# INTRODUCTION

The goal of this project is to develop an algorithm which can be used to predict house prices. In fact, several parameters can be taken into account in the evaluation of a property, such as the characteristics of the house, the street and the town in which it is located, the surrounding shops, the price of neighboring houses, etc.

In order to have accurate estimates, it is important to have a good database and to find the best features that have the greatest influence on the selling price of houses. To do so, part of this project will be devoted to finding a good database containing the characteristics and sales prices of several houses in a given region, followed by the analysis of this database in order to derive the most relevant parameters for the definition of a model and the training of the algorithm.

In addition, in order to obtain the best prediction engine, the performance and the predictive power of different algorithms such as linear regression, decision tree and random forest will be evaluated in this project.

# STATE OF THE ART

Several projects in this direction have already been carried out, whether to predict house prices or any other good. It must be said that the prediction of costs or values is a fairly widespread field of application for artificial intelligence. As part of this project, our implementation will be based on the work published in 2017 by [Eric Kim](https://www.kimanalytics.com/single-post/2017/09/11/Predicting-House-Prices-with-Machine-Learning) on predicting house prices using machine learning. In his publication, he discusses some points of interest of such a prediction tool. These include for example for people who would like to buy a house, whether the price offered corresponds to its value, which characteristics influence the price of houses. This would allow them to review some of their requirements for lower prices. Another useful would be for anyone who wants to invest in real estate and make a big profit from it. This would allow them to invest the most money on the features that drive up house prices the most.

# ABOUT THE DATA

As specified in the "State of the Art" section, our project will be based on the one led by Eric Kim on the same subject and in this regard, we will use the same database, the [Ames data set from Ames, Iowa](https://ww2.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt).

This database is composed of 2930 observations, 82 variables representing the characteristics including for example, the street, the house style, the year built, the type of road access to property etc. Hence, from a strictly quantitative point of view, there are enough data and parameters in the database to carry out the project properly

## DATA QUALITY

## HYPOTHESIS

In this section, we will try to make hypotheses on the characteristics that can most influence house prices

Very often, the geographical region is an important characteristic in the value of a property; however, as all our data comes from the same city, the characteristics of the location will not be exploited.

PREDICTIVE POWER COMPARAISON

In this section, we will compare the predictions obtained with linear regression algorithms with those obtained with decision trees and random forests.

UNITS TESTS

## 