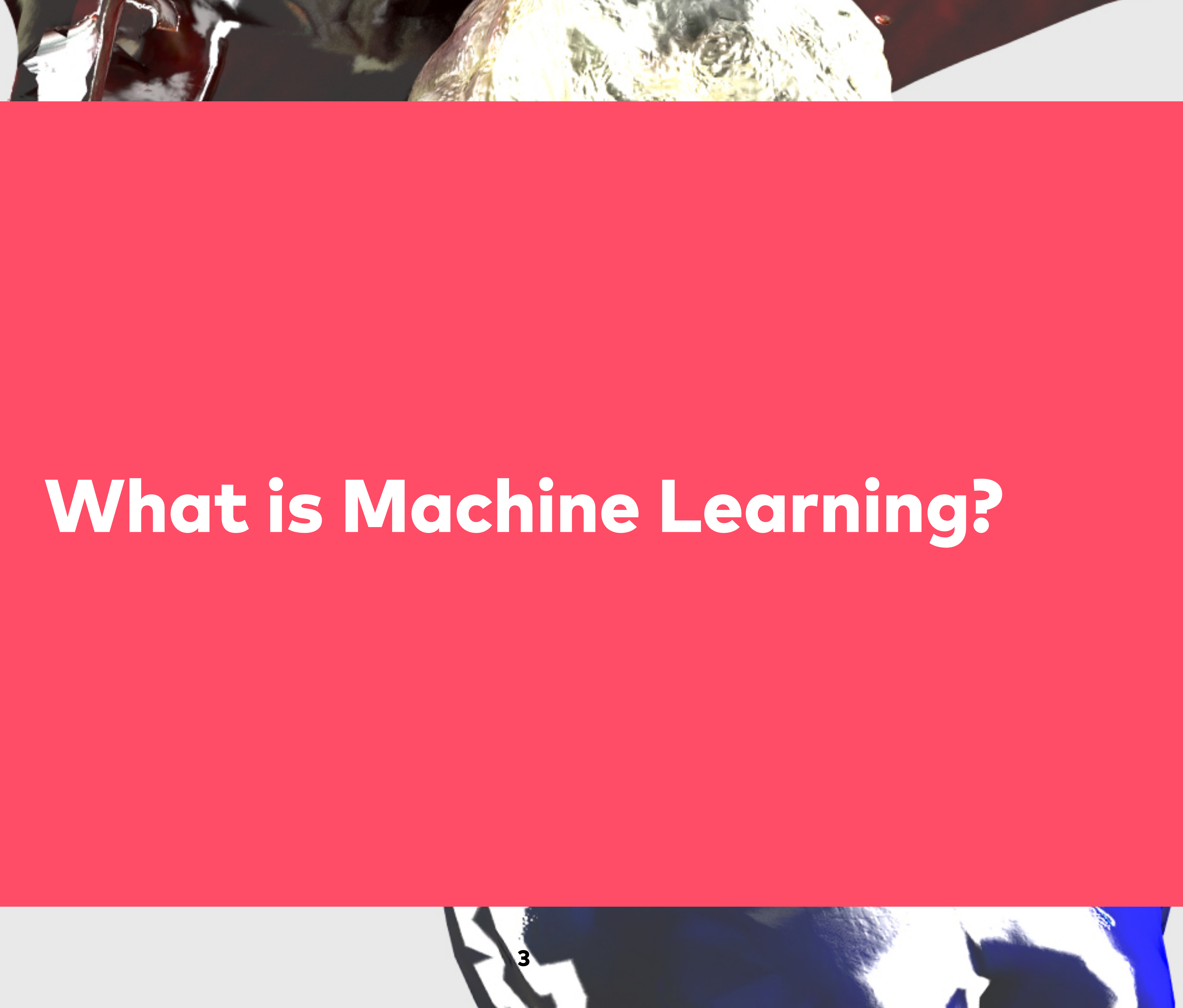
INNOQ

www.innoq.com

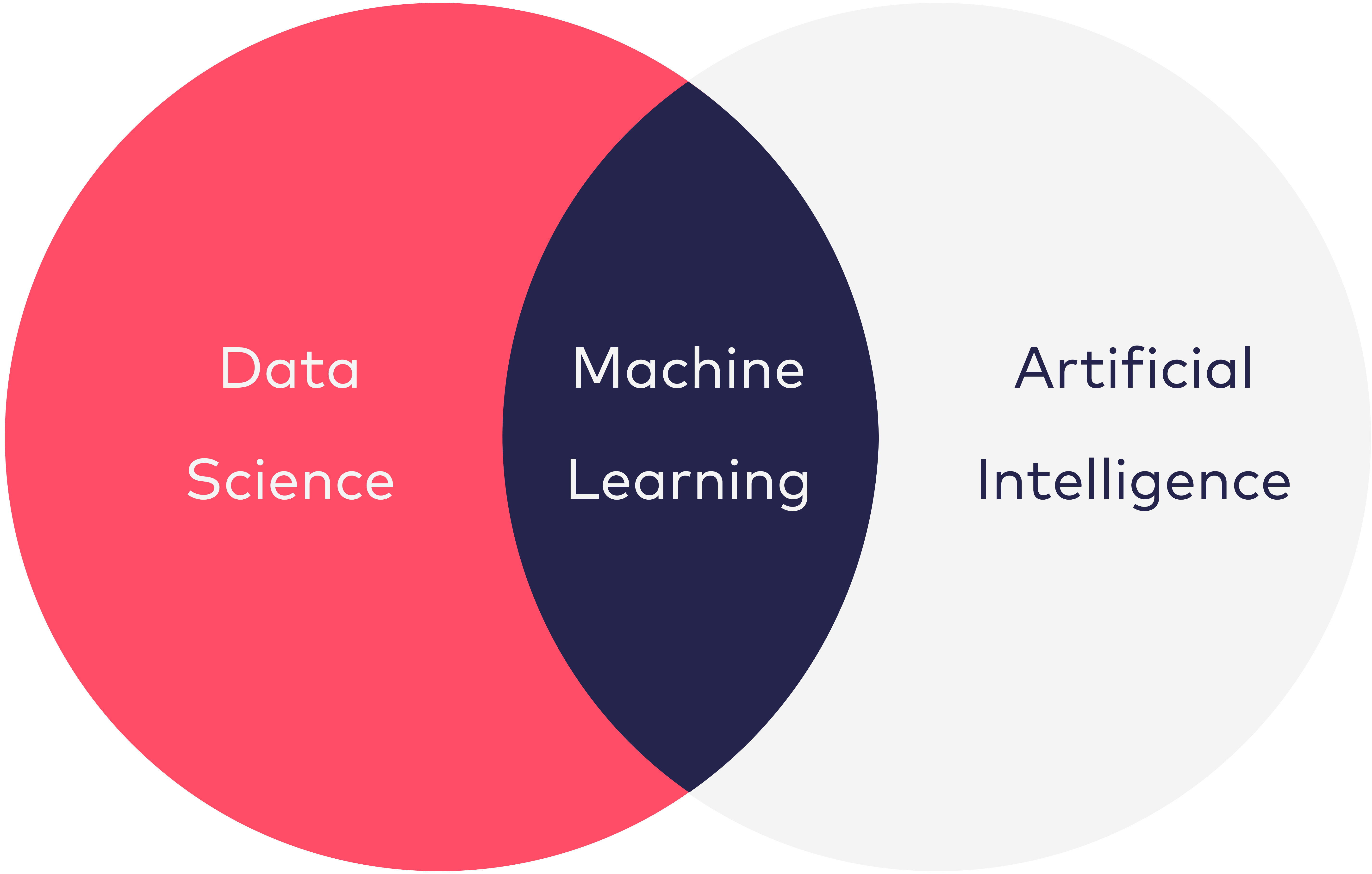
Who are we?

Christine Koppelt

Sonja Scheungrab



What is Machine Learning?


Data
Science
Machine
Learning
Artificial
Intelligence

Definition 1:

»Field of study that gives computers the ability to learn without being explicitly programmed.«

(Arthur Samuel, 1959)

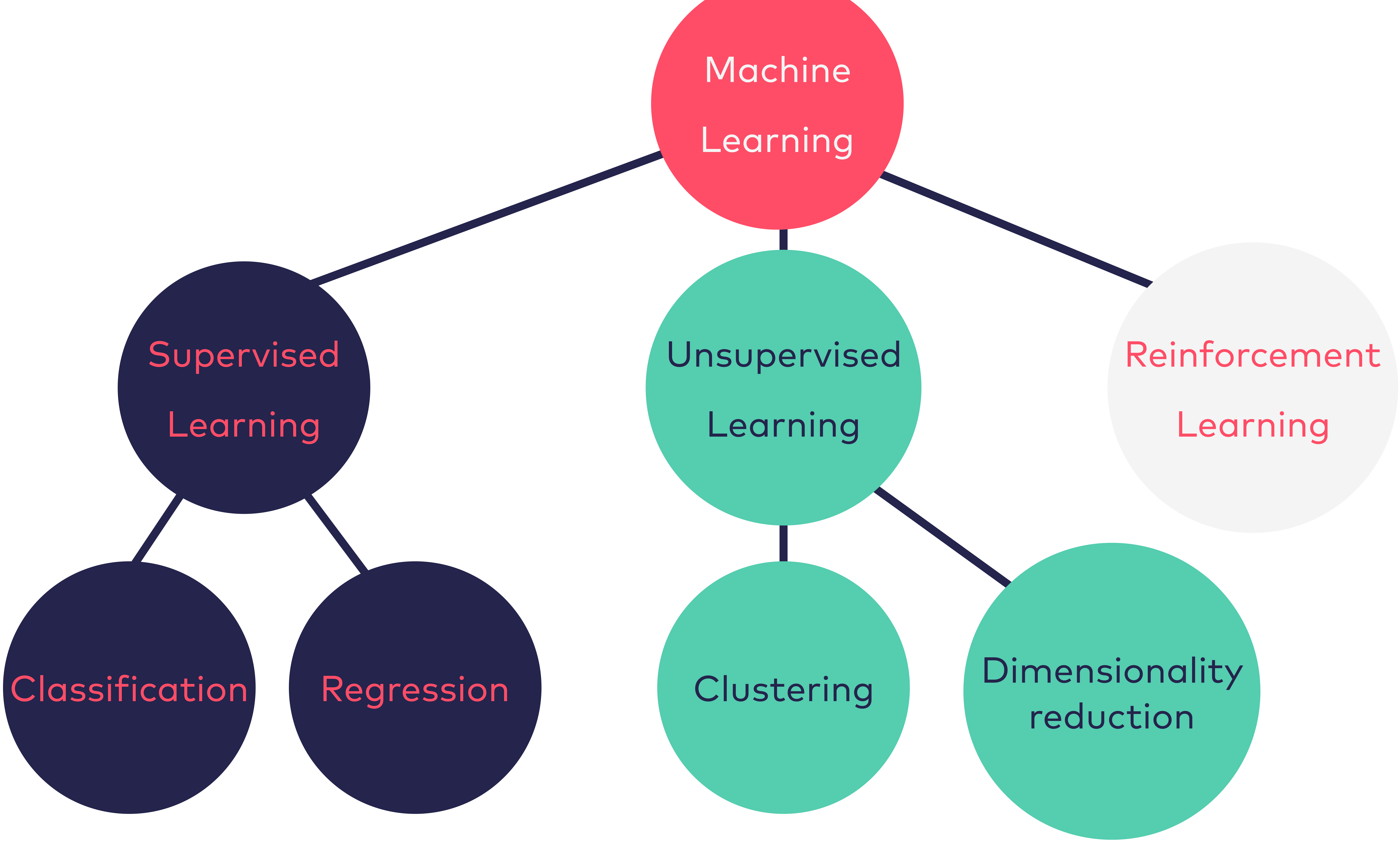
Definition 2:

»Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.«

(Tom Mitchell, 1998)

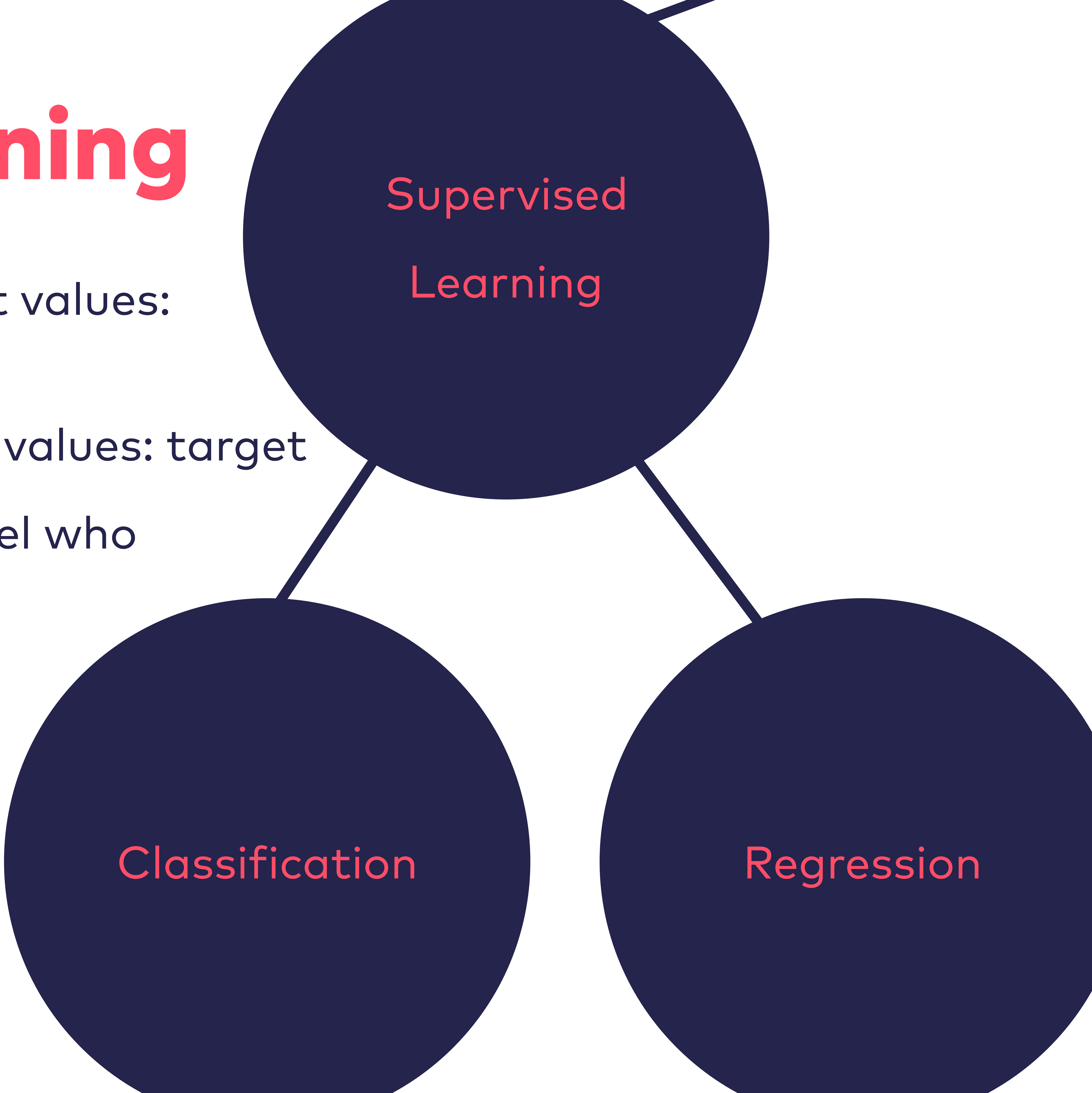
Plan for the Workshop

- Overview about the field of machine learning
- Focus on the practical work with useful frameworks
- Intuition about the background
- Analyze data together
- Experiment with different classification algorithms on that data.


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Classification
Regression
Clustering
Dimensionality reduction

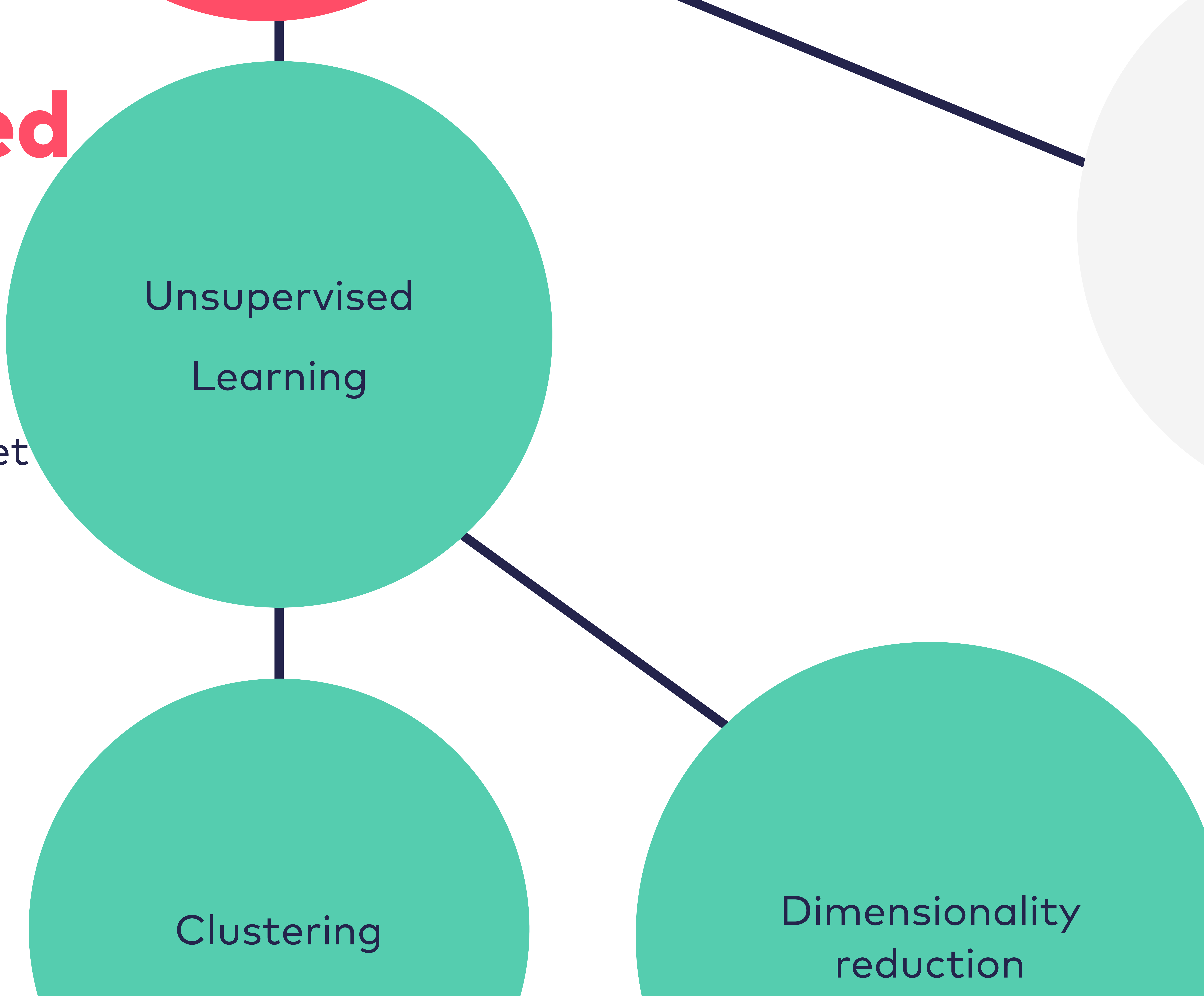
Supervised Learning

- Dataset with input and output values: "Correct answer" is given
- Input values: Features, output values: target
- We are "teachers" of the model who point out errors



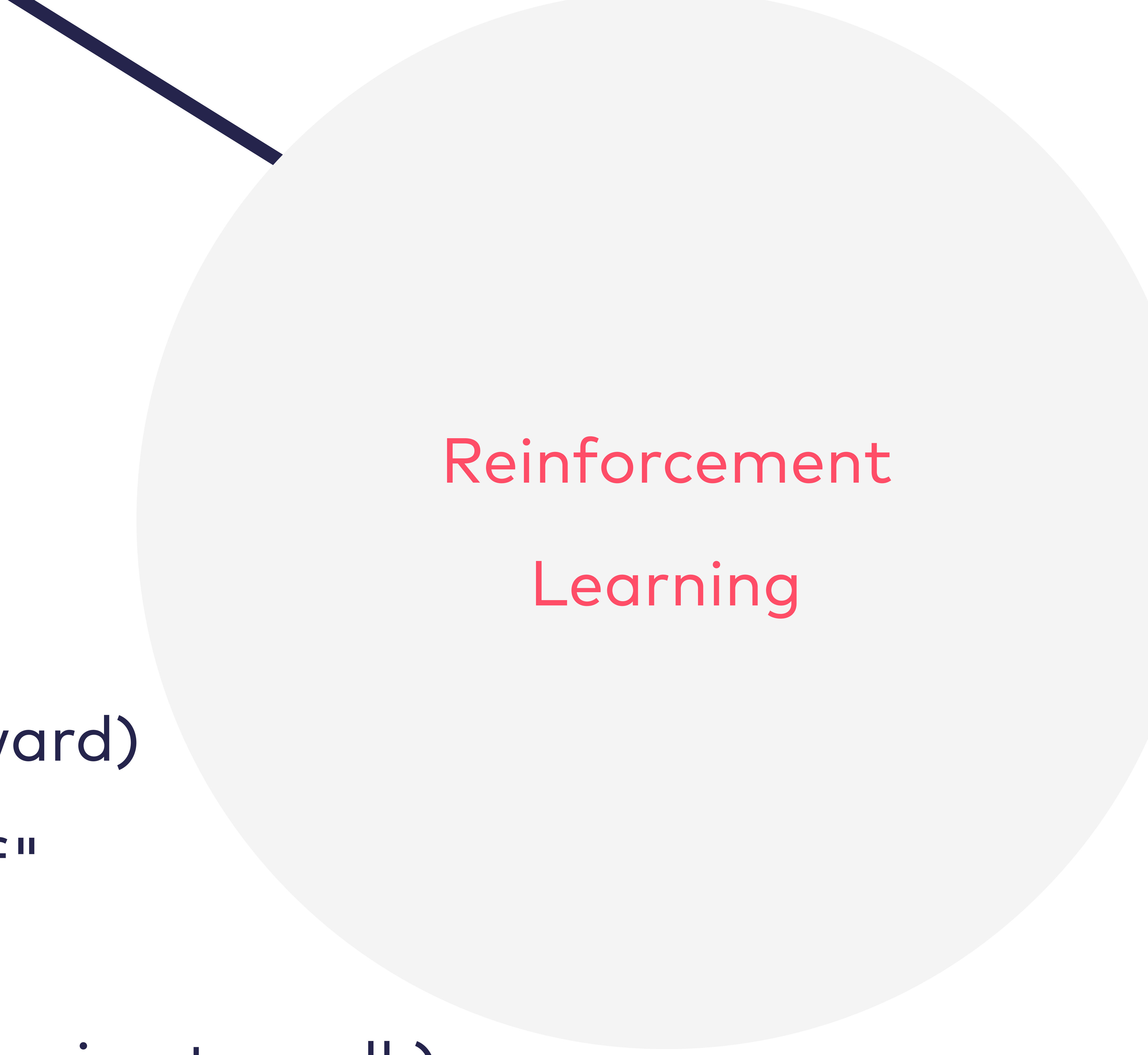
Unsupervised Learning

- No target values
- Find structure in dataset
- e.g clustering the data



Reinforcement Learning

- We don't tell our model **what** it did wrong, only **if** it did good or bad
- System of reward (punishment = negative reward)
- With that guidance the model learns "by itself" trying to maximize the reward
- Takes up "natural" learning strategies (e.g. learning to walk)



Reinforcement Learning

Classification

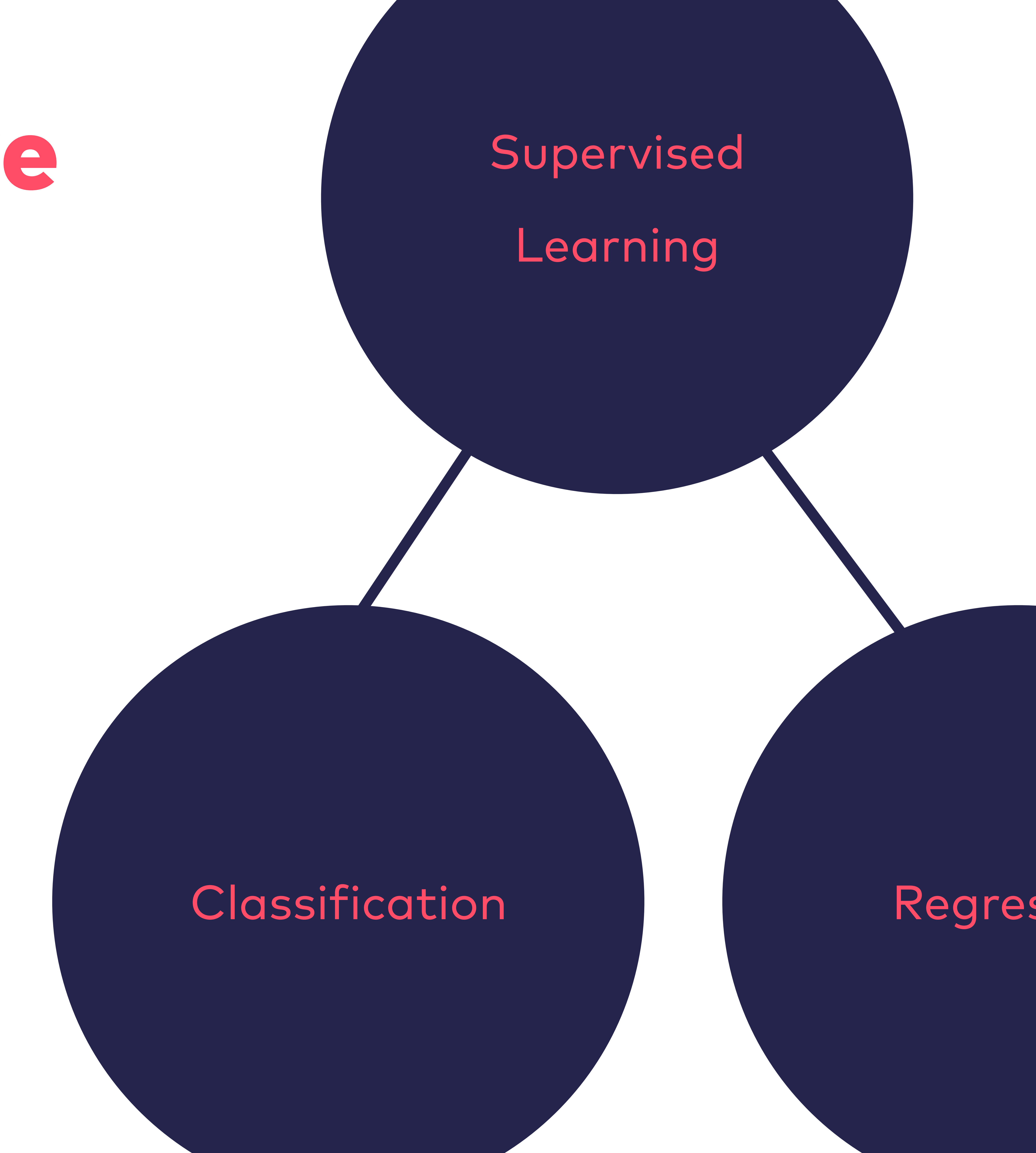
- Predict results in a discrete output
- Map input variables into two or more discrete categories: 🐱🐶🐭

Regression

- Predict results within a continuous output
- Map input variables to some continuous function
- House prizes, stock market…

The usual procedure

1. Choose your model
2. Instantiate your model with hyperparameters
3. Place your data in pattern matrix and a target vector = shape
4. Train the model on the data = fit
5. Check the trained model on unknown (=test) data = predict



Learning Process

- Examine why errors happened during training

size = 70 cm, pattern = stripes >> 🐶**?** 🐶⛔**!** 🐯✅

- Adapt the parameters/weights in order to get a better result

pattern = stripes >> 🐯

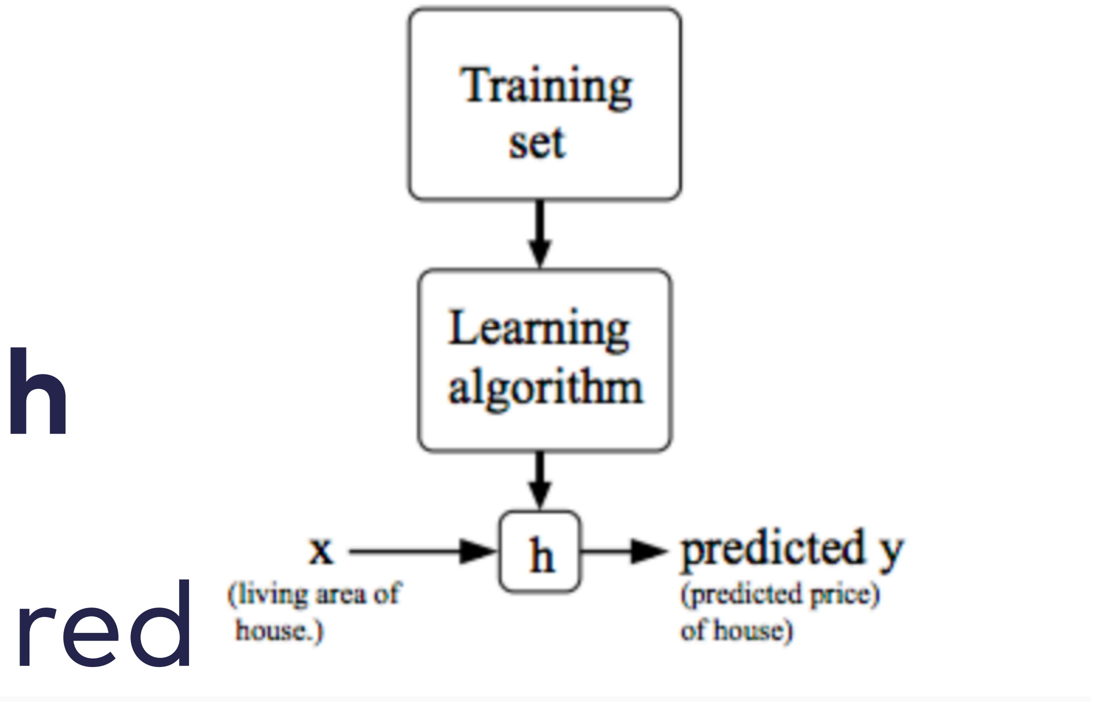
- Goal: Find a good general model we can use on unknown data

size = 70 cm, pattern = stripes >> 🐯

What happens exactly?

Maths magic ✨

- Goal: find a "good" hypothesis function **h**



Training set

Learning algorithm

x (living area of house.)

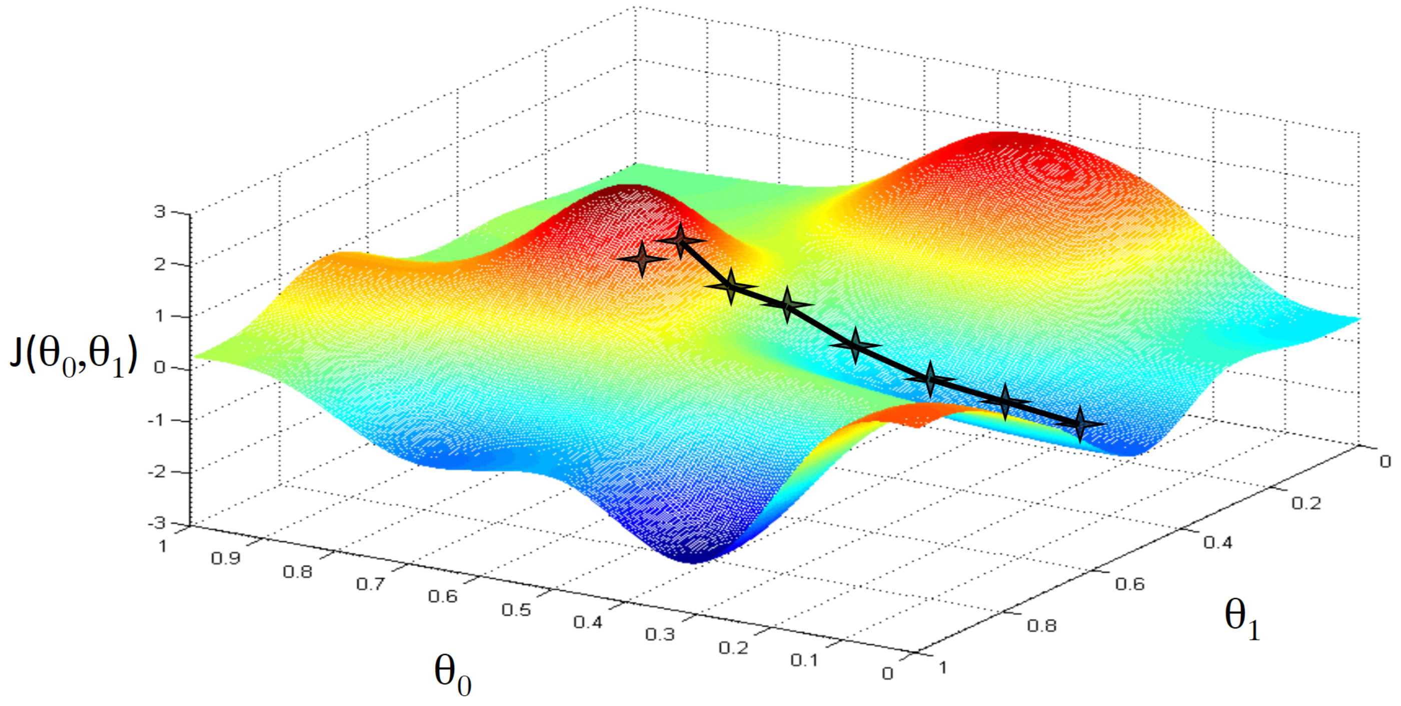
h

predicted y (predicted price) of house)

- The accuracy of our hypothesis is measured with a **cost function** like the **squared error function** that takes an average difference of all the results of the hypothesis with inputs from x's (= y "hat") and the actual output y's

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}_i - y_i\right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i\right)^2$$

- We minimize the error with **gradient descent:** We tal the derivative of our cost function, which tells us hov to update our theta-weights

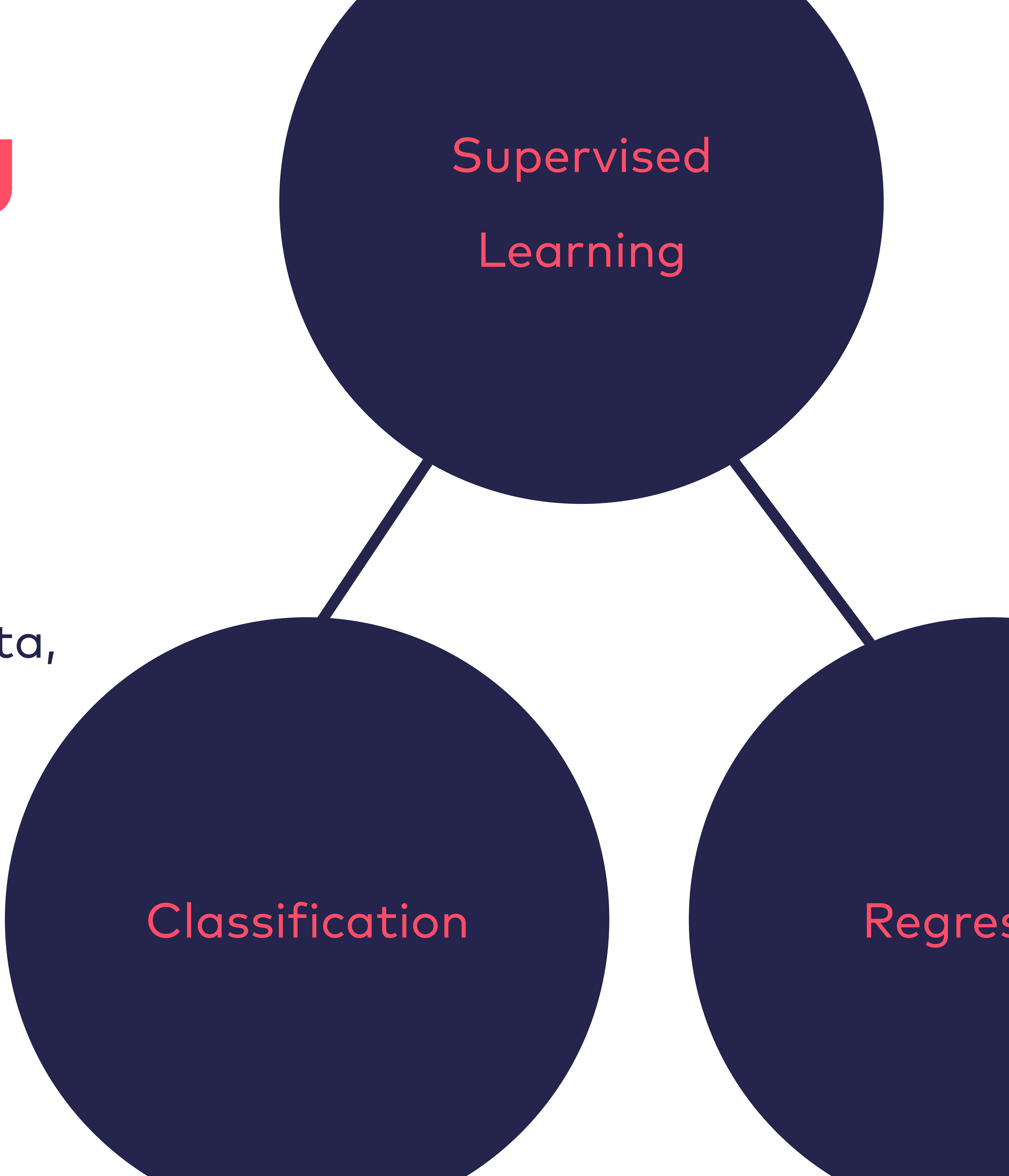


In other words

- Predict …
- Check, how bad we did …
- Minimize our error …
- Repeat.

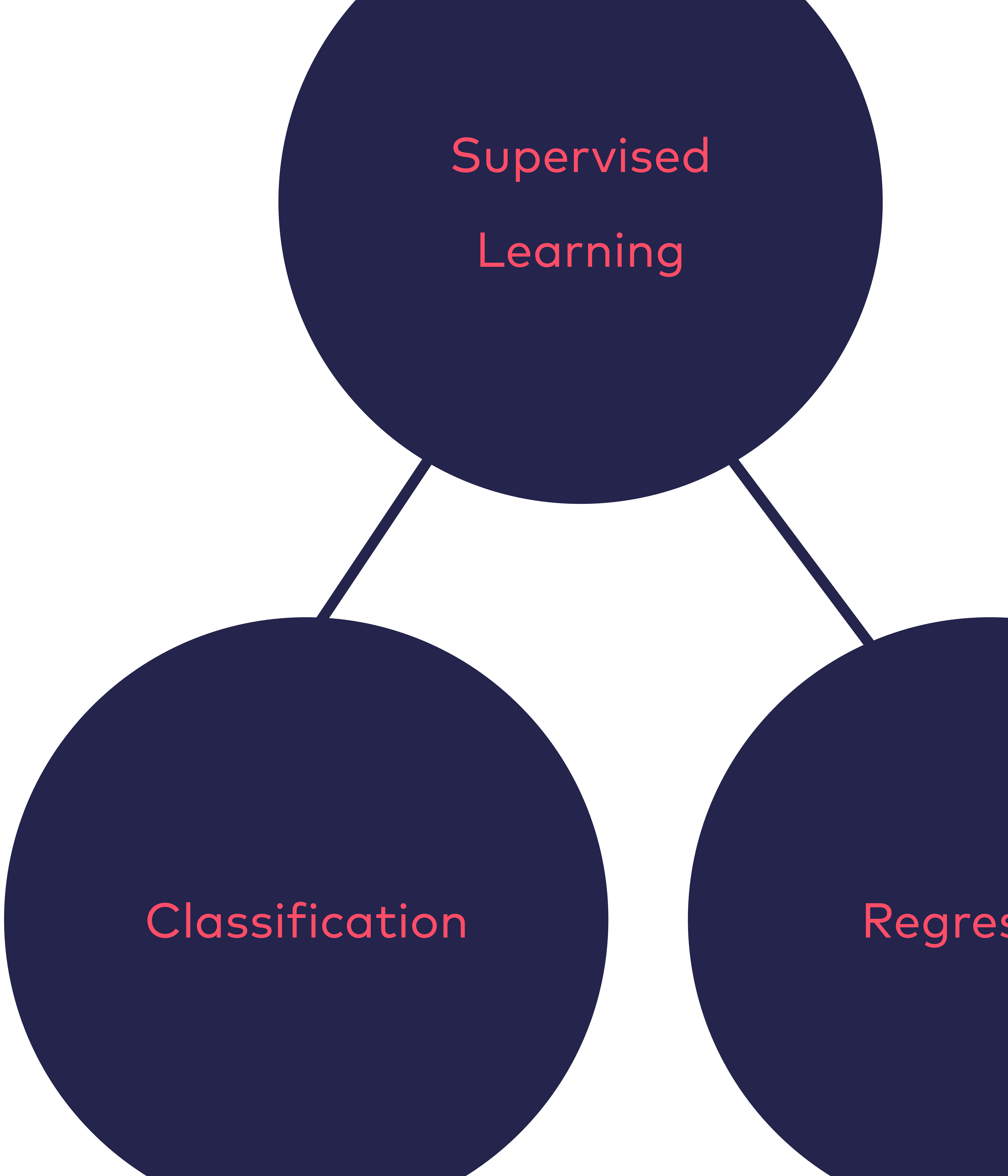
Training and Testing

- Split dataset in training and test sets
- e.g. 80 % training data, 20 % test data
- Remove the output/target values of test data
- "Train" the model with the trainings data, until we have a satisfying algorithm (good match with our data targets)
- Test the trained model on the unknown test data

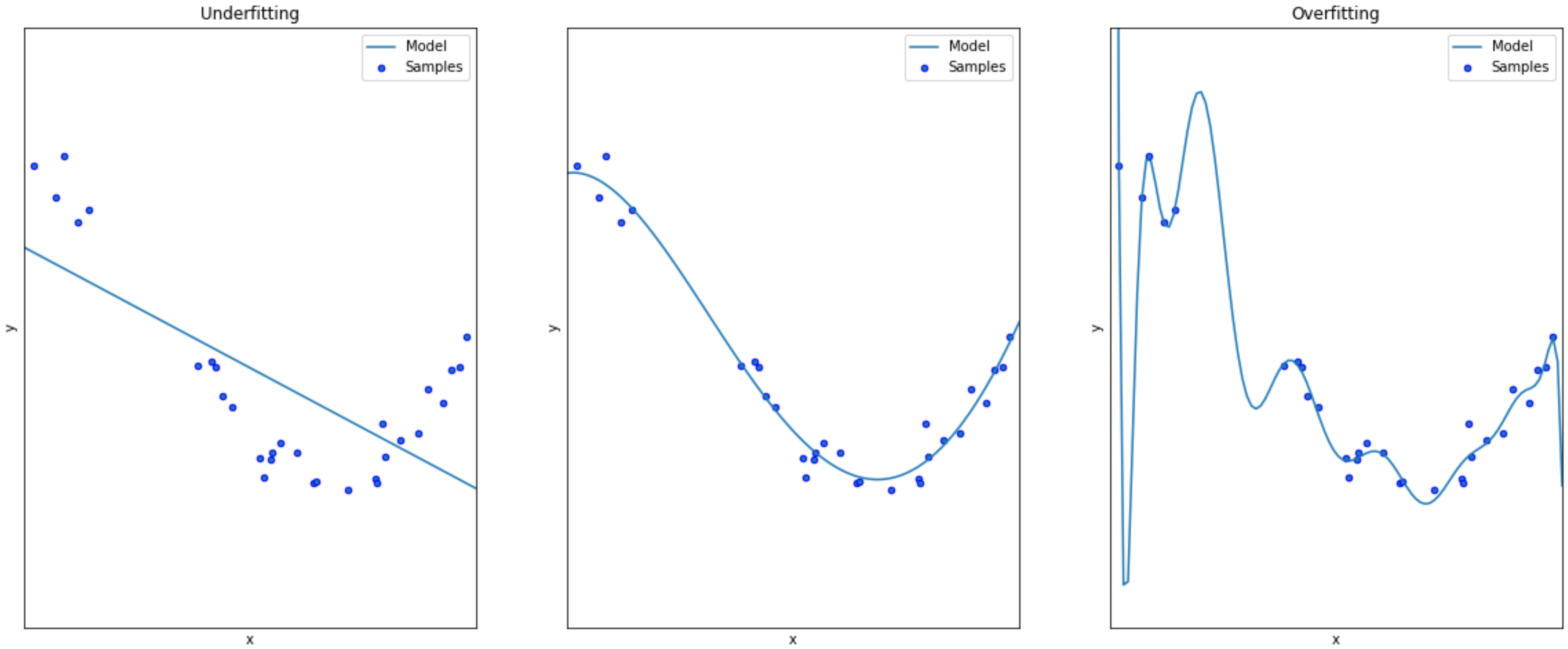


Problems

- Overfitting
- Underfitting
- Bad data quality
- Small data quantity

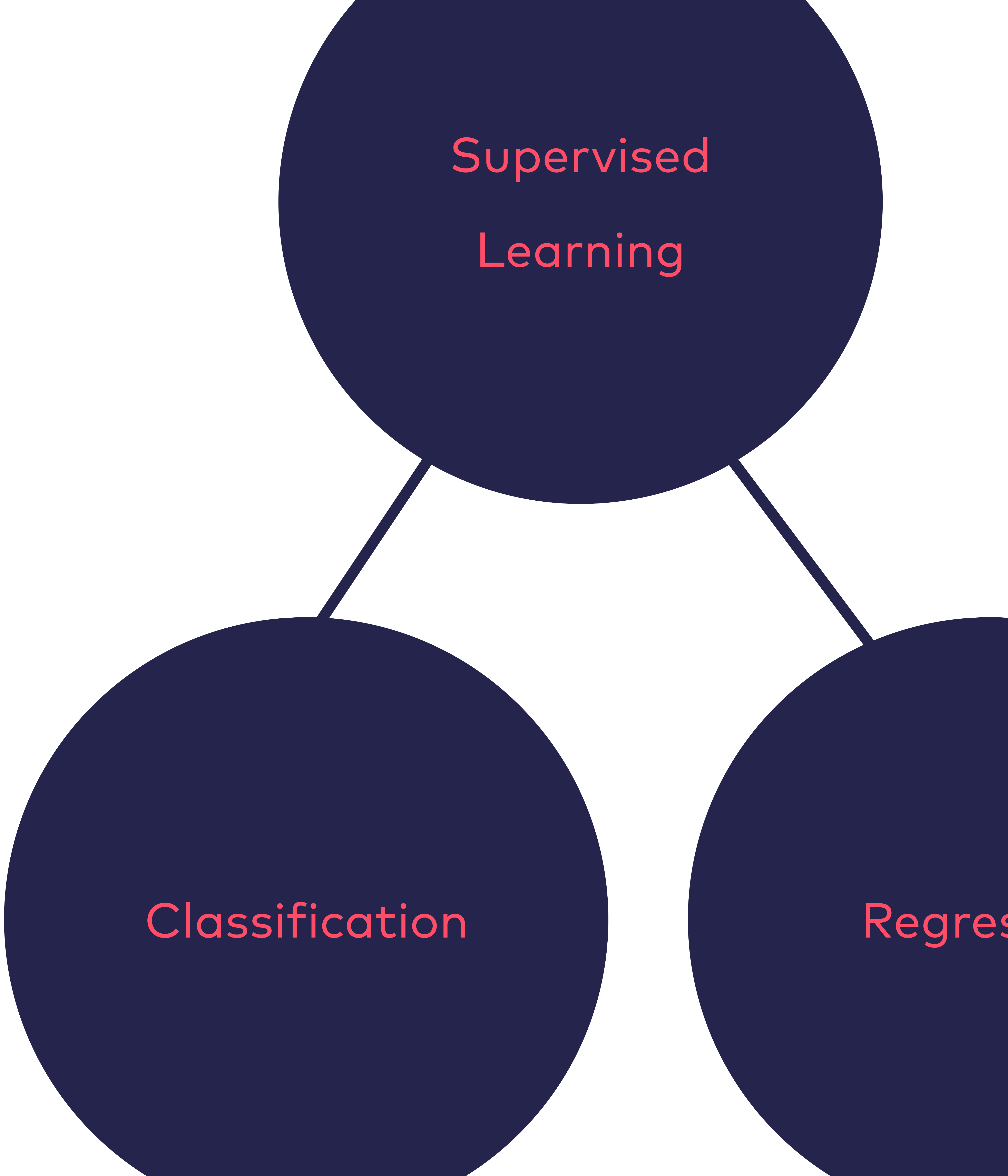


Underfitting and Overfitting



Classification Algorithms

- Logistic Regression
- K-nearest neighbor
- Support Vector Machine
- Decision Tree / Random Forest


Supervised Learning
Classification
Regres