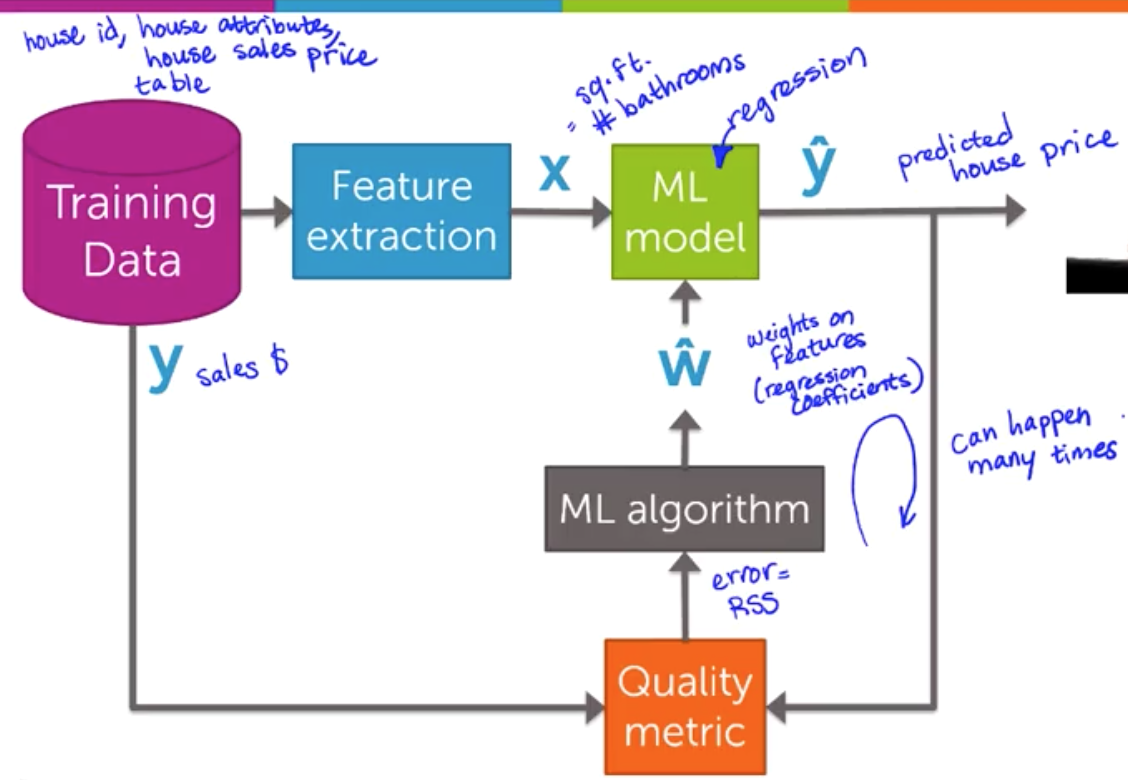
**CASE STUDY 1**

**Regression: Prediction House Prices**

*ML Pipeline:*



*General Idea*:

* Given: set of features
* Model: how observations change when values of features change
* NOTE: can use regression to model importance of features

*House Example*:

* Features (X): size, bedrooms, bathrooms etc.
* Observation (Y): house sales price

*Linear Regression*:

* Fit a line through the data
  + **Function:** fw(x) = w0 + w1x
  + **Parameters:** w = [w0, w1]
* Cost of using a given line:
  + **Residual sum of squares (RSS)** – how far is each observation from line
    - RSS(w0, w1) =
    - Minimizing parameters:
* Predicting house price:
  + Use function
* Note: you can add higher order variables to this – treating another feature as x2

*Evaluating Regression Models*:

* **Overfitting**: model is too specified to data and doesn’t generalize well to new data
* **How to choose model order/complexity?**
  + Simulate predictions
    - Remove some houses
    - Fit model on remaining
    - Predict held out houses and compare
* **Training Set**: houses used to fit model
* **Test Set**: houses model is tested on
* **Training/Test Error**: RSS on training/test set
* **Training/Test Curve**: how do training/test error vary as a function of model complexity
  + Plot model complexity (linear, quadratic…) on X axis
  + Error on Y axis
  + Training Error: decreases
  + Test Error: may decrease then increase
* **Adding Other Features** – are there other features that might affect the observation
  + How many features should you use? – UPCOMING