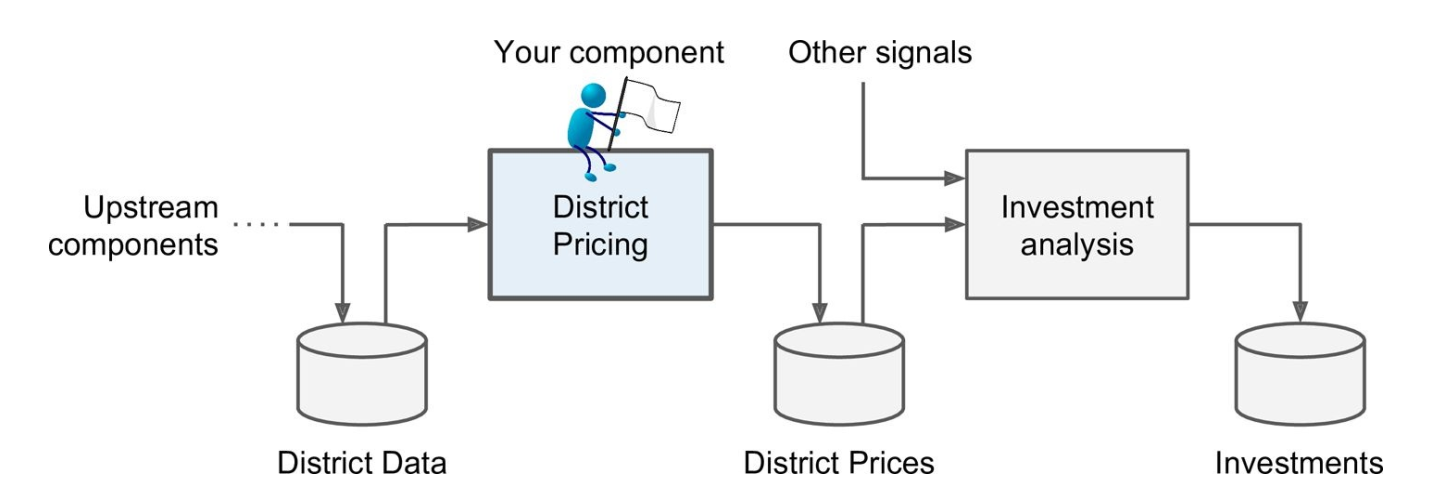
**End To End Machine Learning Problem**

**8 Steps to solve a problem that use Machine Learning Algorithm:**

1. Frame the problem:

The first question to ask your boss is what exactly is the business objective.

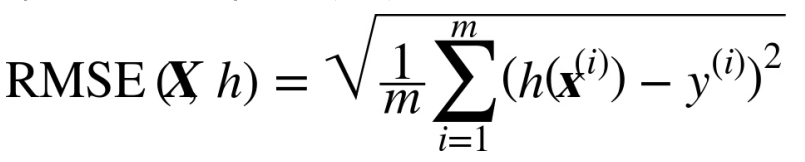


The next question to ask is what the current solution looks like (if any)

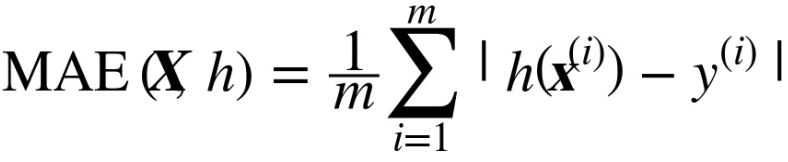
* Okay, with all this information you are now ready to start designing your system. First, you need to frame the problem: is it supervised, unsupervised, or Reinforcement Learning? Is it a classification task, a regression task, or something else? Should you use batch learning or online learning techniques?
* It is clearly a typical supervised learning task since you are given labelled training examples (each instance comes with the expected output, i.e., the district’s median housing price). Moreover, it is also a typical regression task, since you are asked to predict a value. More specifically, this is a multivariate regression problem since the system will use multiple features to make a prediction (it will use the district’s population, the median income, etc.)

1. **Select a Performance Measure:**

Your next step is to select a performance measure. A typical performance measure for regression problems is the Root Mean Square Error (RMSE).



Even though the RMSE is generally the preferred performance measure for regression tasks, in some contexts you may prefer to use another function. For example, suppose that there are many outlier districts. In that case, you may consider using **the Mean Absolute Error** (MAE).



* **Computing the root of a sum of squares (RMSE)** corresponds to the Euclidian norm: it is the notion of distance you are familiar with. It is also called the **ℓ2 norm**, noted ∥ · ∥2 (or just ∥ · ∥).
* **Computing the sum of absolutes (MAE) corresponds to the ℓ1 norm**, noted ∥ · ∥1. It is sometimes called the Manhattan norm because it measures the distance between two points in a city if you can only travel along orthogonal city blocks.
* More generally, the **ℓk norm** of a vector v containing n elements is defined as ||v||k =. **ℓ0 just gives the cardinality of the vector.**

**🡺** The **higher the norm index**, the more it **focuses** on **large values and neglects small ones**. This is why the RMSE is more sensitive to outliers than the MAE. But when outliers are exponentially rare (like in a bell-shaped curve), the RMSE performs very well and is generally preferred.

1. **Check the Assumptions:**

Lastly, it is good practice to list and verify the assumptions that were made so far (by you or others); this can catch serious issues early on.