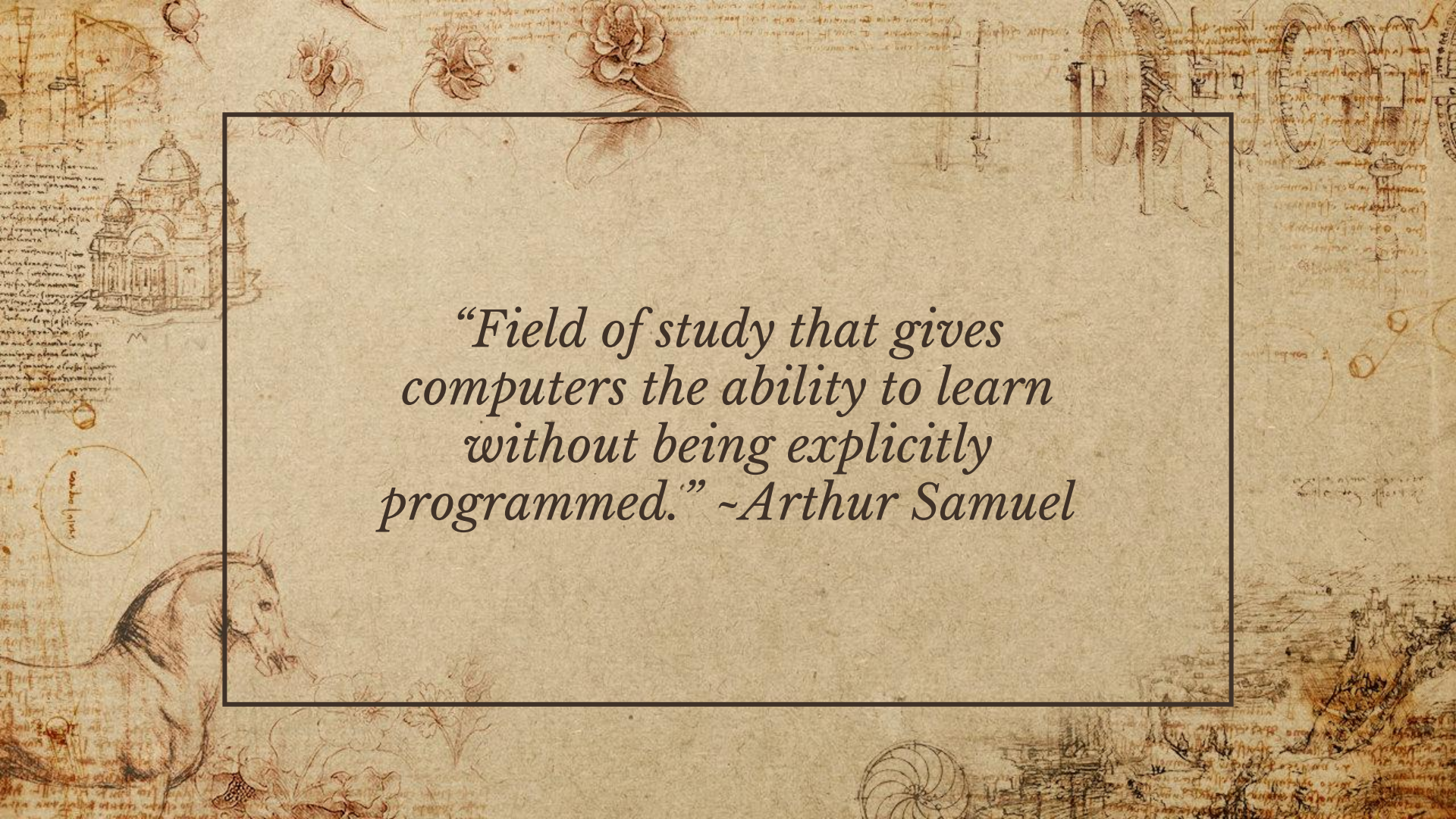


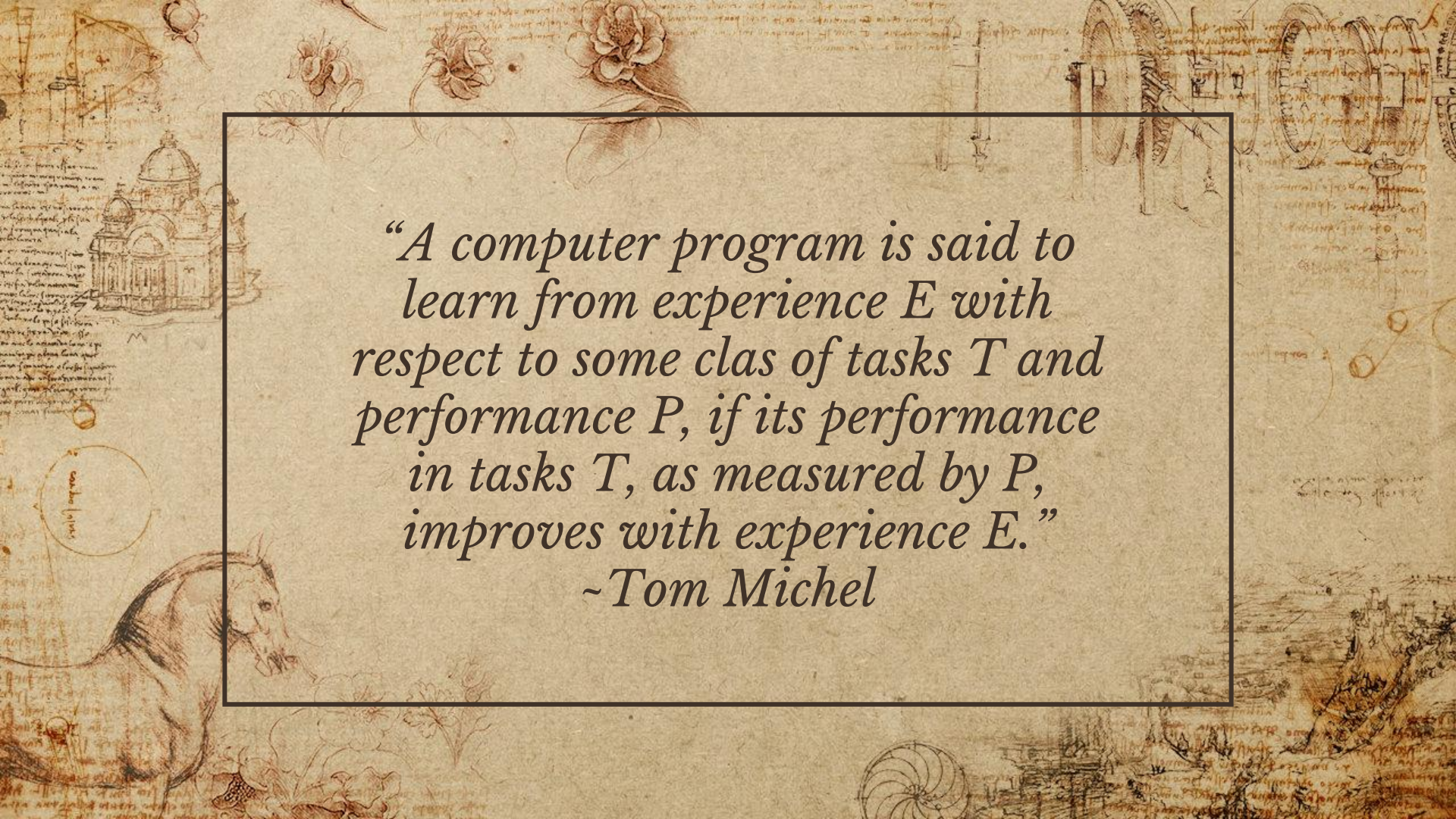
MACHINE LEARNING CRASH COURSE

Week 1



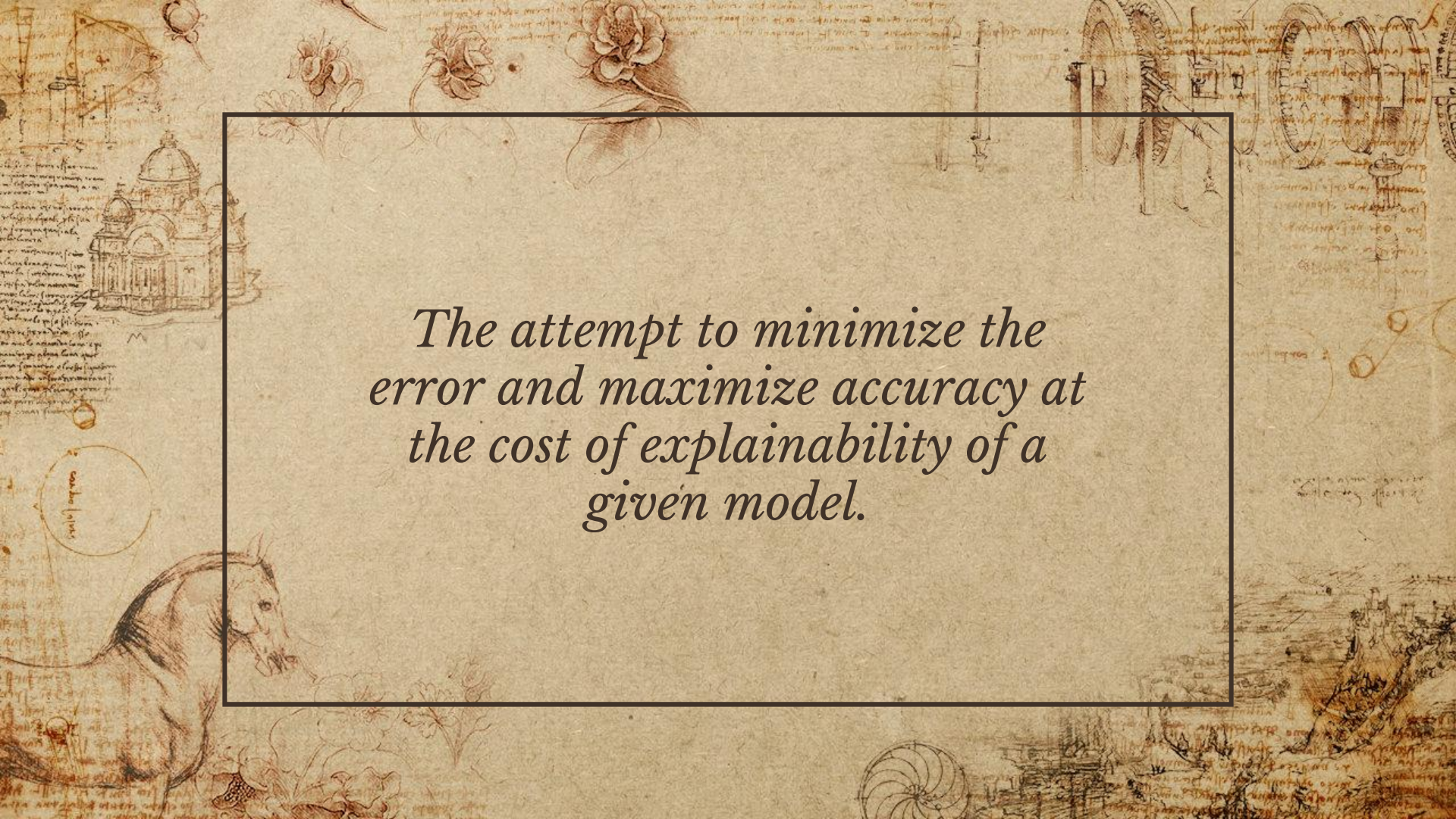
The background is a detailed reproduction of a page from Leonardo da Vinci's notebooks. It features various sketches in brown ink on aged, yellowish paper. At the top, there are several anatomical drawings of flowers and leaves. To the left, a small sketch of a domed building is visible. In the bottom left corner, there is a sketch of a horse's head and neck. The right side of the page contains more mechanical sketches, including what looks like a gear or a pulley system. The overall style is that of a historical scientific or artistic manuscript.

*“Field of study that gives
computers the ability to learn
without being explicitly
programmed.” ~Arthur Samuel*

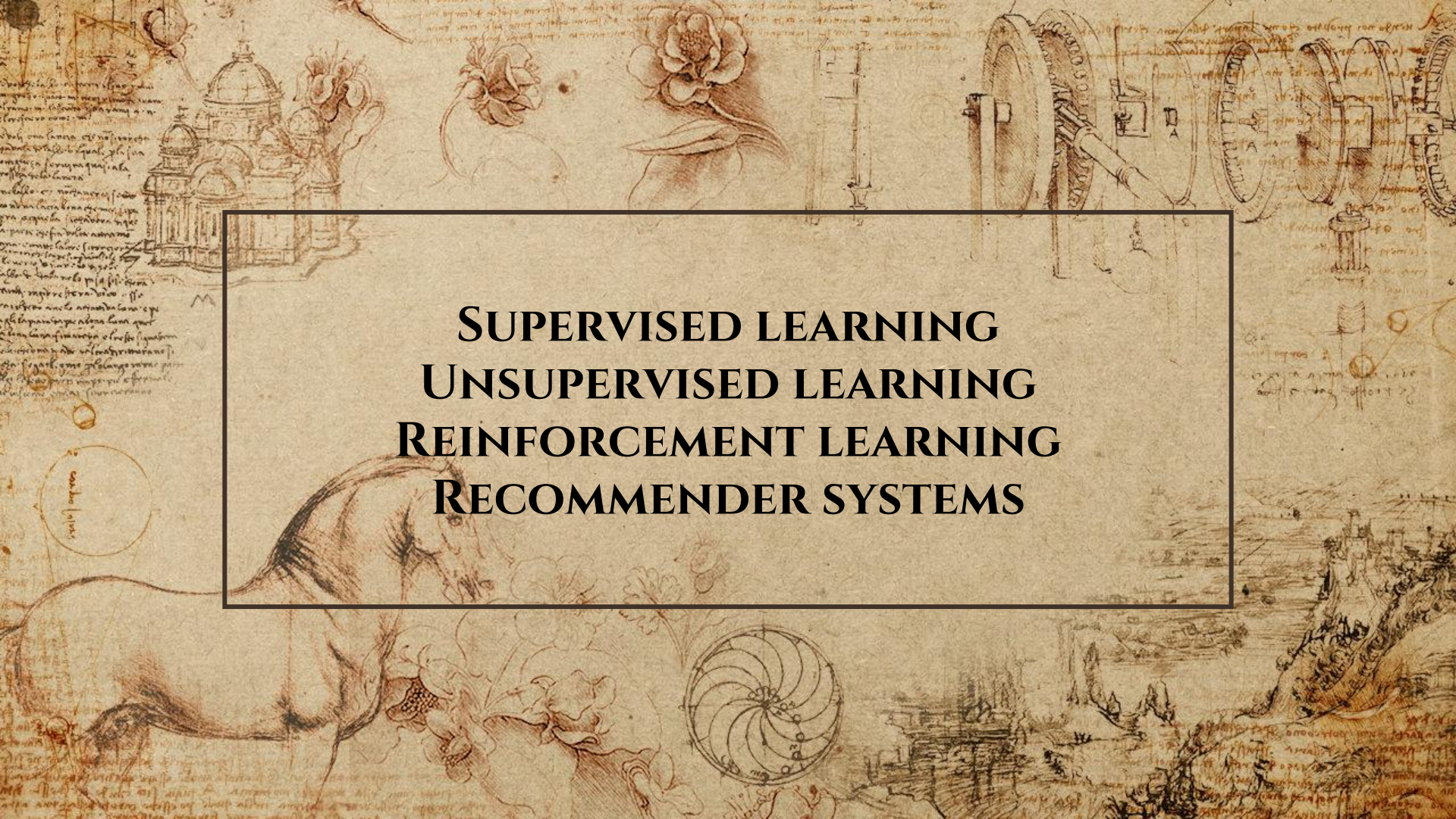
The background of the image is a detailed reproduction of a page from Leonardo da Vinci's notebooks. It features various sketches in brown ink on aged, yellowish paper. At the top, there are floral designs and mechanical components. On the left, a dome-shaped building is sketched. In the bottom left corner, a horse is depicted in profile. The right side contains more mechanical drawings, including what appears to be a gear or pulley system. The overall aesthetic is that of a historical scientific and artistic manuscript.

“A computer program is said to learn from experience E with respect to some class of tasks T and performance P , if its performance in tasks T , as measured by P , improves with experience E .”

~Tom Michel

The background is a detailed reproduction of a page from Leonardo da Vinci's notebooks. It features various sketches in brown ink on aged, yellowish paper. At the top, there are drawings of flowers and mechanical gears. On the left, a large dome-shaped building is sketched. At the bottom left, a horse is shown in profile. The right side contains more mechanical diagrams and handwritten text in Leonardo's characteristic mirror-image script. A large, empty rectangular box with a thin black border is centered on the page, containing the main text.

*The attempt to minimize the
error and maximize accuracy at
the cost of explainability of a
given model.*

The background of the slide is a detailed, sepia-toned reproduction of Leonardo da Vinci's 'Vitruvian Man' drawing. The central figure of the man is partially obscured by a white rectangular box containing text. Surrounding the central figure are various sketches: a dome-like structure in the upper left, a mechanical device with gears in the upper right, a horse's head in the lower left, and a spiral-like mechanical component in the lower center. The entire background is filled with faint, handwritten text in Italian, characteristic of Leonardo's notebooks.

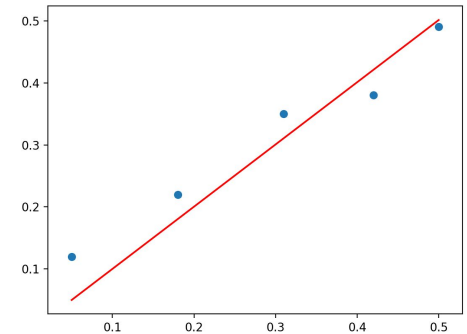
SUPERVISED LEARNING
UNSUPERVISED LEARNING
REINFORCEMENT LEARNING
RECOMMENDER SYSTEMS

SUPERVISED LEARNING

- ❖ Regression Problem
- ❖ Classification Problem

UNSUPERVISED LEARNING

- ❖ Clustering Problem
- ❖ Cocktail Party Problem



TOPICS (W1)

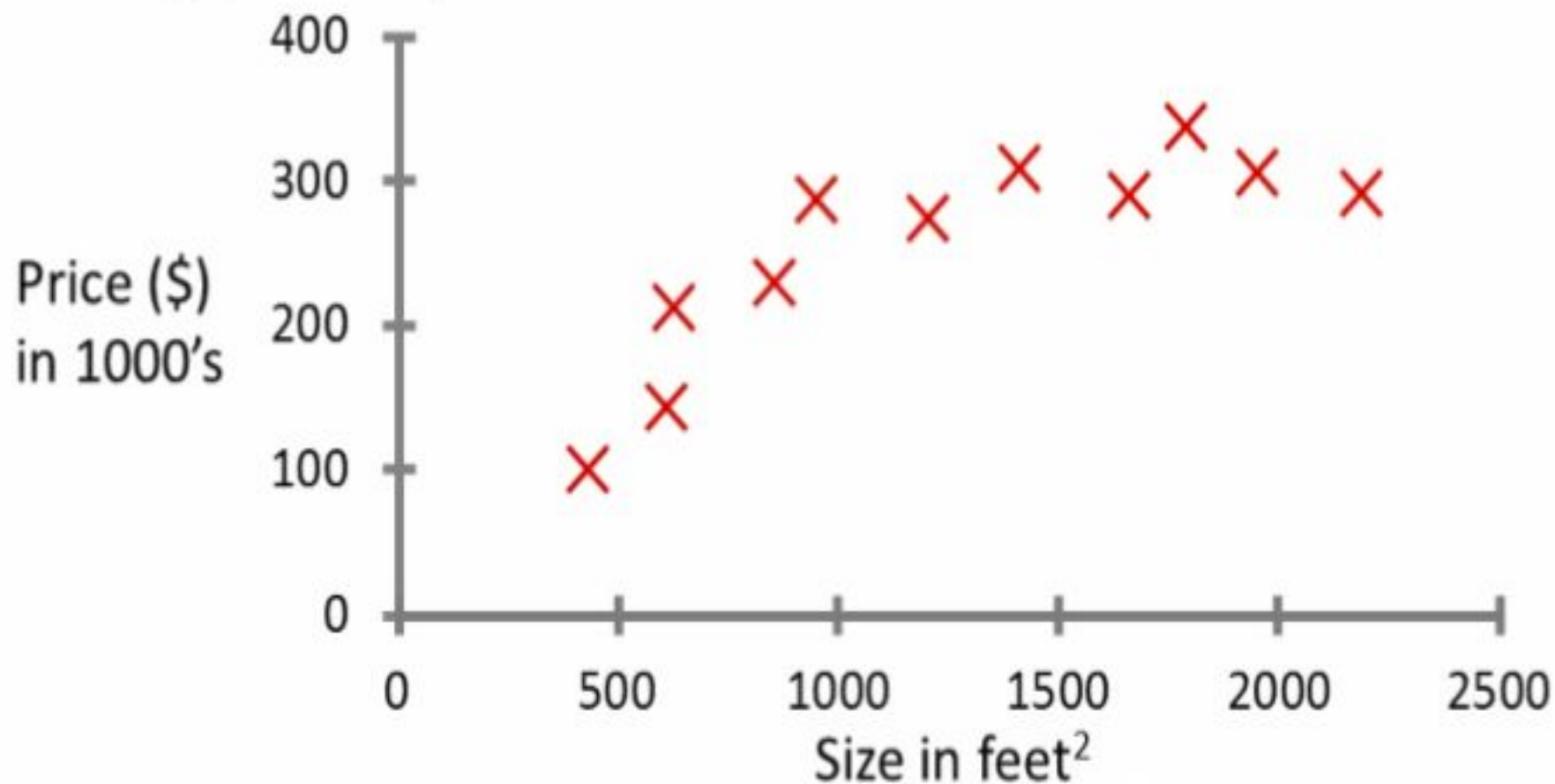
❖ Linear Regression (Univariate)

- ❖ Model Representation
- ❖ Cost function
- ❖ Gradient descent

❖ Linear Algebra

- ❖ Matrices and Vectors
- ❖ Matrix & Vector Addition and Multiplication
- ❖ Matrix Inverse and Transpose

Housing price prediction.



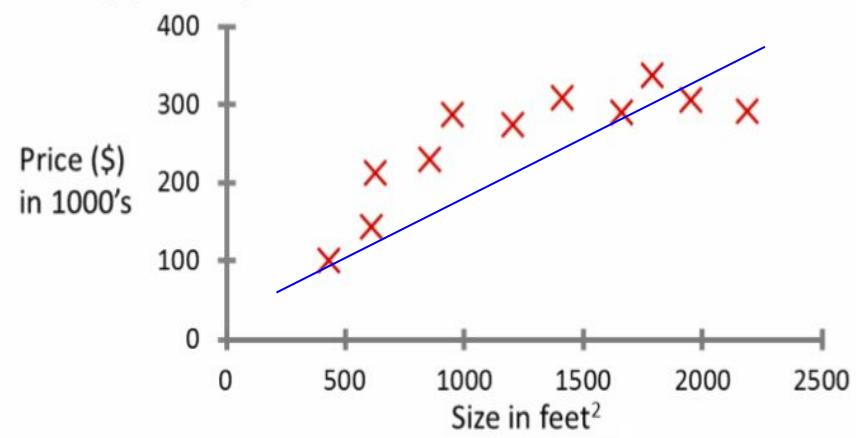
Size in feet ² (x)	Price (\$) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

} m = 47

- Number of training examples (m)
- Input features (x)
- Output features (y)
- lth index to the training data

Goal

Housing price prediction.



MODEL REPRESENTATION (UNIVARIATE)

- Define a hypothesis (h)
- Whereas :
 - Theta (sub-0) is the zero condition
 - Theta (sub-1) is the gradient
- x is some arbitrary variable
- The function uses parameters learned by the system to give a prediction

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Initial Condition

Gradient

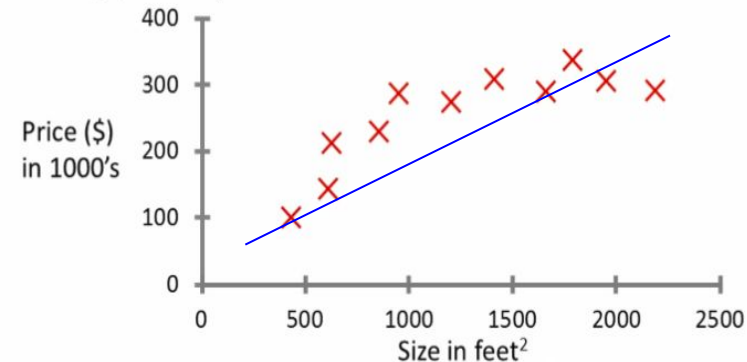
REFINED OBJECTIVE

- Using different values for Theta, define a cost function to best fit the line to the data
- Generate parameters such that:
 - The hypothesis (h) is very close to the actual (y) value
- Minimize the squared difference between $h(x)$ and y for every sample

More Formally

$$\frac{1}{2m} \sum_{i=1}^m \left(h_{\theta}(x^i) - y^i \right)^2$$

Housing price prediction.



COST FUNCTION

- Minimize the cost function for all the training data

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2$$

EXAMPLE 1

- Minimize the cost function
 - Assume $\theta_0 = 0$

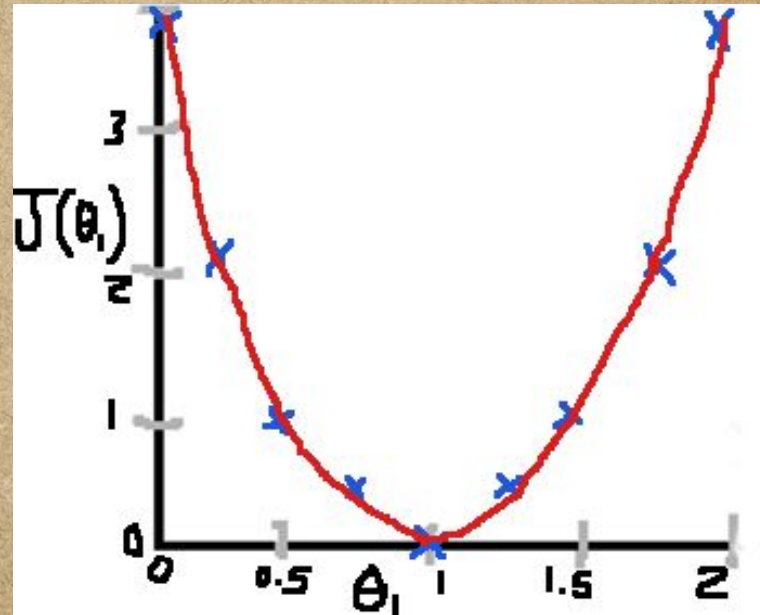
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$\theta_0 = 0$$

$$h_{\theta}(x) = \theta_1 x$$

$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2$$

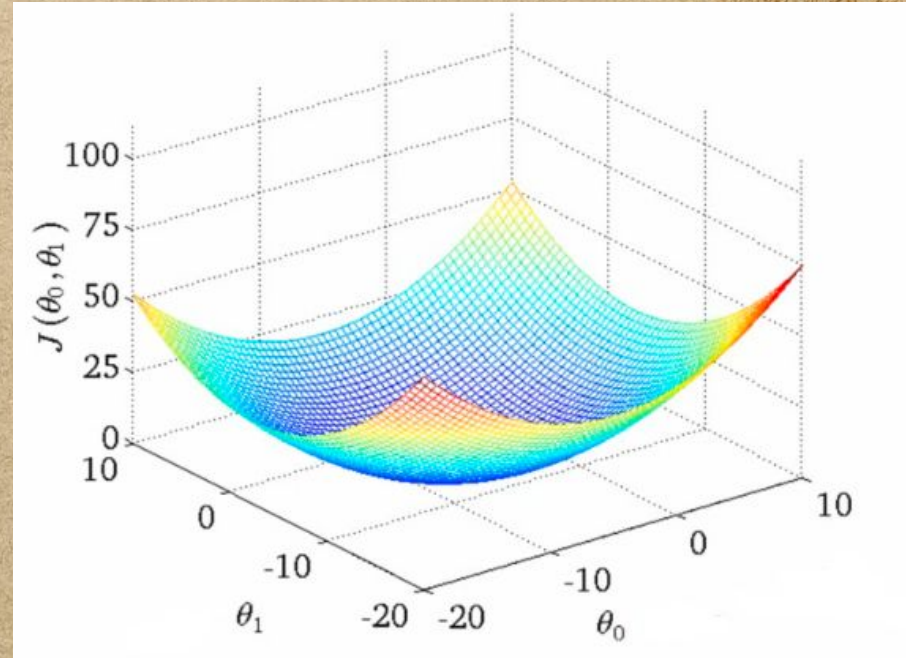
θ_1 is the gradient



EXAMPLE 2

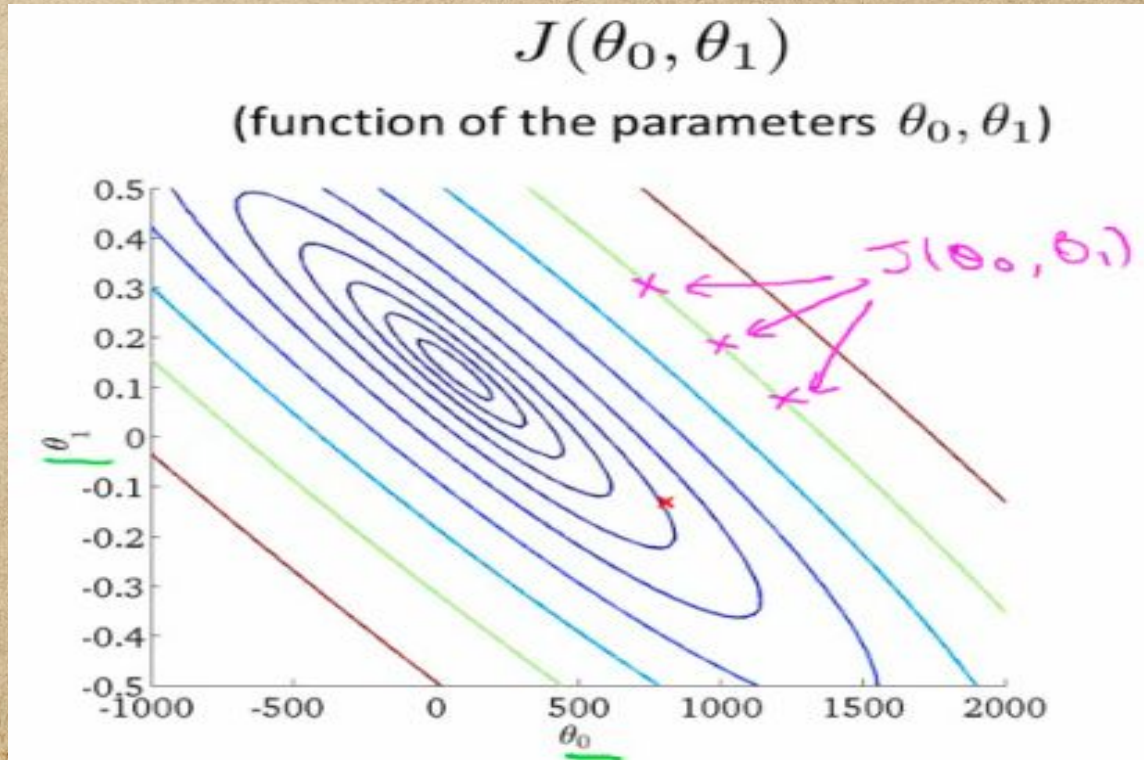
- Minimize the cost function
 - Assume $\theta_0 = 0$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2$$



EXAMPLE 2 (continued)

- Use a contour plot to visualize the cost function



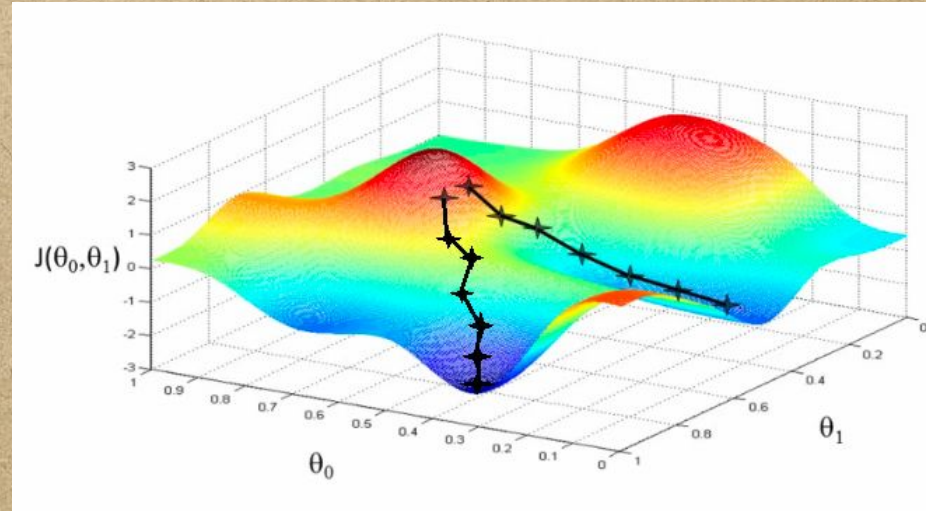
GRADIENT DESCENT

- Used to minimize the cost function with N parameters given by:

$$J(\theta_0, \theta_1, \dots, \theta_n)$$

Algorithm

- Make an initial guess
- Change the parameters in order to reduce the cost function at each step
- Repeat until convergence at a local minimum



GRADIENT DESCENT (CONTINUED)

More Formally: (repeat the following step)

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

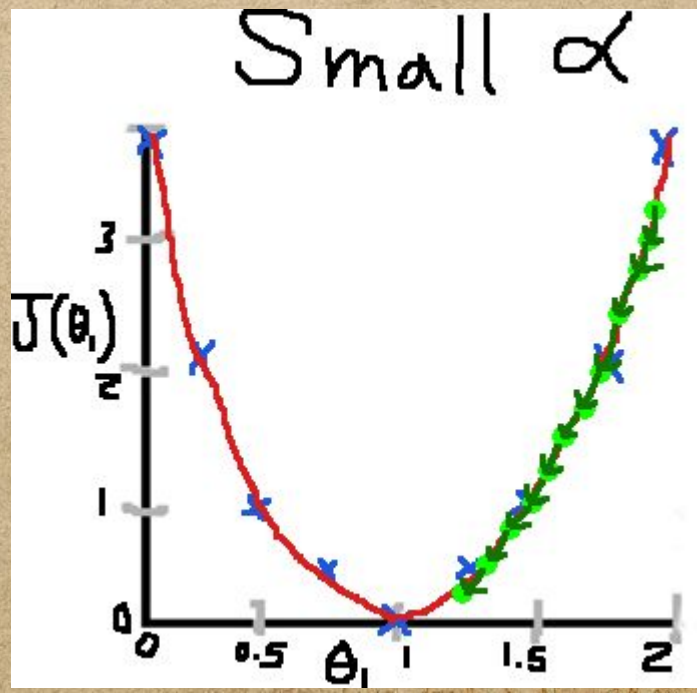
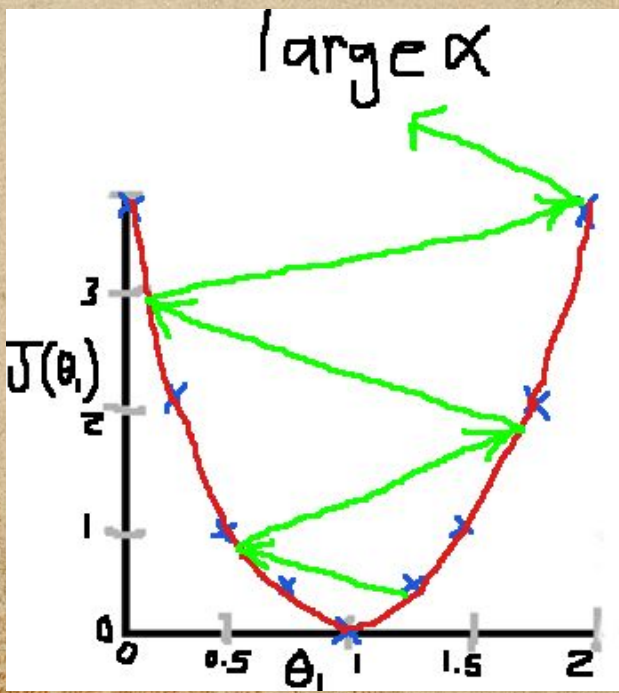
(for $j = 0$ and $j = 1$)

- Simultaneously update the parameters

- α (alpha) is the learning rate
 - Controls the step size

Choosing an Optimal Learning Rate

Consider $J(\theta_1)$

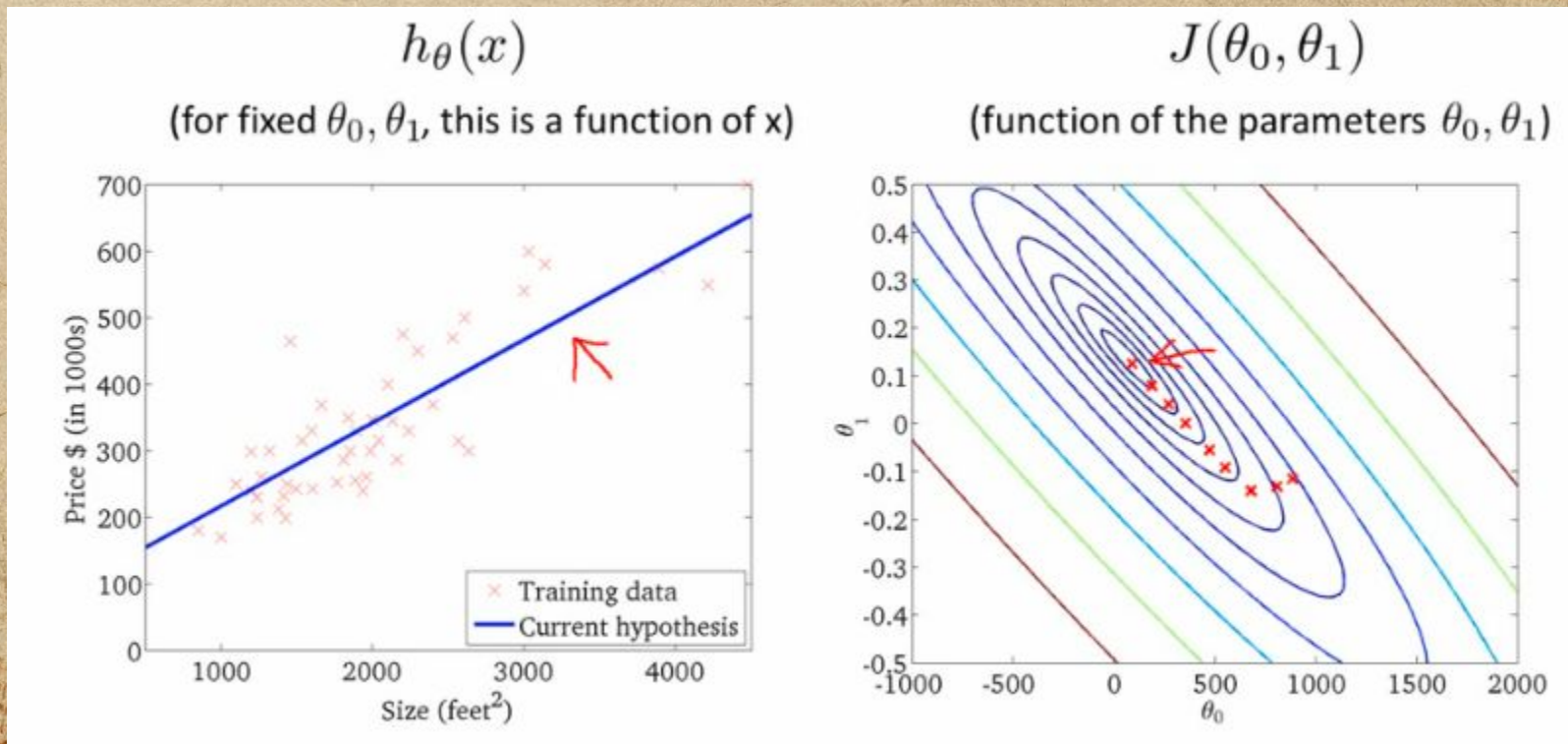


Derivation of the Cost Function

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) &= \frac{\partial}{\partial \theta_j} \cdot \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2 \\ &= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^m (\theta_0 + \theta_1 x^i - y^i)^2 \\ &= \left(\frac{\partial}{\partial \theta_j} h_{\theta}(x^i) \right) \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)\end{aligned}$$

Derivation of the Cost Function (continued)

The linear regression function is always a convex function with one minimum



Fun Fact

Batch Gradient Descent: Iterating over all the training data at every step.

Numerical solutions exist (**Normal Equations**), but don't scale as well.

SYSTEM OF LINEAR EQUATIONS

Recall: $y = mx + b$

$$y = 3x_1 + 2x_3$$

$$y = 2x_1 - 2x_3$$

$$y = x_2 + x_3$$

MATRICES

System of Equations

$$\begin{aligned}Y &= 3x_1 + 2x_3 \\Y &= 2x_1 - 2x_3 \\Y &= x_2 + x_3\end{aligned}$$

Coefficient Matrix

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$$

MATRICES

Identity Matrix (3x3)

Properties:

- 1s along the diagonal
- Same number of rows and columns

Facts:

- It is the optimal end form of row reduced augment matrices
- It is used as one way of calculating a matrix's inverse

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

MATRICES

System of Equations

$$\begin{aligned} 3x_1 + 2x_3 &= 0 \\ 2x_1 - 2x_3 &= 0 \\ x_2 + x_3 &= 0 \end{aligned}$$

Augmented Matrix

$$\left[\begin{array}{ccc|c} 3 & 0 & 2 & 0 \\ 2 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 \end{array} \right]$$

Matrix Operations

- Swap 2 rows
- Multiple a row by a scalar
- Add 2 rows together to re-write that row

MATRICES

$$\left[\begin{array}{ccc|c} 3 & 0 & 2 & 0 \\ 2 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 \end{array} \right]$$



$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

Matrix Operations to RREF:

- $r1 + r2 \rightarrow r1$
- $(1 / 5) * r1 \rightarrow r1$
- $(- 2) * r1 + r2 \rightarrow r2$
- $(- 1 / 2) * r2 \rightarrow r2$
- $r1 \longleftrightarrow r3$
- $(- 1) * r3 + r2 \rightarrow r2$

MATRICES

MATRIX ADDITION

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$$

+

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$$

=

$$\begin{bmatrix} 6 & 0 & 4 \\ 4 & 0 & -4 \\ 0 & 2 & 2 \end{bmatrix}$$

MATRICES

MATRIX MULTIPLICATION

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$$

3x3

$$\begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix}$$

3x1

=

$$\begin{bmatrix} 9 \\ 6 \\ 2 \end{bmatrix}$$

3x1

VECTORS

EXAMPLES

Properties:

- $N \times 1$ matrix

Facts:

- Can be used to express a system of linear equations

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

or

$$\begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix}$$

3x1

3x1

$$\underline{A} \underline{x} = \underline{b}$$

MATRICES

TRANSPOSE

$$A = \begin{bmatrix} 6 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

2x3



$$A^T = \begin{bmatrix} 6 & 0 \\ 4 & 0 \\ 0 & 2 \end{bmatrix}$$

3x2


MATRICES

INVERSE

$$[A|I] = \left[\begin{array}{ccc|ccc} 3 & 0 & 2 & 1 & 0 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array} \right]$$

Matrix Operations to RREF:

- $r1 + r2 \rightarrow r1$
- $(1/5) * r1 \rightarrow r1$
- $(-2) * r1 + r2 \rightarrow r2$
- $(-1/2) * r2 \rightarrow r2$
- $r1 \longleftrightarrow r3$
- $(-1) * r3 + r2 \rightarrow r2$


$$[I|A^{-1}] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 1 & 0 & -0.2 & 0.3 & 1 \\ 0 & 0 & 1 & 0.2 & -0.3 & 0 \end{array} \right]$$

APPLICATION

House sizes:

$$\begin{cases} 2104 \\ 1416 \\ 1534 \\ 852 \end{cases}$$

Matrix

$$\begin{bmatrix} 1 & 2104 \\ 1 & 1416 \\ 1 & 1534 \\ 1 & 852 \end{bmatrix}$$

\times

Matrix

$$\begin{bmatrix} -40 \\ 0.25 \end{bmatrix} \begin{bmatrix} 200 \\ 0.1 \end{bmatrix} \begin{bmatrix} -150 \\ 0.4 \end{bmatrix}$$

$=$

$$\begin{bmatrix} 486 \\ 314 \\ 344 \\ 173 \end{bmatrix} \begin{bmatrix} 410 \\ 342 \\ 353 \\ 285 \end{bmatrix} \begin{bmatrix} 692 \\ 416 \\ 464 \\ 191 \end{bmatrix}$$

Prediction
of 1st
h₀

Predictions
of 2nd
h₀

Have 3 competing hypotheses:

1. $h_{\theta}(x) = -40 + 0.25x$

2. $h_{\theta}(x) = 200 + 0.1x$

3. $h_{\theta}(x) = -150 + 0.4x$

SOURCES

- http://www.holehouse.org/mlclass/01_02_Introduction_regression_analysis_and_gr.html
- <https://www.mathsisfun.com/algebra/matrix-inverse-row-operations-gauss-jordan.html>
- cs.oswego.edu/~kzeller