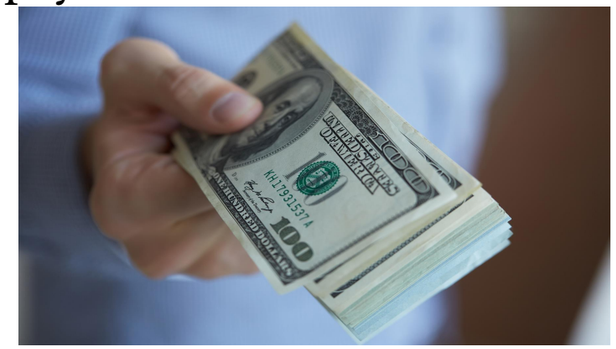
Supervised Learning

Setting

Problem: Given the loan history of customers, predict if a new customer will payback loan or not.



FEATURES

LABEL

Name	Age	Gender	Education	Salary	Married	Location	Decision
John	29	M	M.S	$12500	Yes	Urban	YES
Mary	27	F	Ph.D	$30000	No	Rural	YES
Smith	36	M	B.S	$4500	Yes	Rural	NO
Hari	*28*	*M*	*B.S*	*$16000*	*No*	*Urban*	???

Past customer history

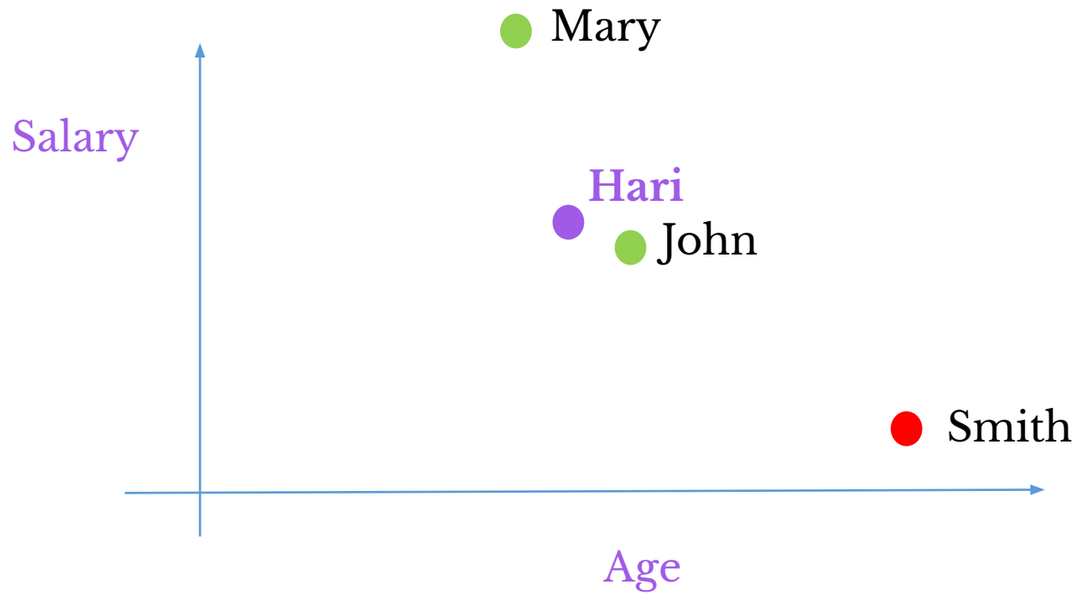
Hari → [ ] → YES / NO

Feature, label pairs of customers

Feature of new customer Hari → [ ] → YES / NO

Algorithms – K-NN

Name	Age	Salary	Decision
John	29	$12500	YES
Mary	27	$30000	YES
Smith	36	$4500	NO
Hari	*28*	*$16000*	???

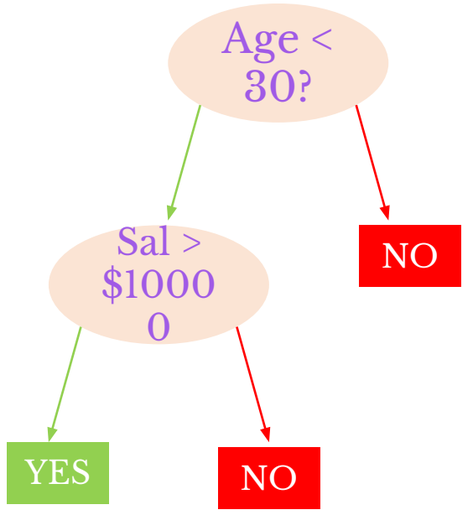


Algorithm: Decide based on the decision of the majority of the **k-nearest neighbours**

Most common issues: 1. Choosing k.
2. Choosing distance measure
3. Storing all data points.

Algorithms – Decision Trees

Name	Age	Salary	Decision
John	29	$12500	YES
Mary	27	$30000	YES
Smith	36	$4500	NO
Susan	24	$6000	NO
Hari	*28*	*16000$*	???



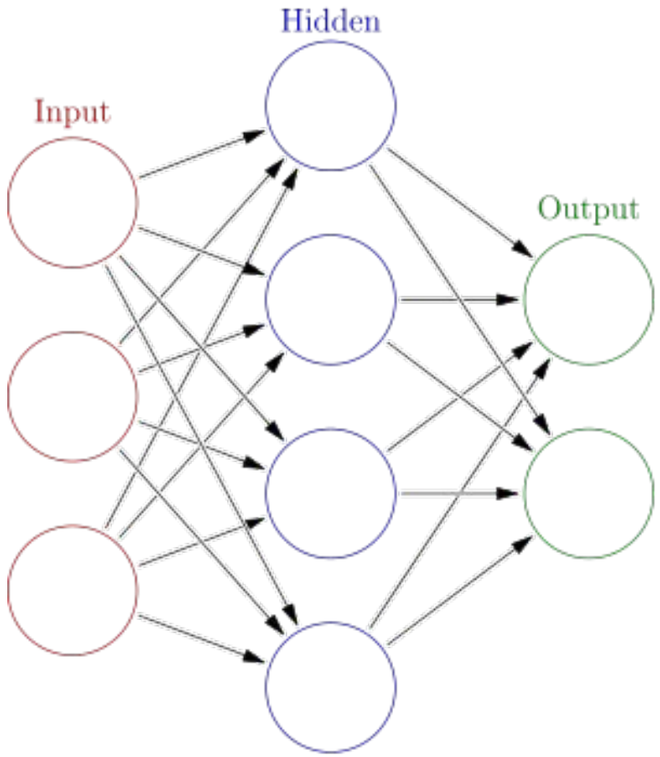
Algorithm

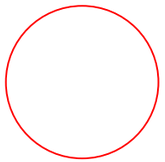
- Find most informative feature, threshold.
- Split data according to chosen feature
- Repeat

Measures of impurity
- **Gini index**
- **Information gain** etc.

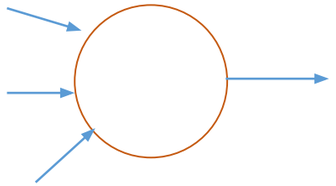
Most common issues:
1. Choosing depth of tree.
2. Choosing information measure, threshold.

Neural Networks

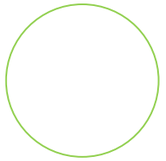




Feature value

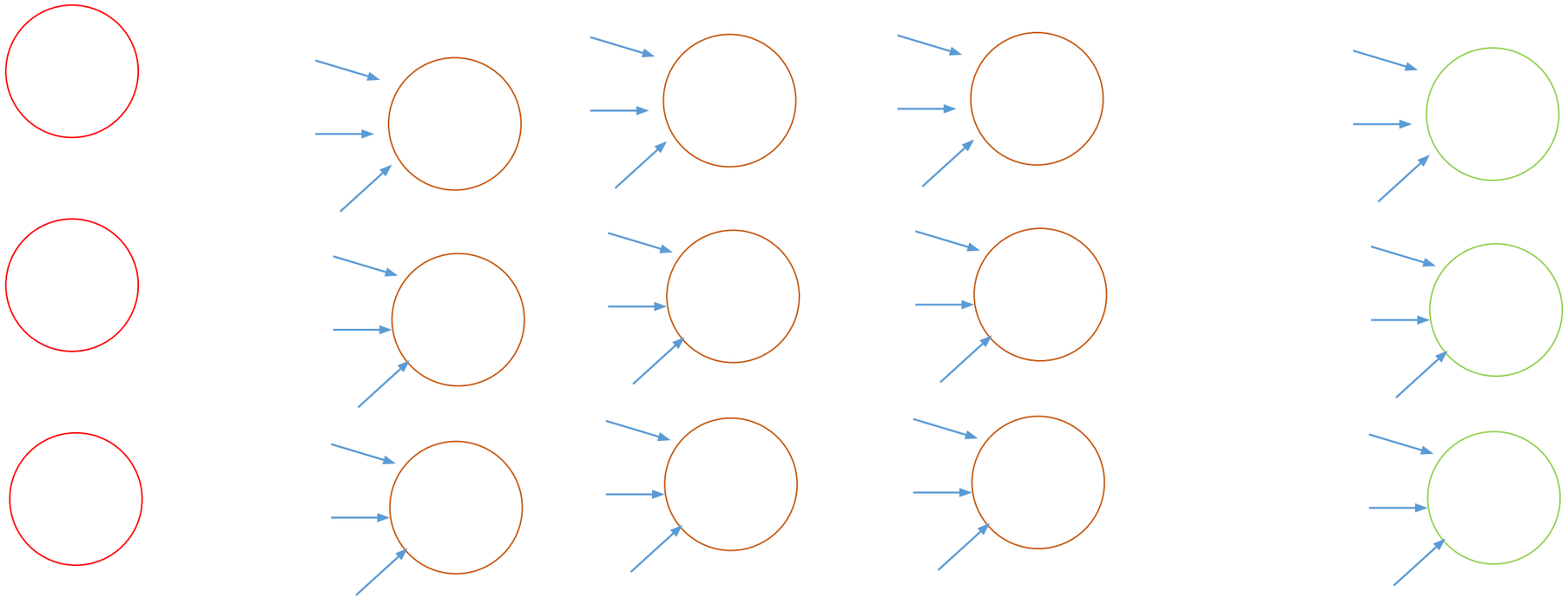


Aggregation of input values

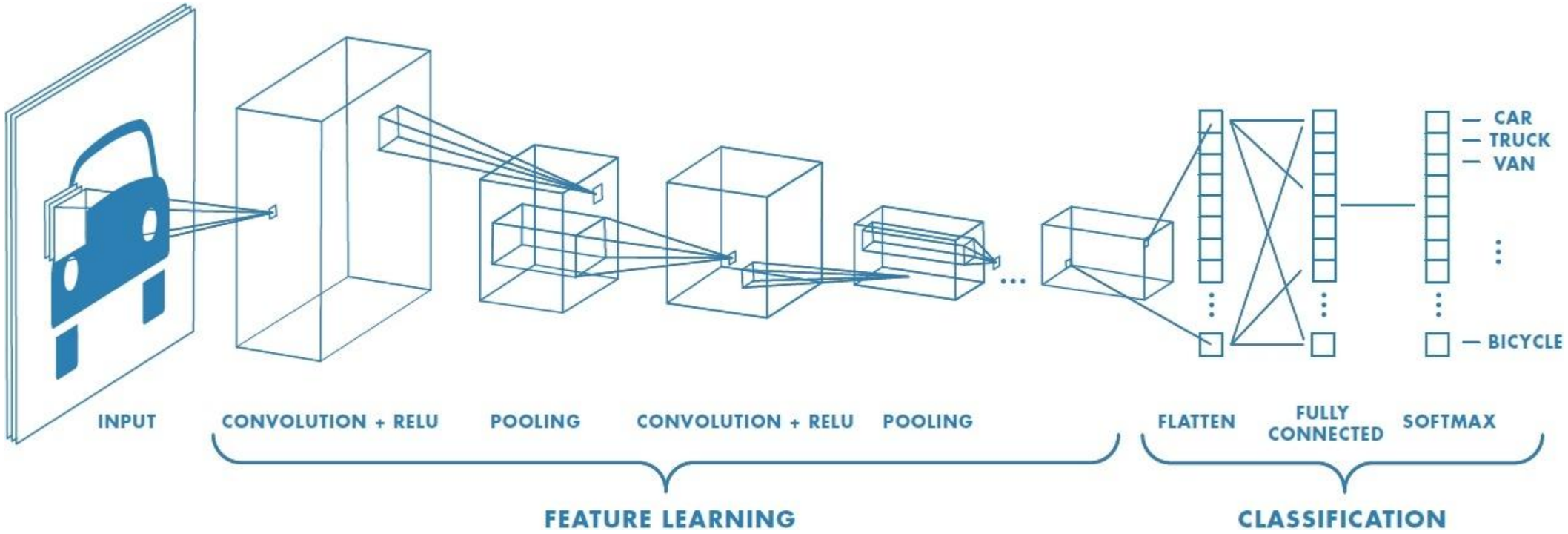


Label probability output

Multi Layer Neural Networks



Convolutional Neural Network

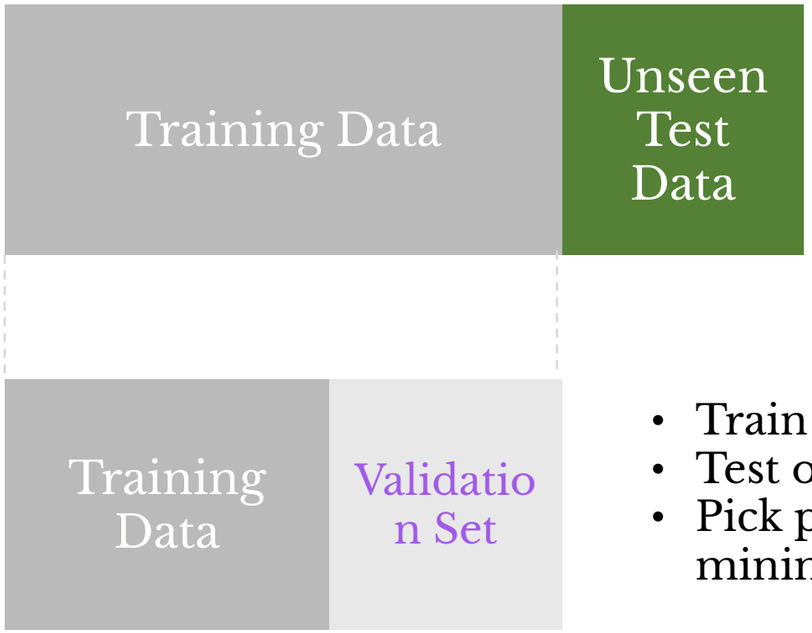


Source: mathworks.com

Over/Under fitting

- Effect of K in K-NN
- Effect of depth in decision tree
- Effect of number of hidden layers in neural networks

Model Selection: Cross Validation



- Train on Training data
- Test on validation data
- Pick parameters (like k, depth etc) that minimizes validation error.

- K-fold cross validation typically improves performance.

Up next:

Online Learning

Experts Problem

At the end of each day, the stock status is known
after a decision (to invest or not) is made.

What algorithm to decide? And how many mistakes?