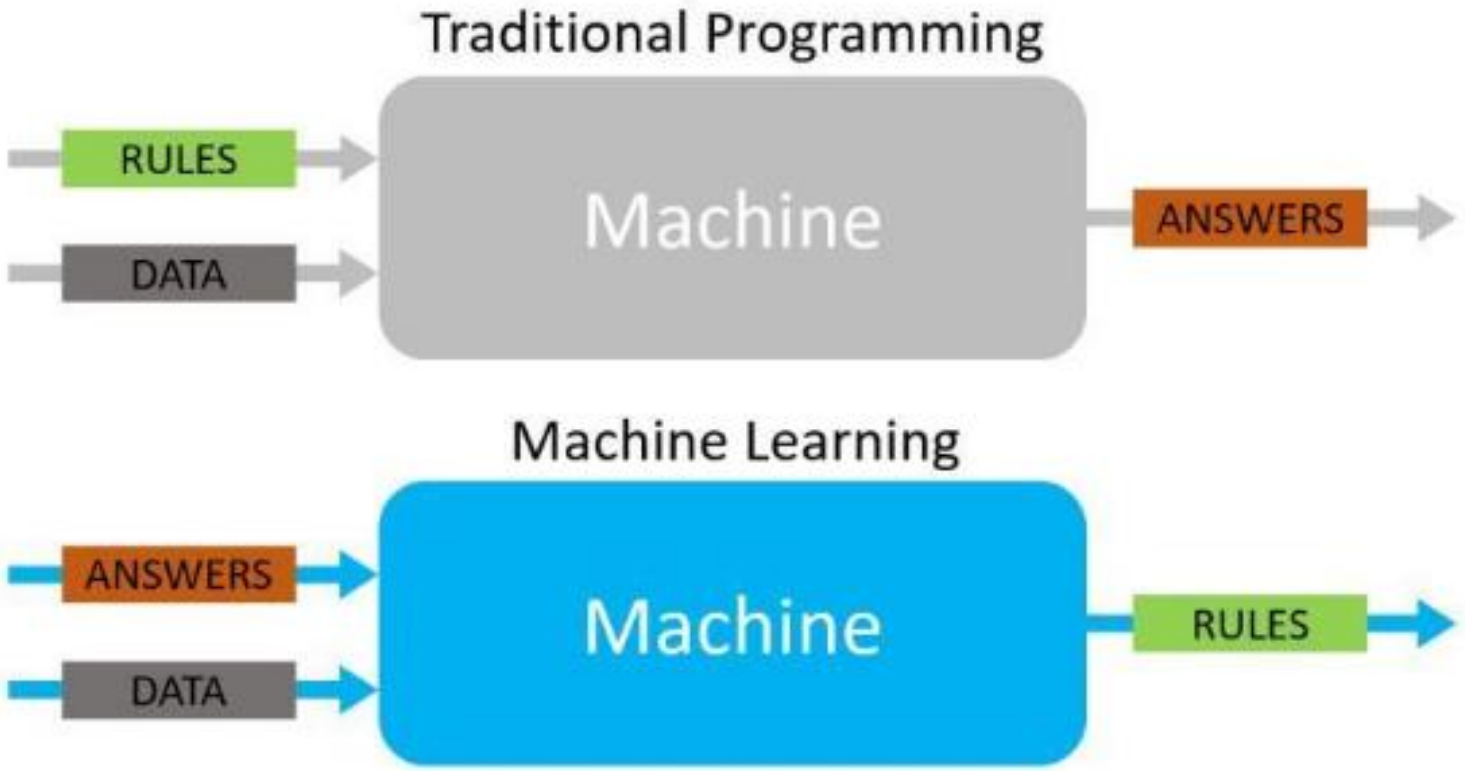
IR4
Fondamentaux de l'IA

Module 2 Leçon 1
Le Machine Learning

ML Bontemps

Le Machine Learning



Le ML est utile …

Problème complexe

- Le problème n'est pas modélisable simplement.
- Il comprend un grand nombre de paramètres.

Environnement changeant

- Quand il faut modifier/ajuster en permanence les règles de fonctionnement ou les algorithmes.
- Dans un environnement évolutif.

Big data

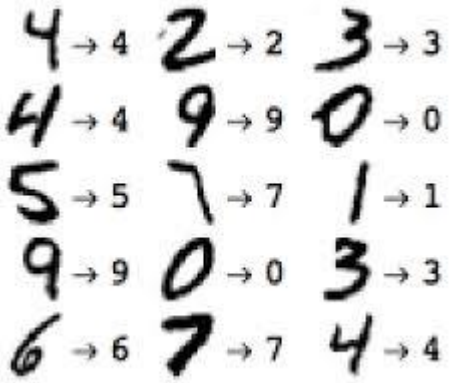
- Grand volume de données.
- Données très diverses.
- Rapidité d'acquisition des données.

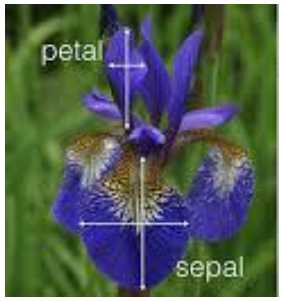
Un projet ML

- Définition du besoin: que veut-on résoudre ?
- Les données sont la source de tout : leur qualité est essentielle
  - Identifier les données nécessaires et leur lieu de stockage.
  - Différentes sources de données peuvent être utilisées.
- Le Data Scientist choisit l'algorithme adéquat.
- Le modèle est mis à la disposition de tous les utilisateurs.
  - Automatisation du maximum de tâches et de transferts de données.
  - Maintenance du modèle et mise à jour avec les nouvelles données.

Result = prediction / classification / recommendation

- Predict
  - Yields in production
  - Machine breakdown
  - Risks / trust in banking operations or customers
  - + Anomaly detection





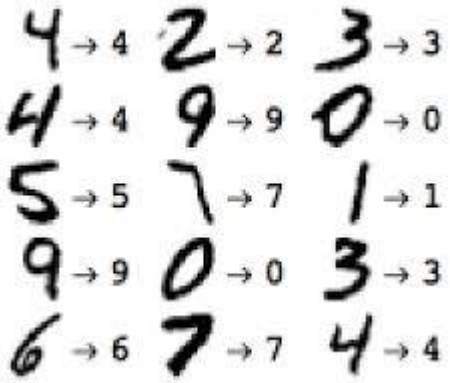
- Classify
  - Documents
  - Mails : spam sorting
  - Images
    - For images themselves
    - For supervision and process monitoring
    - For manual scripting recognition

- Recommendation
  - Products to customers
  - Chatbots

Spam filter :
1. Preparing the text data.
2. Creating word dictionary.
3. Feature extraction process.
4. Training the classifier.

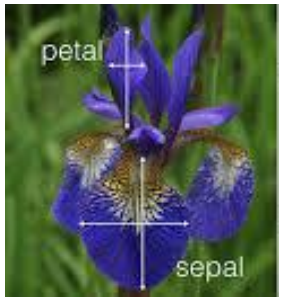
Result = prediction / classification / recommendation

- Predict
  - Yields in production
  - Machine breakdown
  - Risks / trust in banking operations or customers
  - + Anomaly detection



- Recommendation
  - Products to customers
  - Chatbots

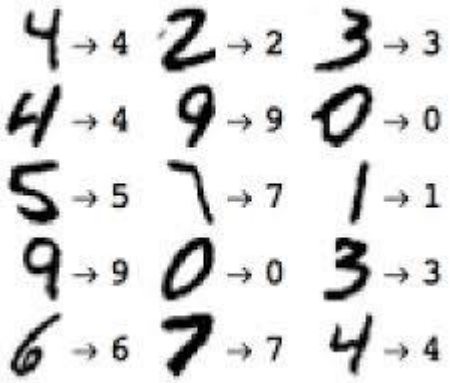
- Classify
  - Documents
  - Mails : spam sorting
  - Images
    - For images themselves
    - For supervision and process monitoring
    - For manual scripting recognition



Document classification :
1. Preparing the text data + tokenization, lemmatization.
2. Creating word dictionary.
3. Feature extraction process : terms frequency.
4. Training the classifier : clustering.

Result = prediction / classification / recommendation

- Predict
  - Yields in production
  - Machine breakdown
  - Risks / trust in banking operations or customers
  - + Anomaly detection



- Recommendation
  - Products to customers
  - Chatbots

- Classify
  - Documents
  - Mails : spam sorting
  - Images
    - For images themselves
    - For supervision and process monitoring
    - For manual scripting recognition

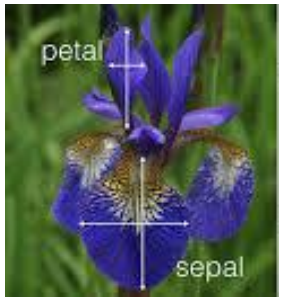
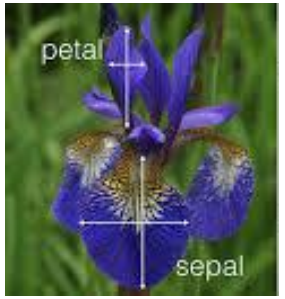


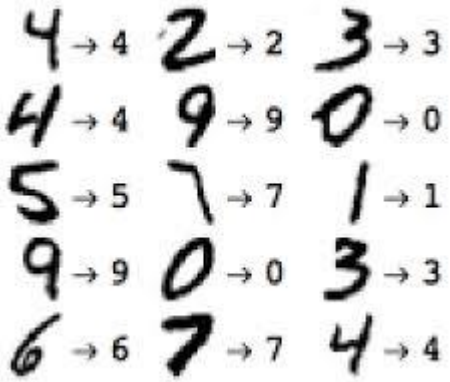
Image recognition :
1- Take a patch of the image
2- Slide the patch along the image

OCR = Optical Character Recognition
1- Text detection in the image
2- Characters segmentation
3- Characters classification
4- Eventually, spelling correction

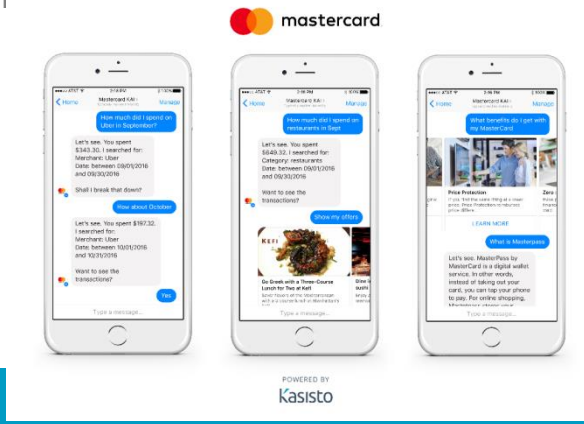
Result = prediction / classification / recommendation

- Predict
  - Yields in production
  - Machine breakdown
  - Risks / trust in banking operations or customers
  - + Anomaly detection





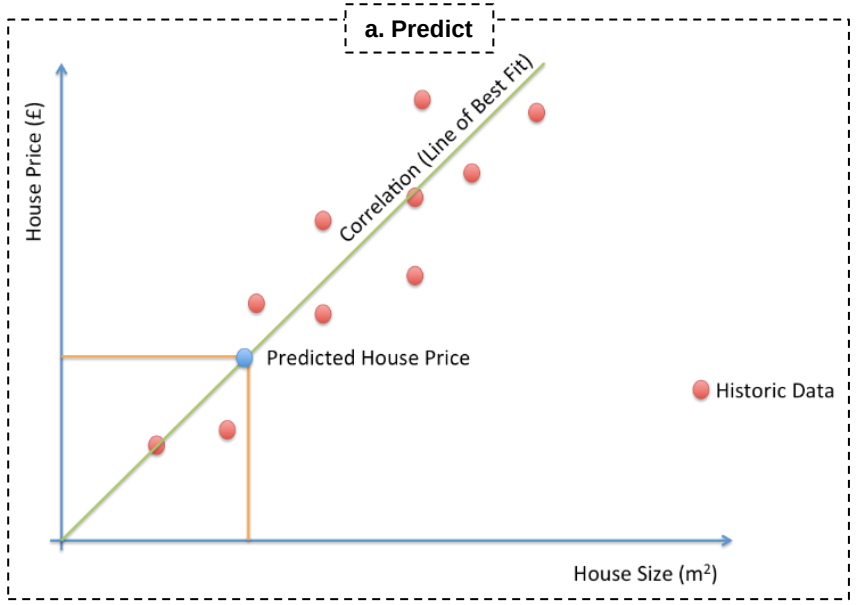
- Classify
  - Documents
  - Mails : spam sorting
  - Images
    - For images themselves
    - For supervision and process monitoring
    - For manual scripting recognition



- Recommendation
  - Products to customers
  - Chatbots

Supervised learning

- Predictive model : Y= Θ X.
- Labelled data



Examples:

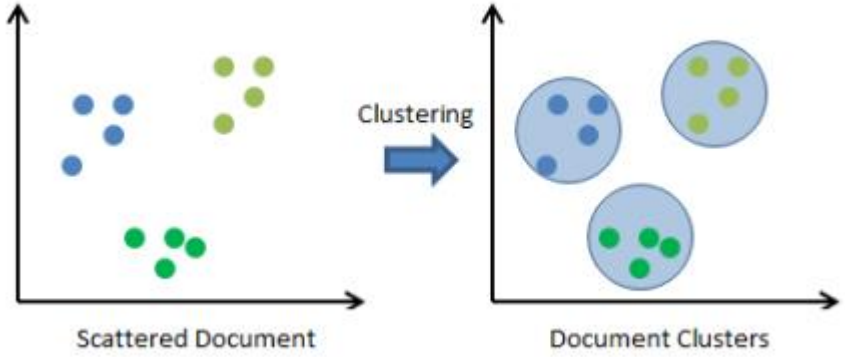
- Regression
- Decision Tree
- Random Forest
- Logistic regression

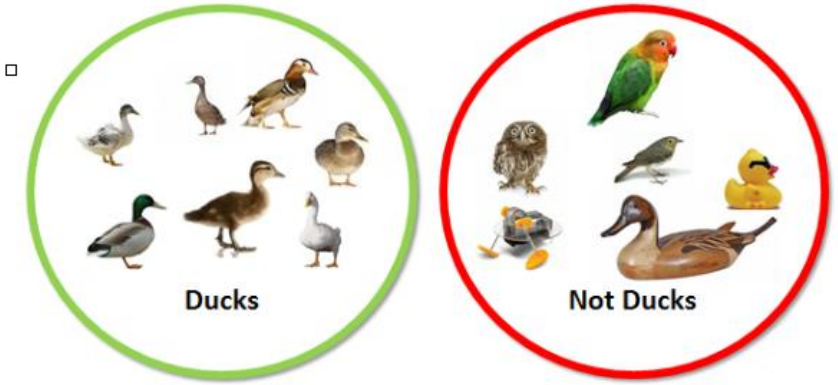
Unsupervised learning

- →No output categories or labels

- Pattern detection :
  identify useful associations within data

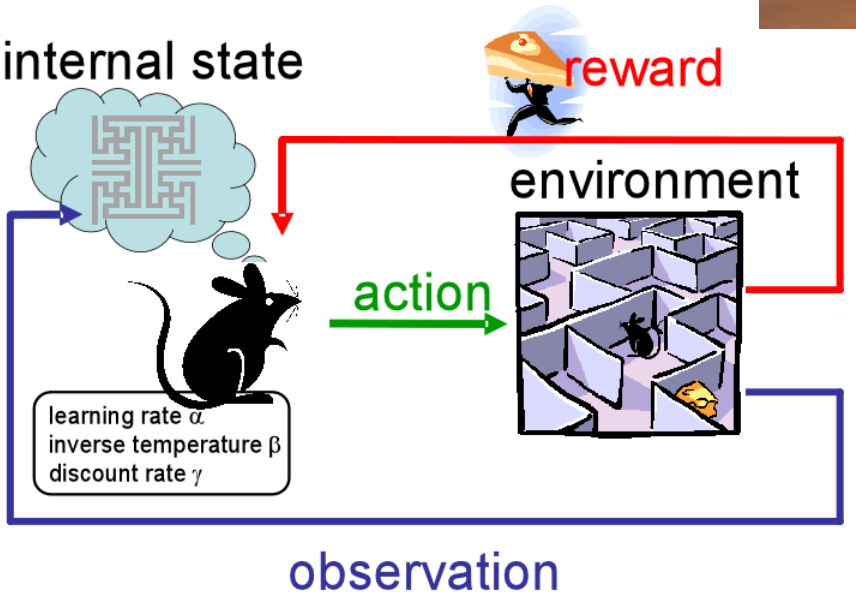
- Descriptive modeling : dividing a dataset into homogeneous groups = clustering



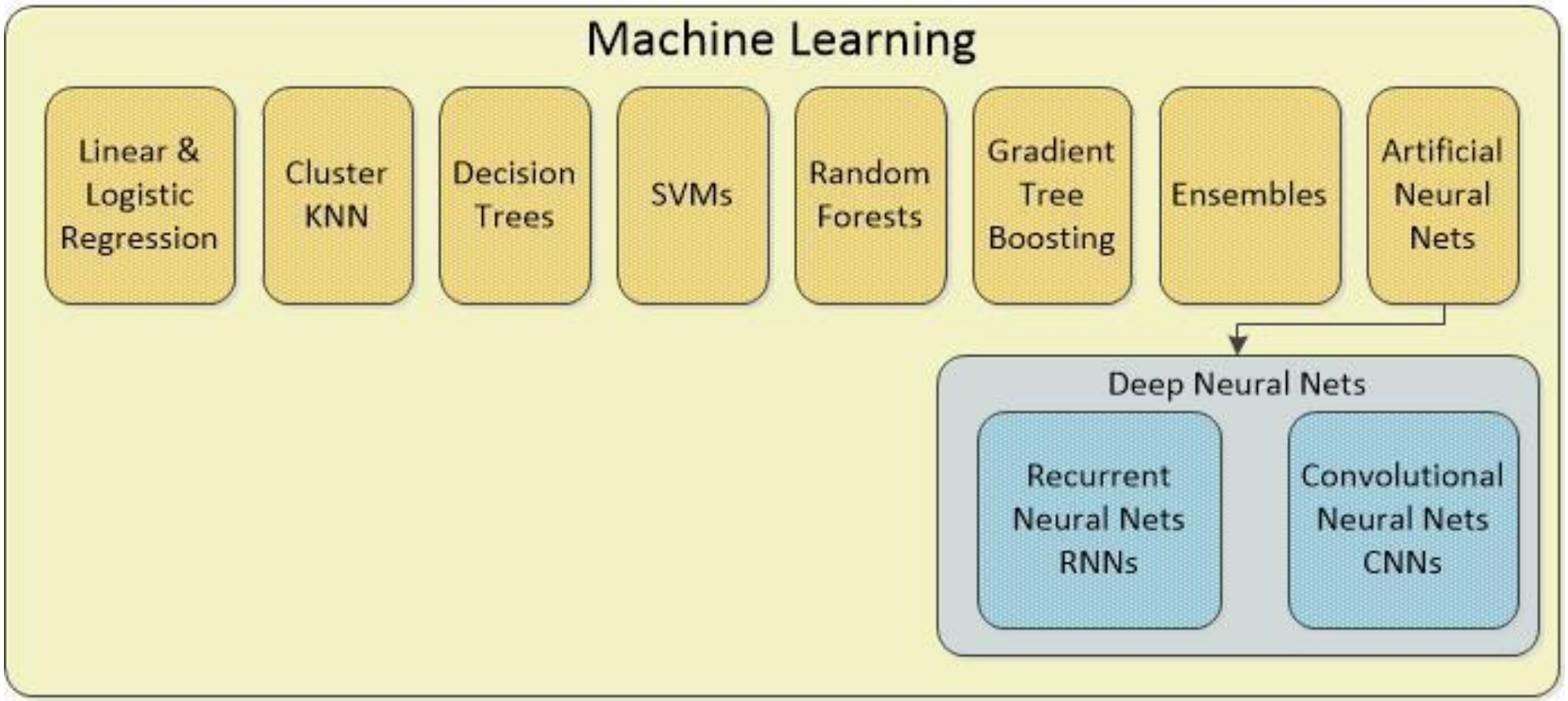


Reinforcement learning

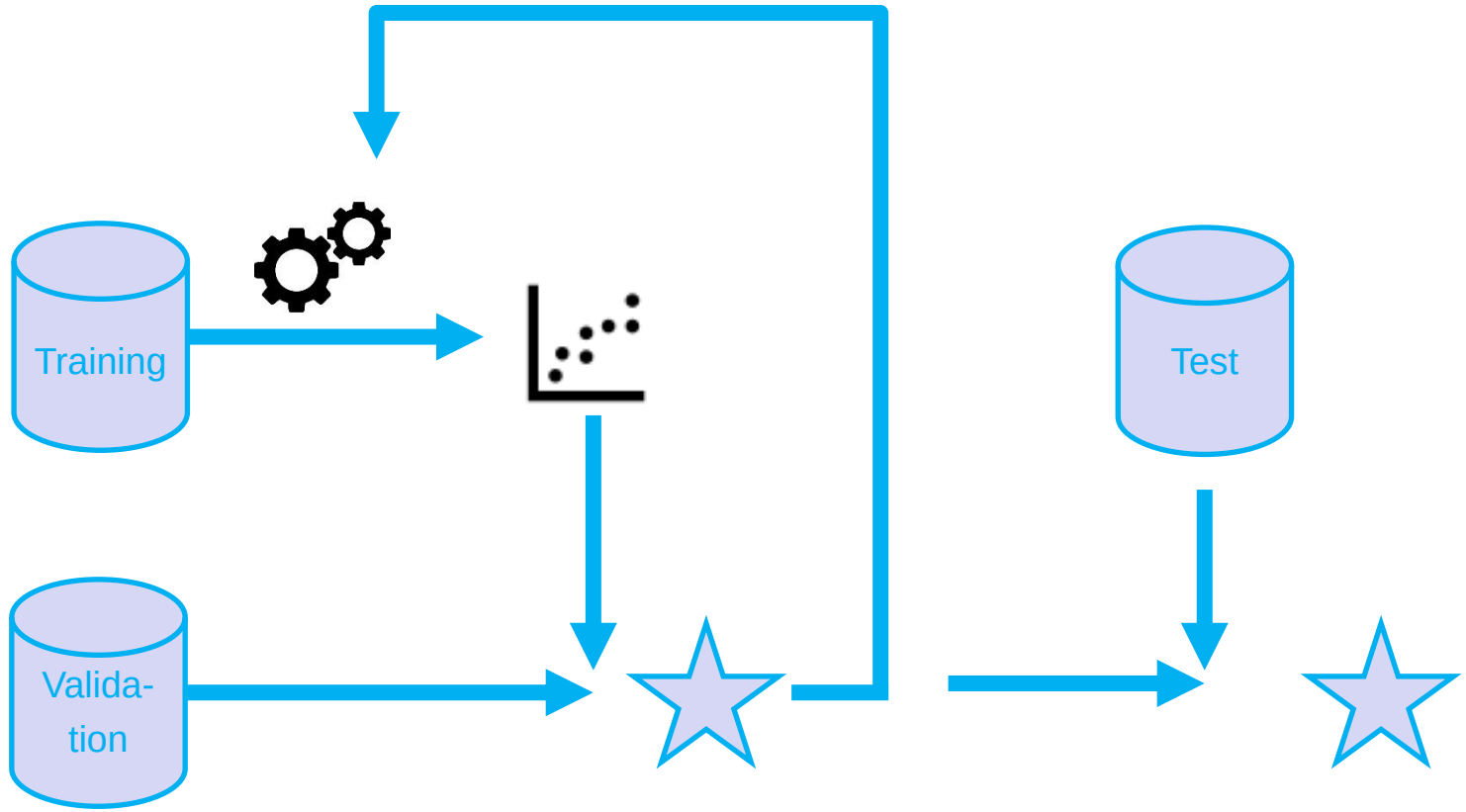
- The machine is trained to make specific decisions.
- Driverless cars
- Self navigating vaccum cleaners
- Example: Google Deepmind Lab



Common ML algorithms



Training, Validation, Test sets



Training Dataset:
The sample of data used to fit the model.

Validation Dataset:
The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as a skill on the validation dataset is incorporated into the model configuration.

Test Dataset:
The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.