Northeastern University
Khoury College of
Computer Sciences

DS 4400
Machine Learning and Data Mining I

Zohair Shafi

Spring 2026

Wednesday | January 7, 2026

Today's Outline

1. Introductions
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

Today’s Outline

1. **Introductions**
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

Introductions

About Me

Zohair Shafi
(he/him)

- B.E. in Computer Science from P.E.S University (2019)
- Performance Engineer at Akamai Technologies (2019-2021)
- Ph.D. at Northeastern University (2021-2026)
    - Advised by Prof. Tina Eliassi-Rad
    - I work at the intersection of Machine Learning and Network Science
    - I’ve worked on graph machine learning for combinatorial optimization problems, gene co-expression networks, adversarial robustness, explainability & fairness and reasoning in LLMs

Introductions

Teaching Assistants

Zaiba Amla

Wanrou Yang

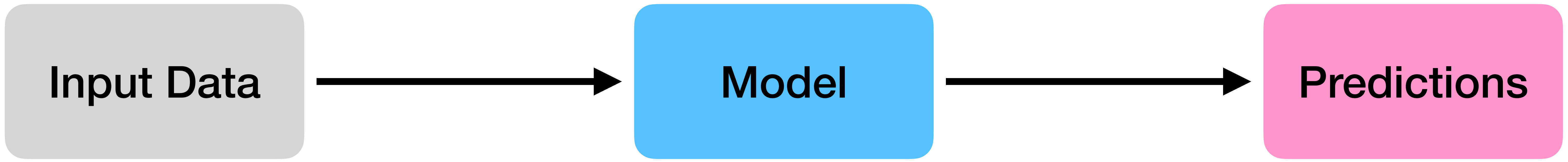
Today's Outline

1. Introductions
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

Today's Outline

1. Introductions
2. **What is Machine Learning (ML)**
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

What is Machine Learning?



What is Machine Learning?

Input Data

Model

Predictions

Bedrooms

Sq. Ft.

Zip Code

Let’s look at a concrete example

What is Machine Learning?

Input Data

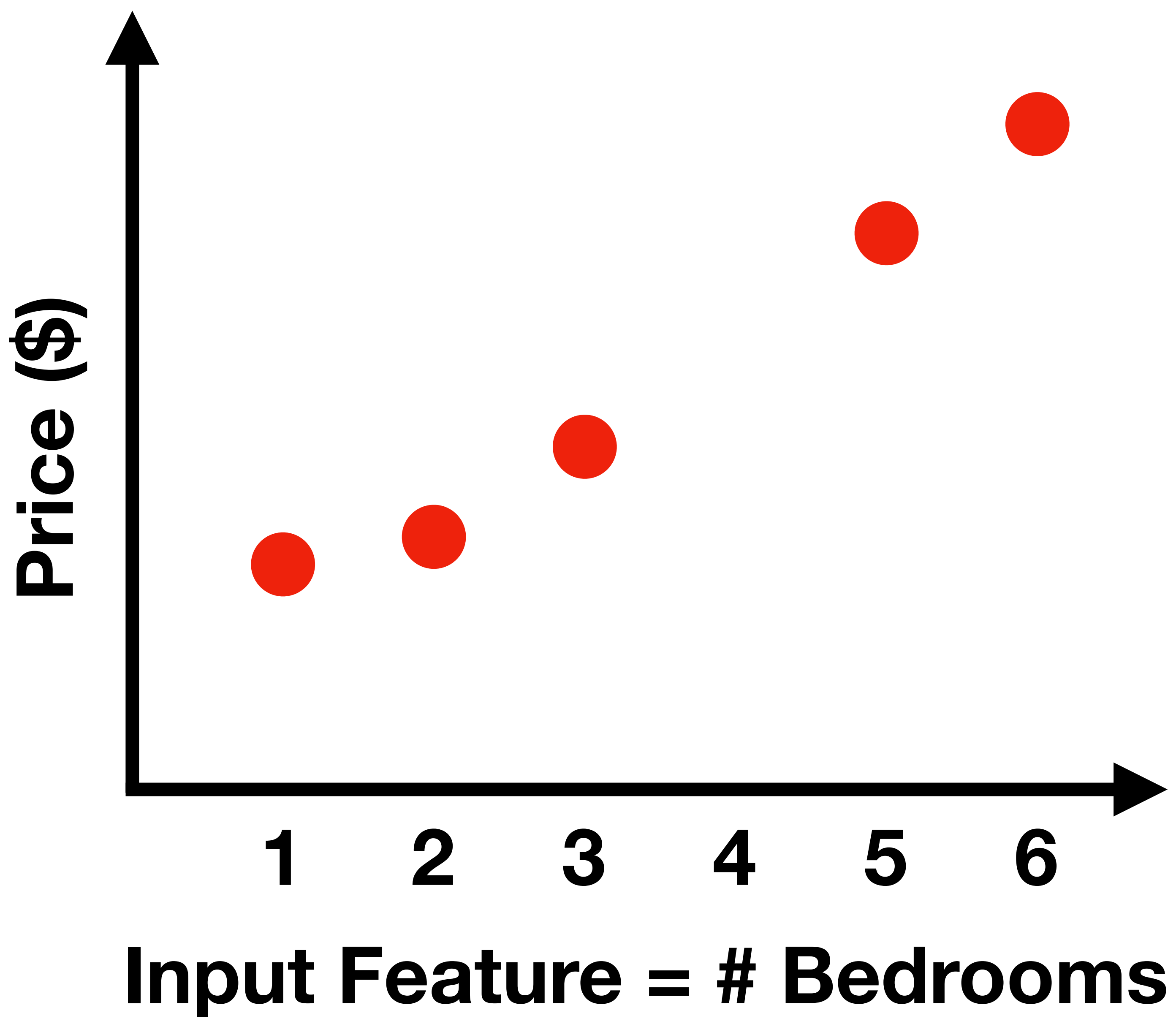
Model

Predictions

Bedrooms

Sq. Ft.

Zip Code



Price	# Bedrooms
2000	1
2100	2
2400	3
3000	5
3500	6

Let’s look at a concrete example

What is Machine Learning?

Input Data

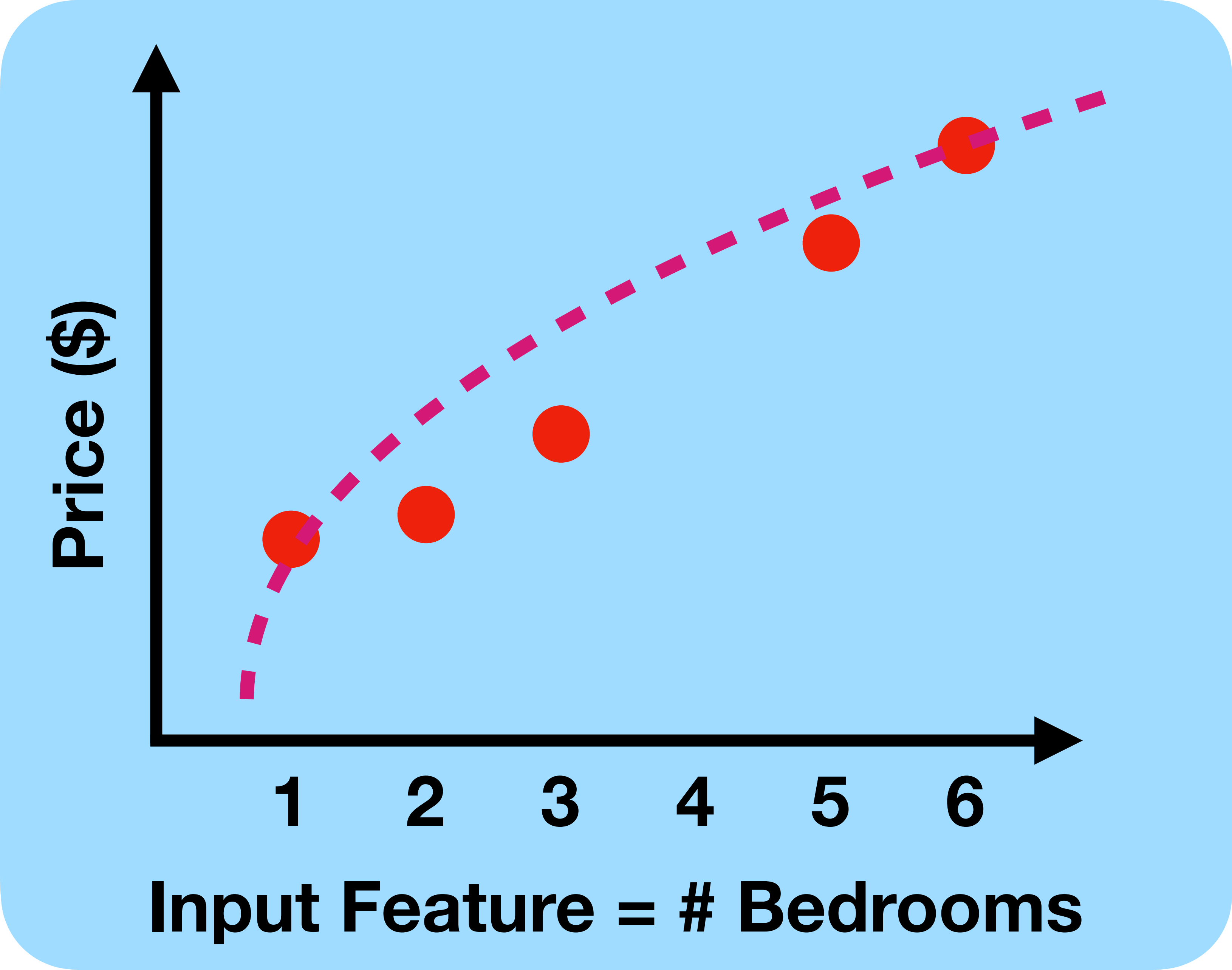
Model

Predictions

Bedrooms

Sq. Ft.

Zip Code



Price	# Bedrooms
2000	1
2100	2
2400	3
3000	5
3500	6

Let’s look at a concrete example

What is Machine Learning?

Input Data

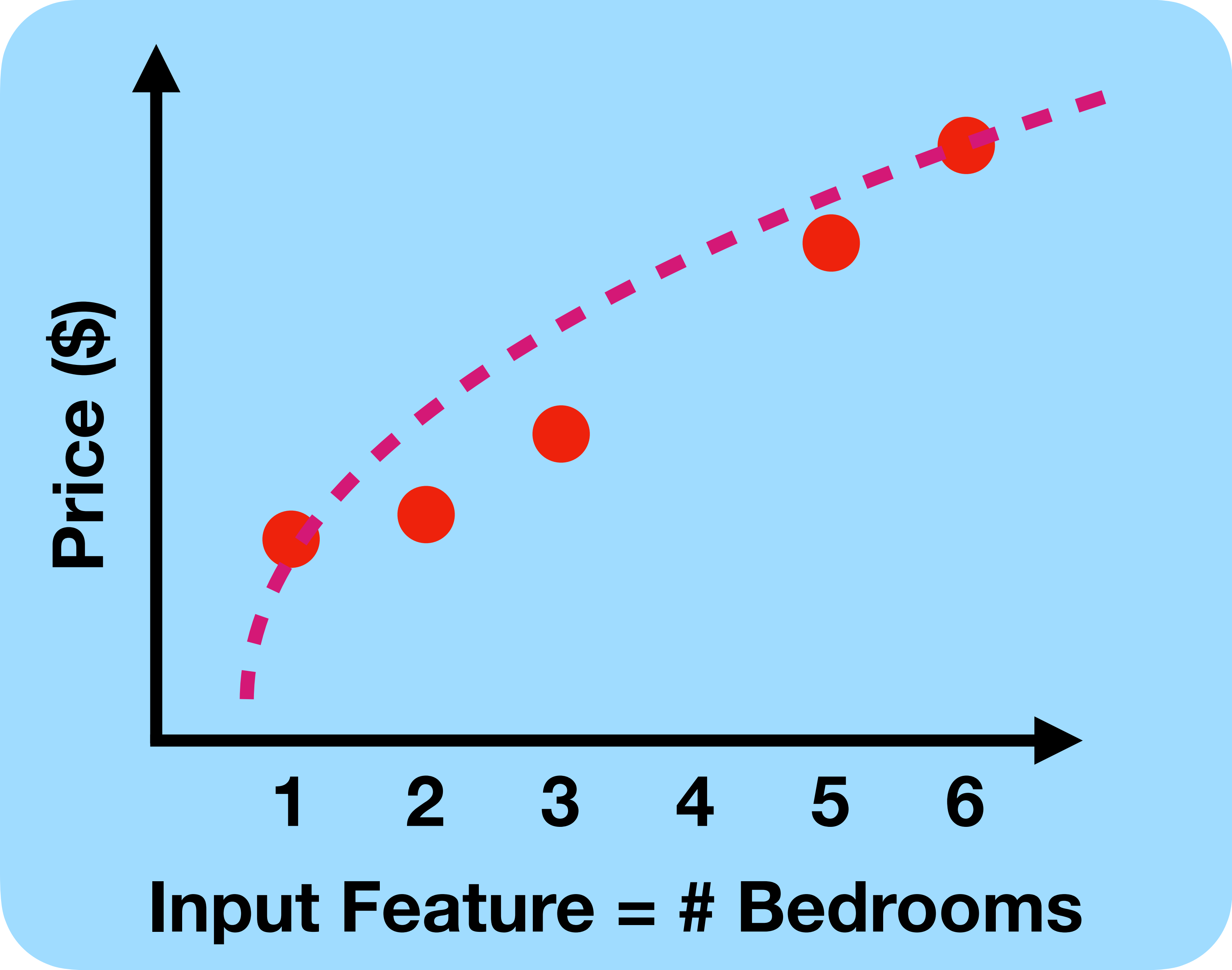
Model

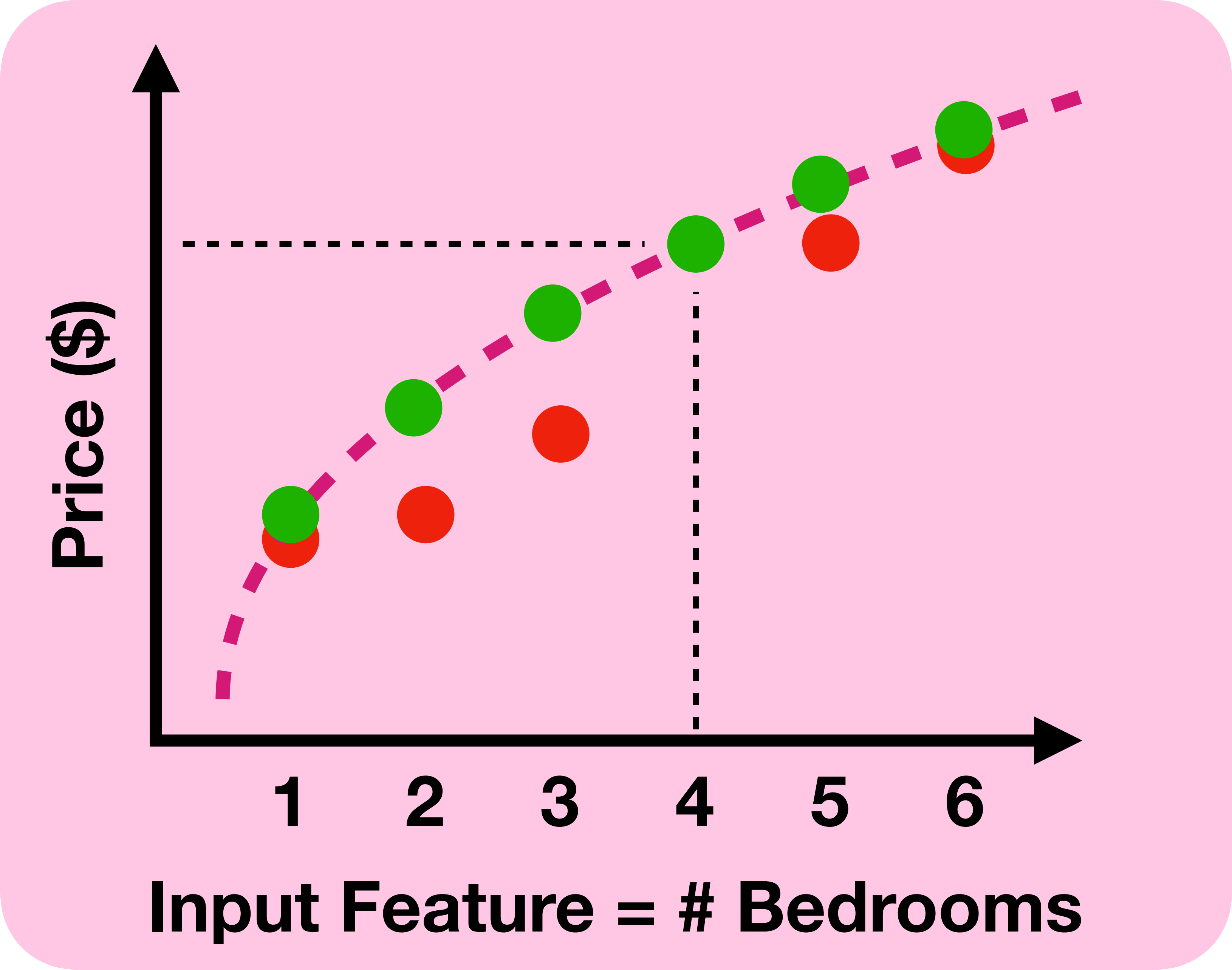
Predictions

Bedrooms

Sq. Ft.

Zip Code





Let’s look at a concrete example

What is Machine Learning?

Input Data

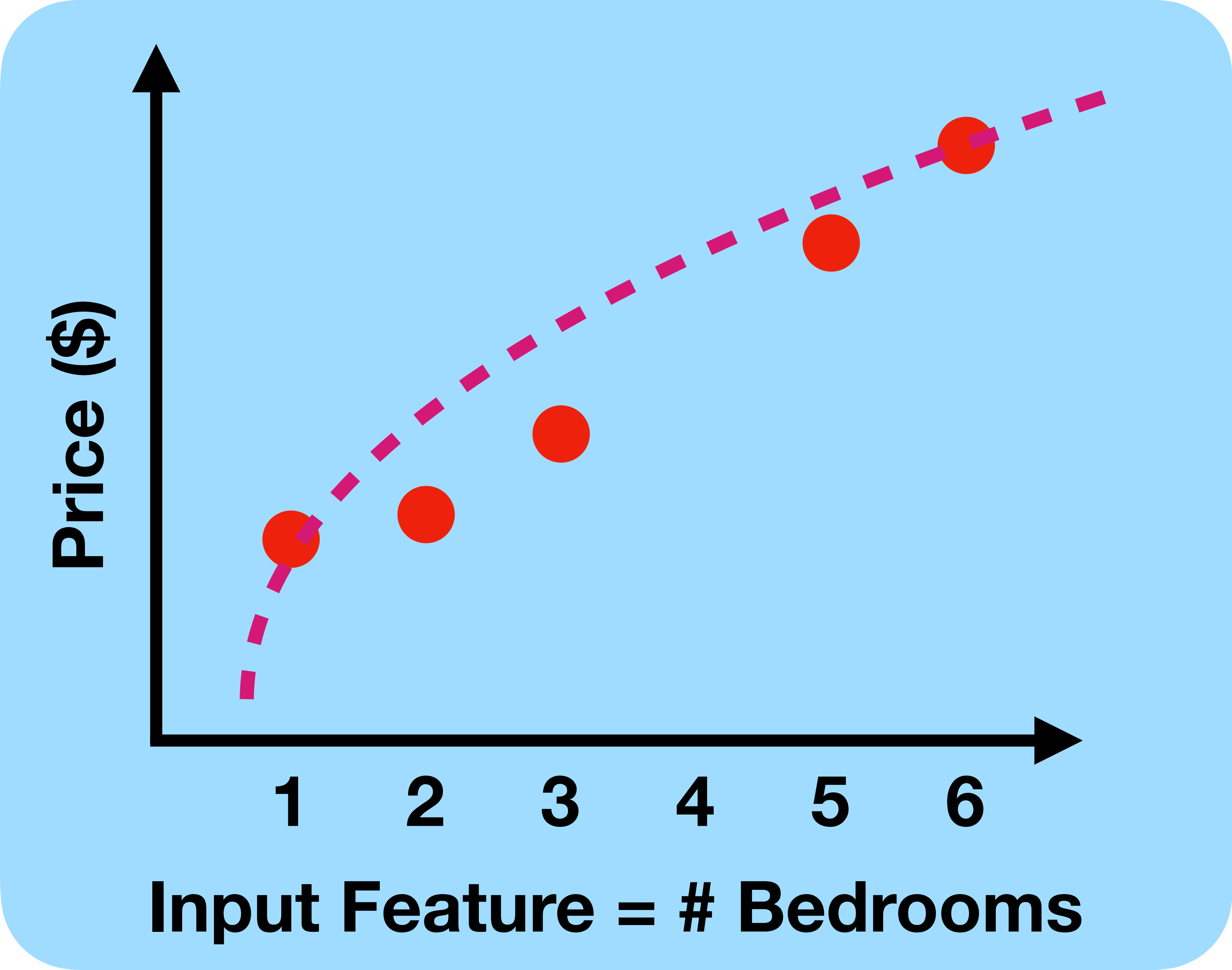
Model

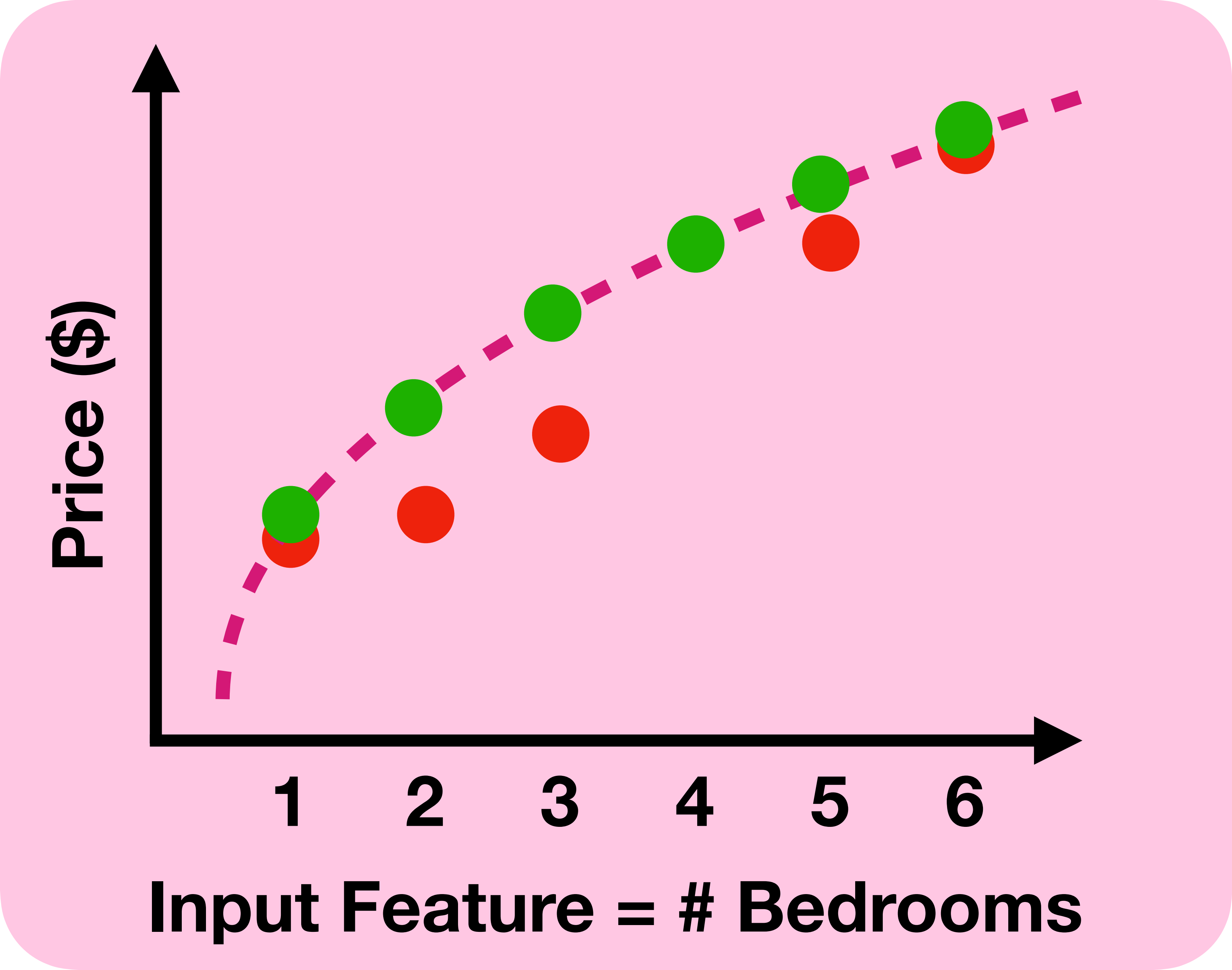
Predictions

Bedrooms

Sq. Ft.

Zip Code





What if we change the input data feature?

What is Machine Learning?

Input Data

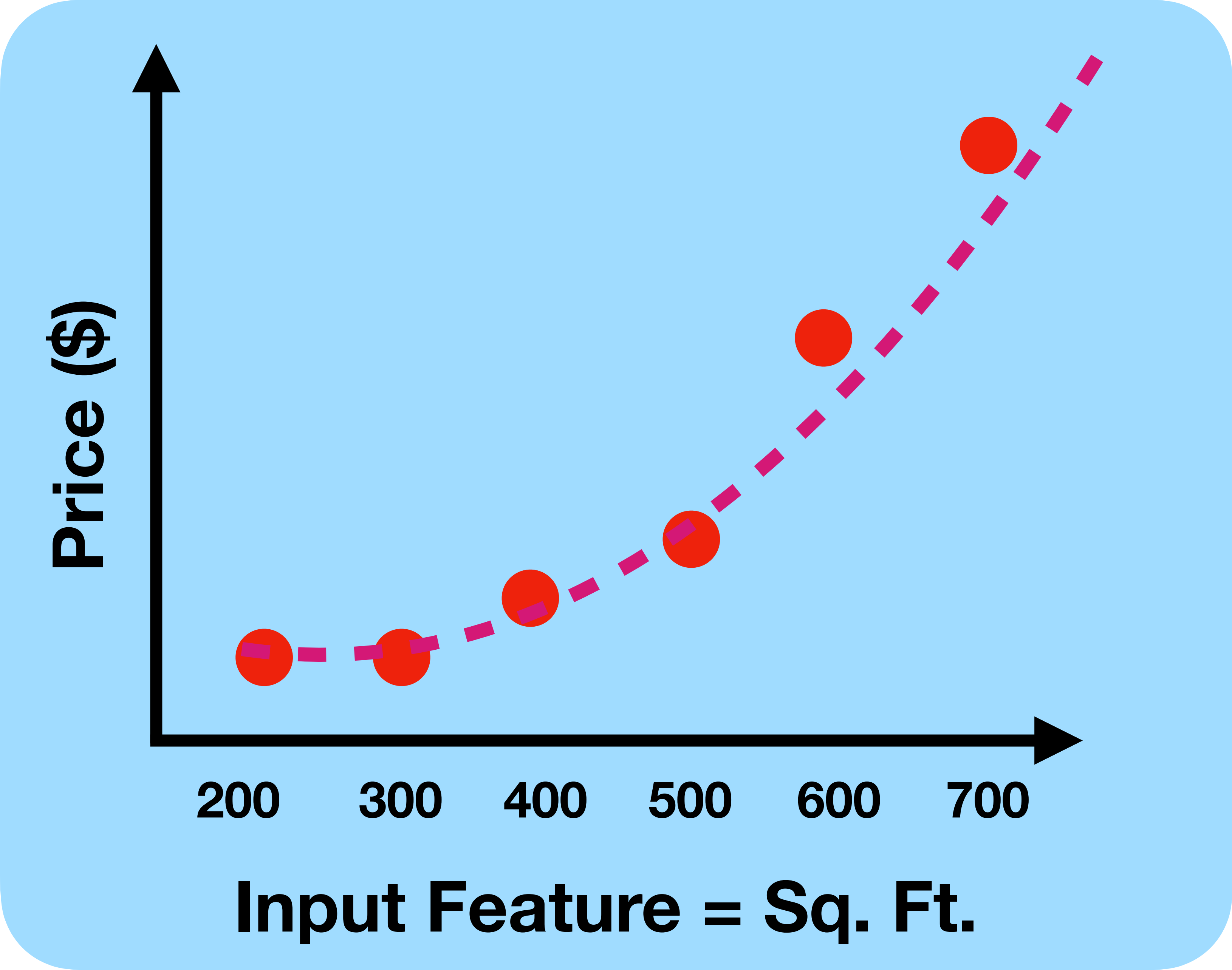
Model

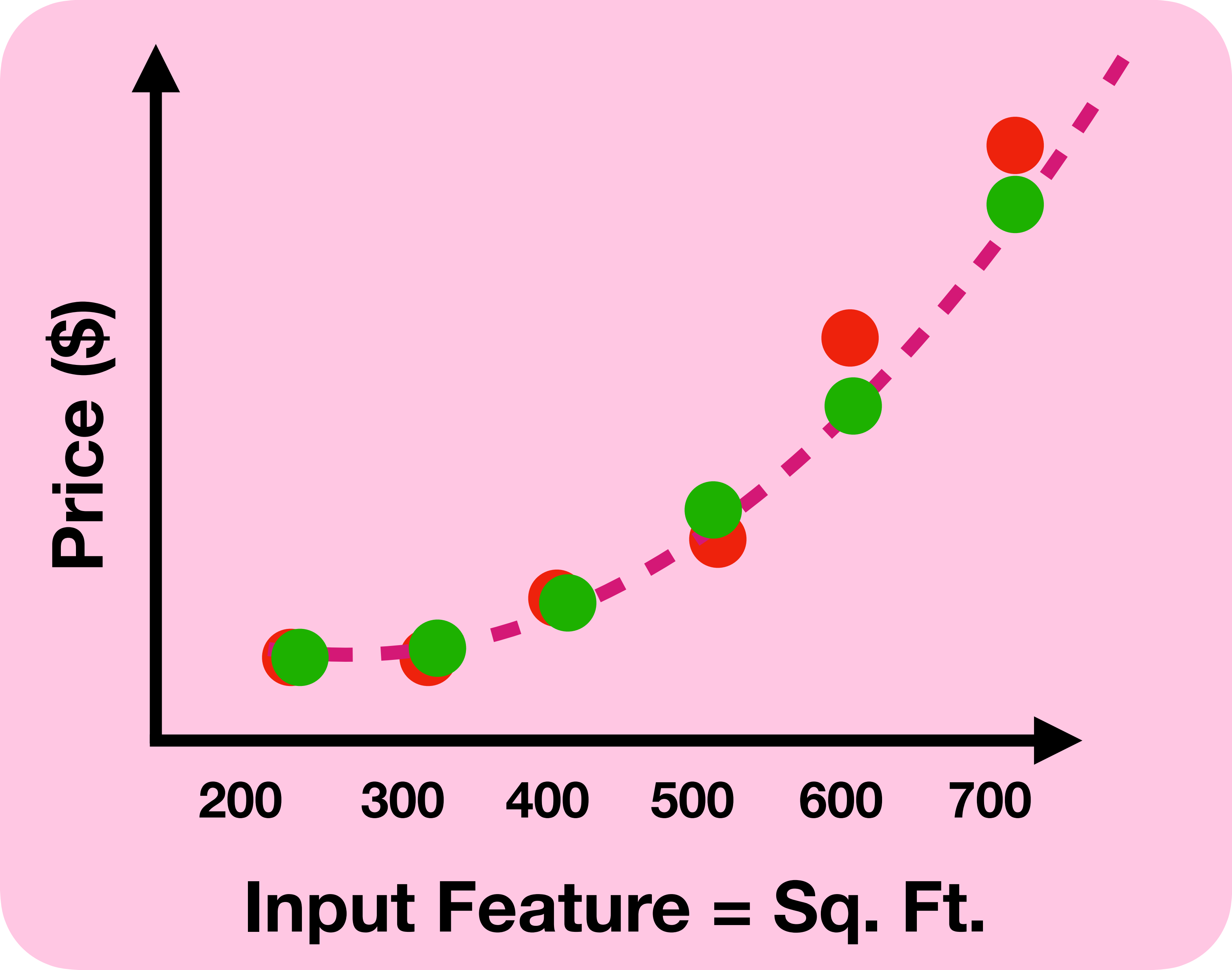
Predictions

Bedrooms

Sq. Ft.

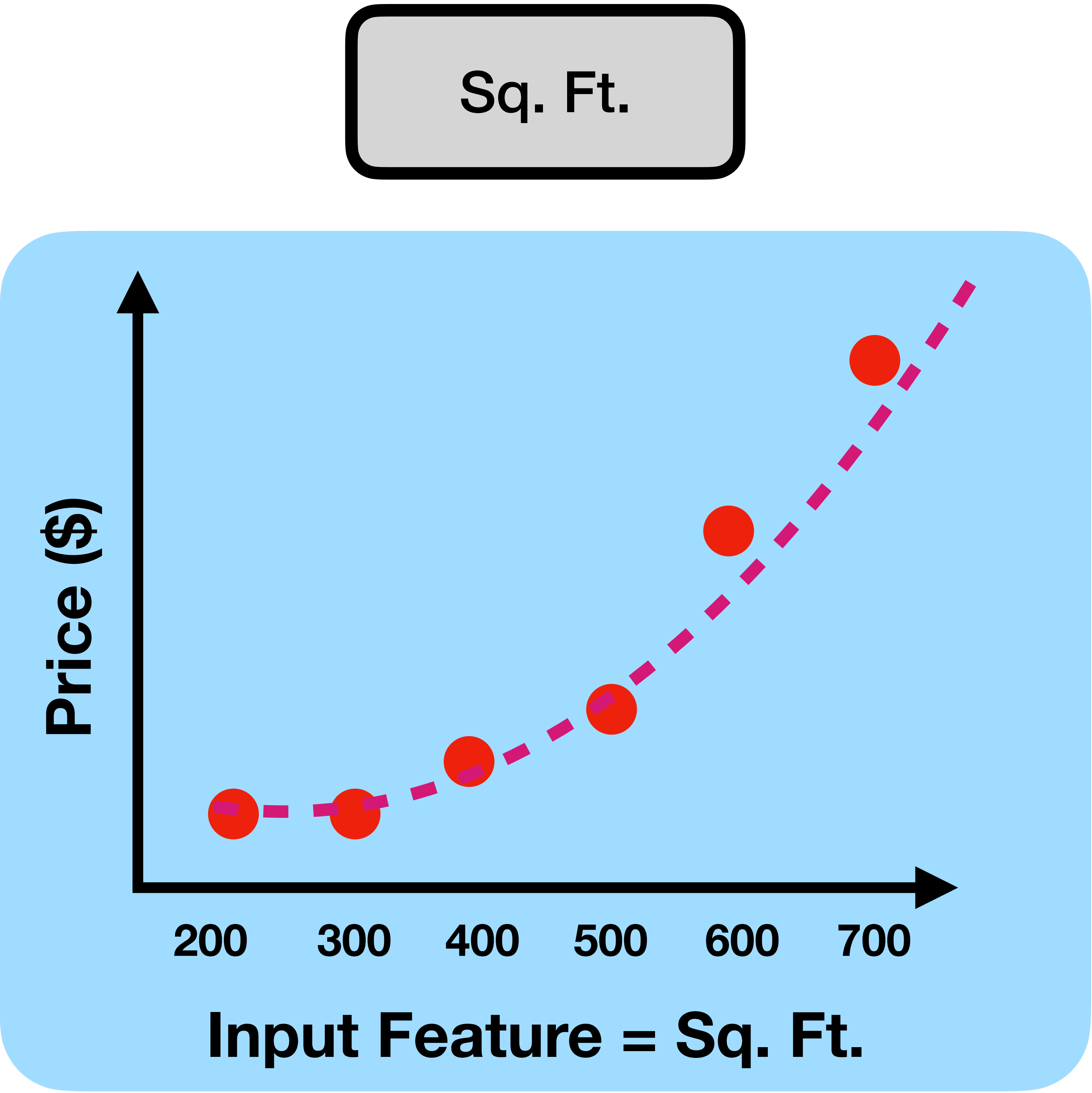
Zip Code

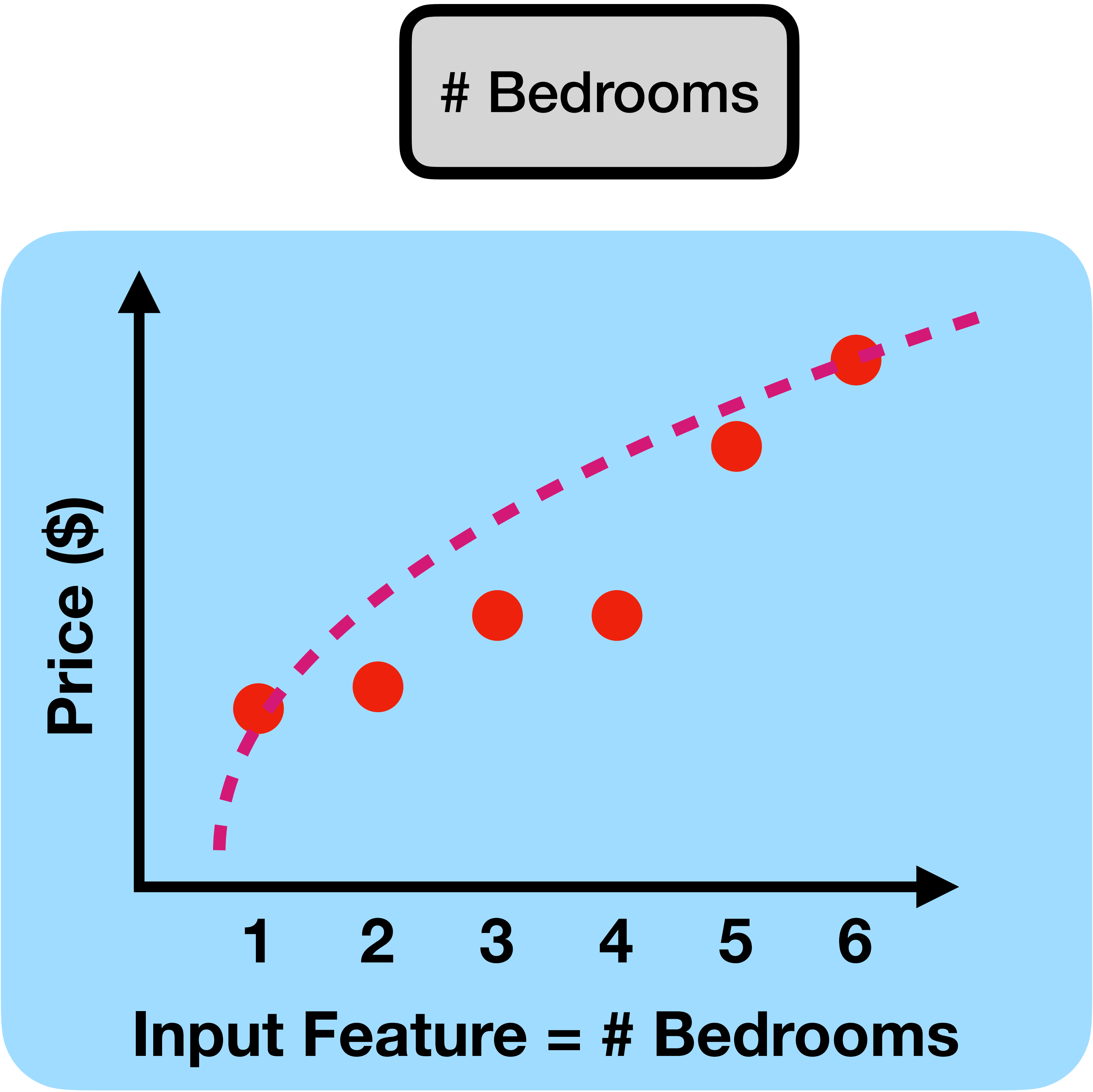




What if we change the input data feature?

What is Machine Learning?

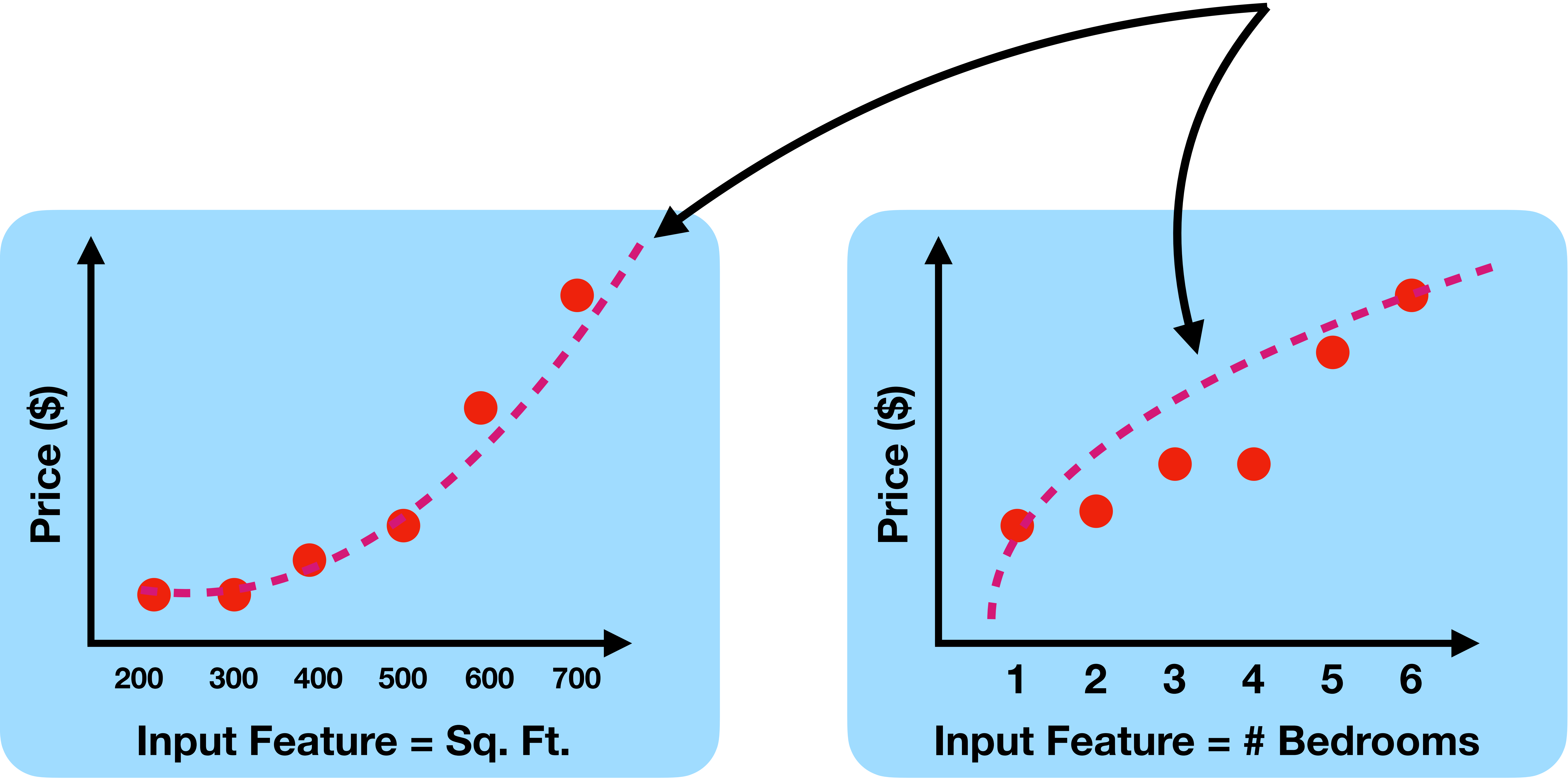




Notice that the curve learned for **Sq. Ft.** is very different from the curve learned for **# of Bedrooms**

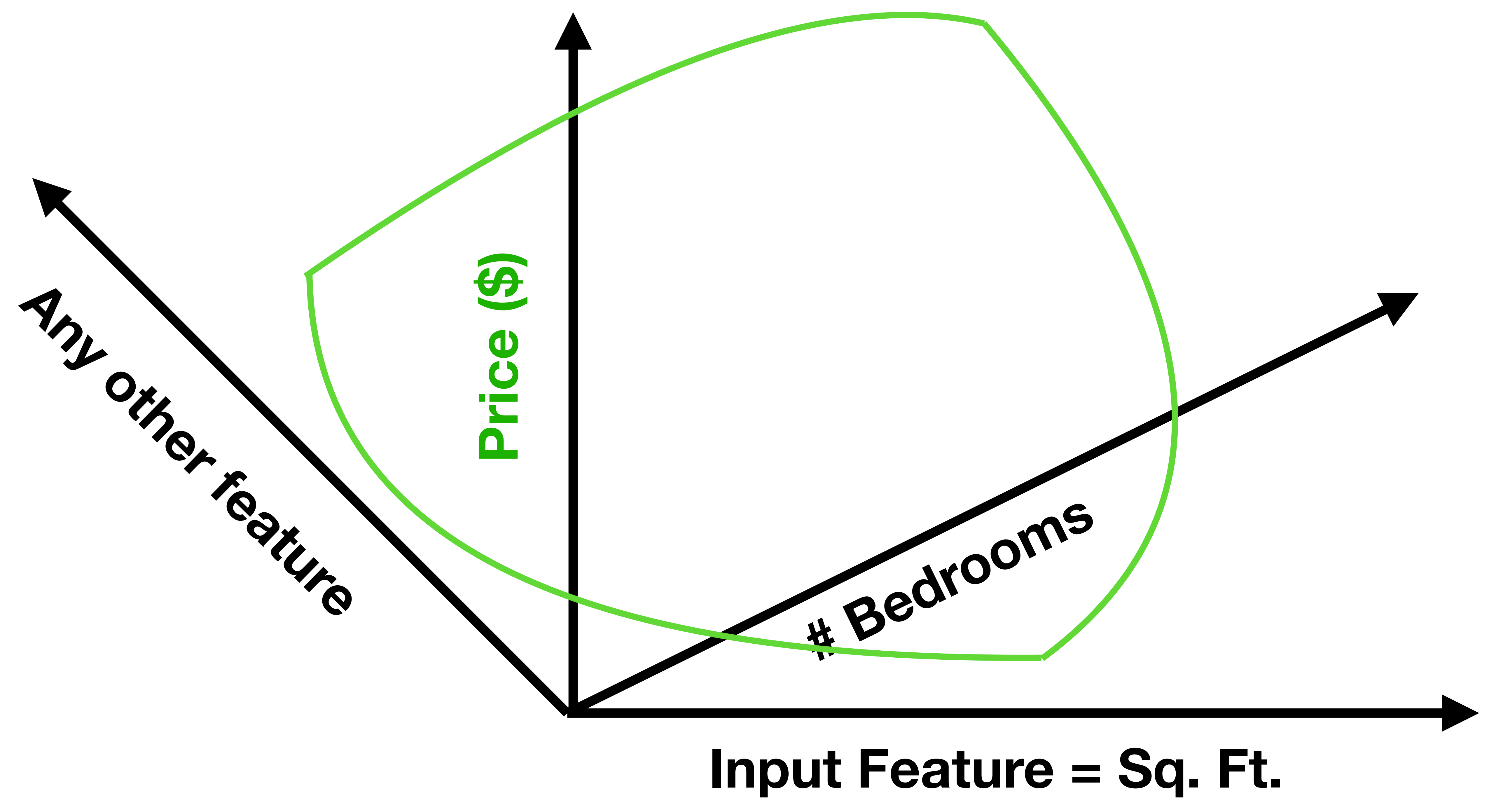
What is Machine Learning?

Machine Learning is the task of trying to learn these curves



What is Machine Learning?

Machine Learning is the task of trying to learn these curves
This task gets harder when you have **multiple** input features



Today’s Outline

1. Introductions
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

Today’s Outline

1. Introductions
2. What is Machine Learning (ML)
3. **Course outline & Logistics**
4. What is Machine Learning (a little more detail)

Course Objectives

Types of ML

- Supervised vs Unsupervised
- Classification vs Regression
- Generative AI

ML Algorithms

- Linear Regression, Spline Regression
- SVM, Decision Trees, Naive Bayes, Ensembles
- Neural Networks

Applications

- Fairness and Ethics
- Explainability
- Security

Course Outline

Probability and Linear Algebra Review (~1 Week)

Linear Regression and Regularization (~2 Weeks)

Classification (~5 Weeks)

- Linear classifiers: logistic regression, LDA
- Non-linear classifiers: kNN, decision trees, SVM, Naive Bayes
- Ensembles: random forests, boosting and bagging

Neural Networks and Deep Learning (~2 Weeks)

- Backpropagation, gradient descent
- Various NN architectures

Applications (~2 Weeks)

- Fairness and Ethics in AI
- Security and Privacy

Course Information

Course Website: https://zohairshafi.github.io/pages/sp26_ds4400.html
Course calendar and slides posted after each lecture

Canvas:
Assignments and grades posted here

Gradescope:
Assignent Submissions
Accessed via Canvas

Emails:
Please ensure all emails to instructor/TA's have [sp26_ds4400] in the subject line.
This helps attend to emails faster.

Course Schedule

Class Hours:

Monday and Wednesday | 02:50 PM - 04:30 PM | Snell 033

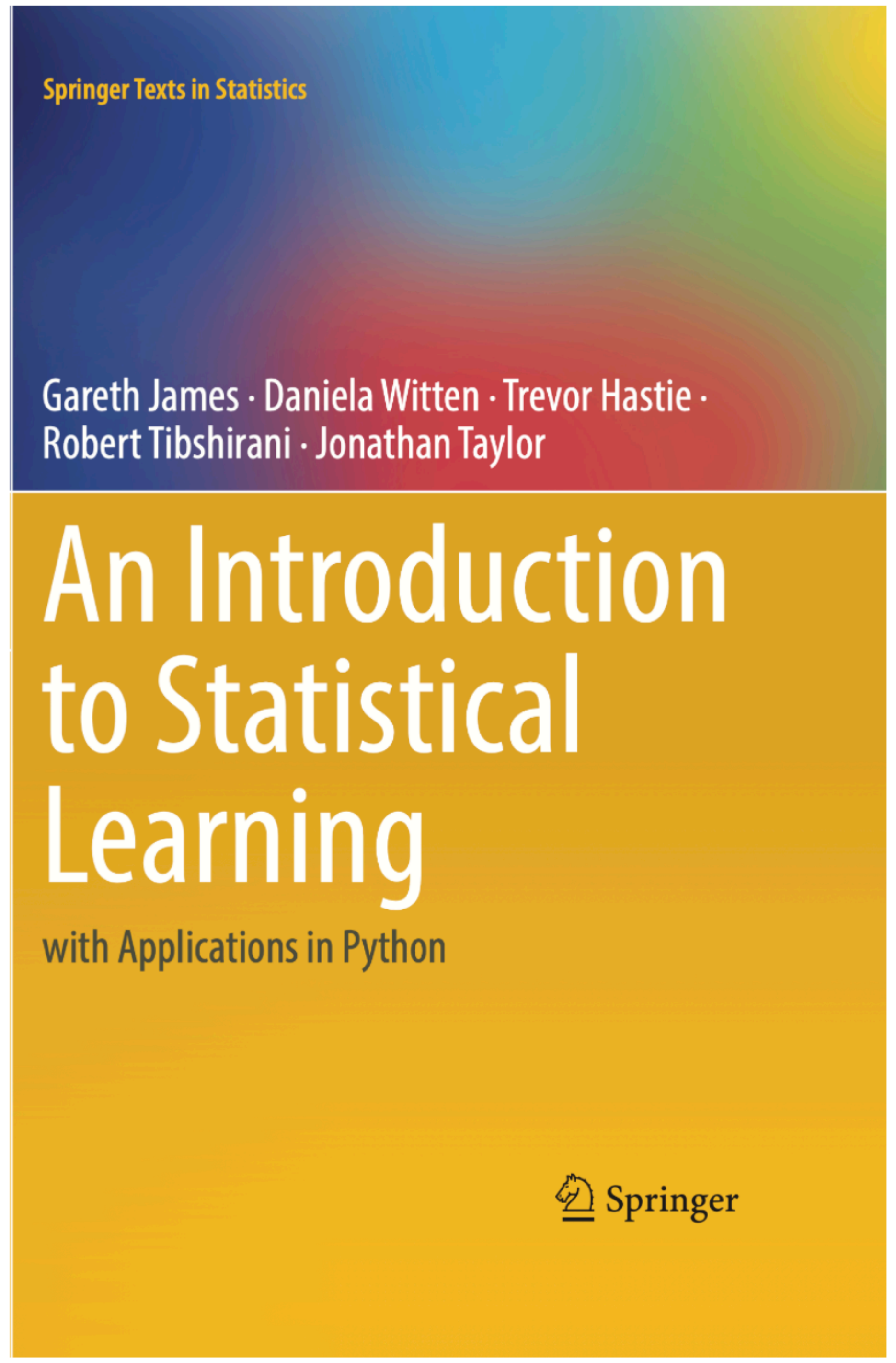
Office Hours:

Wanrou Yang: 1:30 PM - 3:00 PM - Tuesday (Location: TBD)

Zaiba Amla: 1:00 PM - 2:30 PM - Wednesday (Location: TBD)

Zohair Shafi: 1:30PM - 2:30 PM - Monday and Wednesday (Location: TBD)

Resources



Textbook:
An Introduction to Statistical Learning
https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html

Other Resources:

- **Elements of Statistical Learning (Trevor Hastie, Rob Tibshirani, and Jerry Friedman,) Second Edition, Springer, 2009**
- **Pattern Recognition and Machine Learning (Christopher Bishop) Springer, 2006**
- **Dive into Deep Learning (A. Zhang, Z. Lipton, and A. Smola)**
- **Lecture notes by Andrew Ng from Stanford**

Policies

Your Responsibilities

- Please be on time, attend classes, and take notes
- Participate in interactive discussion in class
- Submit assignments / programming projects on time

Late Days for Assignments

- 5 total late days, after that loose 20% for every late day
- Assignments are due at 11:59pm on the specified date
- We will use Gradescope for submitting assignments
- No need to email for late days

Grading

Assignments - 12.5%

8 assignments and programming exercises based on studied material in class

Theory and practical assignments with Jupyter Notebooks

Midterm Exam - 20%

Tentative date: Wednesday, February 18

Final Exam - 25%

Scheduled during finals week

Class participation - 5%

Academic Integrity

- Homework is done individually.
- Rules
  - Can discuss with colleagues or instructors
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet/LLMs
  - Use python packages, but not directly code for ML analysis written by someone else
  - No LLM usage.
- **No cheating will be tolerated.**
  - Any cheating will automatically result in grade F and report to the university administration
  - http://www.northeastern.edu/osccr/academic-integrity-policy/

Today's Outline

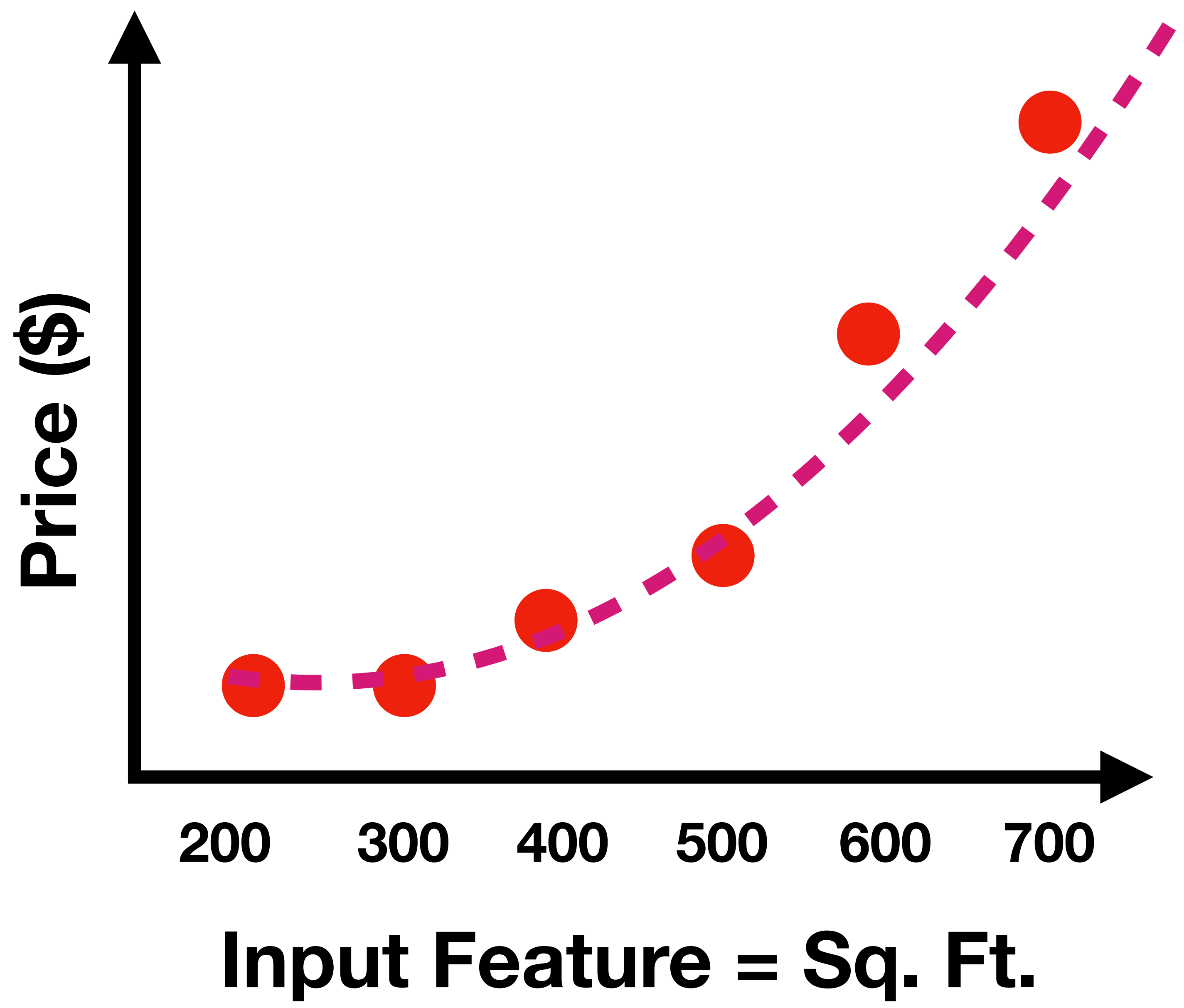
1. Introductions
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. What is Machine Learning (a little more detail)

Today's Outline

1. Introductions
2. What is Machine Learning (ML)
3. Course outline & Logistics
4. **What is Machine Learning (a little more detail)**

What is Machine Learning?

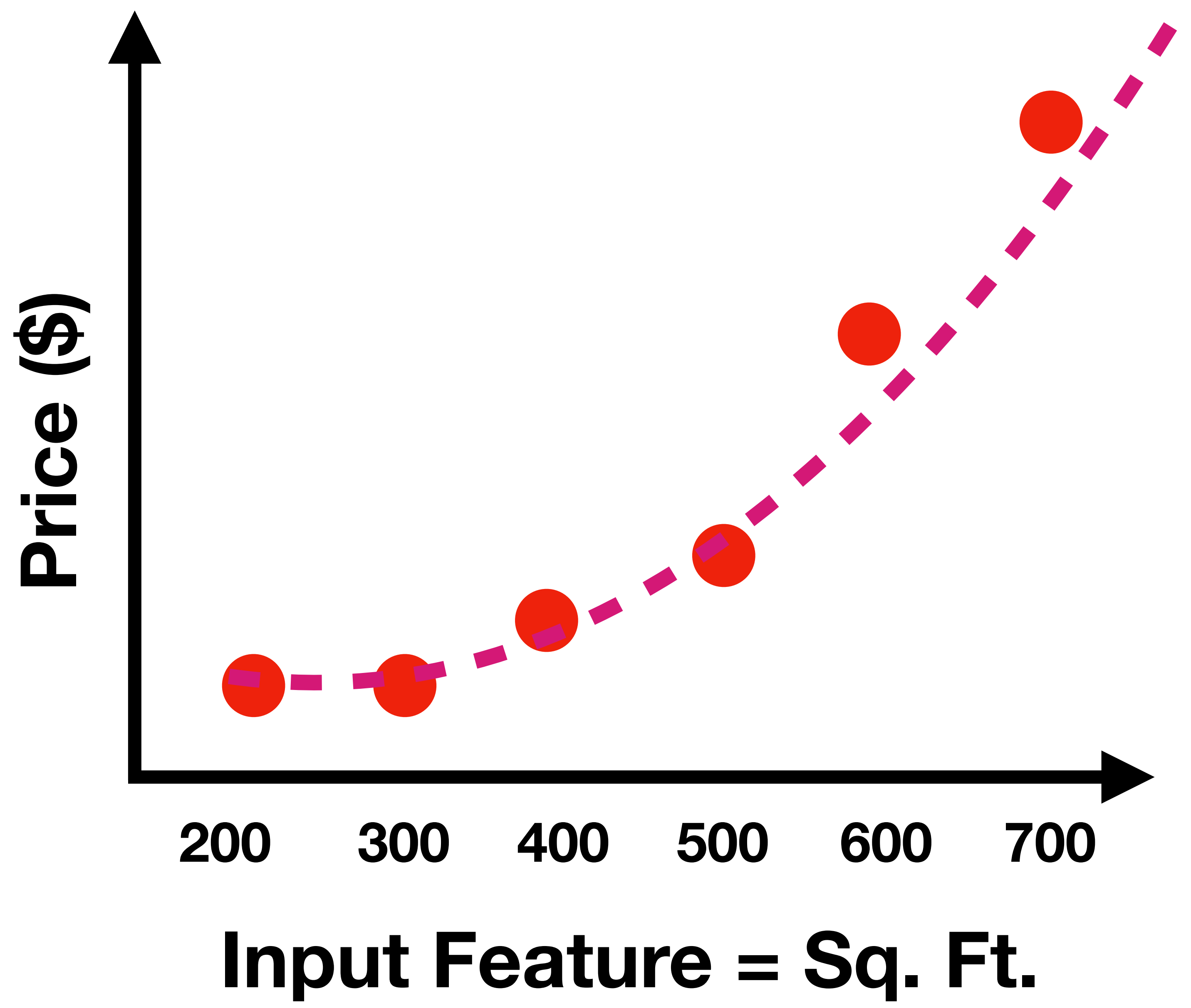
Machine Learning is the task of trying to learn these curves



Price	# Bedrooms
2000	1
2100	2
2400	3
2450	4
3000	5
3500	6

What is Machine Learning?

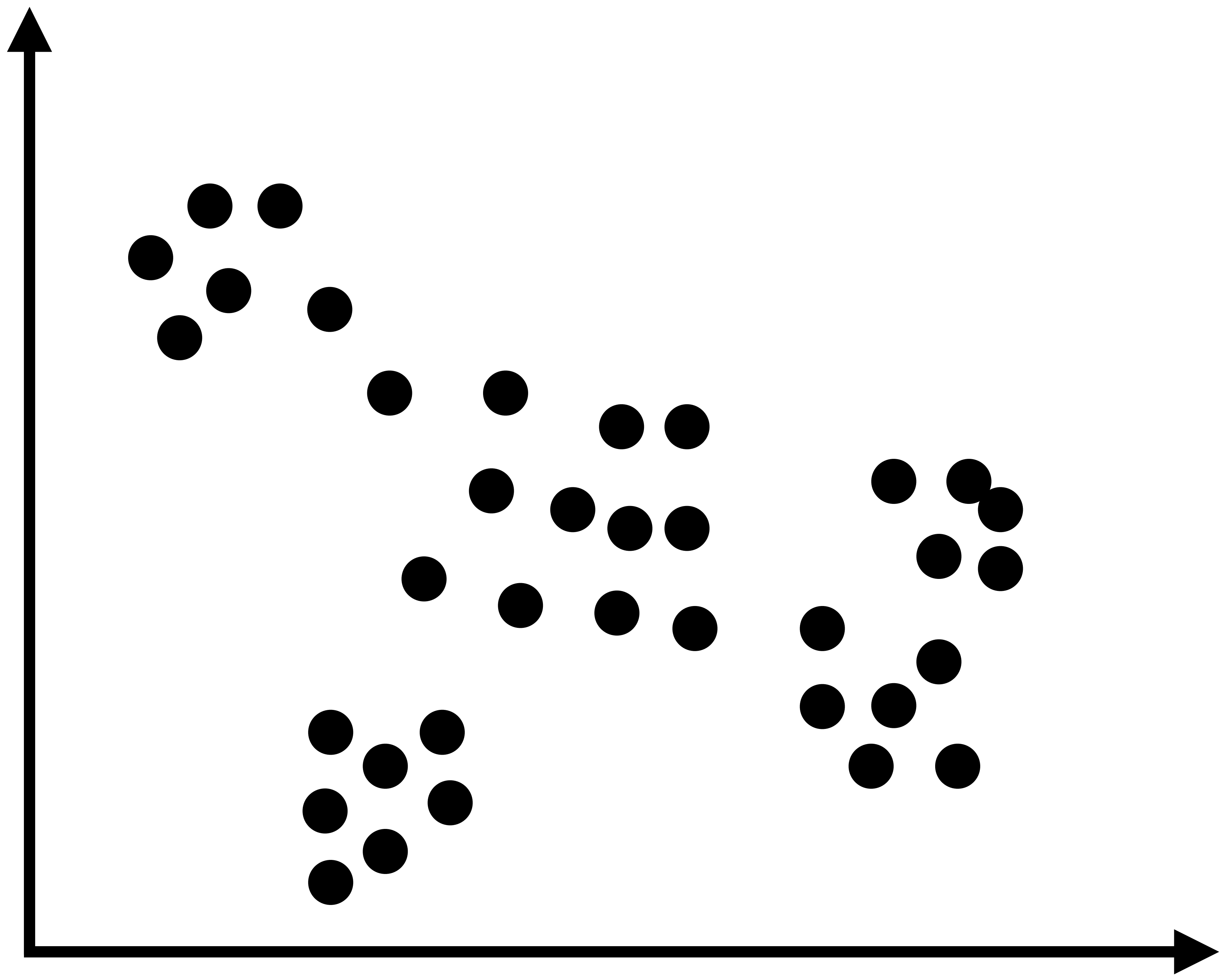
This type of ML is called **Supervised** Learning



Price	# Bedrooms
2000	1
2100	2
2400	3
2450	4
3000	5
3500	6

What is Machine Learning?

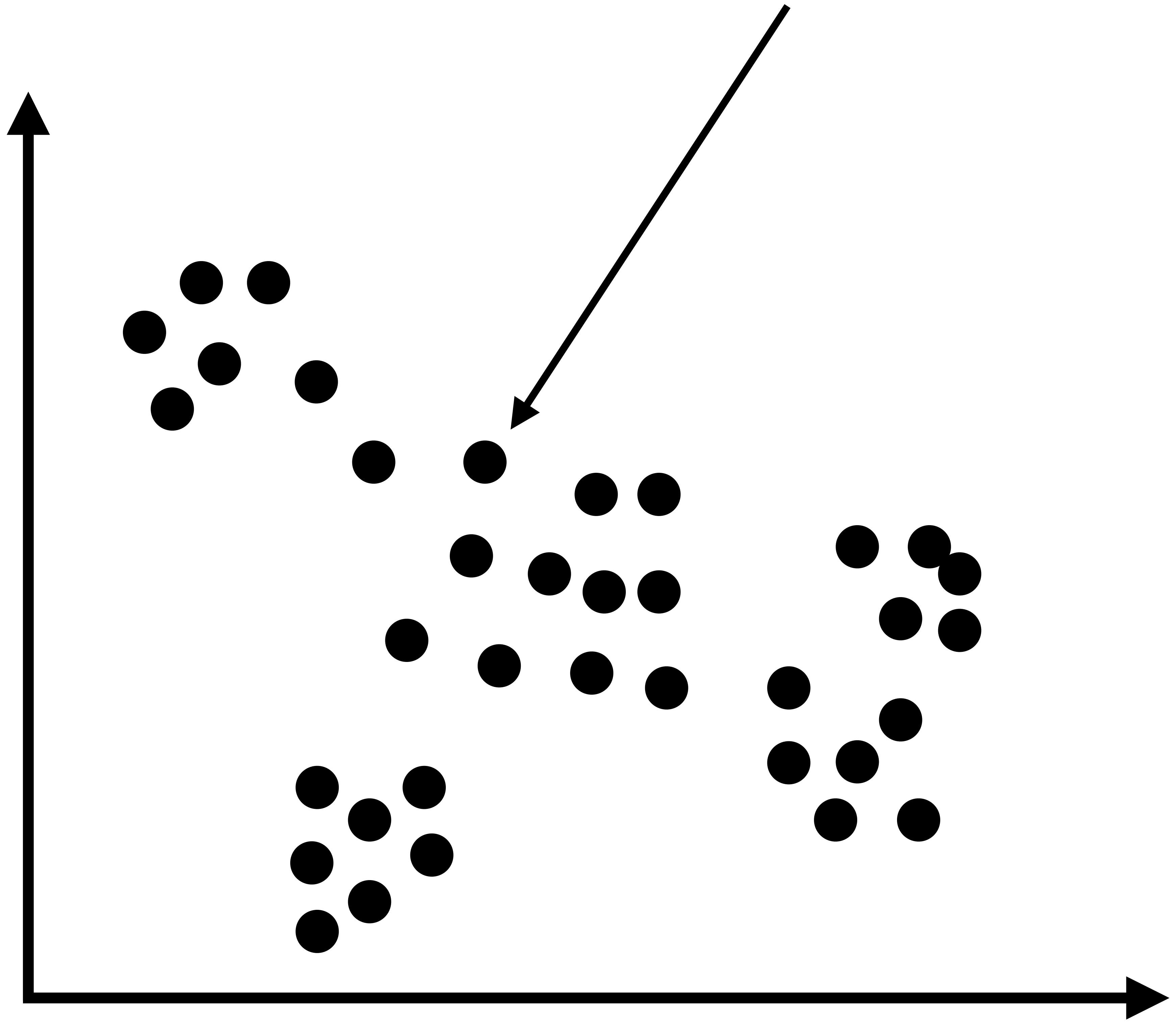
What about when you do **not** have training labels to learn from?



What is Machine Learning?

What about when you do **not** have training labels to learn from?

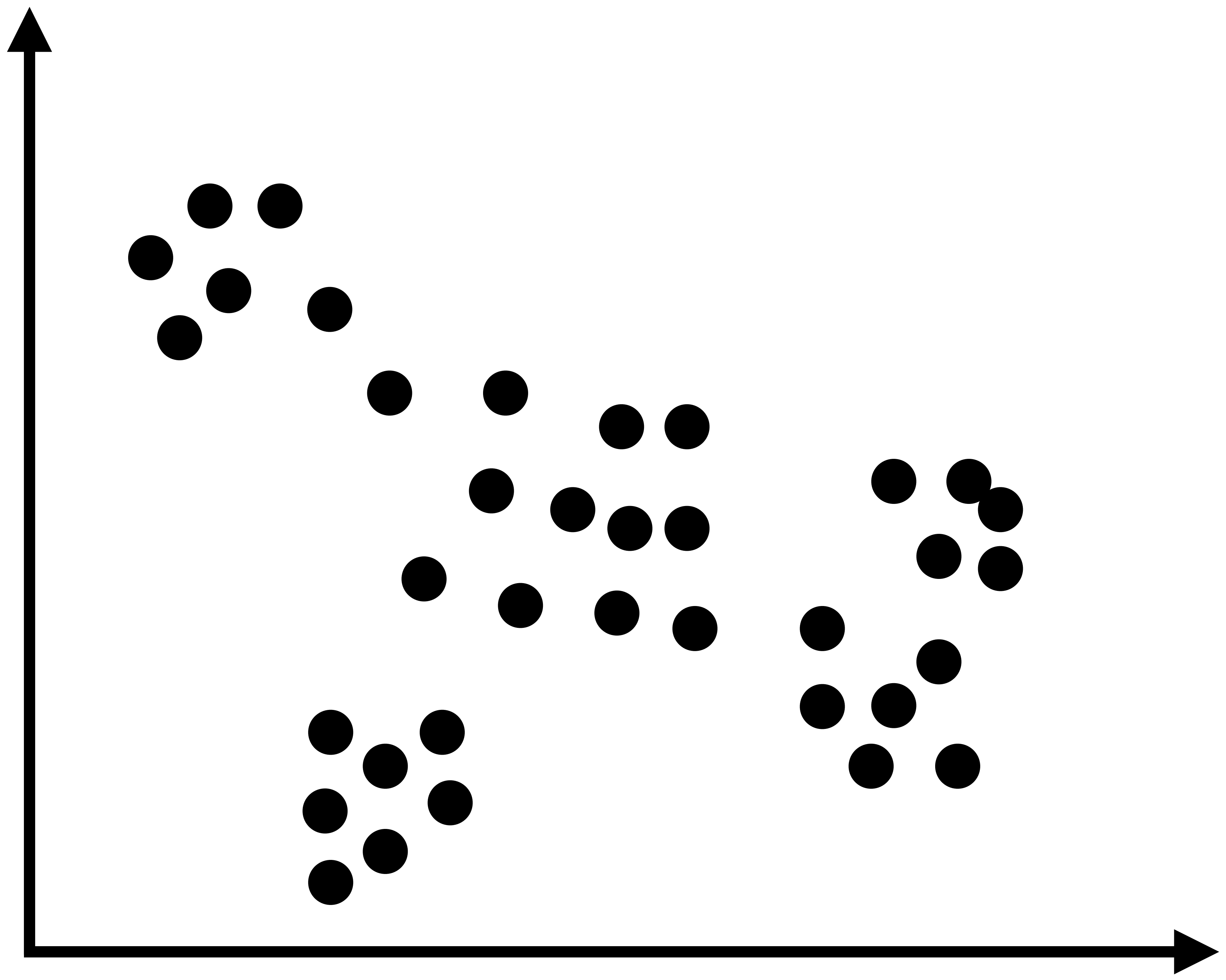
Lets assume that **each data point** here is a song



What is Machine Learning?

What about when you do **not** have training labels to learn from?

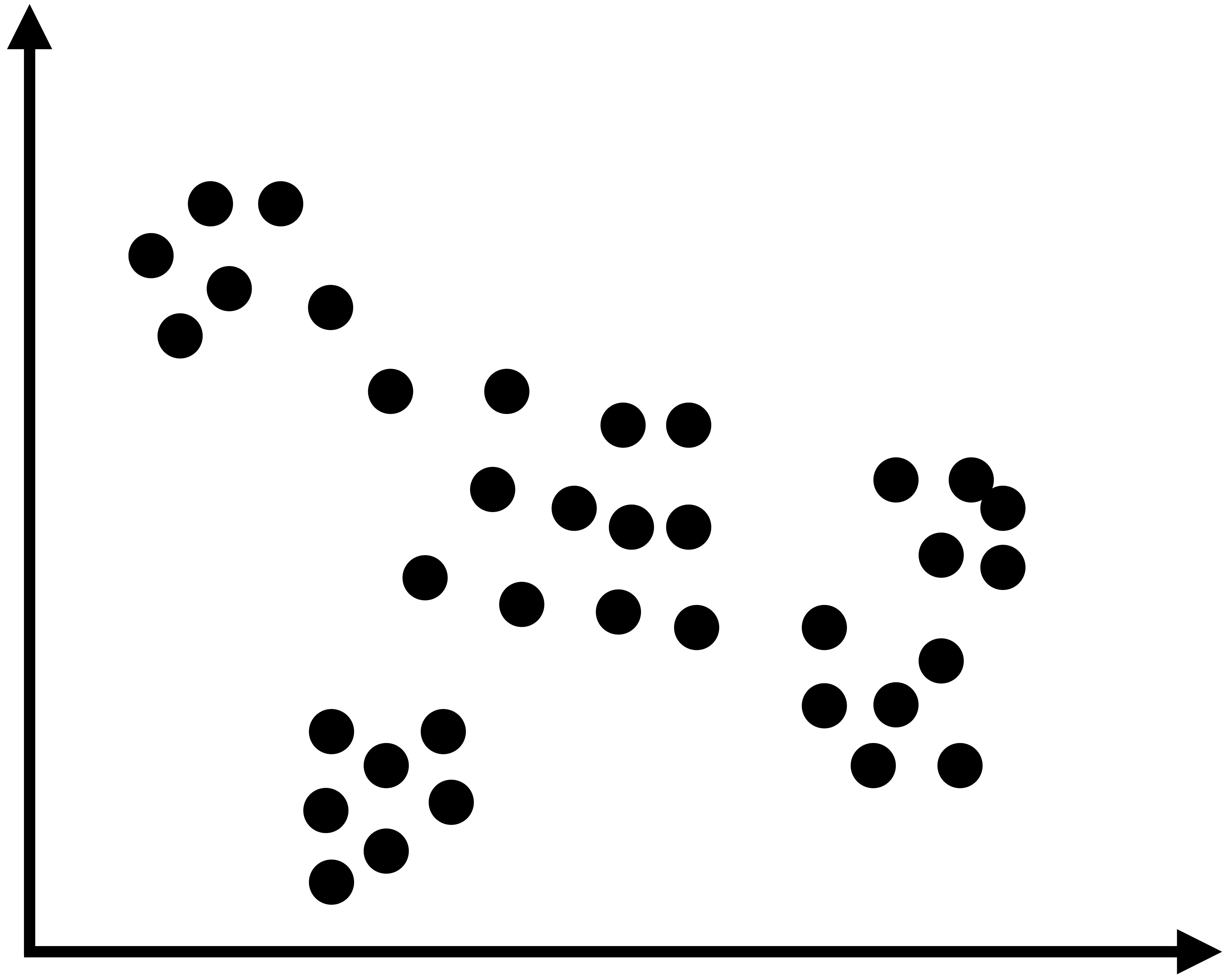
Lets assume that each data point here is a song
How would you **learn** from this data?



What is Machine Learning?

What about when you do **not** have training labels to learn from?

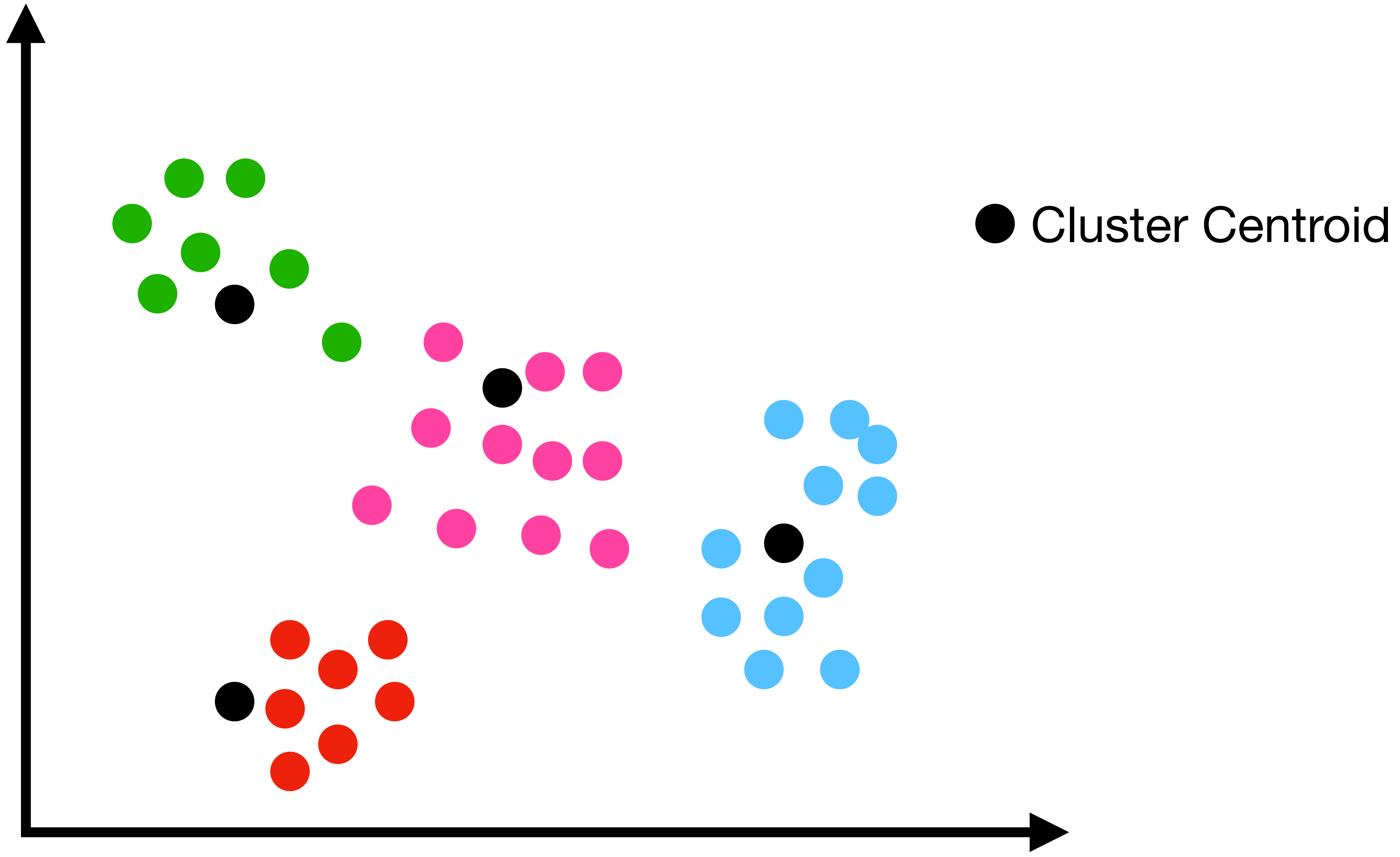
This is where **Unsupervised Learning Algorithms** come in.
For example, we can use Clustering algorithms to chunk this data into groups



What is Machine Learning?

What about when you do **not** have training labels to learn from?

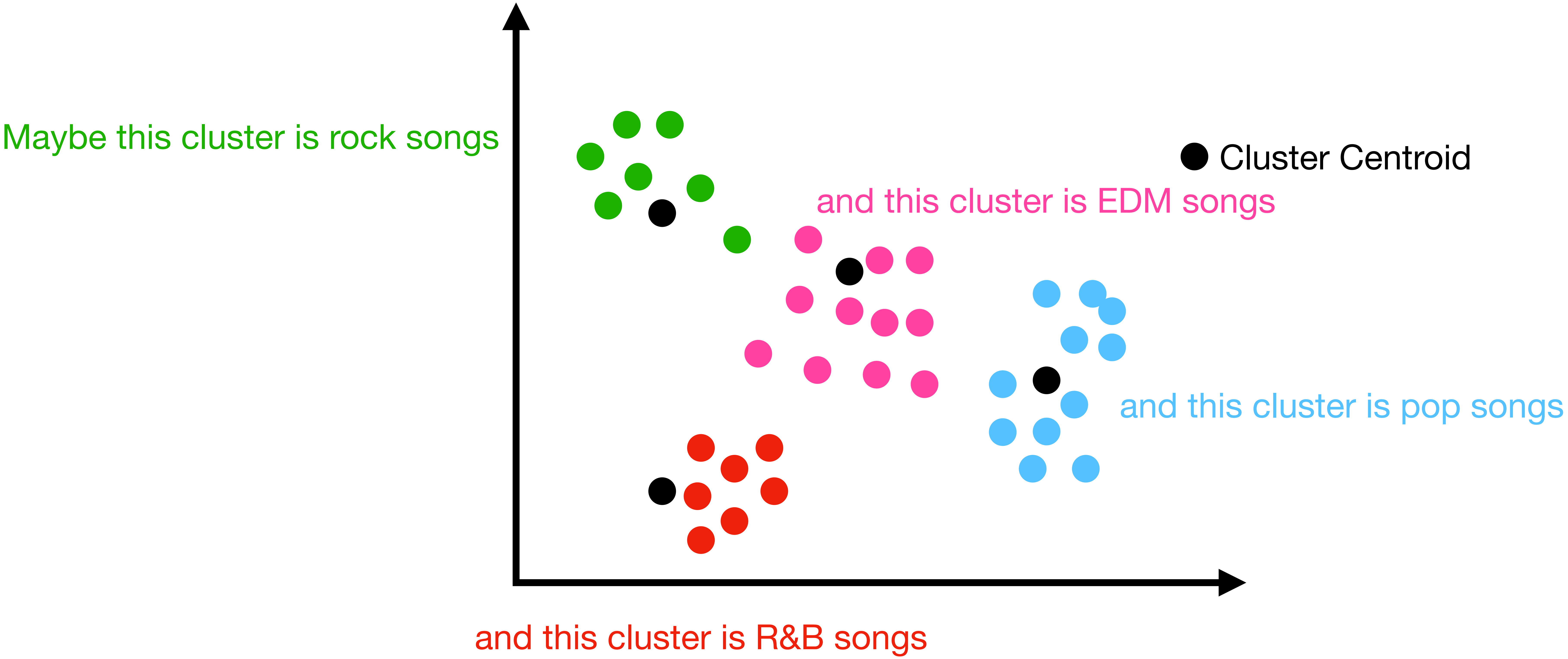
This is where **Unsupervised Learning Algorithms** come in.
For example, we can use Clustering algorithms to chunk this data into groups



What is Machine Learning?

What about when you do **not** have training labels to learn from?

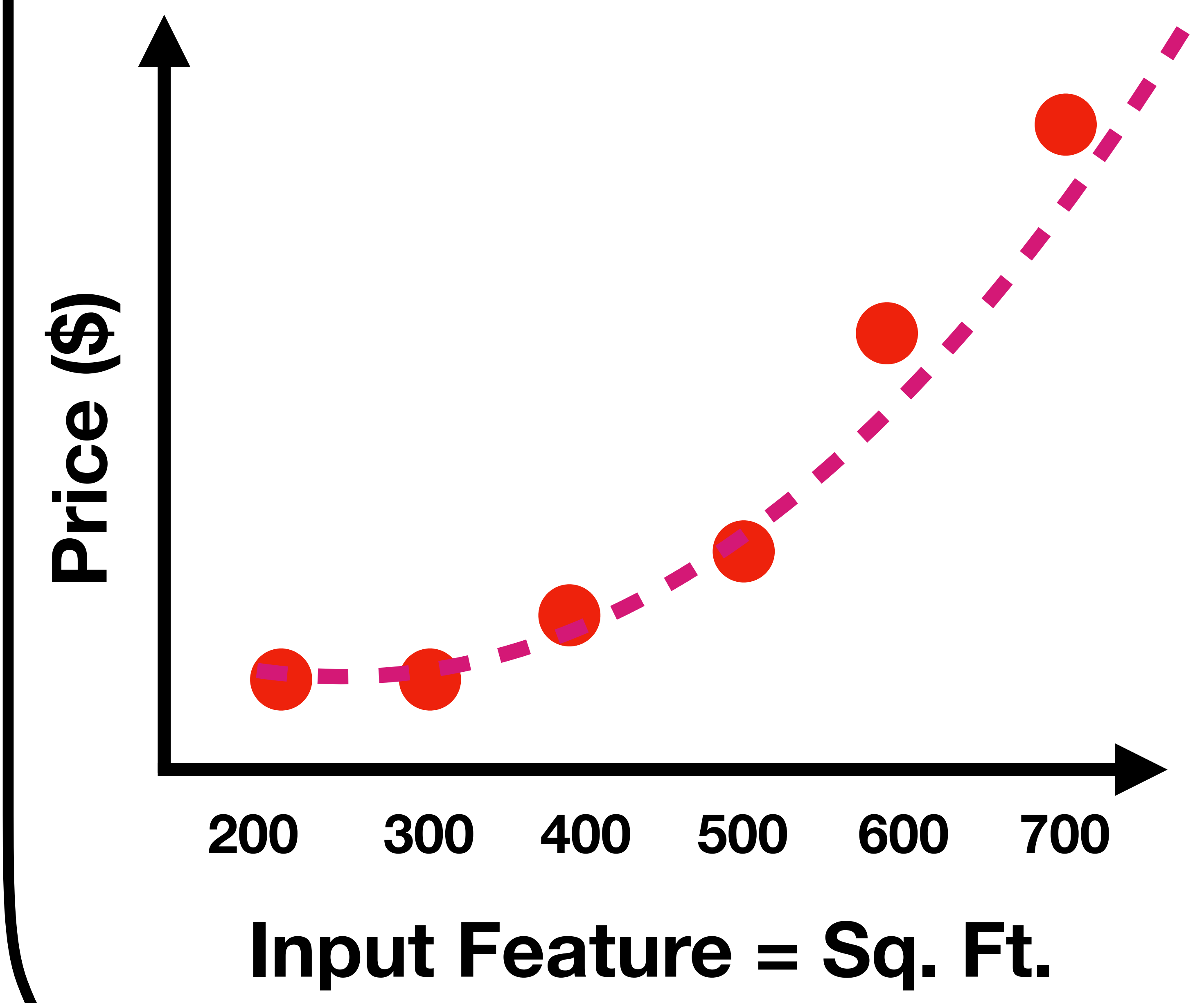
This is where **Unsupervised Learning Algorithms** come in.
For example, we can use Clustering algorithms to chunk this data into groups



What have we learned so far?

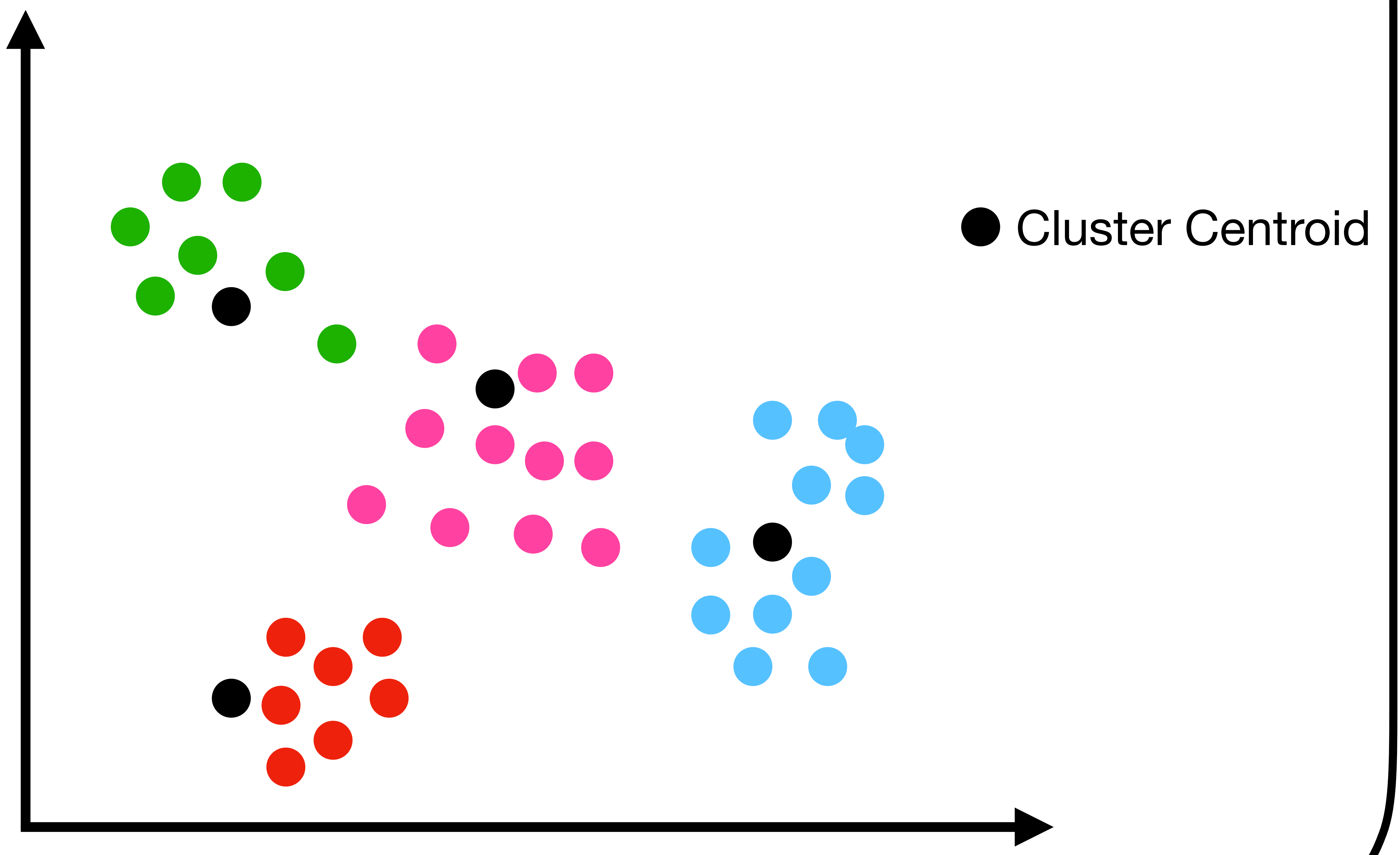
ML can be split into

Supervised Learning



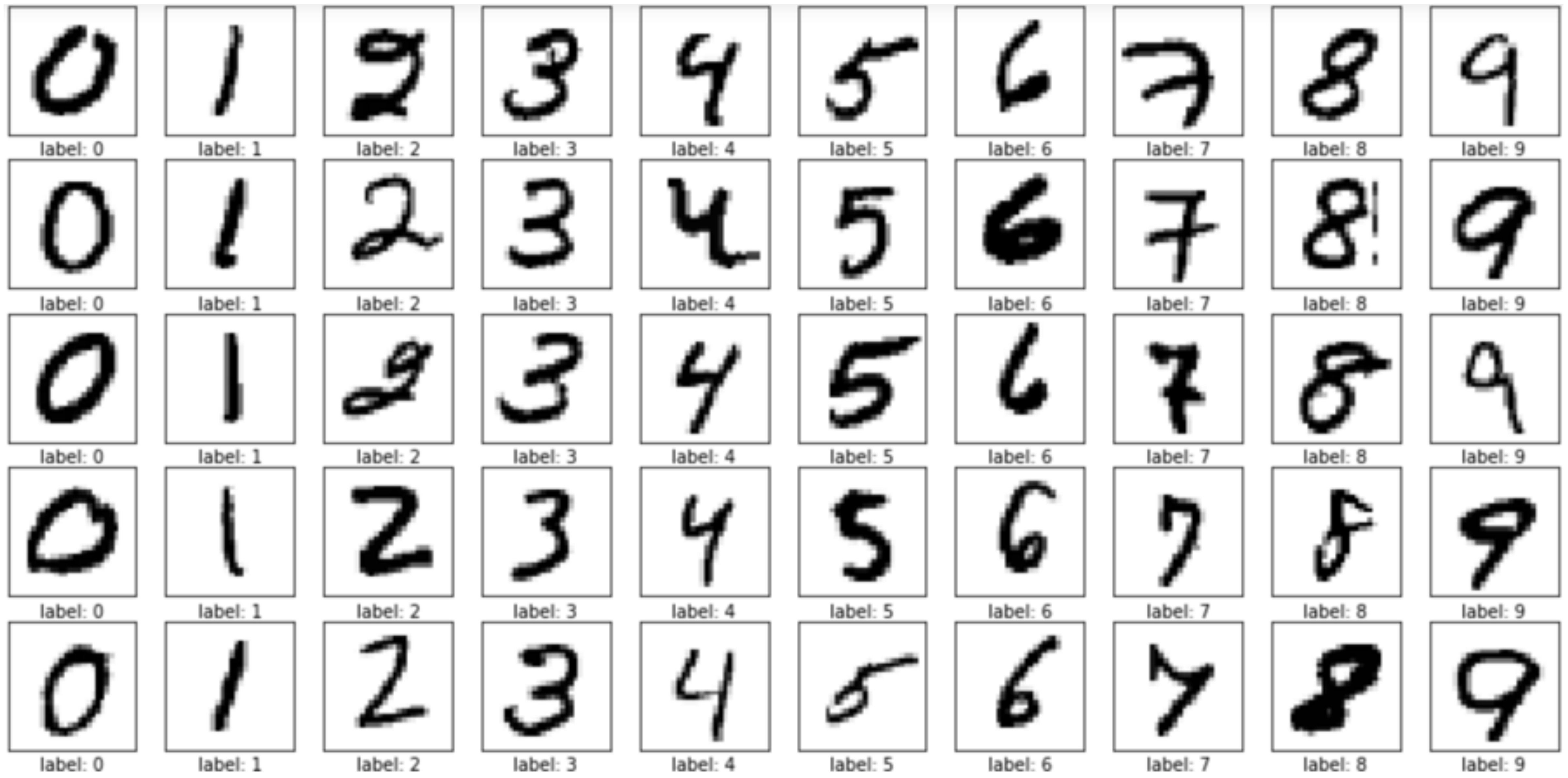
Price	# Bedrooms
2000	1
2100	2
2400	3
2450	4
3000	5
3500	6

Unsupervised Learning



Let’s look at some concrete examples

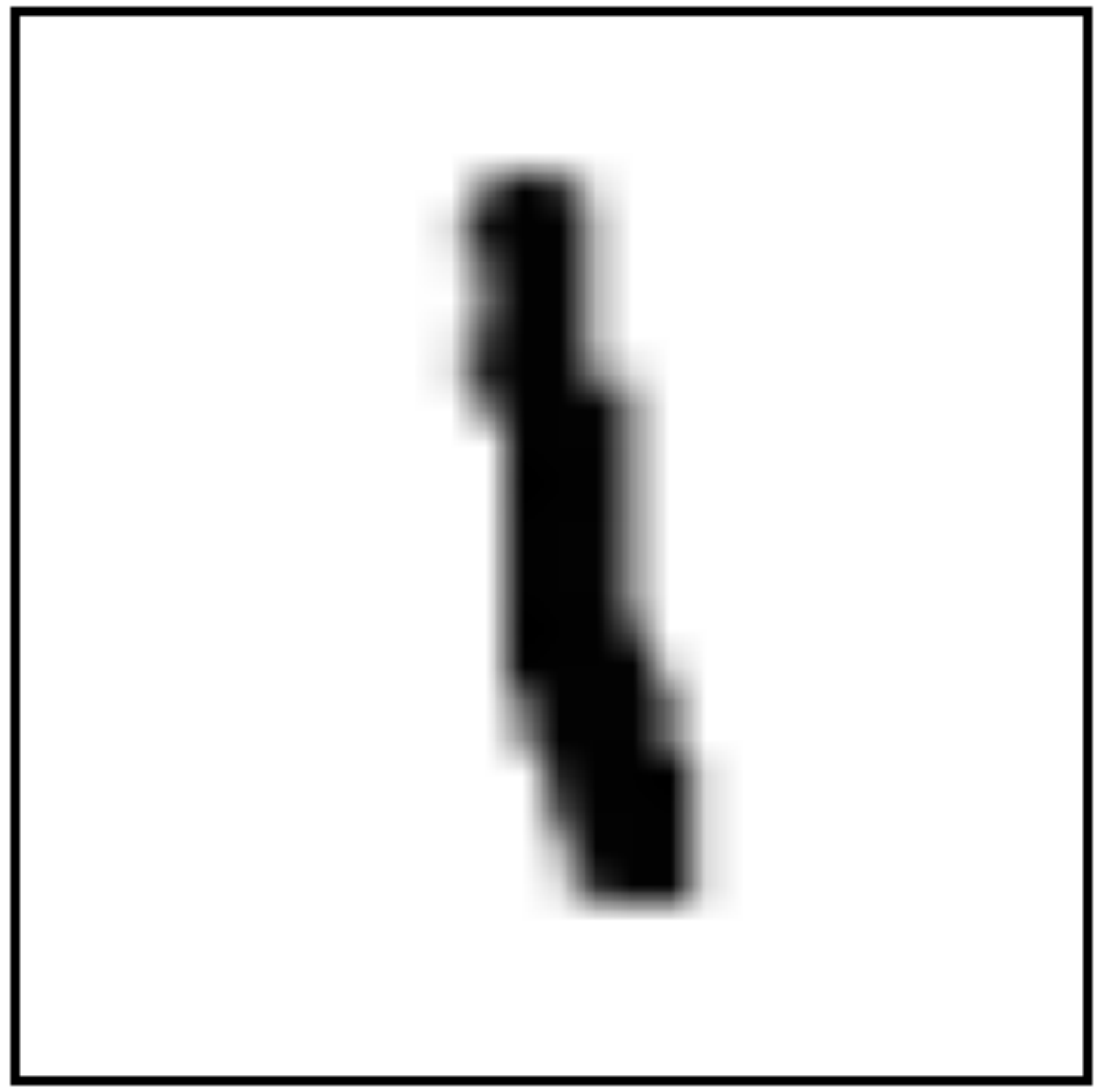
Supervised Learning - Classification



- MNIST Dataset
  - Handwriting Recognition
  - Each image is an array of 28 x 28 pixels
  - One of the first commercial and widely used ML systems for zip code detection and other checks

Let’s look at some concrete examples

Supervised Learning - Classification

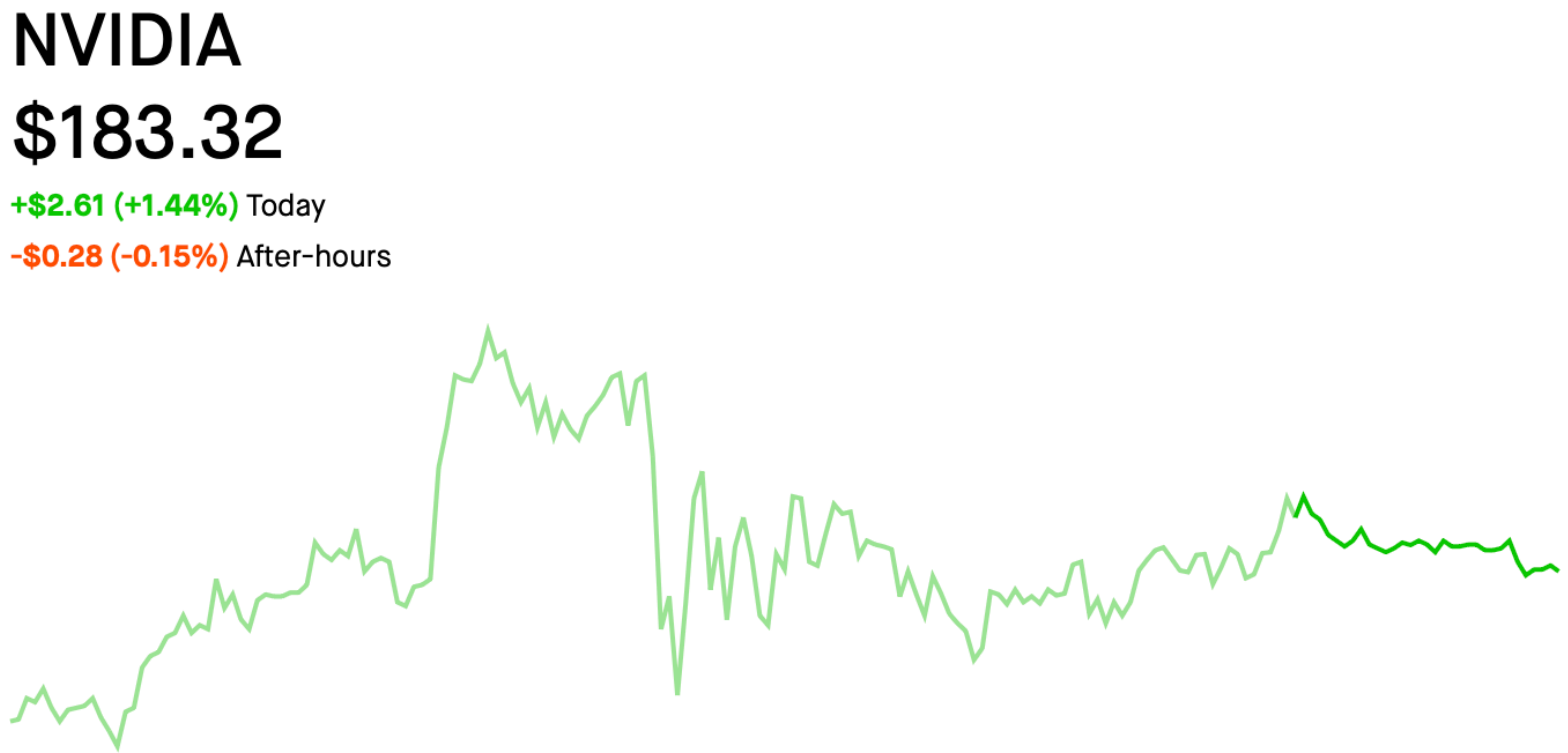


$$\approx \begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

- MNIST Dataset
- Handwriting Recognition
- **Each image is an array of 28 x 28 pixels**
- One of the first commercial and widely used ML systems for zip code detection and other checks

Let’s look at some concrete examples

Supervised Learning - Regression

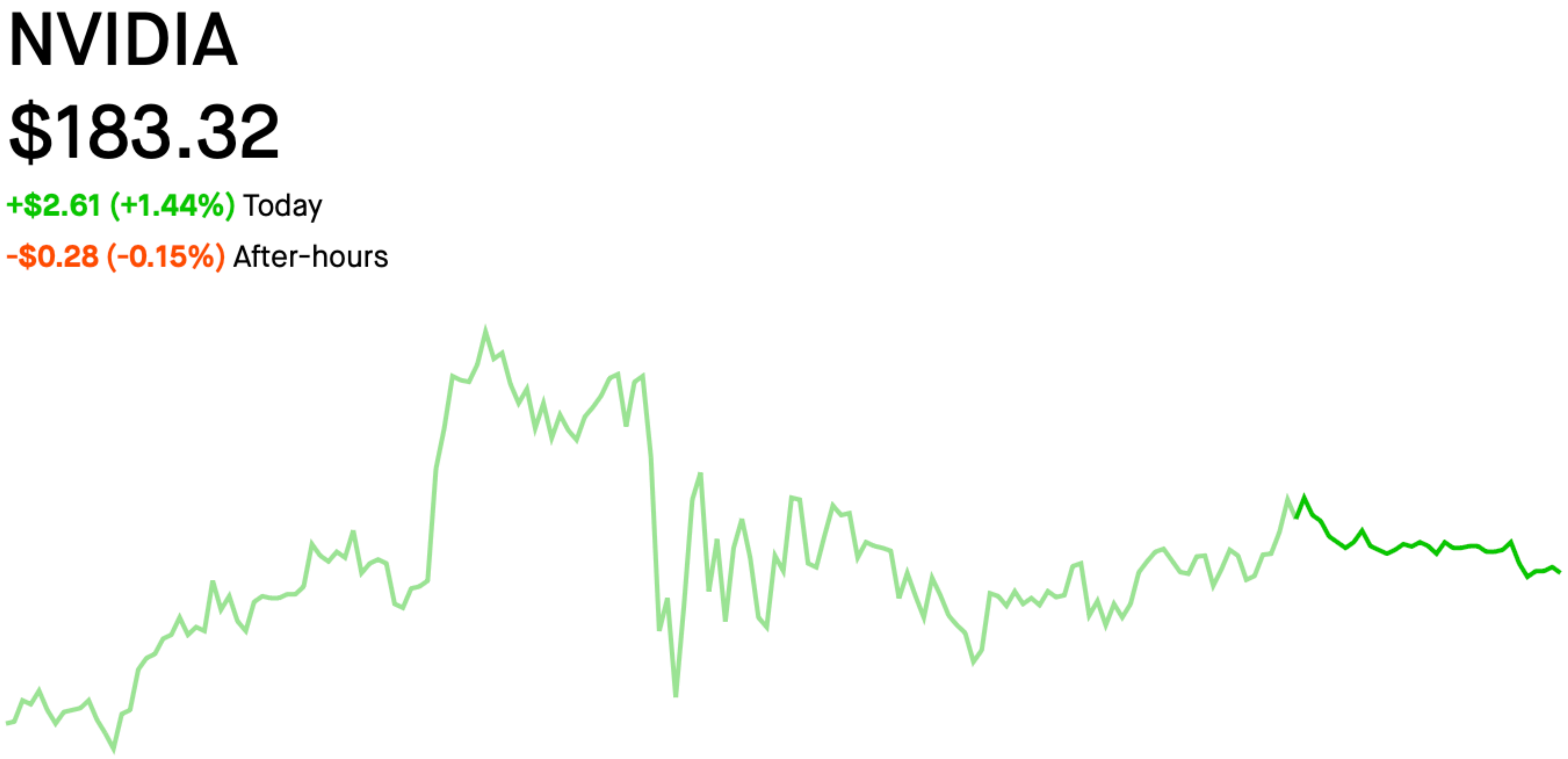


- Stock Price Prediction
  - Given some input features, predict the price of the stock at a future time
  - Given stock price of **other** companies, predict price of a given company of interest
  - Predict a **real-valued** number instead of a class

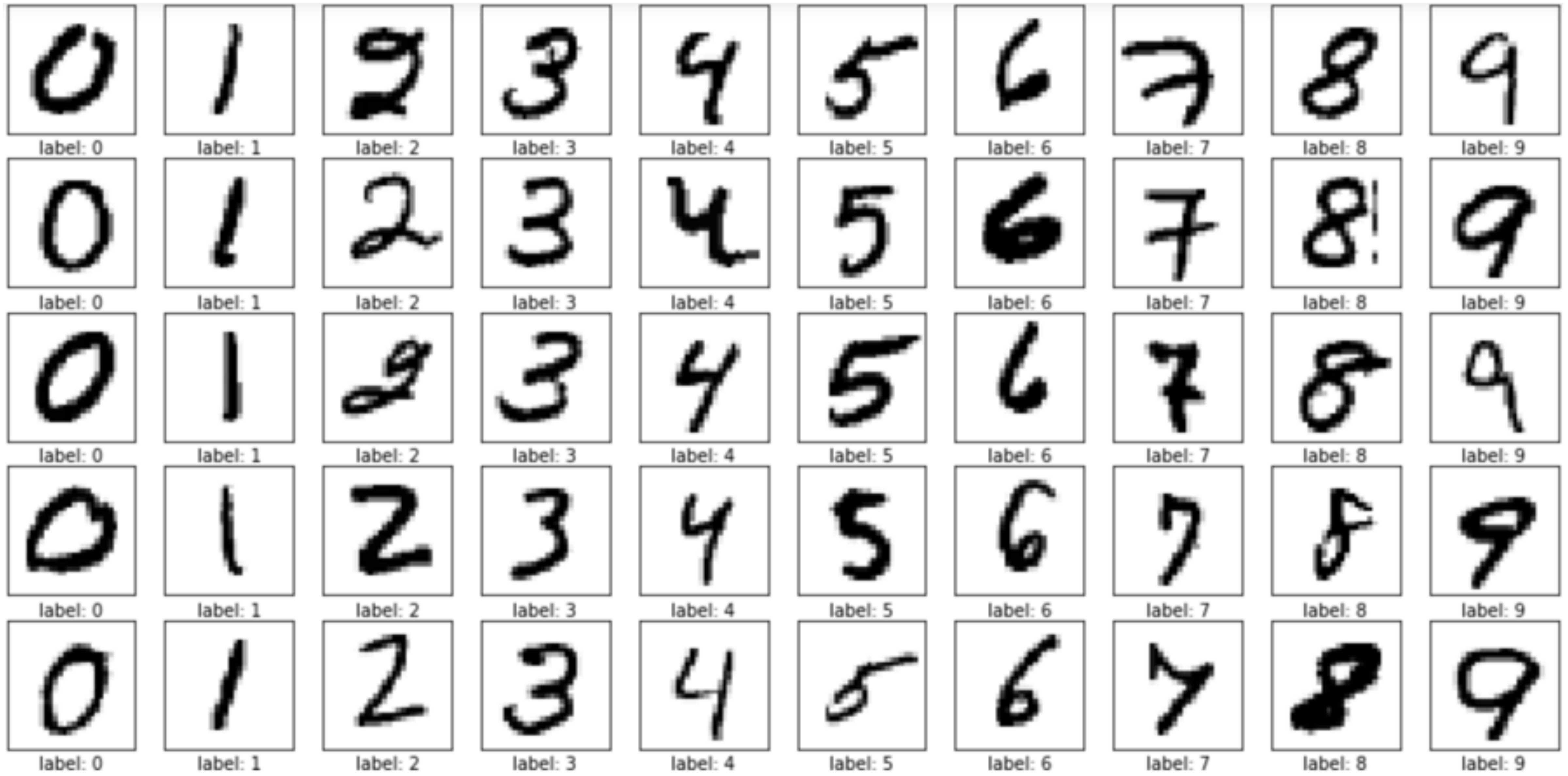
Let’s look at some concrete examples

Regression vs Classification

Supervised Learning - **Regression**



Supervised Learning - **Classification**



Let’s look at some concrete examples

Other Supervised Learning Examples

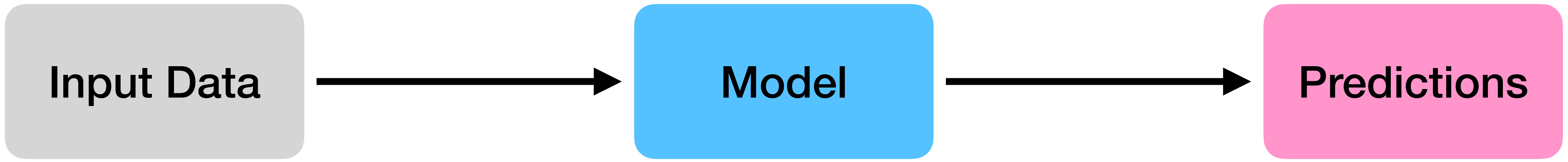
- Spam Classification
  - Is the email you received spam or not?
  - Is the attachment safe?
- Weather prediction
  - **Question**: Is this **classification** or **regression**?
- Image classification
  - What objects are in the image?
  - Where is each object in the image?

Let’s look at some concrete examples

Some Unsupervised Learning Examples

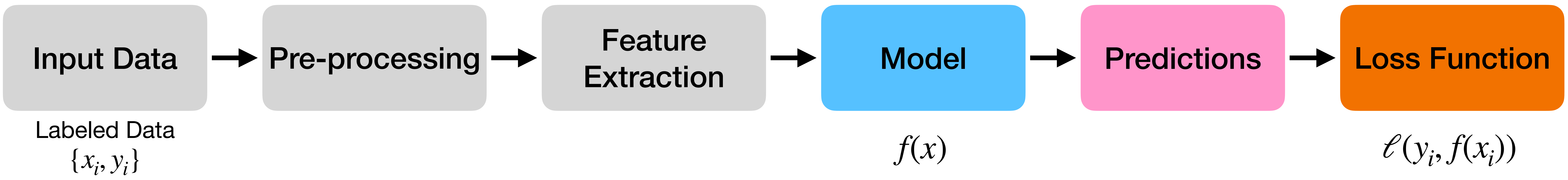
- **Clustering**
  - Group similar points into clusters
  - Example: k-means clustering, hierarchical clustering, density based clustering
- **Dimensionality Reduction**
  - Project input data into lower dimensional space
  - Example: Principle Component Analysis (PCA)
- **Feature Learning**
  - Find low dimensional feature representations
  - Think of this as a “learned” PCA
  - Example: Autoencoders

What does the overall pipeline look like?



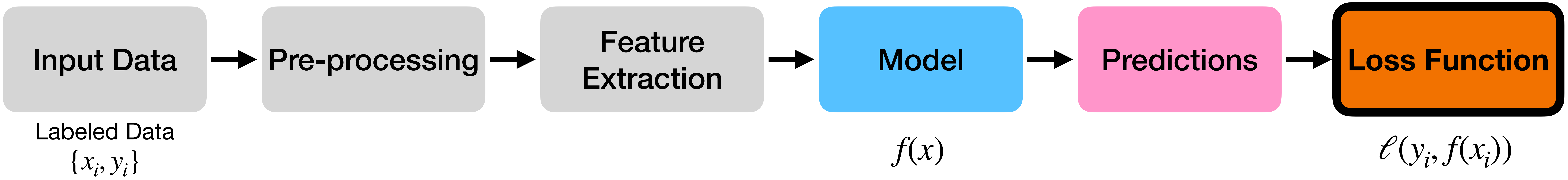
What does the overall pipeline look like?

Training Pipeline:

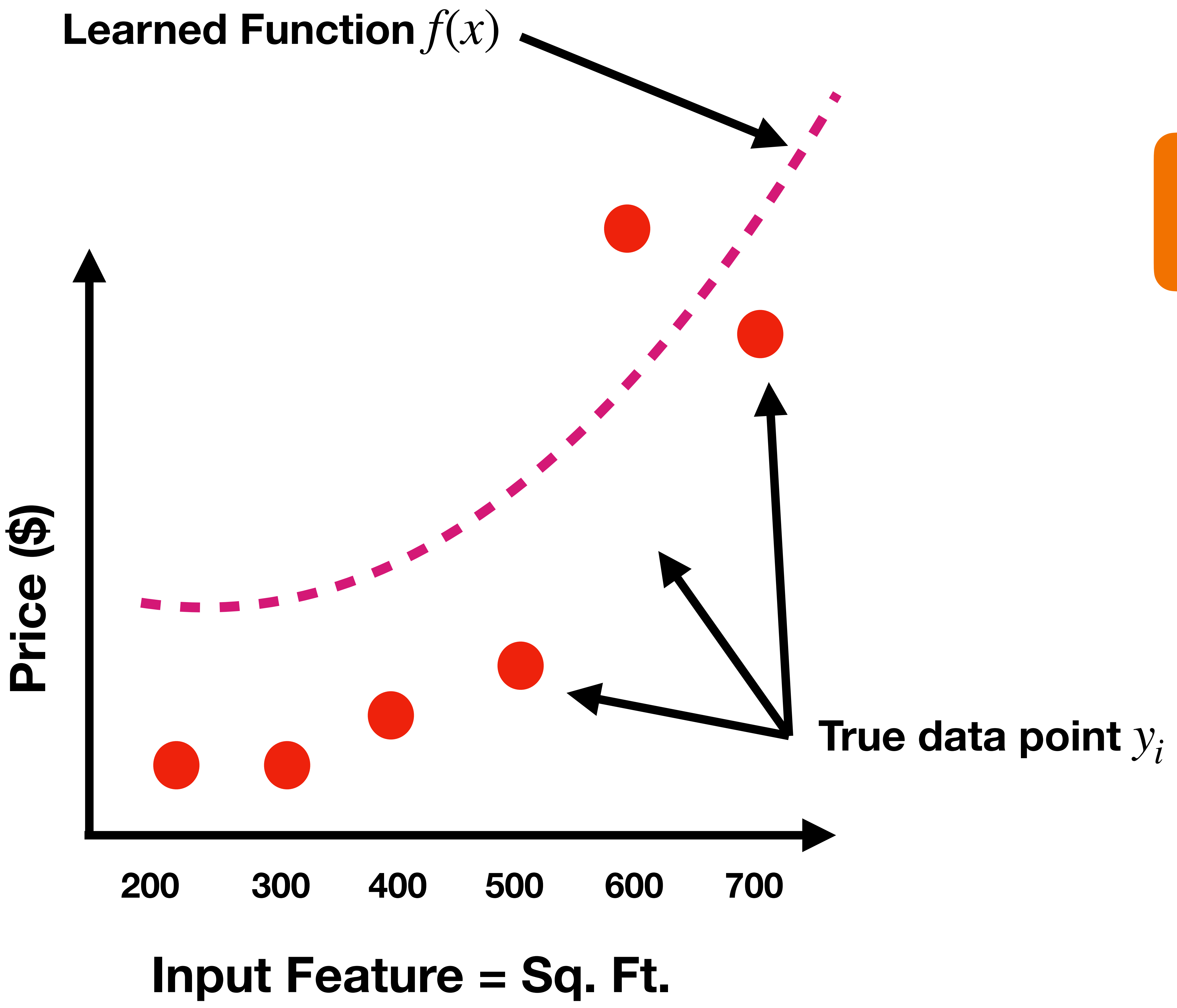


What does the overall pipeline look like?

Training Pipeline:

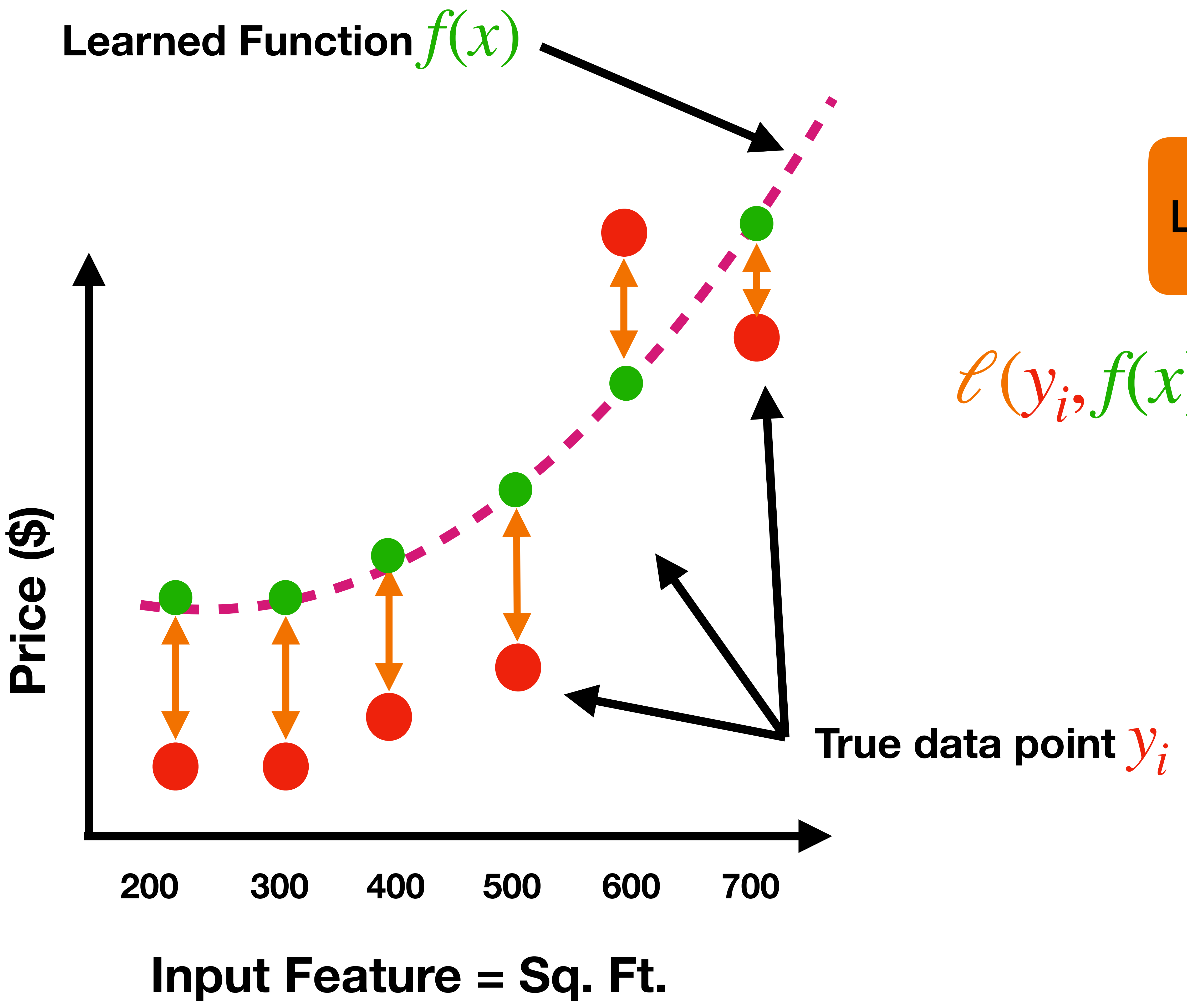


What is a loss function?



$\ell(y_i, f(x_i))$

What is a loss function?

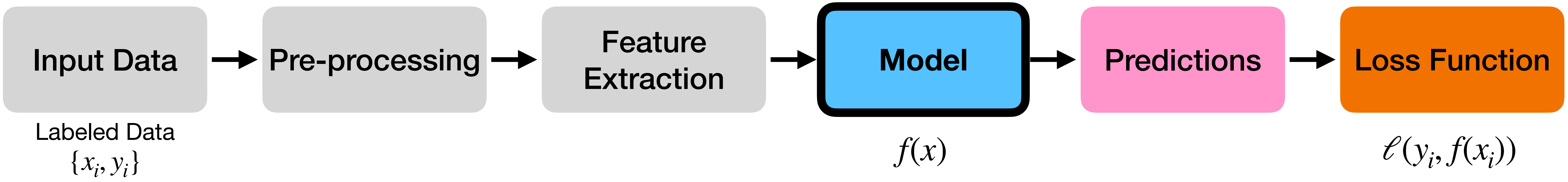


Loss Function

$$\ell(y_i, f(x)) = \sum y_i - f(x_i)$$

What does the overall pipeline look like?

Training Pipeline:



What are some common models and their loss functions?

- Linear Regression
  - **Goal**: Predict continuous output $\hat{y}$ from input features x
  - **Model**: $\hat{y} = w_0 + w_1x_1 + w_2x_2$
  - **Loss Function**: $\frac{1}{m}\sum_i (y_i - \hat{y}_i)^2$

What are some common models and their loss functions?

- Linear Regression
  - **Goal**: Predict continuous output $\hat{y}$ from input features x
  - **Model:** $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$
  - **Loss Function:** $\frac{1}{m} \sum_i (y_i - \hat{y}_i)^2$

These are the predicted values

What are some common models and their loss functions?

- Linear Regression
  - **Goal**: Predict continuous output $\hat{y}$ from input features x
  - **Model**: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$
  - **Loss Function**: $\frac{1}{m}\sum_i (y_i - \hat{y}_i)^2$

These are the learnable weights/parameters

What are some common models and their loss functions?

- Linear Regression
  - **Goal**: Predict continuous output $\hat{y}$ from input features x
  - **Model**: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$
  - **Loss Function**: $\frac{1}{m} \sum_i (y_i - \hat{y}_i)^2$

These are the input features

What are some common models and their loss functions?

- Linear Regression
  - **Goal**: Predict continuous output $\hat{y}$ from input features x
  - **Model**: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$
  - **Loss Function**: $\frac{1}{m} \sum_i (y_i - \hat{y}_i)^2$

This is the true label

What are some common models and their loss functions?

- Logistic Regression
  - **Goal**: Predict probability of binary class membership (classification)
  - **Model**: $\mathbb{P}(y = 1 \mid x) = \sigma(w_0 + w_1x_1 + w_2x_2)$
  - **Loss Function**: $-\frac{1}{m}\sum_i y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$

This is the sigmoid operator - it caps outputs within a range of 0-1

What are some common models and their loss functions?

- Logistic Regression
  - **Goal**: Predict probability of binary class membership (classification)
  - **Model**: $\mathbb{P}(y = 1 \mid x) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$
  - **Loss Function**: $-\frac{1}{m} \sum_i y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$

Notice how this looks similar to the linear regression model

What are some common models and their loss functions?

- Logistic Regression
  - **Goal**: Predict probability of binary class membership (classification)
  - **Model**: $\mathbb{P}(y = 1 \mid x) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$
  - **Loss Function**: $-\frac{1}{m}\sum_i y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$

But the loss function is now different, this is the binary cross entropy loss

Review Outline

1. Probability
2. Linear Algebra

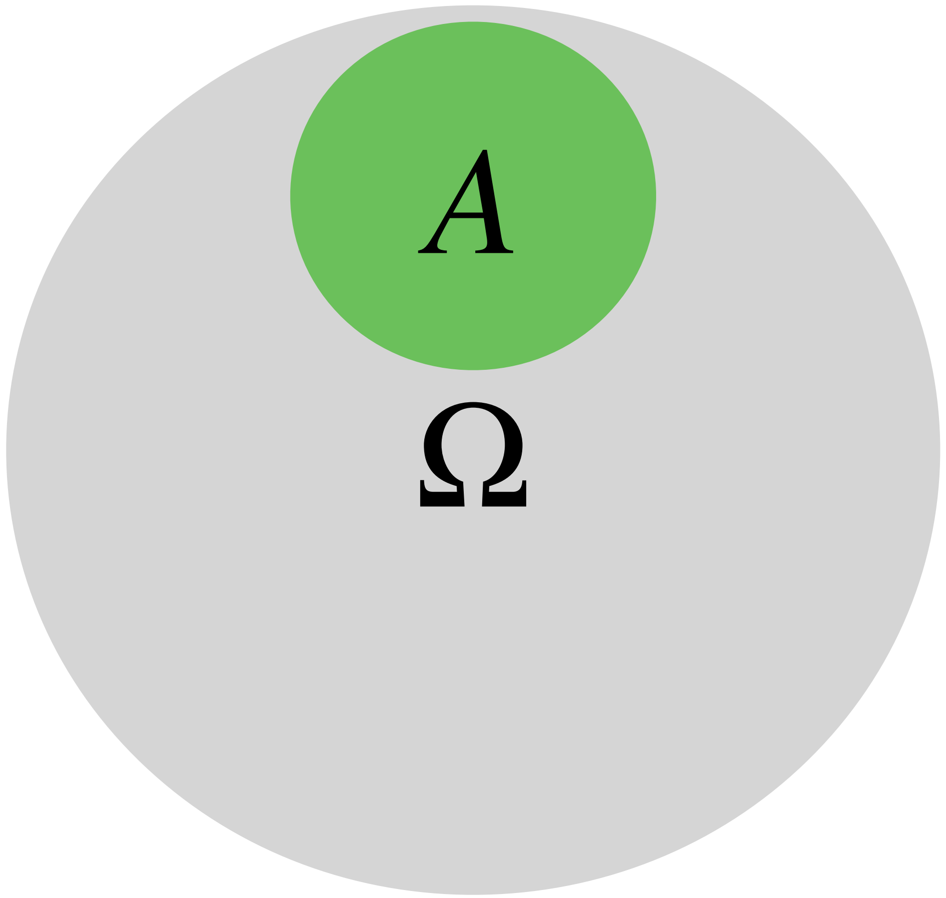
Review Outline

1. **Probability**
2. Linear Algebra

Probability

Basic Concepts

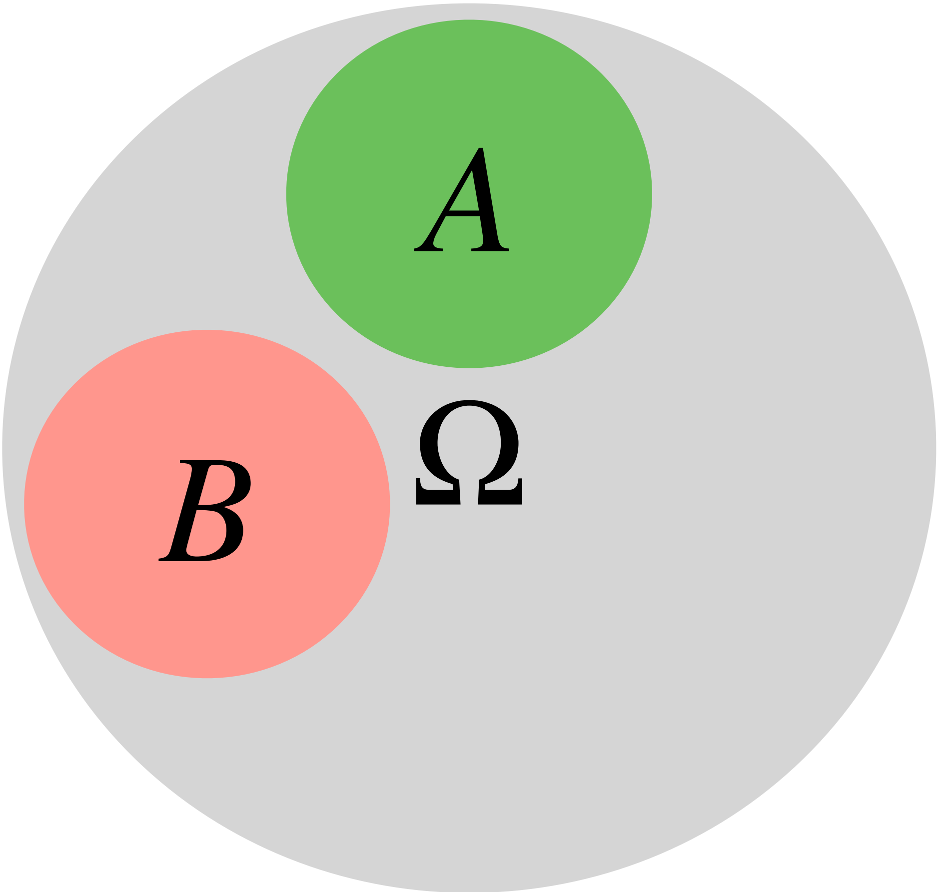
- Sample Space and Events
  - The sample space Ω is the set of all possible outcomes of an experiment.
  - An event A is a subset of the sample space Ω.
  - The probability $P(A)$ is a number between 0 and 1 representing how likely event A is to occur.



Probability

Basic Concepts

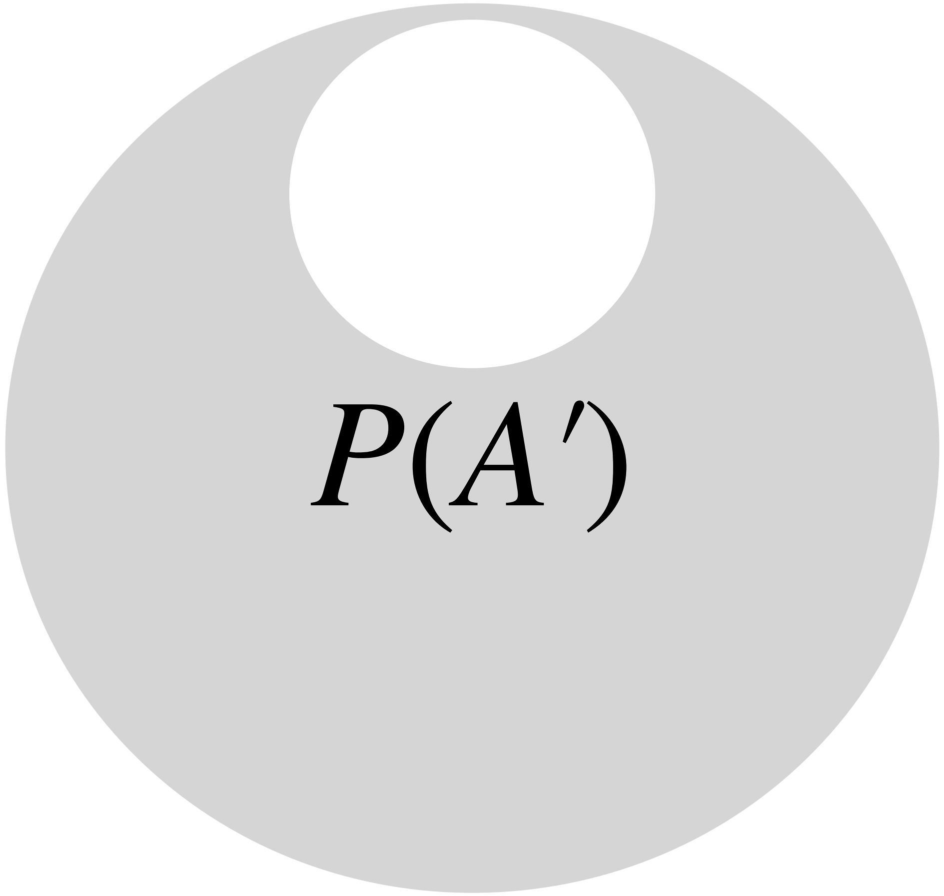
- Sample Space and Events
  - $P(A) \geq 0$
    - Why?
      - Because A is a subset of Ω
  - $P(\Omega) = 1$
  - For mutually exclusive events A and B:
    - $P(A \cup B) = P(A) + P(B)$

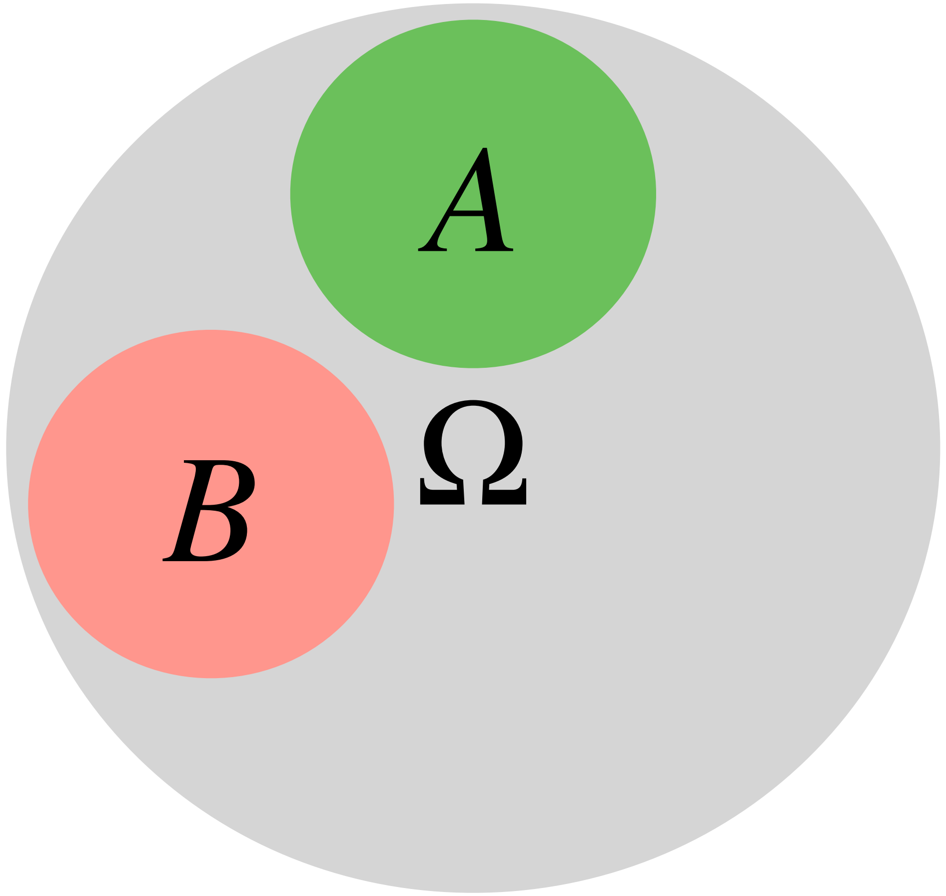


Probability

Basic Concepts

- Sample Space and Events
  - For mutually exclusive events A and B:
    - $P(A \cup B) = P(A) + P(B)$
  - Complement Rule:
    - $P(A') = 1 - P(A)$





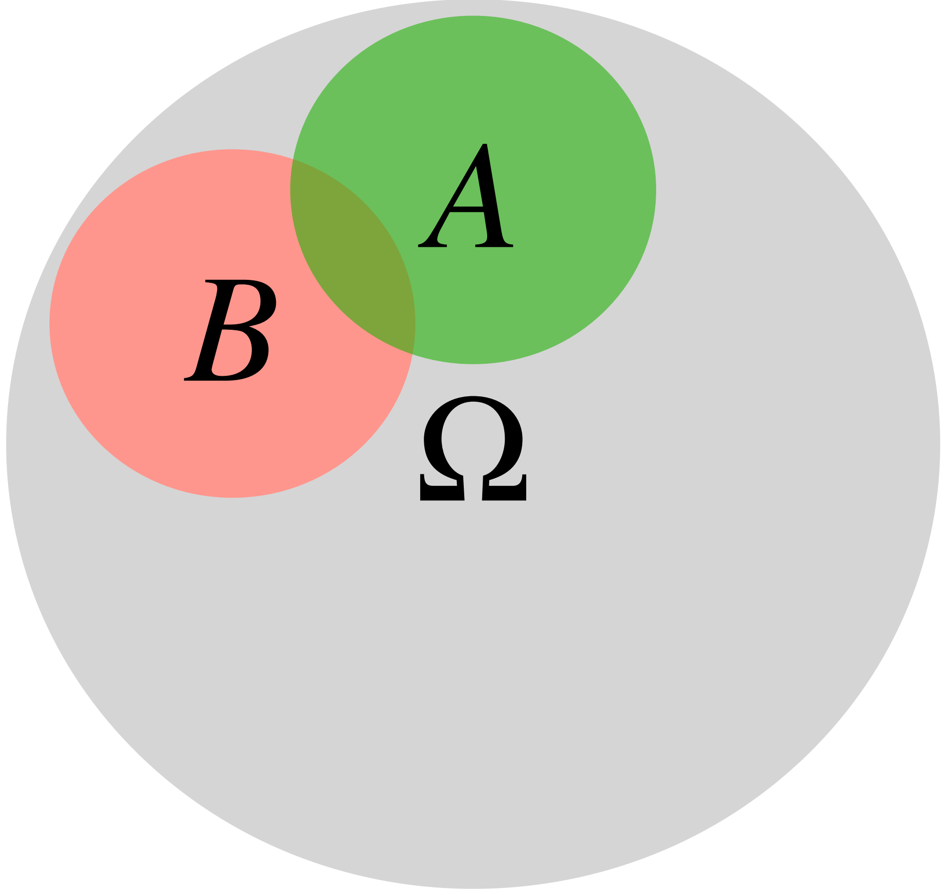
Probability

Conditional Probability

- Probability of A, **given** that B has **already** occurred

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

This is fundamental in machine learning. Why?



Probability

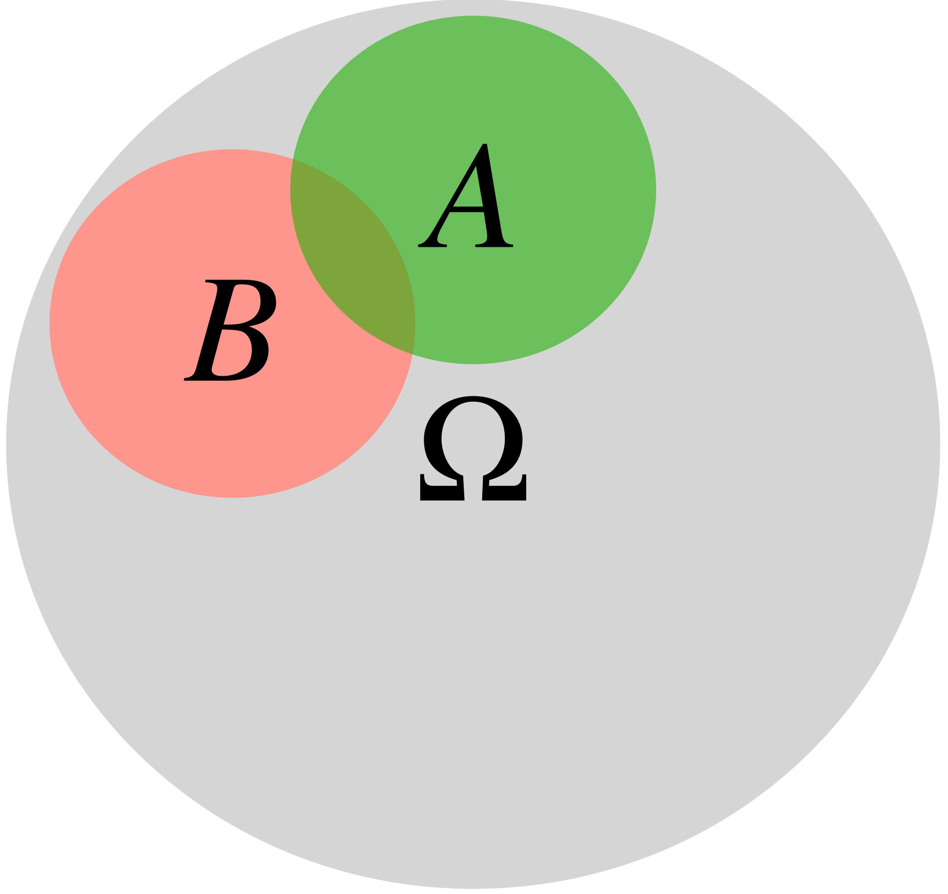
Conditional Probability

- Probability of A, **given** that B has **already** occurred

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

spam email

This is fundamental in machine learning. Why?



Probability

Conditional Probability

- Probability of A, **<u>given</u>** that B has **<u>already</u>** occurred

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

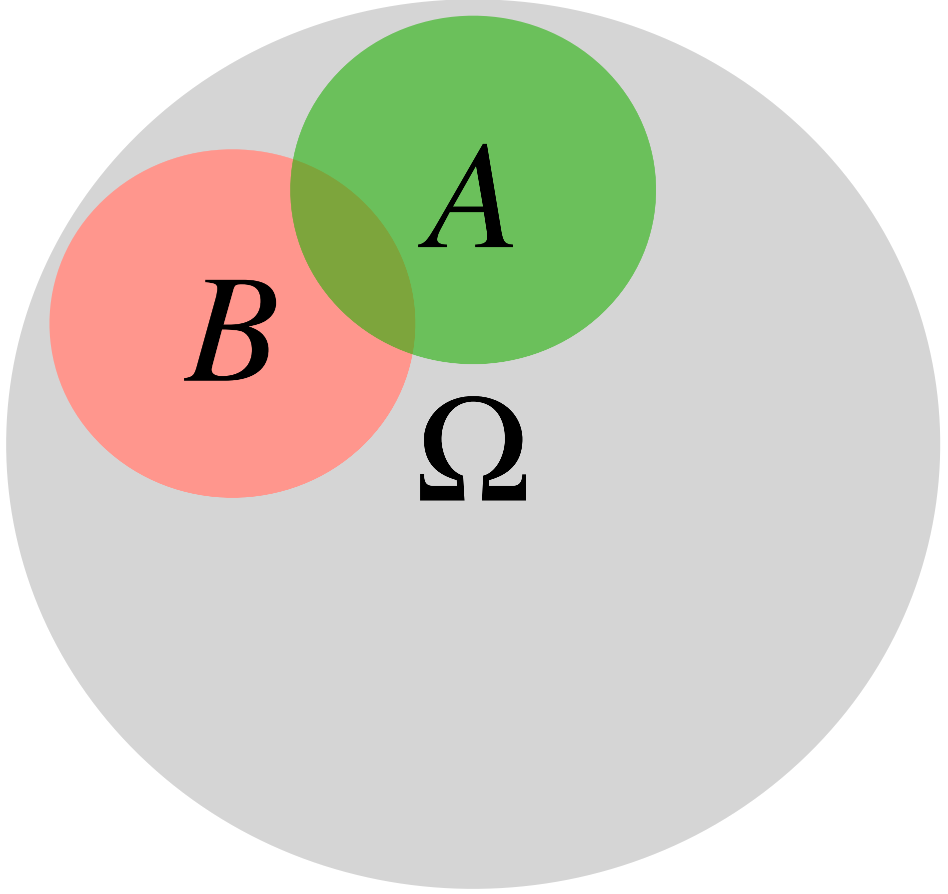
spam email

Two events A and B are **independent** if

$$P(A \cap B) = P(A) \cdot P(B)$$

or

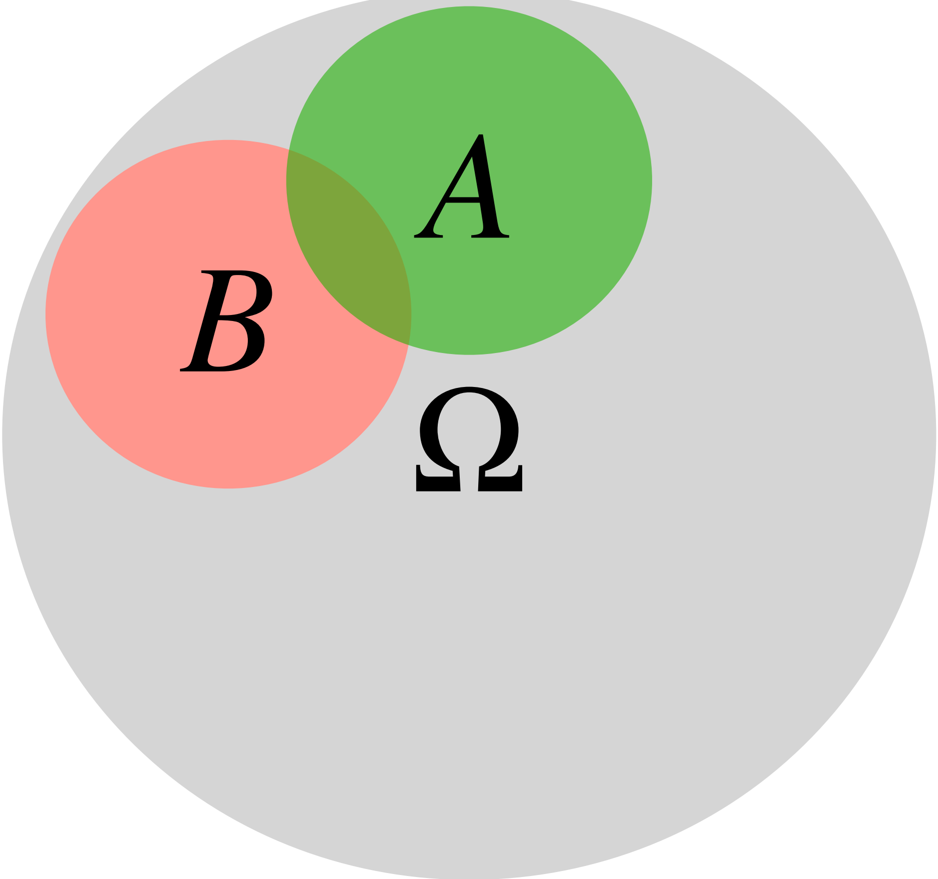
$$P(A \mid B) = \frac{P(A) \cap P(B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$



Probability

Conditional Probability - Bayes' Theorem

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



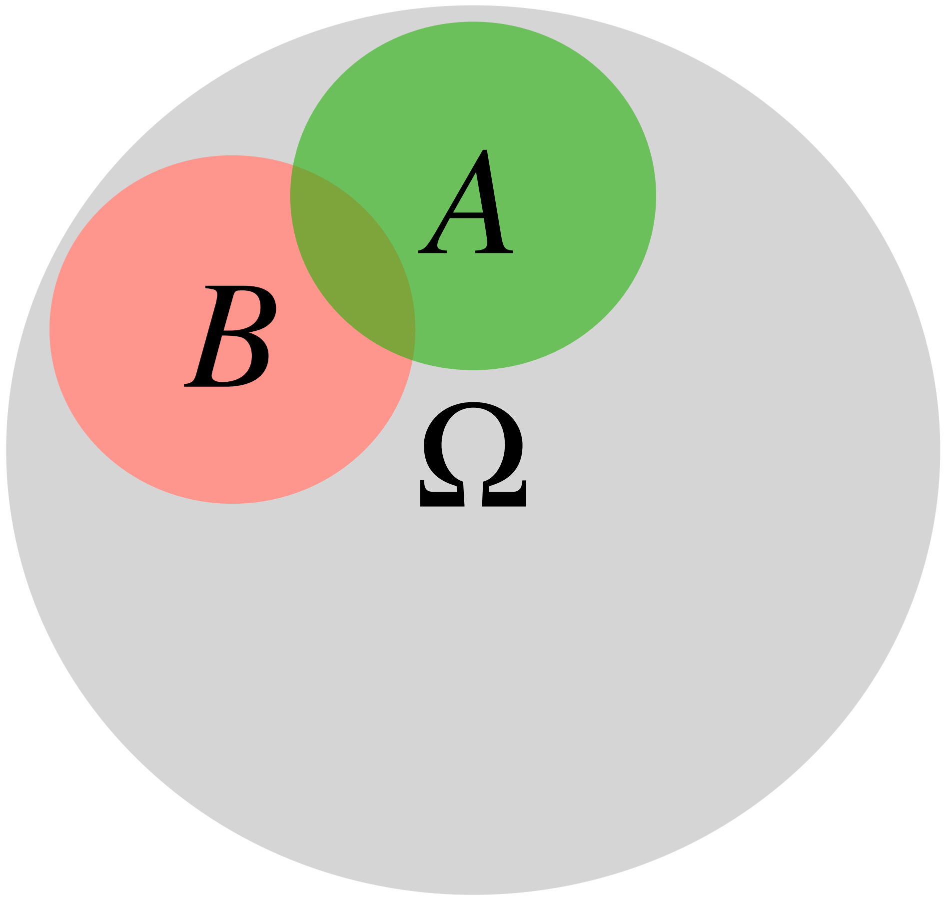
Probability

Conditional Probability - Bayes' Theorem

Prior - what we believe **before** seeing the data B

Or probability of event A occurring before having made any observation about event B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



Probability

Conditional Probability - Bayes' Theorem

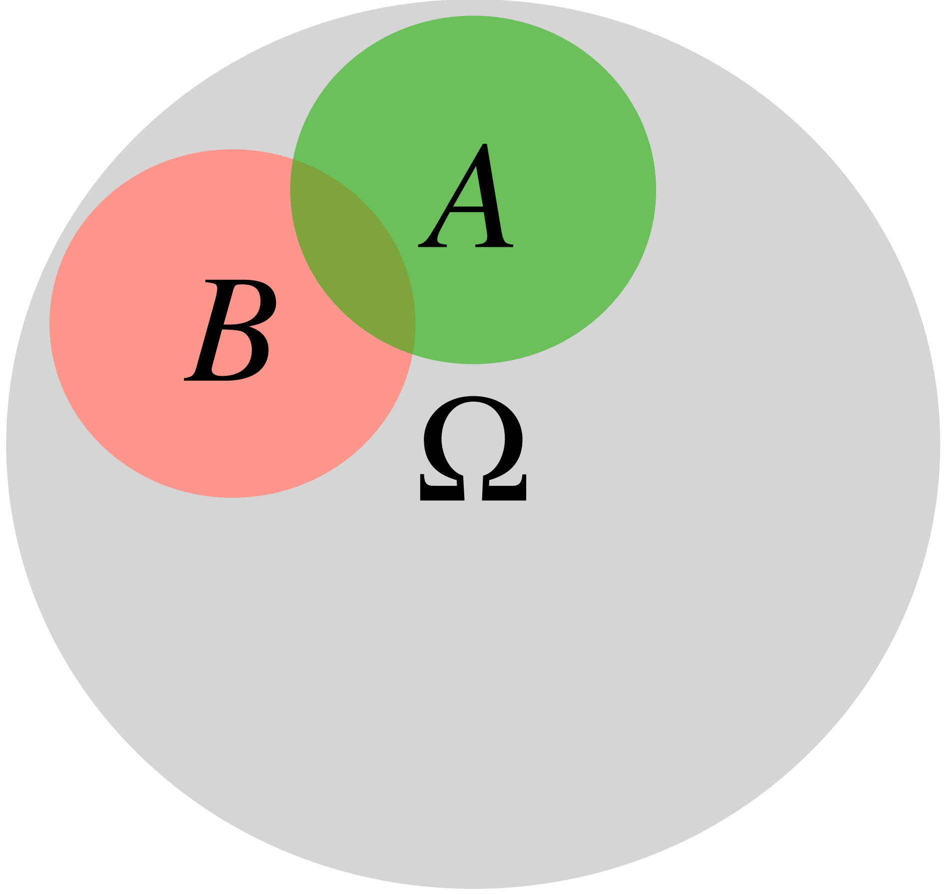
Likelihood - probability of data B given A

Or probability of event B occurring, given event A has already occurred

Prior - what we believe **before** seeing the data B

Or probability of event A occurring before having made any observation about event B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



Probability

Conditional Probability - Bayes' Theorem

Likelihood - probability of data B given A

Or probability of event B occurring, given event A has already occurred

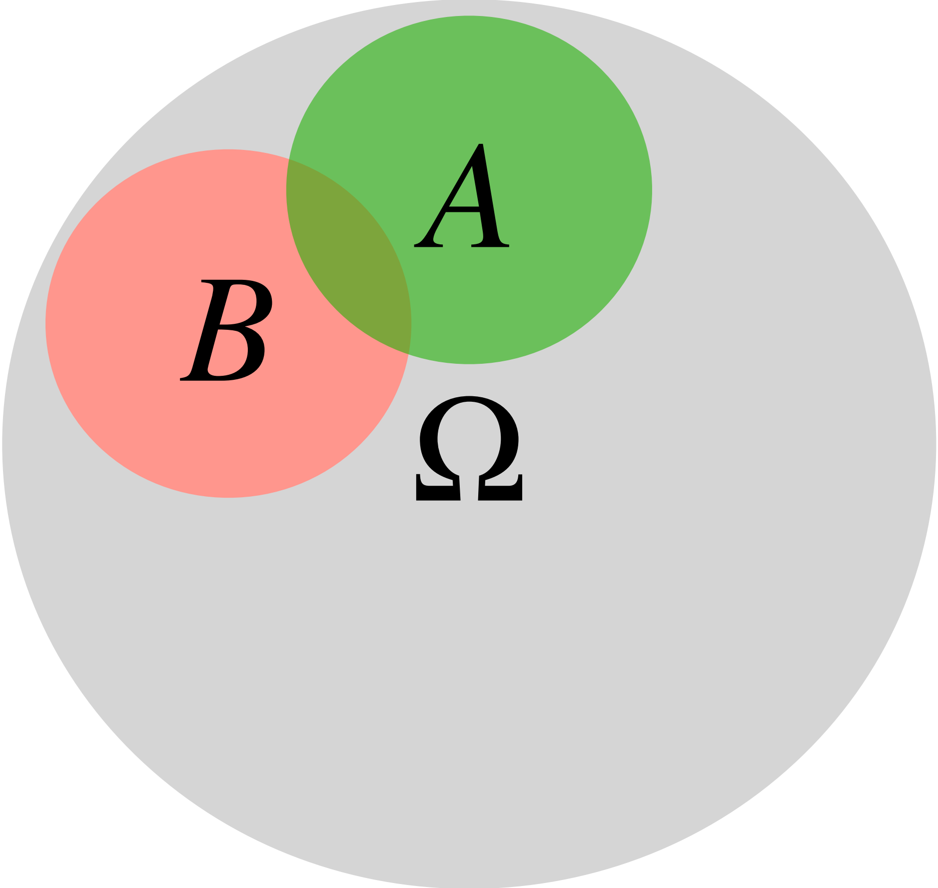
Prior - what we believe **before** seeing the data B

Or probability of event A occurring before having made any observation about event B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Evidence - marginal likelihood or

Or probability of event B occurring before having made any observation about event A



Probability

Conditional Probability - Bayes' Theorem

Likelihood - probability of data B given A

Or probability of event B occurring, given event A has already occurred

Prior - what we believe **before** seeing the data B

Or probability of event A occurring before having made any observation about event B

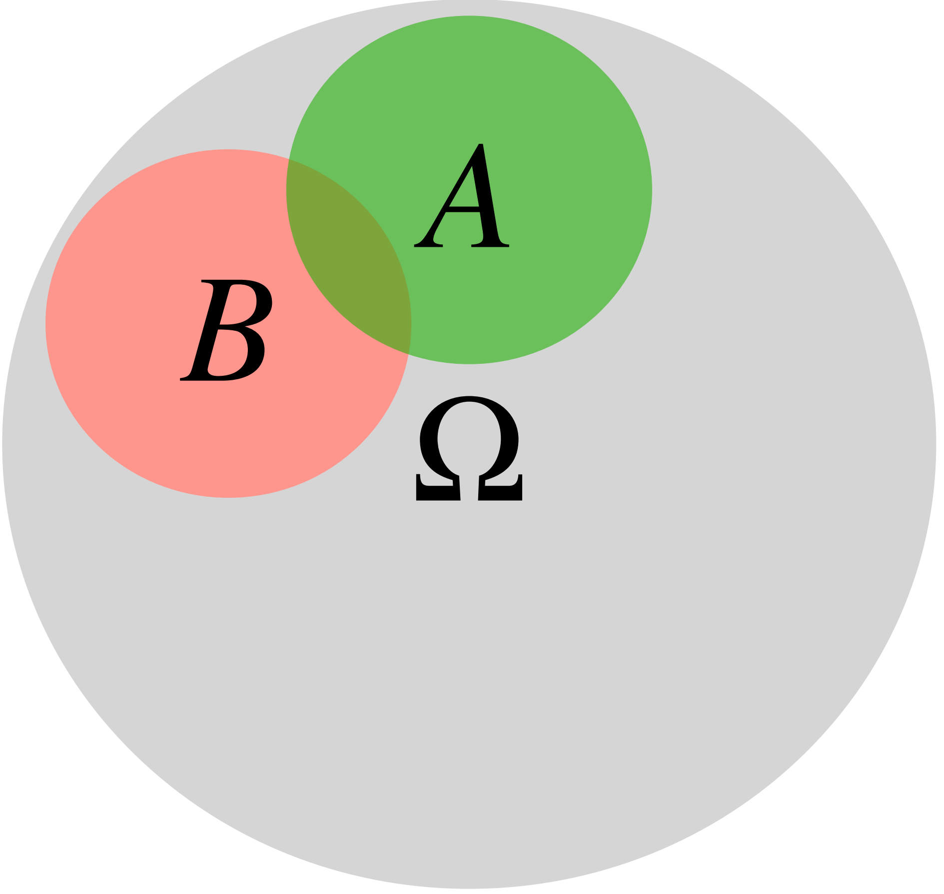
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Posterior - updated belief **after** seeing the data

Or probability of event A occurring **after** having made an observation about event B

Evidence - marginal likelihood

Or probability of event B occurring before having made any observation about event A



Probability

Random Variables and Distributions

- A random variable X is a function that maps **outcomes** in the sample space (Ω) to real numbers.
- Random variables can be **discrete** (taking countable values) or **continuous** (taking any value in an interval).

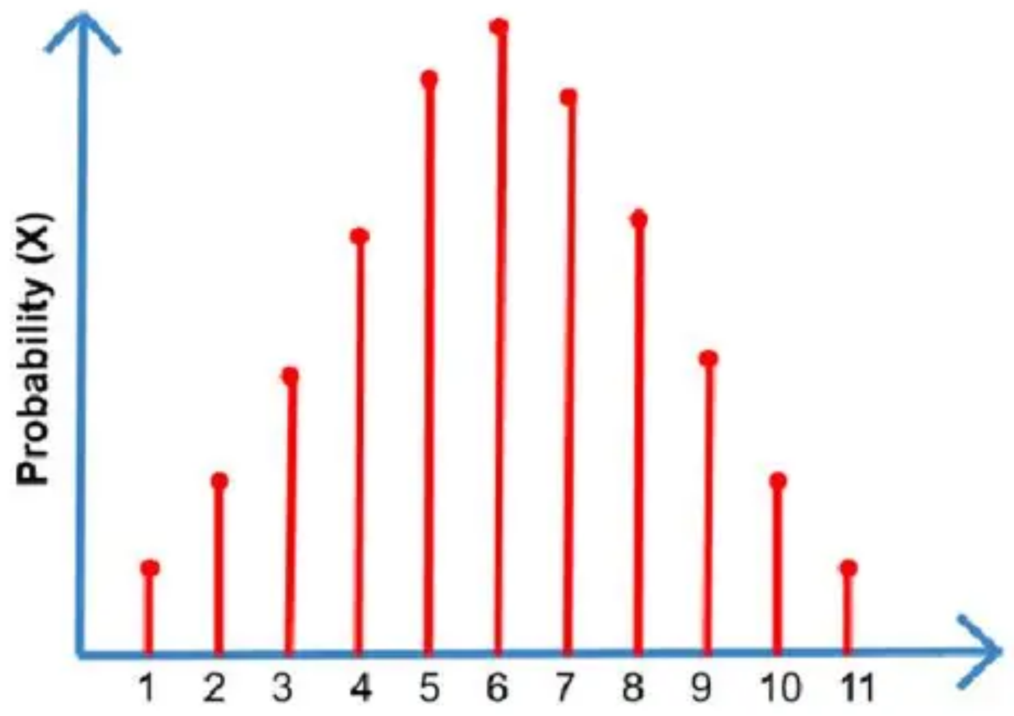
Probability

Random Variables and Distributions

- Probability Mass Function (PMF) - for **discrete** random variables
  - $P(X = x)$ gives the probability that X takes value x.
  - The sum over all possible values equals 1.
- Probability Density Function (PDF) — for **continuous** random variables
  - $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - Note that $f(x)$ itself is not a probability; it can exceed 1.
- Cumulative Distribution Function (CDF)
  - $F(x) = P(X \leq x)$
  - Works for both discrete and continuous variables.

Probability

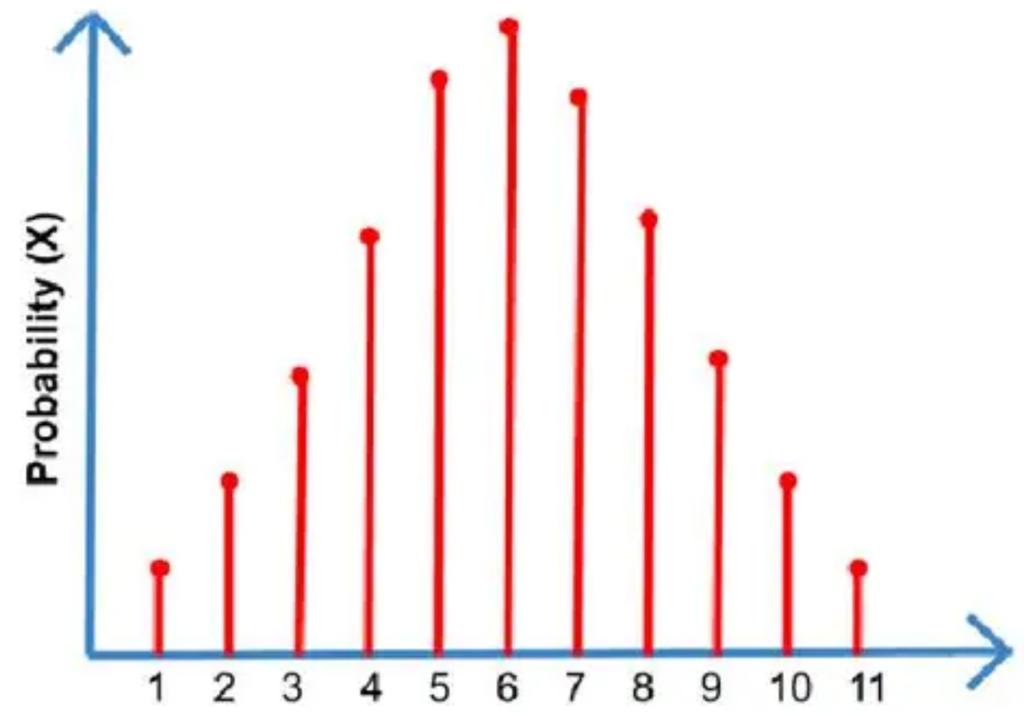
Random Variables and Distributions

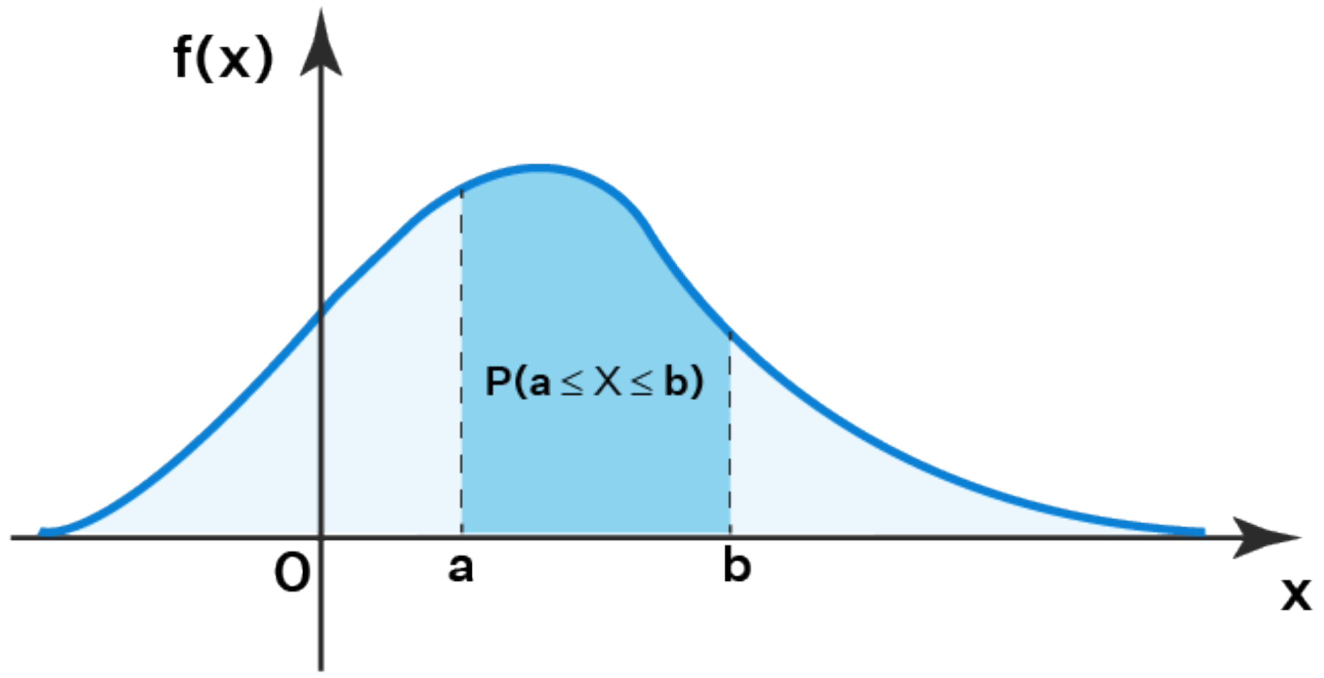


- Probability Mass Function (PMF) - for **discrete** random variables
  - $P(X = x)$ gives the probability that X takes value x.
  - The sum over all possible values equals 1.
- Probability Density Function (PDF) — for **continuous** random variables
  - $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - Note that $f(x)$ itself is not a probability; it can exceed 1.
- Cumulative Distribution Function (CDF)
  - $F(x) = P(X \leq x)$
  - Works for both discrete and continuous variables.

Probability

Random Variables and Distributions



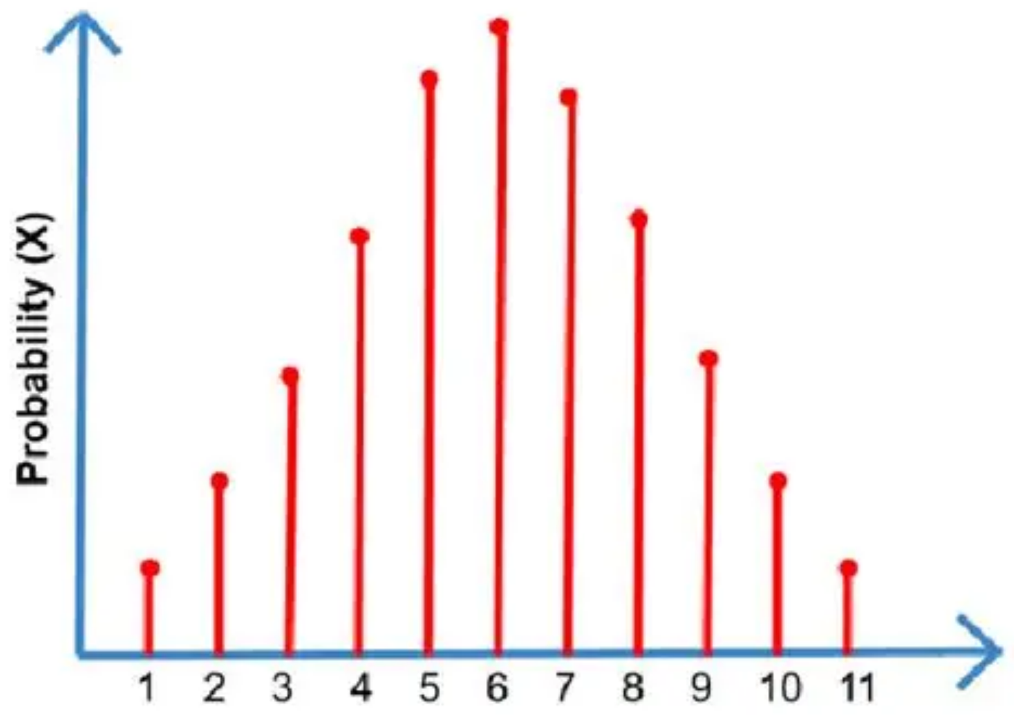


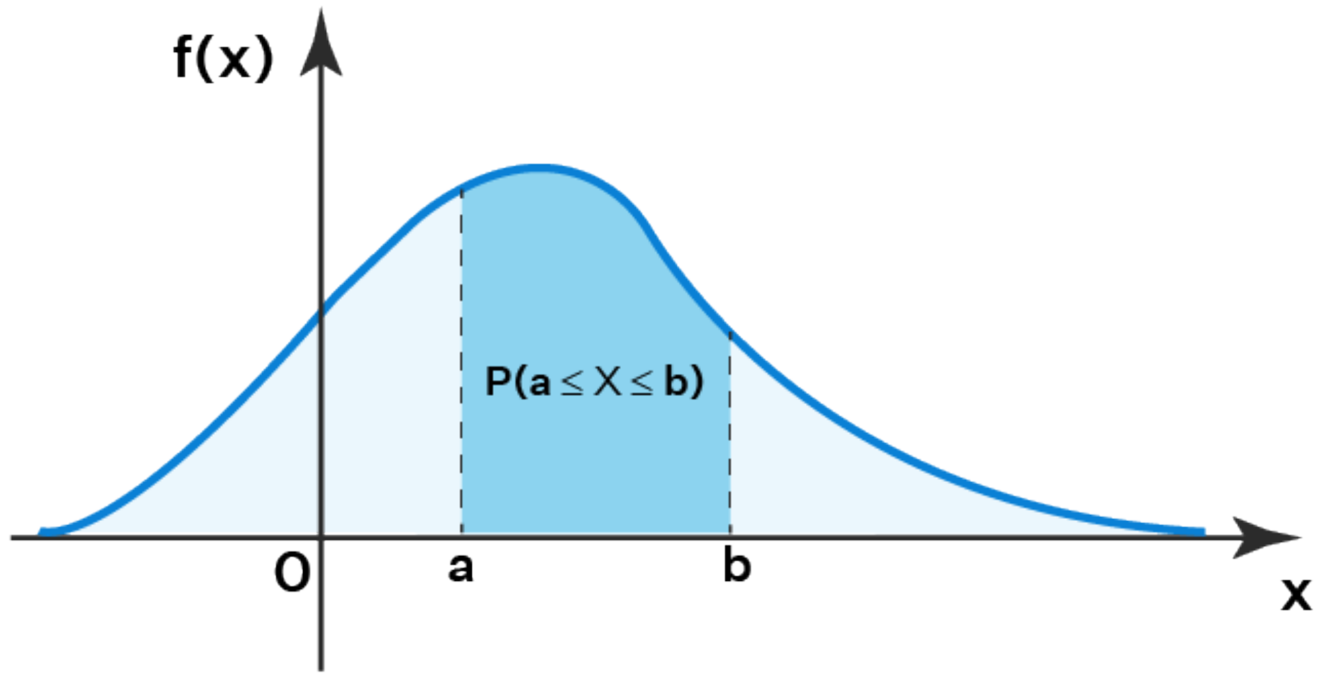
- Probability Mass Function (PMF) - for **discrete** random variables
  - $P(X = x)$ gives the probability that X takes value x.
  - The sum over all possible values equals 1.
- Probability Density Function (PDF) — for **continuous** random variables
  - $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - Note that $f(x)$ itself is not a probability; it can exceed 1.
- Cumulative Distribution Function (CDF)
  - $F(x) = P(X \leq x)$
  - Works for both discrete and continuous variables.

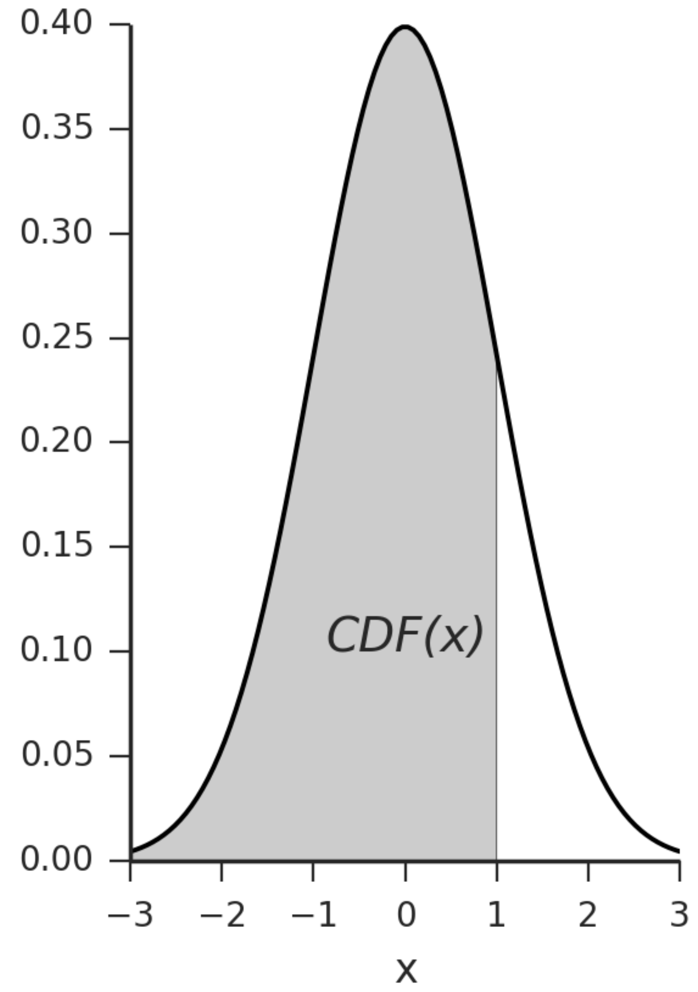
Probability

Random Variables and Distributions

- Probability Mass Function (PMF) - for **discrete** random variables
  - $P(X = x)$ gives the probability that X takes value x.
  - The sum over all possible values equals 1.
- Probability Density Function (PDF) — for **continuous** random variables
  - $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - Note that $f(x)$ itself is not a probability; it can exceed 1.
- Cumulative Distribution Function (CDF)
  - $F(x) = P(X \leq x)$
  - Works for both discrete and continuous variables.







Probability

Key Distributions in ML

- Bernoulli Distribution
- Binomial Distribution
- Gaussian (Normal) Distribution

Probability

Key Distributions in ML

- **Bernoulli Distribution**
  - Models **binary** outcomes (success/failure)
  - $P(X = 1) = p$
  - $p(X = 0) = q = 1 - p$

Probability

Key Distributions in ML

- **Binomial Distribution**
  - Models number of successes in n independent trials
  - $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$

Probability

Key Distributions in ML

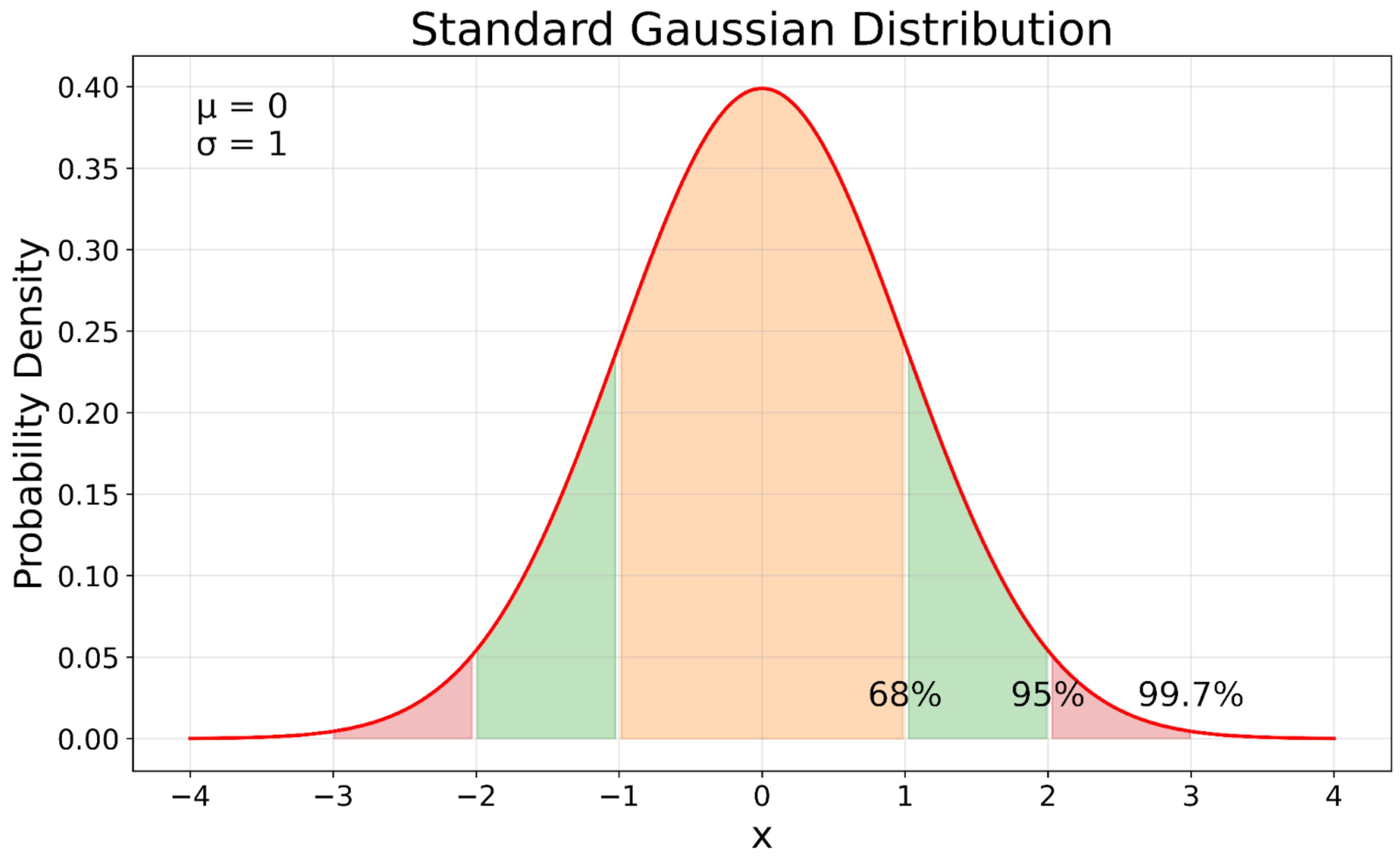
- **Binomial Distribution**
  - Models number of successes in n independent trials
  - $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$

Number of combinations, or ways, of **choosing** k **items** from a total of n items

Probability

Key Distributions in ML

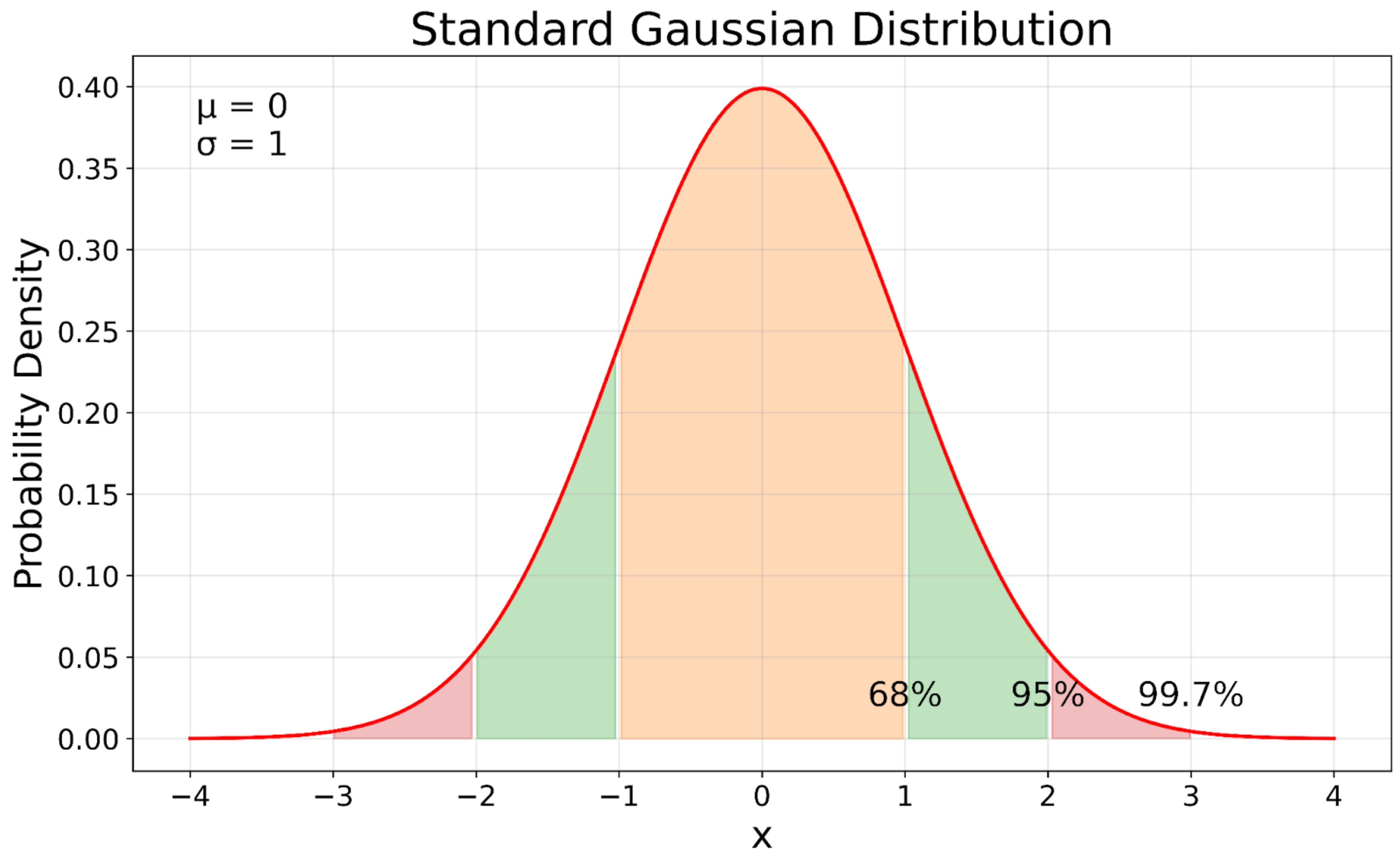
- **Gaussian (Normal) Distribution**
  - One of the most important distributions in ML



$$PDF: f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability

Key Distributions in ML



- **Gaussian (Normal) Distribution**
  - One of the most important distributions in ML

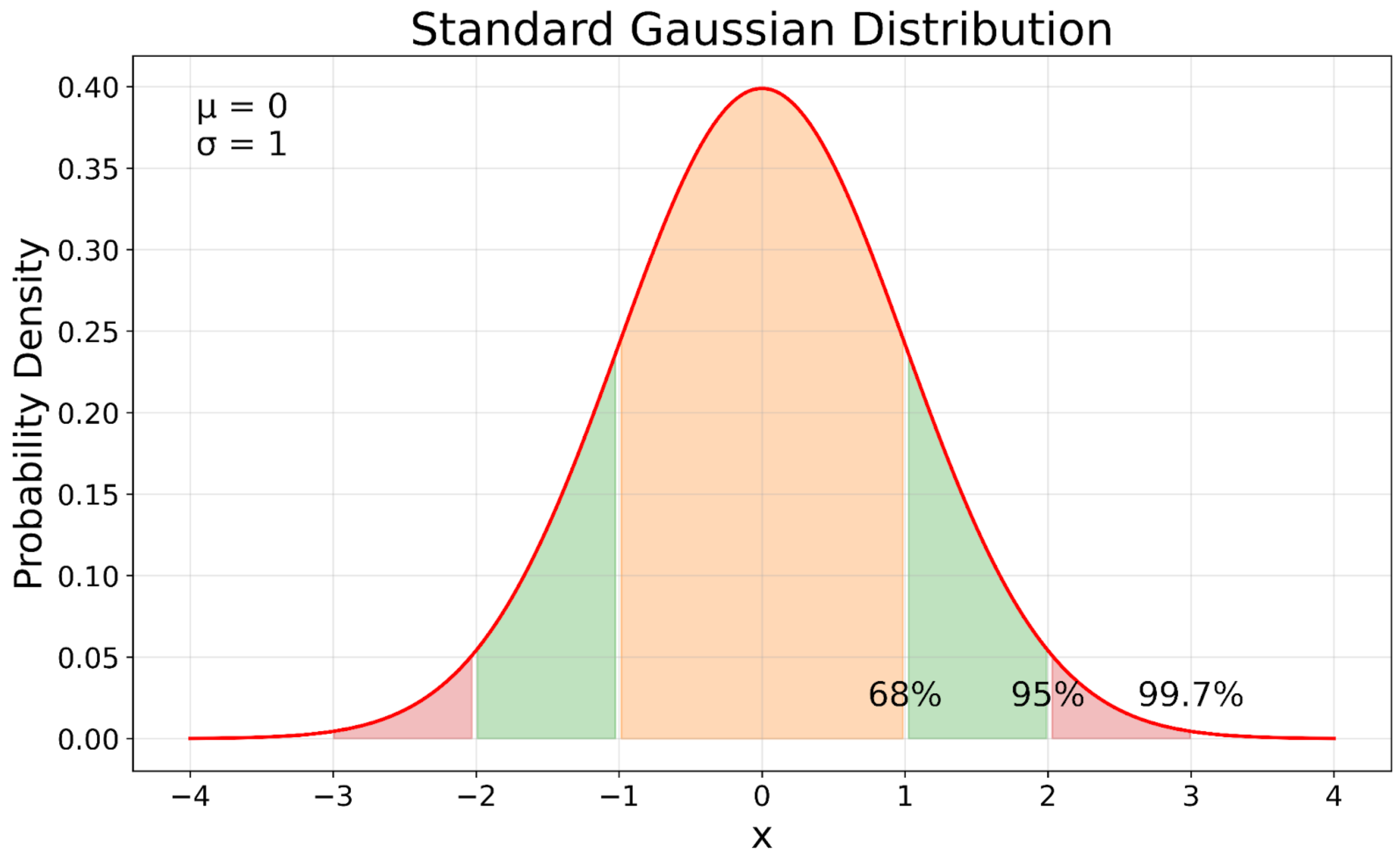
Mean of the distribution

$$PDF : f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability

Key Distributions in ML

- **Gaussian (Normal) Distribution**
  - One of the most important distributions in ML



Mean of the distribution

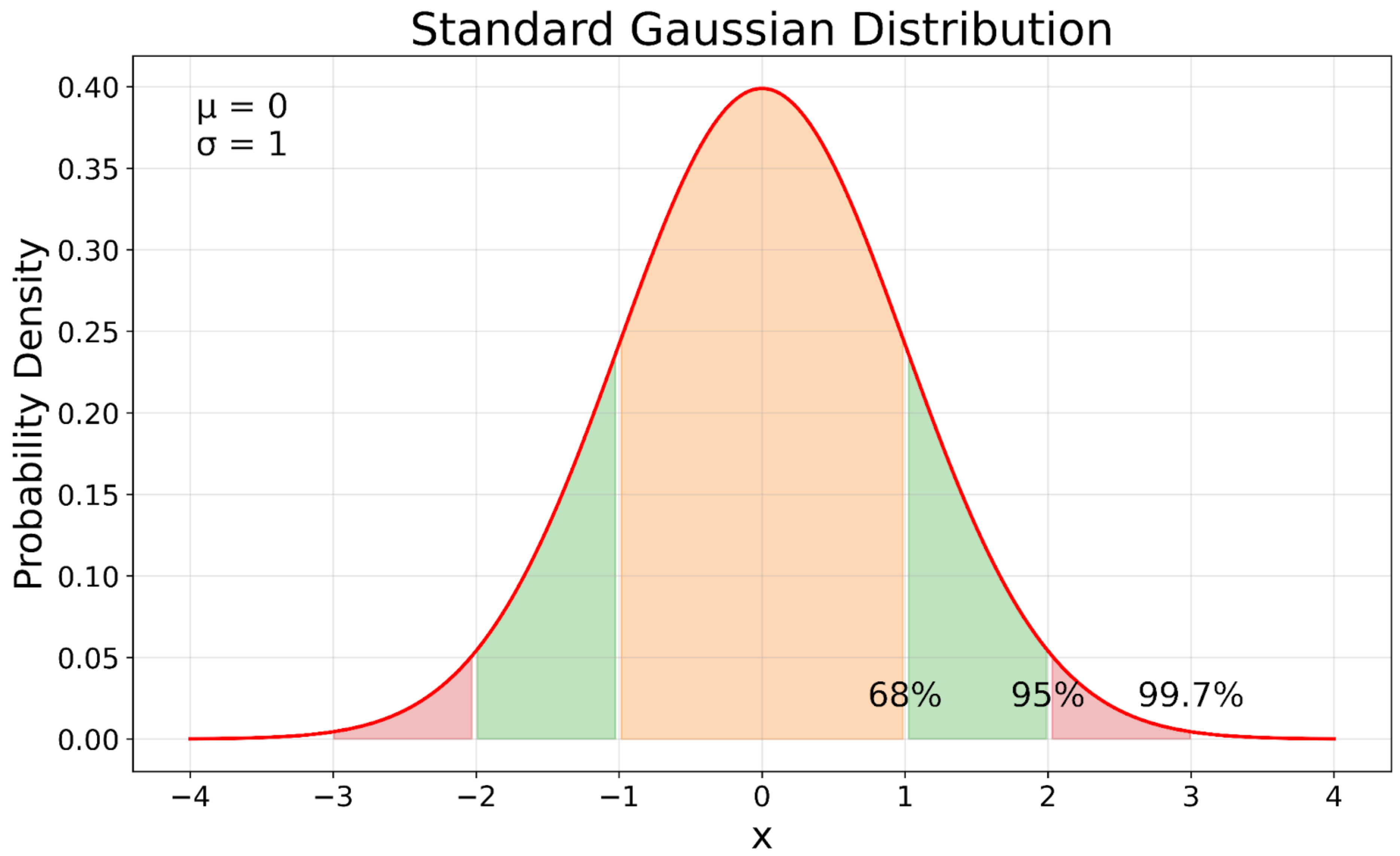
$$PDF : f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Variance of the distribution

Probability

Key Distributions in ML

- **Gaussian (Normal) Distribution**
  - One of the most important distributions in ML



A **standard** normal distribution like the one shown above, has **mean** $(\mu) = 0$ and **variance** $(\sigma^2) = 1$

Mean of the distribution

$$PDF : f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Variance of the distribution

Probability

Expectation

- Expectation of a random variable is also the **mean** value of that variable.
- For a **discrete** random variable X

$$\mathbb{E}[X] = \sum x \cdot P(X = x)$$

- For a **continuous** random variable$X$

$$\int x \cdot f(x) dx$$

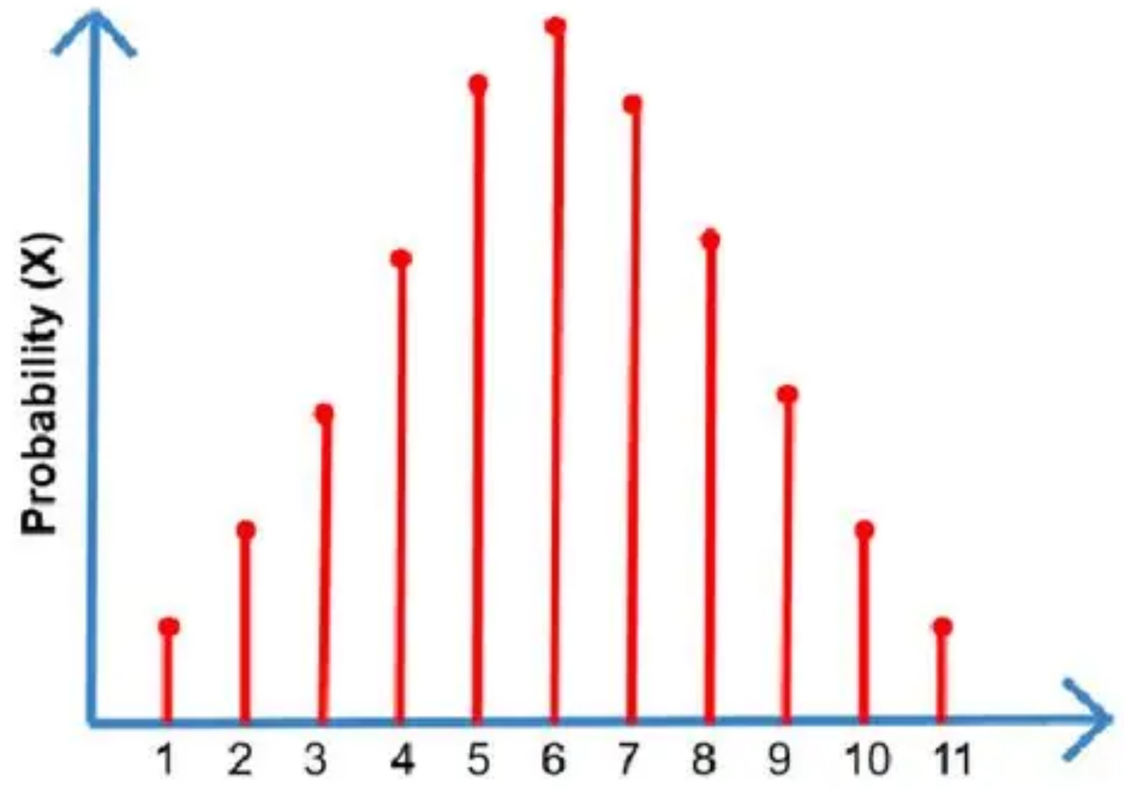
Probability

Expectation

- For a **discrete** random variable X

$$\mathbb{E}[X] = \sum x \cdot P(X = x)$$

Sum over **all possible** values that x can take multiplied by the probability of achieving that value



$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + 3 \cdot \mathbb{P}(X = 3) + \cdots 11 \cdot \mathbb{P}(X = 11)$$

- For a **continuous** random variable X

$$\int x \cdot f(x) dx$$

Probability

Properties of Expectations

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

(linearity)

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

(always true, even if variables are dependent)

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

(this only holds true if X and Y are **independent**)

Probability

Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

Probability

Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

Expectation of the **squared difference** between the **random variable** and the **mean** of the random variable

Probability

Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

Expectation of the **square** of the random variable

Probability

Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

Expectation of the random variable **squared**

Probability

Properties of Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

Probability

Properties of Variance

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[\mathbb{X}])^2$$

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

This is the **covariance** of random variables X and Y

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

How much X moves from its mean $\mathbb{E}[X]$

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

How much Y moves from its mean $\mathbb{E}[Y]$

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Product of means of X and Y

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covariance measures how X and Y vary together.
Positive means they tend to **increase/decrease** together
Negative means one **increases** as the other **decreases**.

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covariance measures how X and Y vary together.
Positive means they tend to **increase/decrease** together
Negative means one **increases** as the other **decreases**.

$$Corr(X, Y) = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covariance measures how X and Y vary together.
Positive means they tend to **increase/decrease** together
Negative means one **increases** as the other **decreases**.

$$Corr(X, Y) = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Standard Deviation: $\sigma_X = \sqrt{Var(X)}$

Probability

Covariance and Correlation

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covariance measures how X and Y vary together.
Positive means they tend to **increase/decrease** together
Negative means one **increases** as the other **decreases**.

$$Corr(X, Y) = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

This simply **normalizes** the Covariance to between -1 and +1

Review Outline

1. Probability
2. Linear Algebra

Review Outline

1. Probability
2. **Linear Algebra**

Linear Algebra

Vectors

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$

Linear Algebra

Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Addition**:
    - $\vec{u} + \vec{v} = [u_1 + v_1, u_2 + v_2, u_3 + v_3, \cdots, u_n + v_n]$

Linear Algebra
Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Addition**:
    - $\vec{u} + \vec{v} = [u_1 + v_1, u_2 + v_2, u_3 + v_3, \cdots, u_n + v_n]$

This is simply the addition of each element of the vector
"Element-wise" addition

Linear Algebra

Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
    - ***<u>Scalar</u>*** **Multiplication**:
        - $c \cdot \vec{u} = [c \cdot u_1, c \cdot u_2, c \cdot u_3, \cdots, c \cdot u_n]$

Linear Algebra

Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - ***Scalar*** **Multiplication**:
    - $c \cdot \vec{u} = [c \cdot u_1, c \cdot u_2, c \cdot u_3, \cdots, c \cdot u_n]$

This is simply the multiplication of each element of the vector with a scalar c

"Element-wise" multiplication

Linear Algebra

Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Inner Product / Dot Product**:
    - $\vec{u} \cdot \vec{v} = \sum u_i \cdot v_i$
      $= u_1 \cdot v_1 + u_2 \cdot v_2 + u_3 \cdot v_3 + \cdots + u_n \cdot v_n$

Linear Algebra

Vector Operations

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Inner Product / Dot Product**:
    - $\vec{u} \cdot \vec{v} = \sum u_i \cdot v_i$
      $= u_1 \cdot v_1 + u_2 \cdot v_2 + u_3 \cdot v_3 + \cdots + u_n \cdot v_n$

Notice that this returns a **scalar** value, not another vector

Linear Algebra

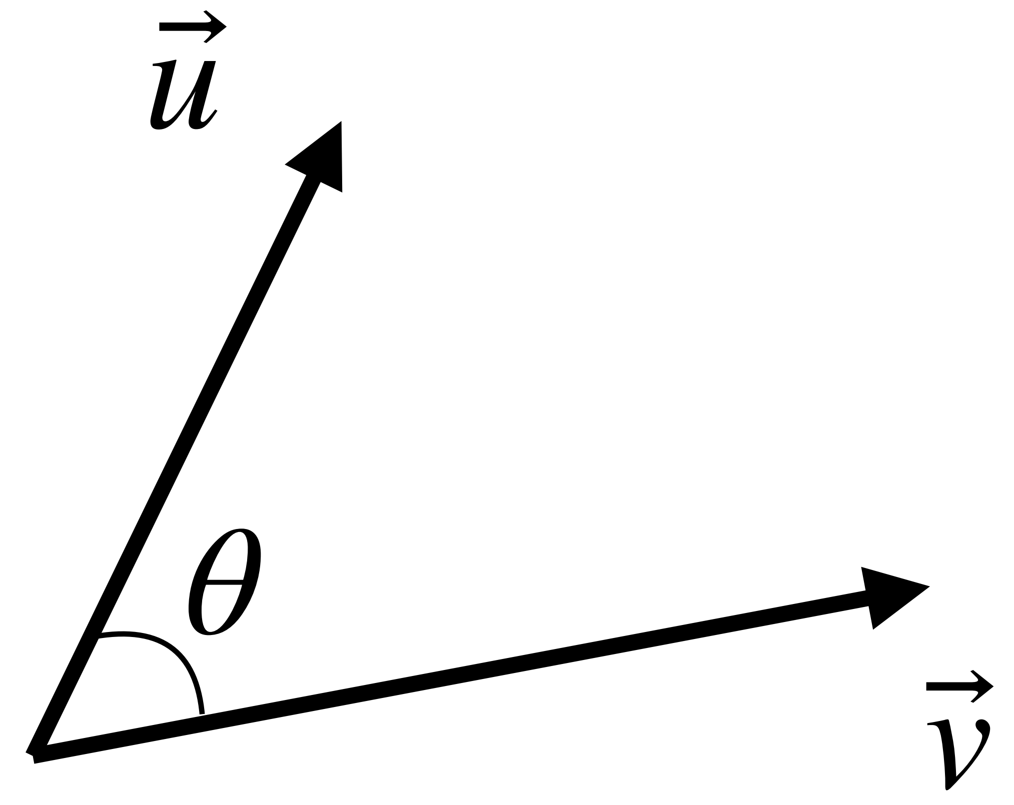
Vector Operations

The dot product also relates to the angle θ between the two vectors as

$\vec{u} \cdot \vec{v} = ||\vec{u}|| \, ||\vec{v}|| cos(\theta)$

- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Inner Product / Dot Product:**
    - $\vec{u} \cdot \vec{v} = \sum u_i \cdot v_i$

      $= u_1 \cdot v_1 + u_2 \cdot v_2 + u_3 \cdot v_3 + \cdots + u_n \cdot v_n$



Notice that this returns a **scalar** value, not another vector

Linear Algebra

Vector Operations

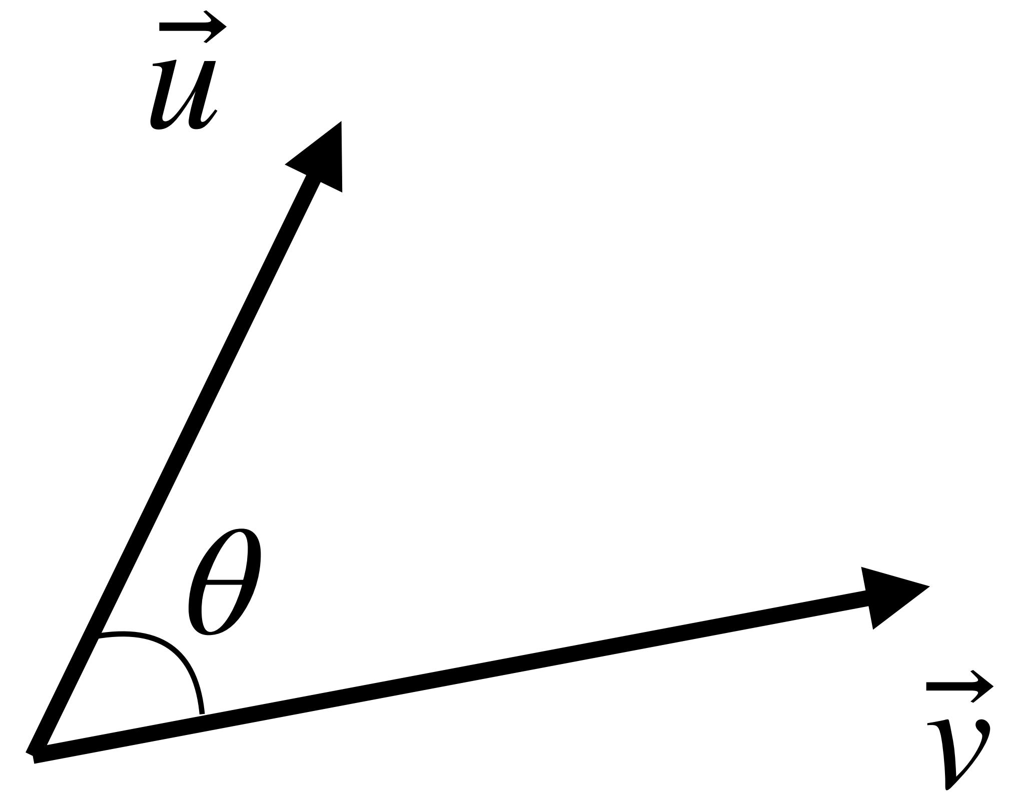
- Lets a vector $\vec{u} = [u_1, u_2, u_3, \cdots, u_n]$
- Then, this vector obeys the following operations:
  - **Inner Product / Dot Product:**
    - $\vec{u} \cdot \vec{v} = \sum u_i \cdot v_i$
      $= u_1 \cdot v_1 + u_2 \cdot v_2 + u_3 \cdot v_3 + \cdots + u_n \cdot v_n$

Notice that this returns a **scalar** value, not another vector

The dot product also relates to the angle θ between the two vectors as

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| cos(\theta)$$

This denotes the **magnitude** of the vector



Linear Algebra
Vector Norms

- L_2 Norm (Euclidean Norm)

$$||\vec{v}||_2 = \sqrt{(\sum v_i^2)}$$

- L_1 Norm (Manhattan Norm)

$$||\vec{v}||_1 = \sum |v_i|$$

- L_∞ Norm

$$||\vec{v}||_\infty = max(v_i)$$

Linear Algebra

Vector Norms

- L_2 Norm (Euclidean Norm)

$$||\vec{v}||_2 = \sqrt{(\sum v_i^2)}$$

- L_1 Norm (Manhattan Norm)

$$||\vec{v}||_1 = \sum |v_i|$$

- L_∞ Norm

$$||\vec{v}||_\infty = max(v_i)$$

These norms appear in regularization tasks later

Linear Algebra

Matrices

- A matrix $A \in \mathbb{R}^{m \times n}$ has m rows and n columns

m rows

$$\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ & & \vdots & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{matrix}$$

n columns

Linear Algebra

Matrix Operations

- **Addition**:
  - Element-wise
  - Same dimensions needed, i.e., to perform $A + B, A, B \in \mathbb{R}^{m \times n}$
- **Scalar Multiplication**:
  - Element-wise
  - This simply multiplies each entry of the matrix by some scalar c
- **Transpose**:
  - Denoted by A^T
  - This simply swaps the rows and columns.
  - If $A \in \mathbb{R}^{m \times n}$, then $A^T \in \mathbb{R}^{n \times m}$

$$\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ & & \vdots & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{matrix}$$

m rows

n columns

Linear Algebra

Matrix Operations

- **Matrix Multiplication**

  Let $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$

  Then the product matrix $(AB)_{i,j} = \sum_k A_{ik}B_{kj}$

i rows

$$\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3j} \\ & & \vdots & & \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} \end{matrix}$$

j columns

Linear Algebra

Matrix Operations

- **Matrix Multiplication**

Let $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$

Then the product matrix $(AB)_{i,j} = \sum_k A_{ik} B_{kj}$

The inner dimensions **must** be the same

i rows

$$\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3j} \\ & & \vdots & & \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} \end{matrix}$$

j columns

Linear Algebra

Matrix Operations

- **Matrix Multiplication**

  Let $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$

  Then the product matrix $(AB)_{i,j} = \sum_k A_{ik} B_{kj}$

$$
\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3j} \\ & & \vdots & & \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} \end{matrix}
$$

i rows

j columns

$$
\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} (a_{11} \cdot b_{11}) + (a_{12} \cdot b_{21}) & (a_{11} \cdot b_{12}) + (a_{12} \cdot b_{22}) \\ (a_{21} \cdot b_{11}) + (a_{22} \cdot b_{21}) & (a_{21} \cdot b_{12}) + (a_{22} \cdot b_{22}) \end{bmatrix}
$$

Linear Algebra

Matrix Operations

- **Matrix Multiplication**

Let $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$

Then the product matrix $(AB)_{i,j} = \sum_{k} A_{ik} B_{kj}$

$$\begin{matrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3j} \\ & & \vdots & & \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} \end{matrix}$$

i rows

j columns

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} (a_{11} \cdot b_{11}) + (a_{12} \cdot b_{21}) & (a_{11} \cdot b_{12}) + (a_{12} \cdot b_{22}) \\ (a_{21} \cdot b_{11}) + (a_{22} \cdot b_{21}) & (a_{21} \cdot b_{12}) + (a_{22} \cdot b_{22}) \end{bmatrix}$$

Linear Algebra

Matrix Operations

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} (a_{11} \cdot b_{11}) + (a_{12} \cdot b_{21}) & (a_{11} \cdot b_{12}) + (a_{12} \cdot b_{22}) \\ (a_{21} \cdot b_{11}) + (a_{22} \cdot b_{21}) & (a_{21} \cdot b_{12}) + (a_{22} \cdot b_{22}) \end{bmatrix}$$

- **Matrix Multiplication**

Let $A \in \mathbb{R}^{i \times j}$ and $B \in \mathbb{R}^{j \times k}$

Then the product matrix $(AB)_{i,j} = \sum_k A_{ik} B_{kj}$

- Not Commutative: $AB \neq BA$

- Associative: $A(BC) = (AB)C$

- Distributive: $A(B + C) = AB + AC$

- Transpose: $(AB)^T = B^T A^T$

Linear Algebra
Special Matrices

- **Identity**
    - $AI = IA = A$
    - I is a matrix where all diagonal entries are 1 and everything else is 0
- **Diagonal**
    - A more general case of I where all diagonal entries are **non-zero** and all non-diagonal elements are zero
- **Symmetric**
    - $A = A^T$
    - For example, covariance matrices are symmetric
- **Orthogonal**
    - $A^TA = AA^T = I$
    - The dot product of each column is zero

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

This is a system of m equations with n unknown parameters

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

$$a_{31}x_1 + a_{32}x_2 = b_3$$

$$a_{41}x_1 + a_{42}x_2 = b_4$$

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

This is a system of m equations with n unknown parameters

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 &= b_1 \\
a_{21}x_1 + a_{22}x_2 &= b_2 \\
a_{31}x_1 + a_{32}x_2 &= b_3 \\
a_{41}x_1 + a_{42}x_2 &= b_4
\end{aligned}
\longrightarrow
\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} =
\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}
$$

Conclusion

- We looked at **supervised** vs **unsupervised** algorithms
- We looked at **regression** vs **classification** problems
- We looked at a few models and their loss functions
- We reviewed probability and linear algebra

Conclusion

- Performance of any learned model depends on
  - **Data**
    - Distribution of data
    - Quality and labelling of data
    - Dimensionality of data
    - Type of data
      - Images vs audio vs graphs
  - **Model**
    - Type of model used
      - Neural Networks vs Decision Tree vs XGBoost
    - Loss functions used
      - Mean Squared Error vs Mean Absolute Error vs Root Mean Squared Error

Looking Ahead

- **Next Class**:
  - Review - Linear Algebra and Calculus, Linear regression
- **Next Couple Weeks**:
  - Linear regression, gradient descent, regularization
- **Goal For This Course:**
  - Give students the ability to read and understand papers on their own