



INTRODUCTION TO MACHINE LEARNING

MUSTAFA ALDEMIR, INTEL TURKEY

AI IS THE NEW ELECTRICITY

«Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years.»

Dr. Andrew Ng



OUTLINE

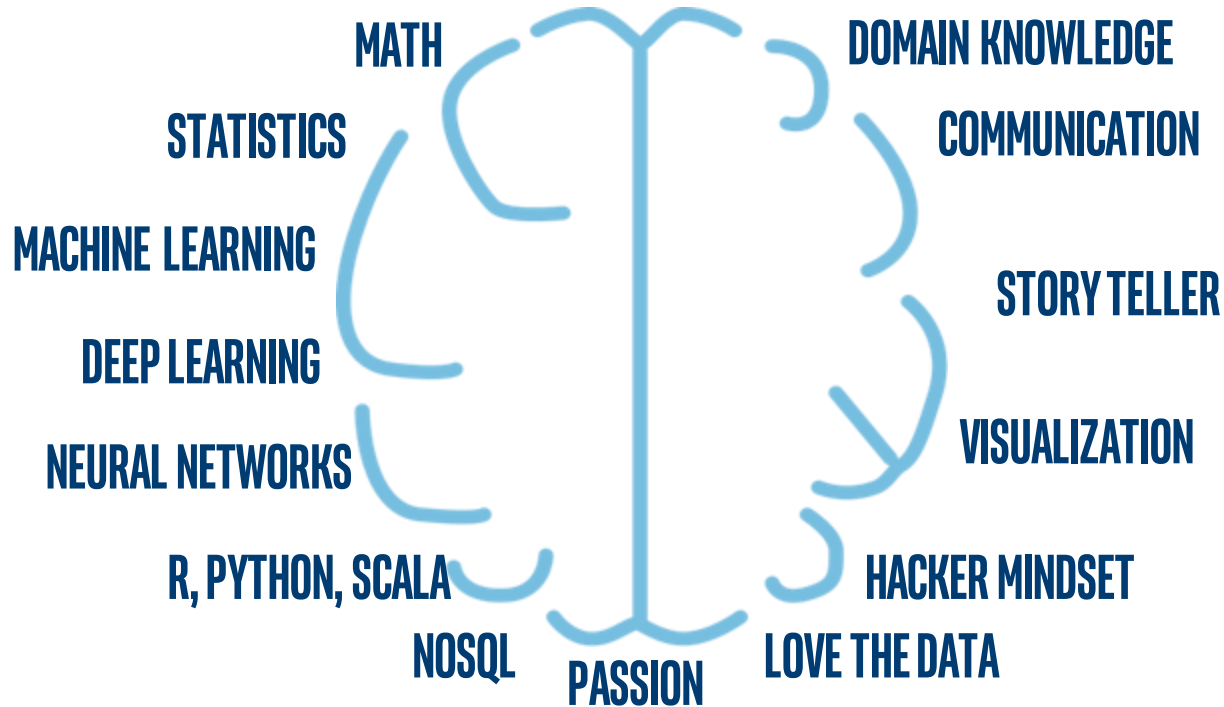
- Introduction to Data Science
- Introduction to Machine Learning
 - Supervised Learning
 - Unsupervised Learning
- Some Implementation
- Q&A

- Introduction to Deep Learning
 - Artificial Neural Networks
 - Convolutional Neural Networks
- Intel Deep Learning Training Tool
 - Installing
 - Using
- Q&A

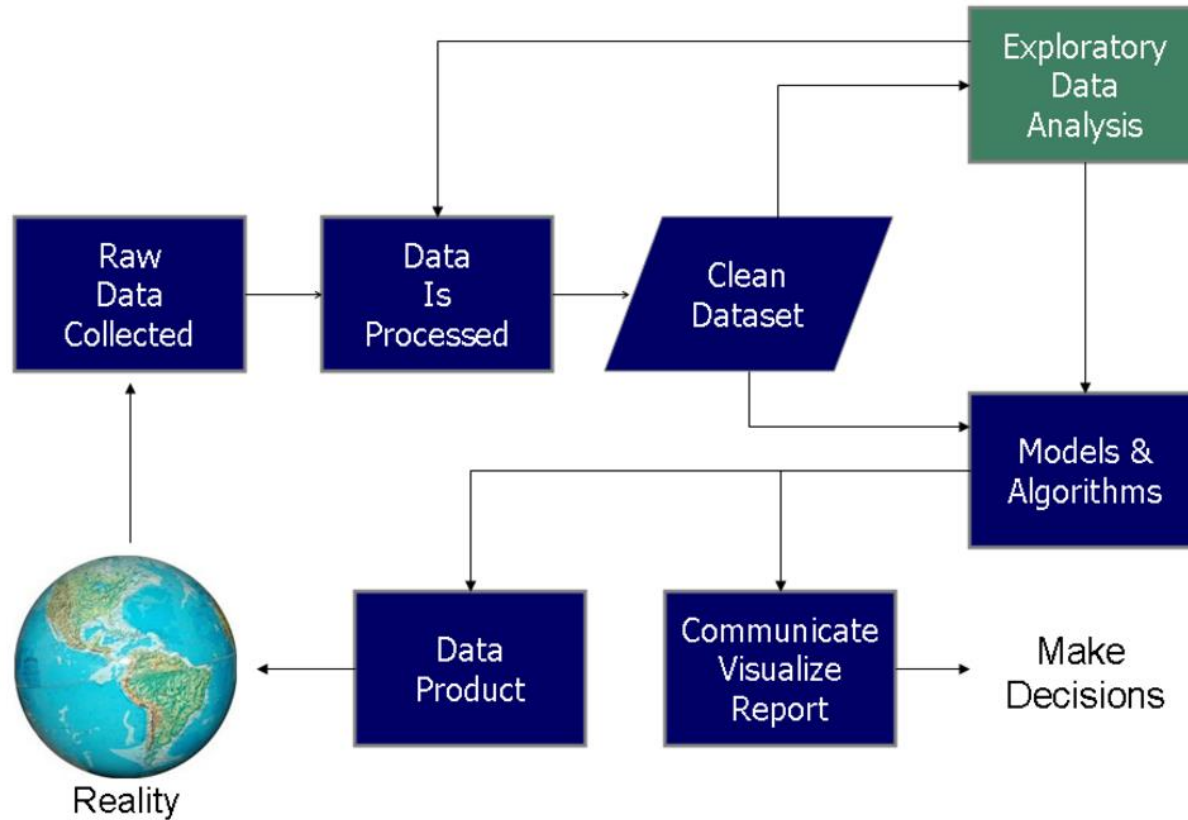
WHAT IS DATA SCIENCE?

The science of extracting knowledge and information from data and requires competencies in both statistical and computer-based data analysis.

HOW TO BECOME A DATA SCIENTIST?



The Data Science Process



Source: https://en.wikipedia.org/wiki/Data_science

DAILY DATA GENERATION IN 2020



1.5GB



3,000GB



4,000GB



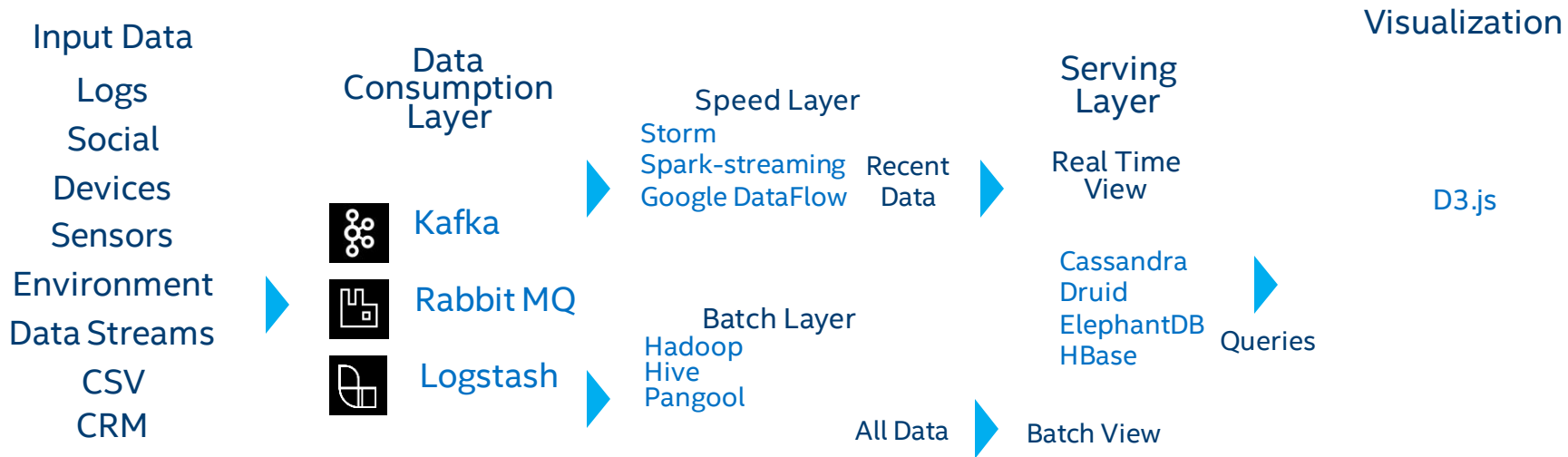
40,000GB



1,000,000GB



DATA SCIENCE - INGESTION TO VISUALIZATION



WHAT IS ARTIFICIAL INTELLIGENCE?

«The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.»

The Oxford Dictionary

AI IS TRANSFORMING INDUSTRIES



CONSUMER

Smart Assistants
Chatbots
Search
Personalization
Augmented Reality
Robots



HEALTH

Enhanced Diagnostics
Drug Discovery
Patient Care
Research
Sensory Aids



FINANCE

Algorithmic Trading
Fraud Detection
Research
Personal Finance
Risk Mitigation



RETAIL

Support Experience
Marketing
Merchandising
Loyalty
Supply Chain
Security



GOVERNMENT

Defense
Data Insights
Safety & Security
Resident Engagement
Smarter Cities



ENERGY

Oil & Gas Exploration
Smart Grid
Operational Improvement
Conservation



TRANSPORT

Automated Cars
Automated Trucking
Aerospace
Shipping
Search & Rescue



INDUSTRIAL

Efficiency Improvement
Factory Automation
Predictive Maintenance
Precision Agriculture
Field Automation



OTHER

Advertising
Education
Gaming
Professional & IT Services
Telco/Media
Sports

EXAMPLES

EARLY ADOPTION

Source: Intel forecast



ARTIFICIAL
INTELLIGENCE

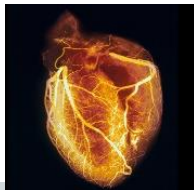


STUDENT DEVELOPER PROGRAM

RECENT CUSTOMER EXAMPLES



HEALTH



Early Tumor Detection

Leading medical imaging company



Early detection of malignant tumors in mammograms



Millions of "Diagnosed" Mammograms



Deep Learning (CNN) tumor image recognition



Higher accuracy and earlier breast cancer detection

Personalized Care

Renowned US Hospital system



Accurately diagnose fatal heart conditions



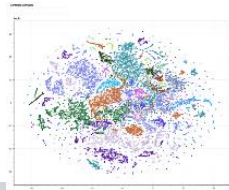
10,000 health attributes used



Saffron memory-based reasoning



Increased accuracy to 94% compared with 54% for average cardiologist



Data Synthesis

Financial services institution with >\$750B assets



Parse info to reduce portfolio manager time to insight



Vast stores of documents (news, emails, research, social)



Deep Learning (RNN w/ encoder/decoder)



Faster and more informed investment decisions

FINANCE



Customer Personalization

Leading Insurance Group



Increase product recommendation accuracy



5 Product Levels
1,353 Products
12M Members



Saffron memory-based reasoning



50% increase in product recommendation accuracy





WHAT IS MACHINE LEARNING

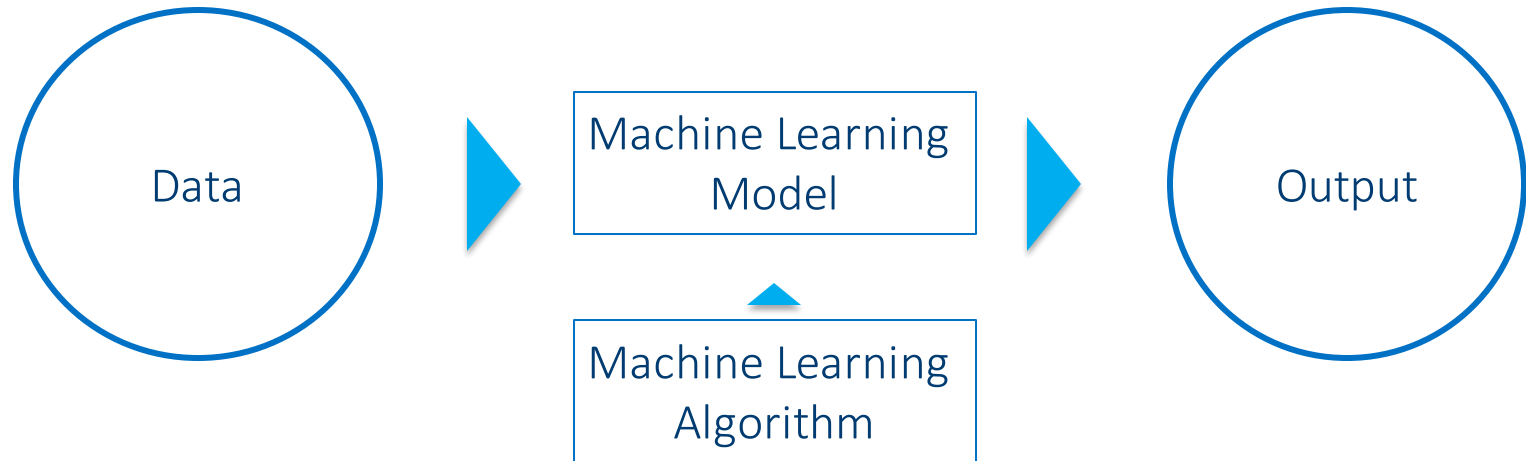
WHAT IS MACHINE LEARNING?

«The field of study that gives computers the ability to learn without being explicitly programmed»

Arthur Samuel, 1959



THE MACHINE LEARNING PIPELINE



TRAINING DATA SET

In order to train the model, we need a Training Dataset. If we have dataset of 100,000 houses sold in Portland this year, we take 75-80% of the data to train the model.

TEST DATA SET

Remaining 20% of the Data - we hide it from the model. That will help understanding how well the model will perform for new Data. That 20% is called a Test Dataset

FRAMEWORKS & LANGUAGES

Top Frameworks



Programming languages



An awesome list: <https://github.com/josephmisiti/awesome-machine-learning>



TYPES OF MACHINE LEARNING

Types of Machine Learning

Supervised Learning

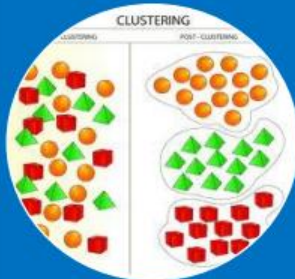
Teach desired behavior with labeled data



Make sense of new data based on prior data

Unsupervised Learning

Make inferences without labeled data



Discover unknown or hidden patterns

Reinforcement Learning

Act in an environment to maximize reward



Build autonomous agents that learn



SUPERVISED LEARNING

WE FEED THE MODEL WITH CORRECT ANSWERS , THE MODEL LEARNS AND FINALLY PREDICTS.

WE FEED THE MODEL WITH “GROUND TRUTH”.

MACHINE LEARNING SOLUTIONS

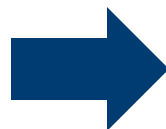
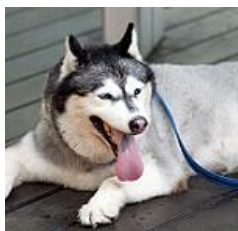
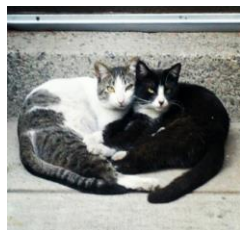
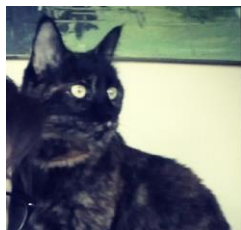
CLASSIFICATION

Predicting a discrete value for an entity with a given set of features.

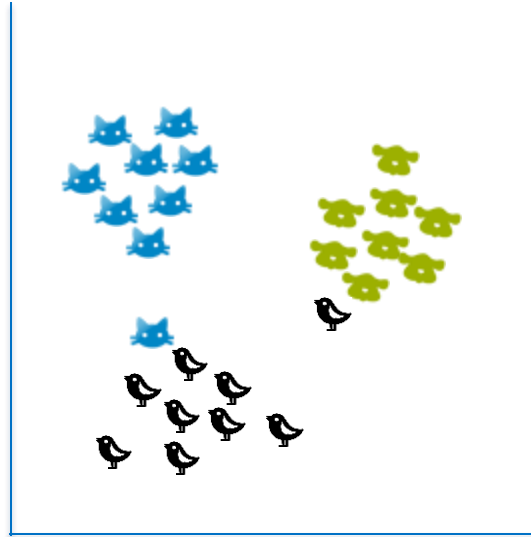
REGRESSION



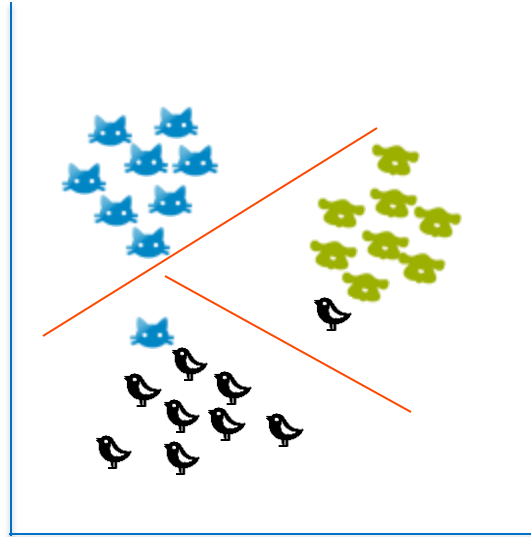
CLASSIFICATION



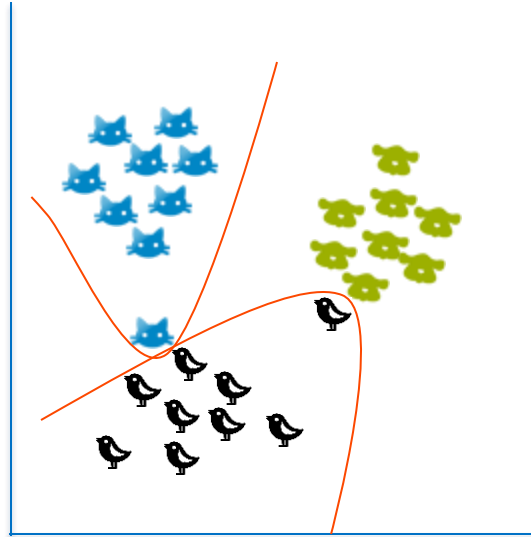
CLASSIFICATION



CLASSIFICATION



CLASSIFICATION





HANDS-ON WORK

Installing Anaconda, Scikit-Learn, Tensorflow & Jupyter

Installing Anaconda

Download Anaconda Python 3.6 version from
<https://www.continuum.io/downloads>

Run the installer it:

Windows & Mac OS: double click

Linux: `cd Downloads`
 `cmod u+x Anaconda3-5.0.0.1-Linux-x86_64.sh`
 `./Anaconda3-5.0.0.1-Linux-x86_64.sh`

Install Required Packages

```
conda update conda
```

```
conda config --add channels intel
```

```
conda create -n workshop intelpython3_core python=3.6
```

```
activate workshop (Windows)
```

```
source activate workshop (Linux & Mac)
```

```
conda install numpy pandas scikit-learn tensorflow jupyter
```

```
pip install tflearn pillow --no-deps
```

Run Jupyter

jupyter notebook

Navigate to

<http://localhost:8888/>





HANDS-ON WORK

Case Study: Iris Dataset

CASE STUDY: IRIS PLANTS

Iris Dataset:

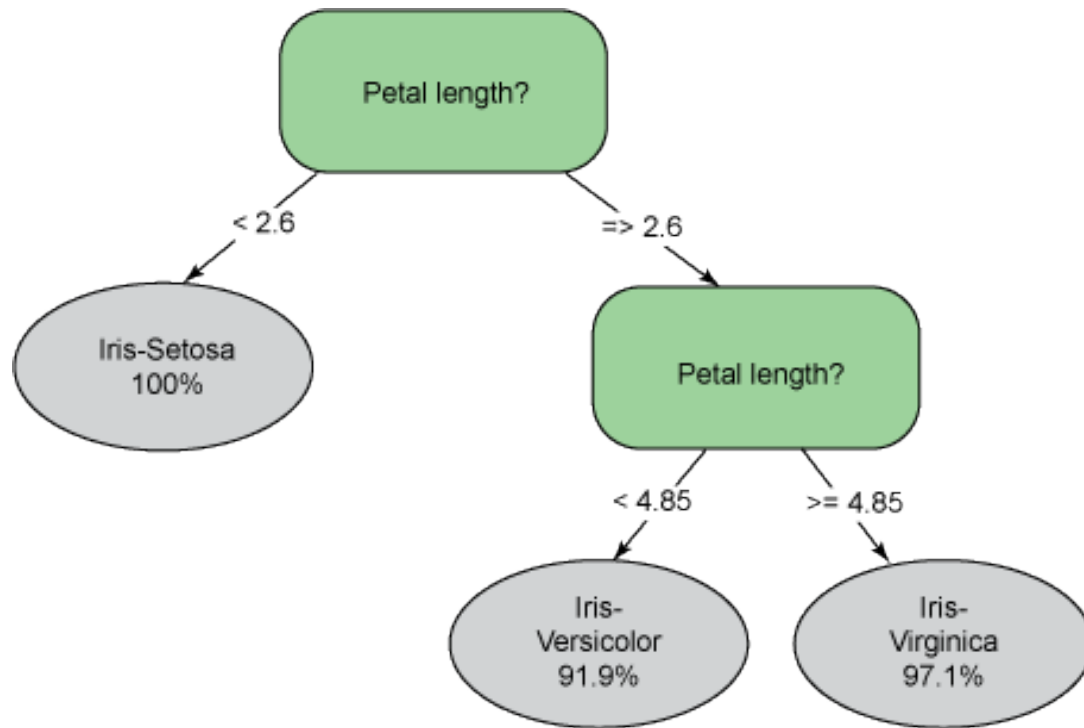
The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Number of Attributes: 4 (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm)

Number of Instances: 150 (50 in each of three classes)

Target: Iris-Setosa, Iris-Versicolour, Iris-Virginica

DECISION TREES



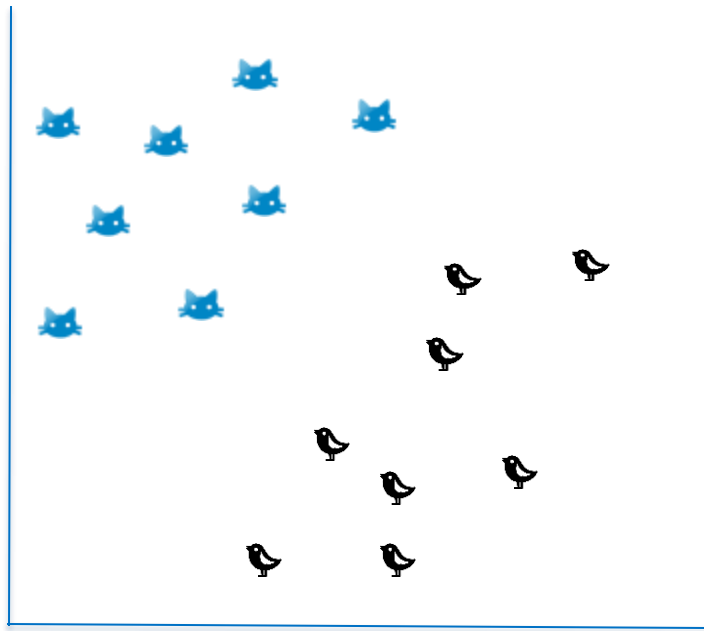
CASE STUDY: IRIS PLANTS

Decision Tree Classification

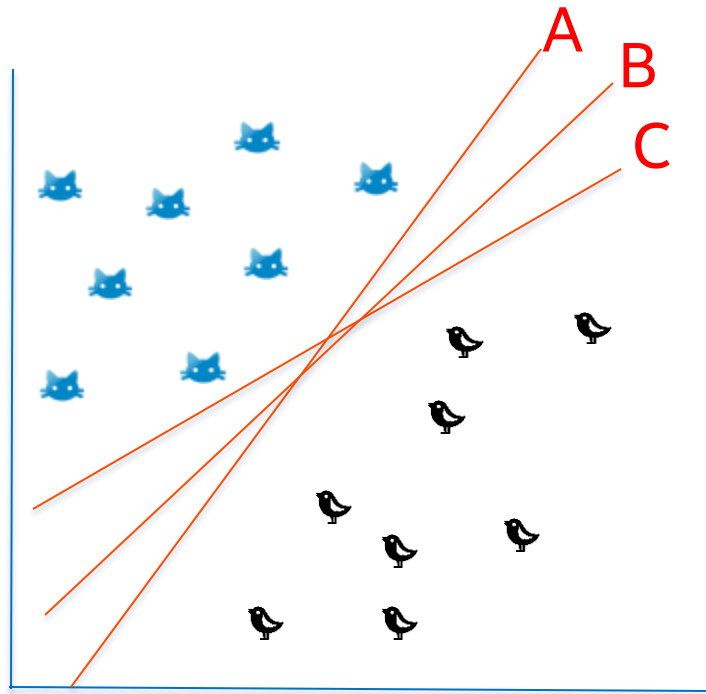
iPython notebook:

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/DecisionTreeClassifier.ipynb>

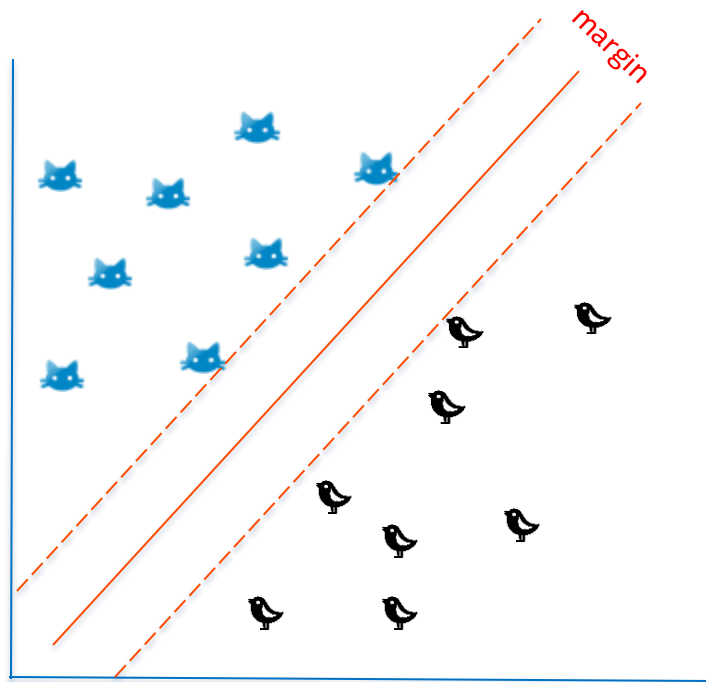
SUPPORT VECTOR MACHINES



SUPPORT VECTOR MACHINES

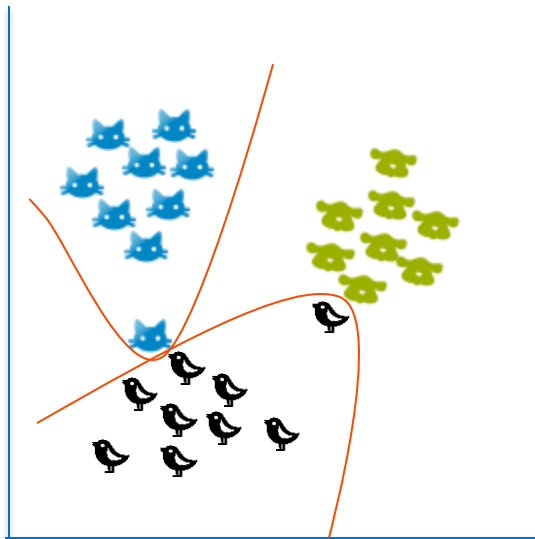


SUPPORT VECTOR MACHINES

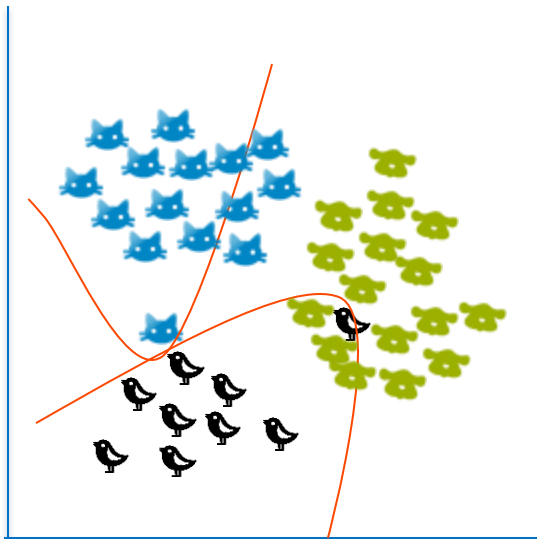


OVERFITTING

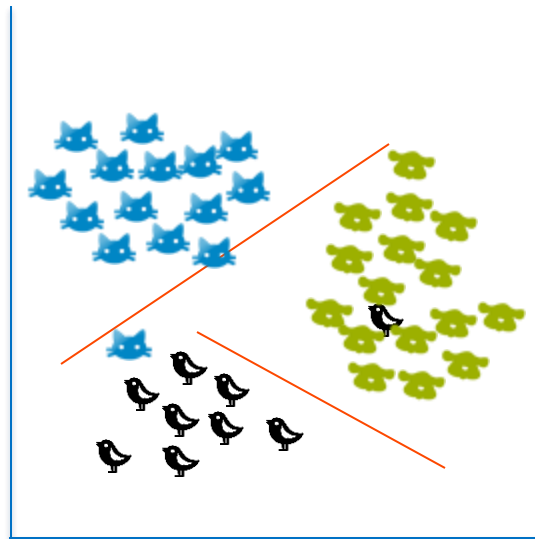
TRAINING



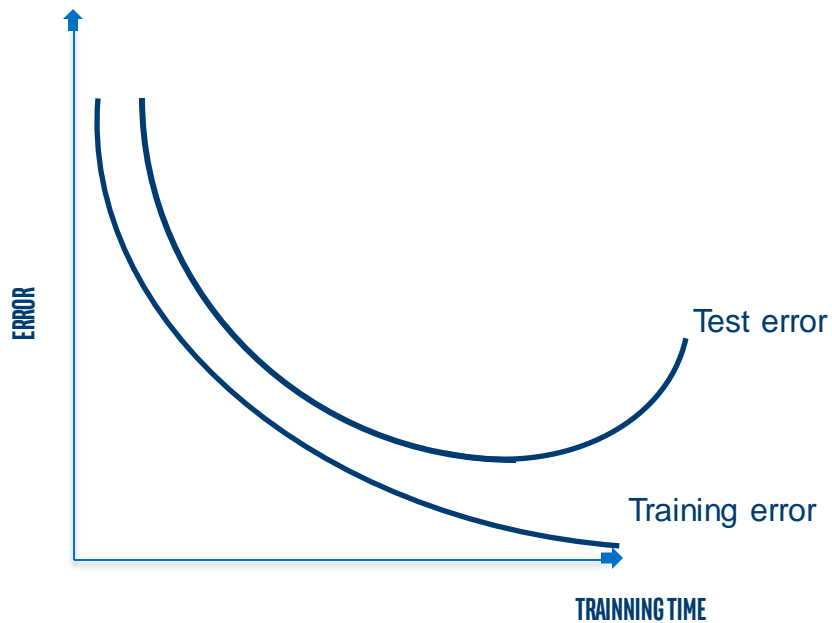
TESTING



TESTING



OVERFITTING



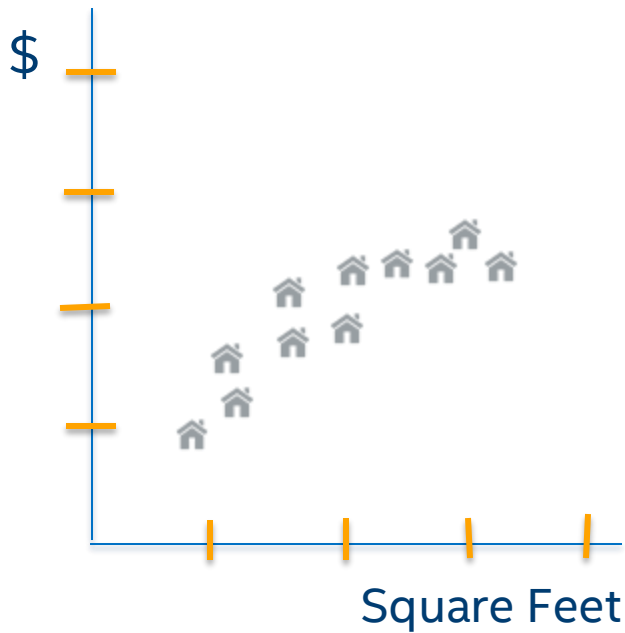
SUPERVISED LEARNING

CLASSIFICATION

REGRESSION

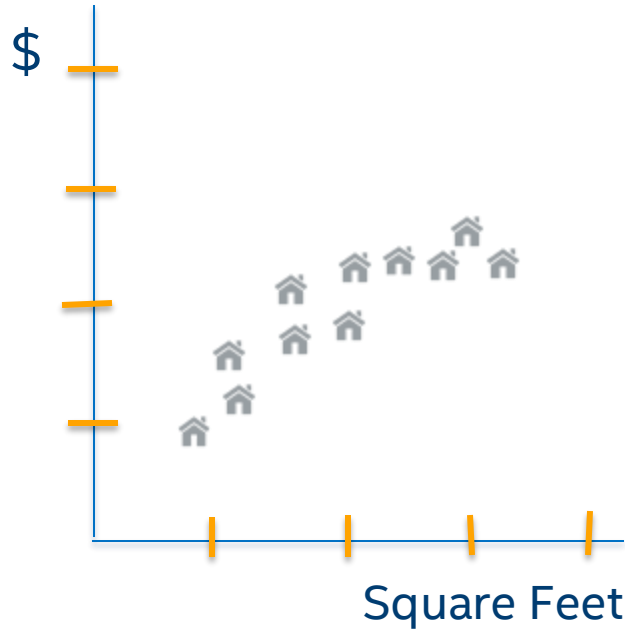
Regression attempts to predict a real numeric value for an entity with a given set of features.



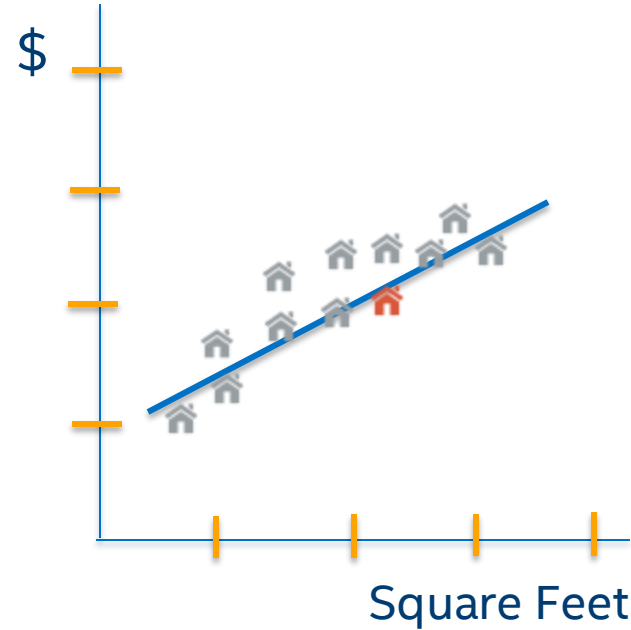


Set of Input vectors having a class or label

You train the model for predicting fair value of a house based on house attributes using historical home sales data. The model build can now predict the fair value of a new home.



Set of Input vectors having a class or label



Classify New data point into one of the already known class

You train the model for predicting fair value of a house based on house attributes using historical home sales data. The model build can now predict the fair value of a new home.



HANDS-ON WORK

Case Study: Regression with Diabetes Dataset

CASE STUDY: DIABETES

Diabetes Dataset:

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

Number of Attributes: 10

Number of Instances: 442

Target: Column 11 is a quantitative measure of disease progression one year after baseline

CASE STUDY: DIABETES

Linear Regression

iPython notebook:

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/LinearRegression.ipynb>



HANDS-ON WORK

Case Study: Regression with CO2 Emission and Global Warming Datasets

CASE STUDY: DIABETES

Carbon Dioxide Emission Dataset:

Number of Attributes: 1

Number of Instances: 260

Target: Column Total is the amount of emission from year 1751 to 2010

Global Warming Dataset:

Number of Attributes: 1

Number of Instances: 272 (recordings from 2 sources are listed together)

Target: Column Mean is the increase of global temperature from year 1880 to 2010



CASE STUDY: DIABETES

Linear Regression

iPython notebook:

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/LinearRegression2.ipynb>

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/CompareRegressionMethods.ipynb>

UNSUPERVISED LEARNING

**DATA IS GIVEN TO THE MODEL. RIGHT ANSWERS ARE NOT PROVIDED TO THE MODEL.
THE MODEL MAKES SENSE OF THE DATA GIVEN TO IT.**



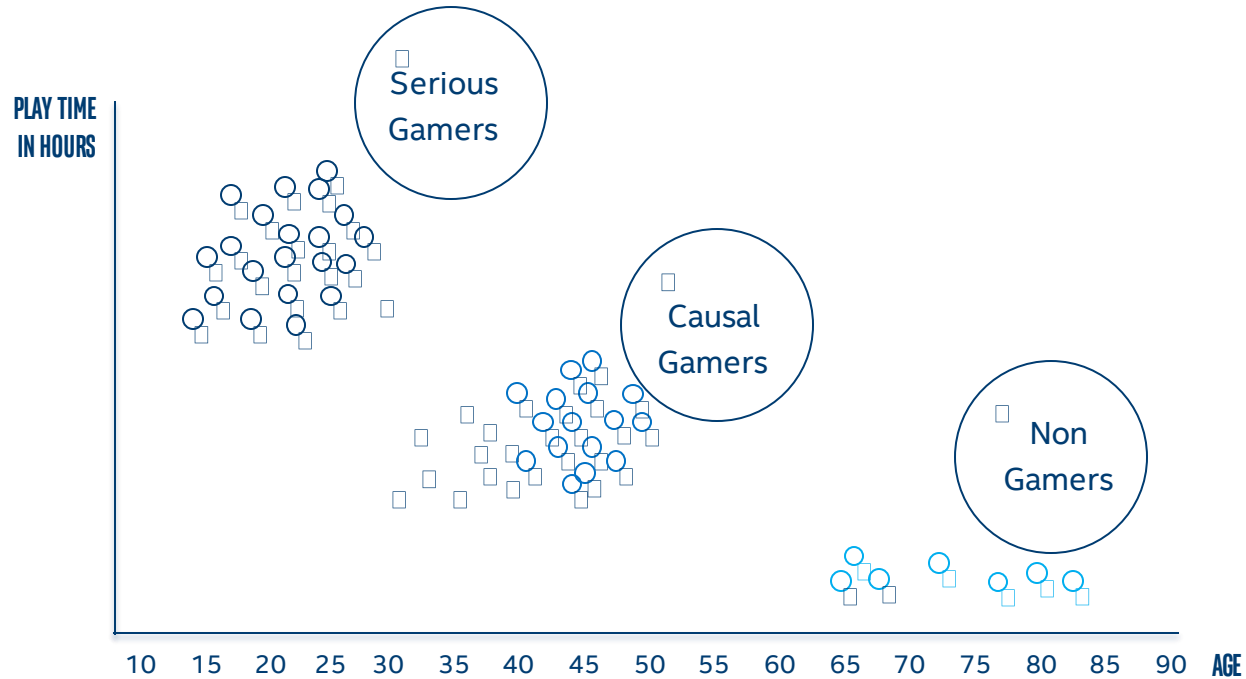
UNSUPERVISED LEARNING

CLUSTERING

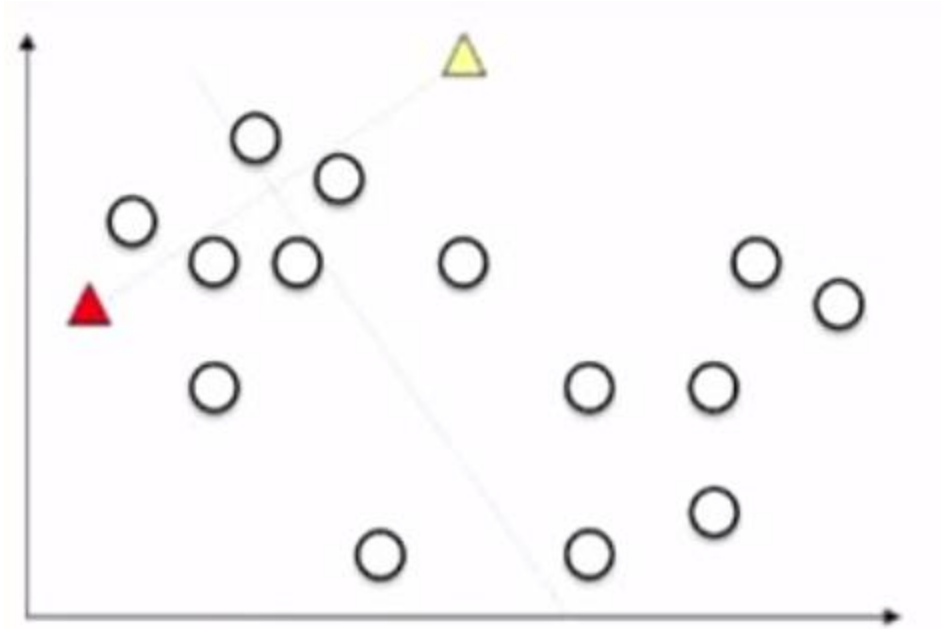
Grouping entities with similar features.
Unsupervised learning.



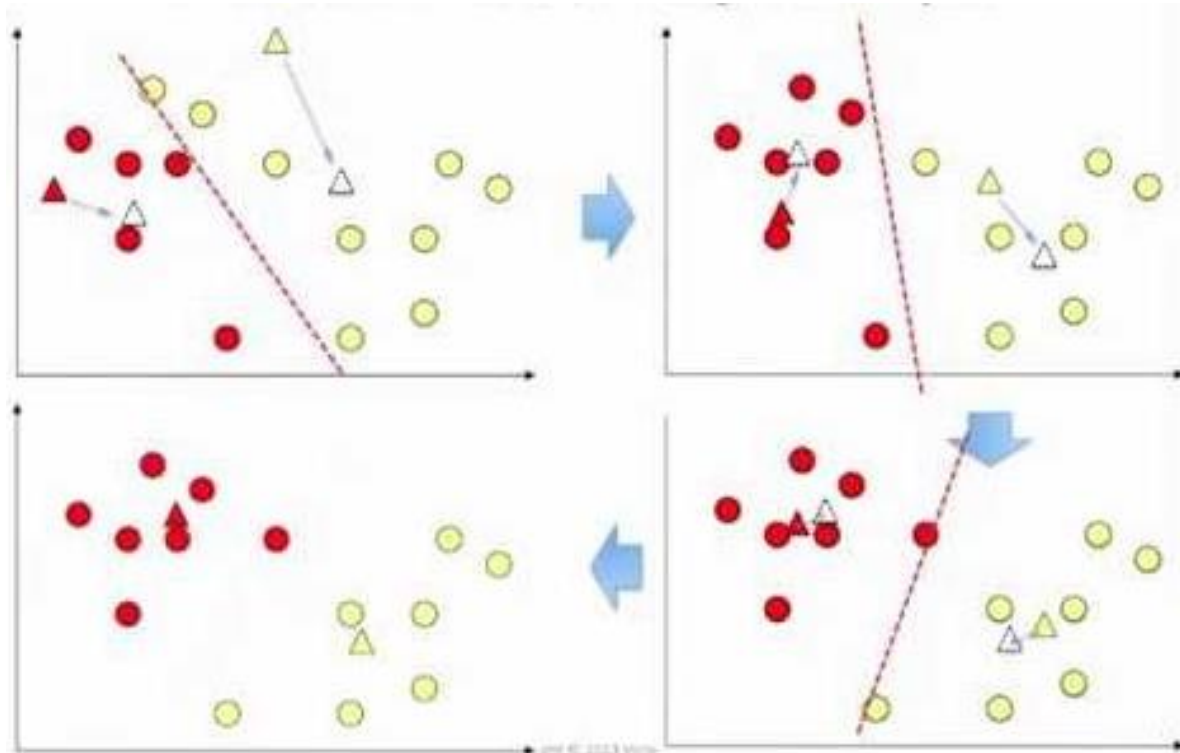
CLUSTERING EXAMPLE: MARKET SEGMENTATION



K-MEANS CLUSTERING



K-MEANS CLUSTERING





HANDS-ON WORK

Case Study: Iris Dataset

CASE STUDY: IRIS PLANTS

Iris Dataset:

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Number of Attributes: 4 (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm)

Number of Instances: 150 (50 in each of three classes)

Target: Iris-Setosa, Iris-Versicolour, Iris-Virginica

CASE STUDY: IRIS PLANTS

K-Means Clustering

iPython notebook:

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/KMeansClustering.ipynb>

<https://github.com/mstfldmr/IntelAIWorkshop/blob/master/KMeansClustering2.ipynb>

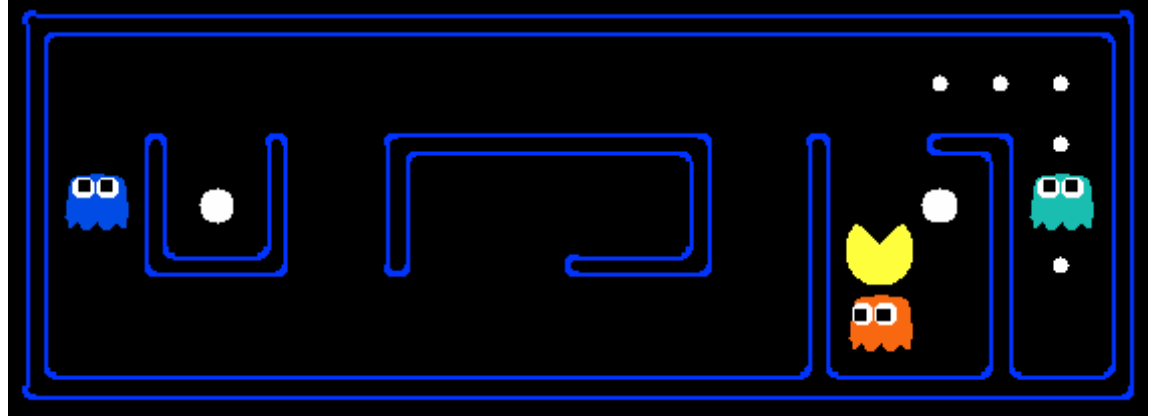
REINFORCEMENT LEARNING

REINFORCEMENT LEARNING IS THE PROBLEM OF GETTING AN AGENT TO ACT IN THE WORLD SO AS TO MAXIMIZE ITS REWARDS.

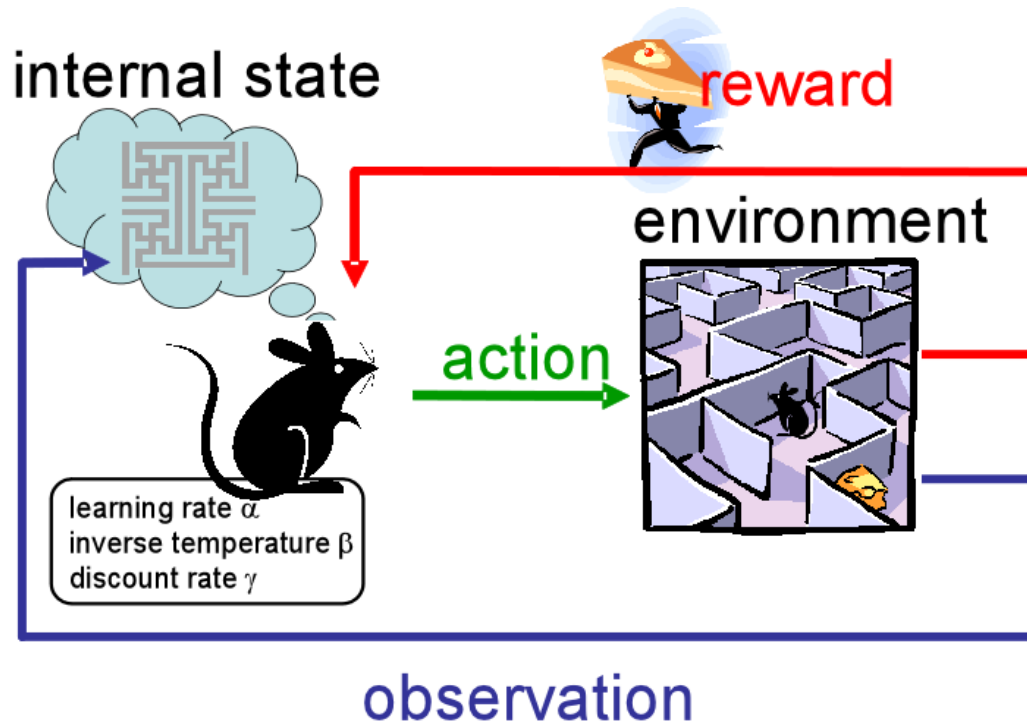


REINFORCEMENT LEARNING

- Robotics
- Healthcare
- Smart cities



REINFORCEMENT LEARNING



SOURCES OF DATASETS

- <https://www.kaggle.com/datasets>
- <http://tinyletter.com/data-is-plural>
- <http://www.kdnuggets.com/datasets>
- <https://github.com/caesar0301/awesome-public-datasets>
- <https://aws.amazon.com/public-datasets>
- <http://opendata.dc.gov>
- <https://www.data.gov/open-gov>



Q&A



Software

STUDENT DEVELOPER PROGRAM