Probability Review

CSCI-P556 Applied Machine Learning
Lecture 8

D.S. Williamson

Agenda and Learning Outcomes

Today's Topics

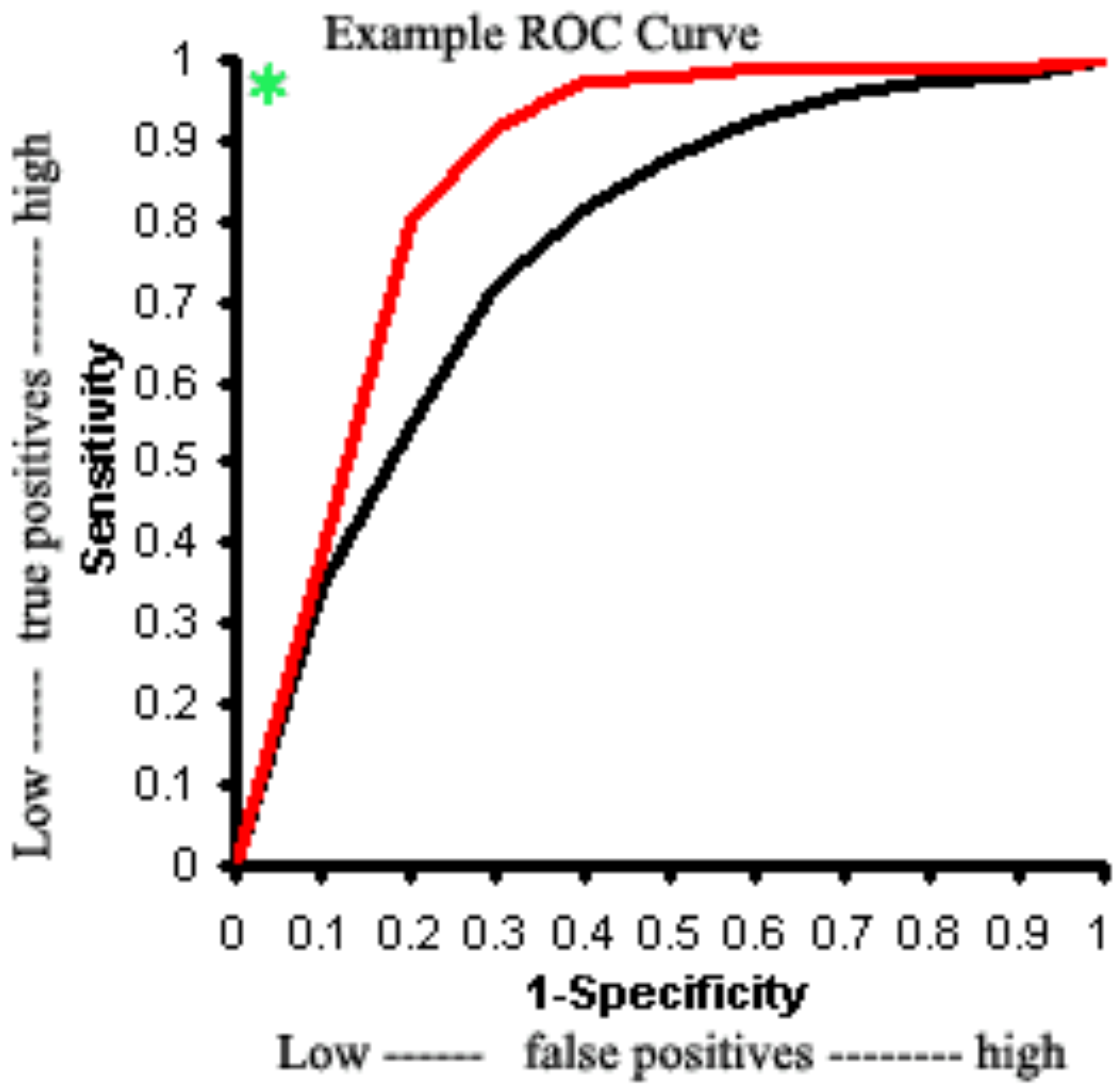
- **Topics**:
  - A slightly different approach for generating ROC curves
  - Measures of performance for regression
  - Probability Review (Part I)

- **Announcements**:
  - Homework 1 has been posted to Canvas. Due on 2/24.
    - Will give time during class today to meet partner
    - Create Repos per Piazza post instructions
  - No class Tuesday (Wellness Day)

ROC Curve Generation

A slightly different perspective

- **<u>Step 1</u>**: Sort predictions on test set
- **<u>Step 2</u>**: Locate a *threshold* between examples with opposite categories
- **<u>Step 3</u>**: Compute TPR & FPR for each *threshold* of Step 2
- **<u>Step 4</u>**: Connect the dots

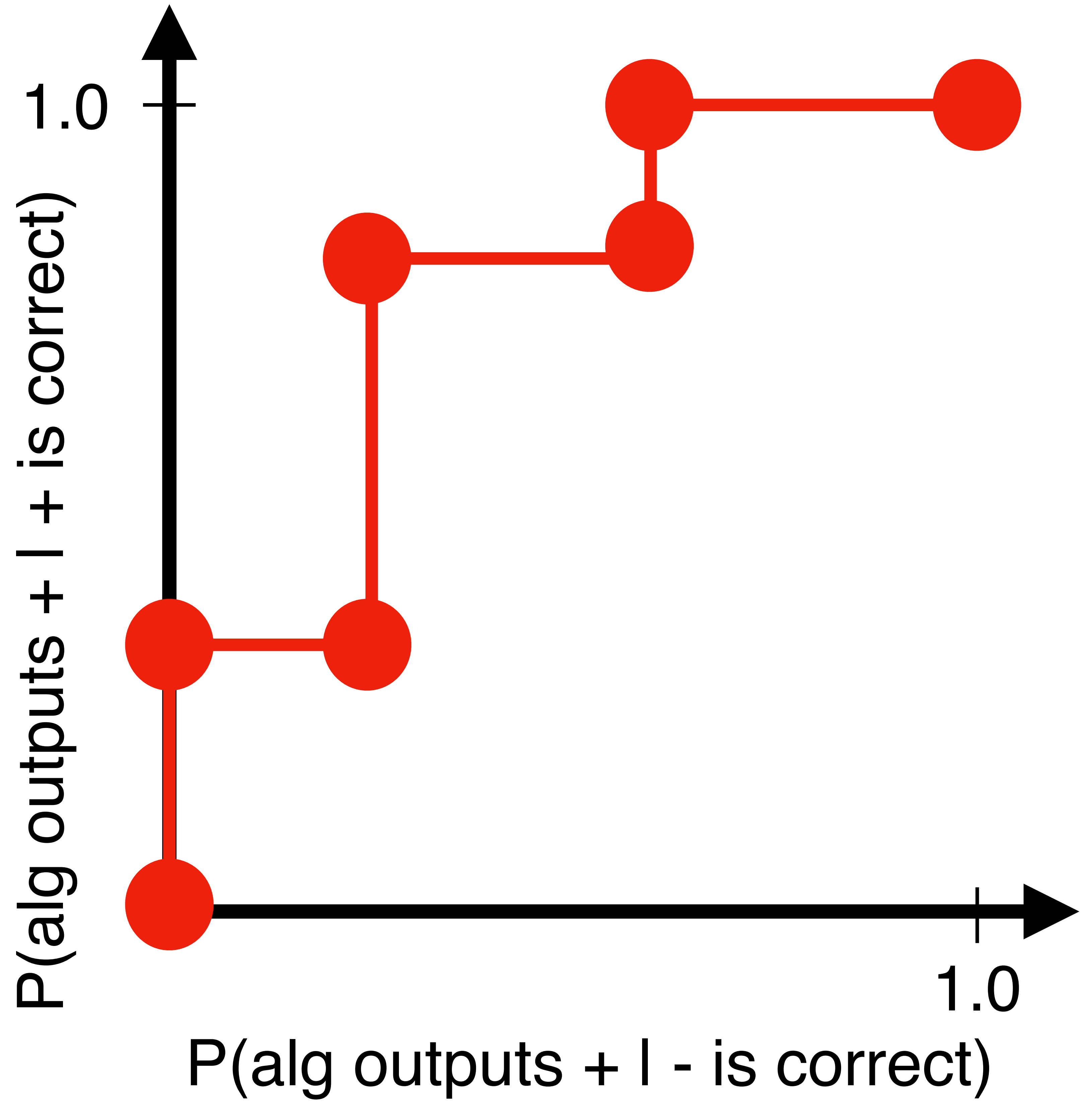


Example: Plotting ROC Curves

A slightly different perspective

- **Step 1**: Sort predictions on test set
- **Step 2**: Locate a *threshold* between examples with opposite categories
- **Step 3**: Compute TPR & FPR for each *threshold* of Step 2
- **Step 4**: Connect the dots

	ML Output Sorted			Correct Category
Thr. #1			TPR = (0/5), FPR = (0/5)	
	Ex. 9	0.99	TPR = (2/5), FPR = (0/5)	+
	Ex. 7	0.98		+
Thr. #2				
	Ex. 1	0.72	TPR = (2/5), FPR = (1/5)	-
Thr. #3				
	Ex. 2	0.70	TPR = (4/5), FPR = (1/5)	+
	Ex. 6	0.65		+
Thr. #4				
	Ex. 10	0.51	TPR = (4/5), FPR = (3/5)	-
	Ex. 3	0.39		-
Thr. #5				
	Ex. 5	0.24	TPR = (5/5), FPR = (3/5)	+
Thr. #6				
	Ex. 4	0.11	TPR = (5/5), FPR = (5/5)	-
	Ex. 8	0.01		-
Thr. #7				



Evaluating Regression Problems

Recall: Types of Labels (or Targets)

Labels are generally divided into two classes

$$D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

- **Regression**: Decimal (Continuous) values are assigned as the label
  - Examples:
    - A person's height or weight to the 3rd decimal place
    - The cost of a home
    - Stock market price
    - Outputting an image of a dog/cat/bear/fish
    - Create musical audio signals
  - It is termed ***regression*** when a supervised learning algorithm learns a mapping from an input to a continuous label

	size [sqft]	age [yr]	dist [mi]	inc [$]	dens [ppl/mi^2]	y
$\mathbf{x}_1$	1250	5	2.85	56,650	12.5	2.35
$\mathbf{x}_2$	3200	9	8.21	245,800	3.1	3.95
$\mathbf{x}_3$	825	12	0.34	61,050	112.5	5.10

Table 3.2: An example of a regression problem: prediction of the price of a house in a particular region. Here, features indicate the size of the house (size) in square feet, the age of the house (age) in years, the distance from the city center (dist) in miles, the average income in a one square mile radius (inc), and the population density in the same area (dens). The target indicates the price a house is sold at, e.g. in hundreds of thousands of dollars.

$y \in \mathbb{R}$

Real number (e.g. decimal or float; 1.232,343,232.4545,…)

Evaluating Regression Models

Ex: Median Housing Price prediction

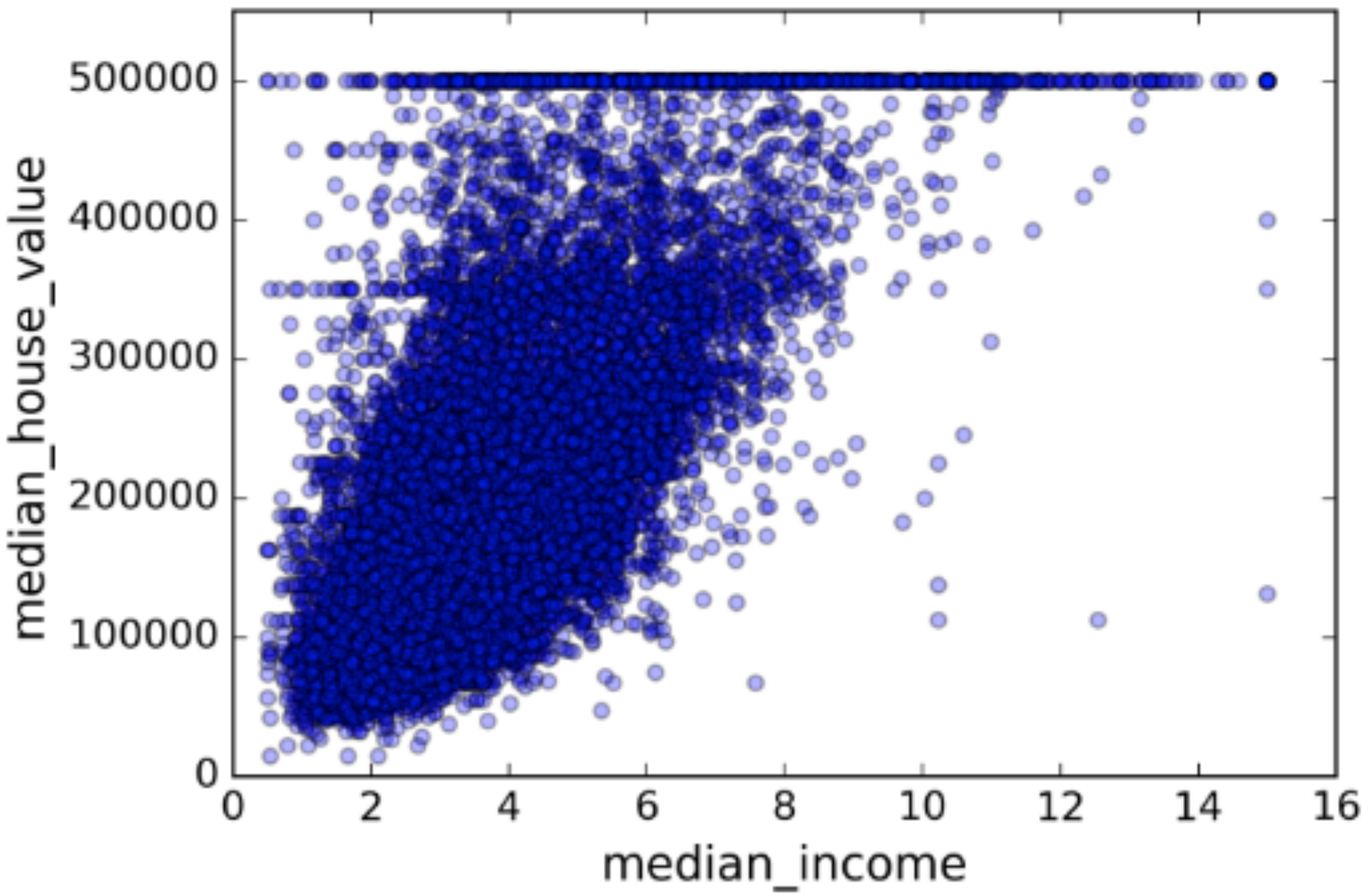
- **Recall**: Suppose you are a Data Scientist at a Housing Corporation. Your boss wants you to build a prediction model of median housing prices in California using their census data

- **Modifications:**

  - Let's use 'Median Income' as the only feature/ attribute

  - Based on relationship between 'Median Income' and 'Median Housing Price', let's use linear regression to perform the prediction

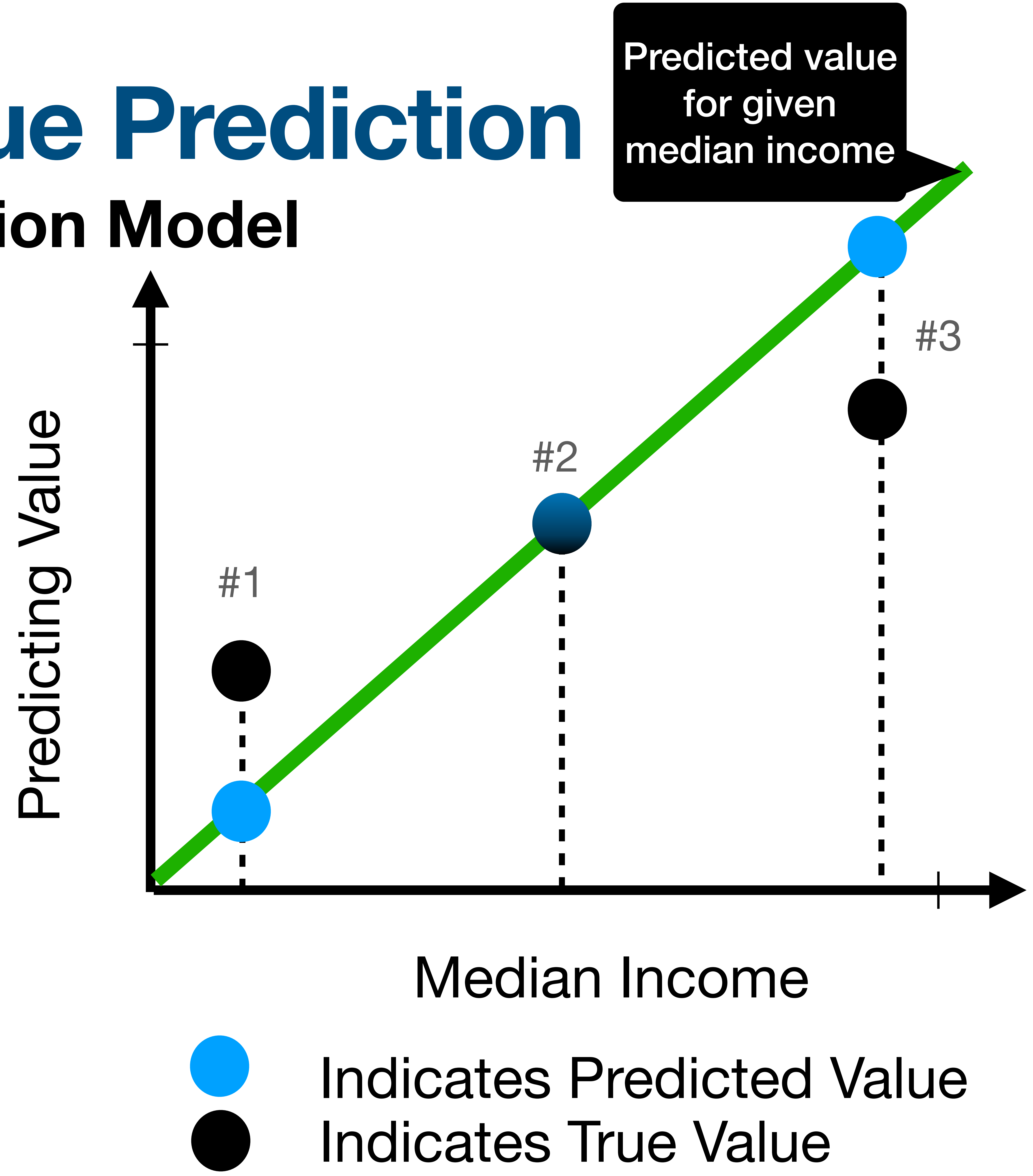
  - Assume linear regression model has been trained



Median House Value Prediction

A simplified Linear Regression Model

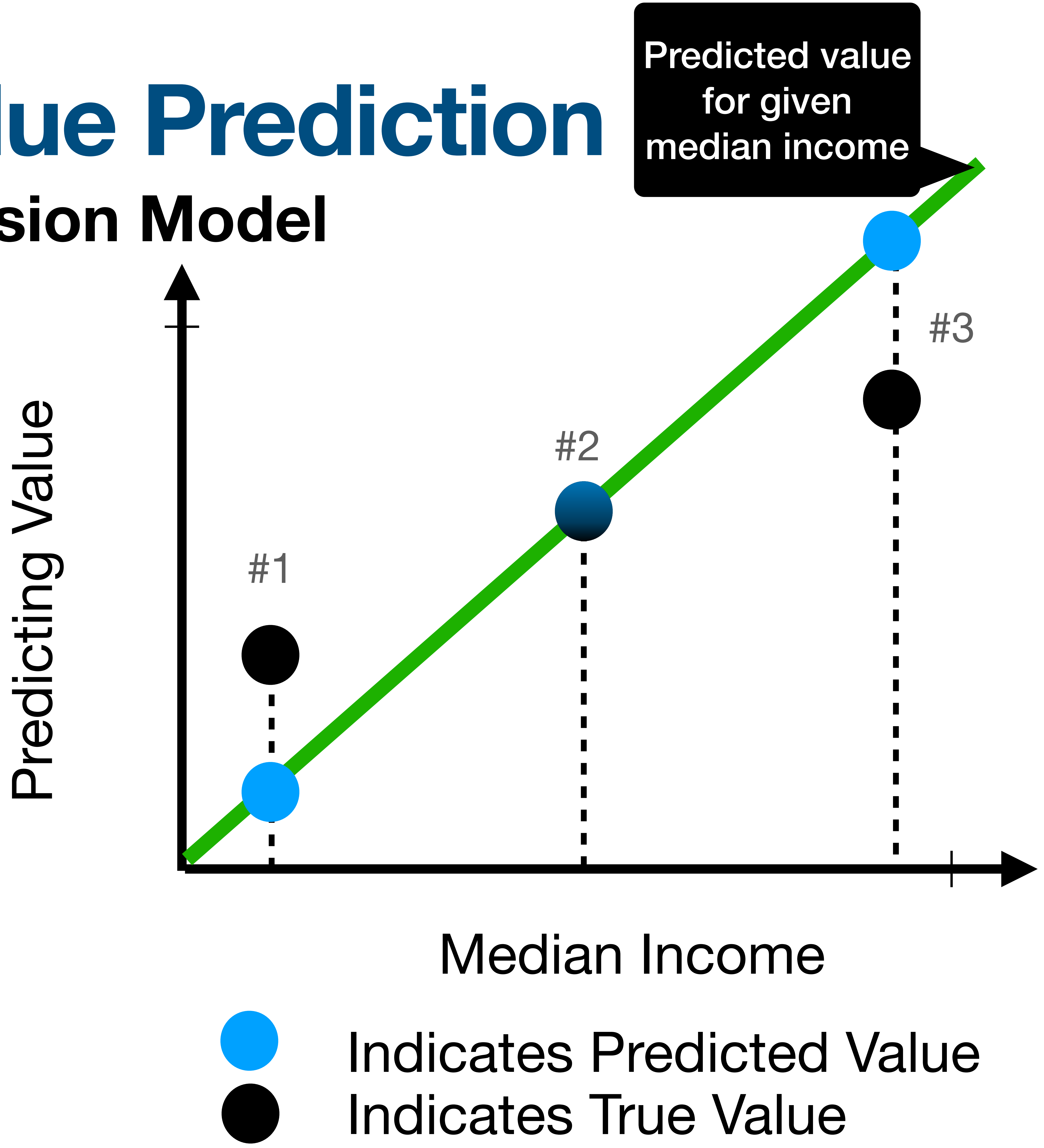
- For a given income, the model outputs the estimated house value
- Consider three districts (represented as points)
- Point#2 is predicted correctly, whereas Pts. #1 and #3 are incorrect



Median House Value Prediction

A simplified Linear Regression Model

- Need to summarize performance over all points/ predictions
- Need a metric for regression similar to accuracy or AUC
- **Two common metrics are:**
    - Mean Absolute Error (MAE)
    - Root Mean-Square Error (RMSE)



Median House Value Prediction

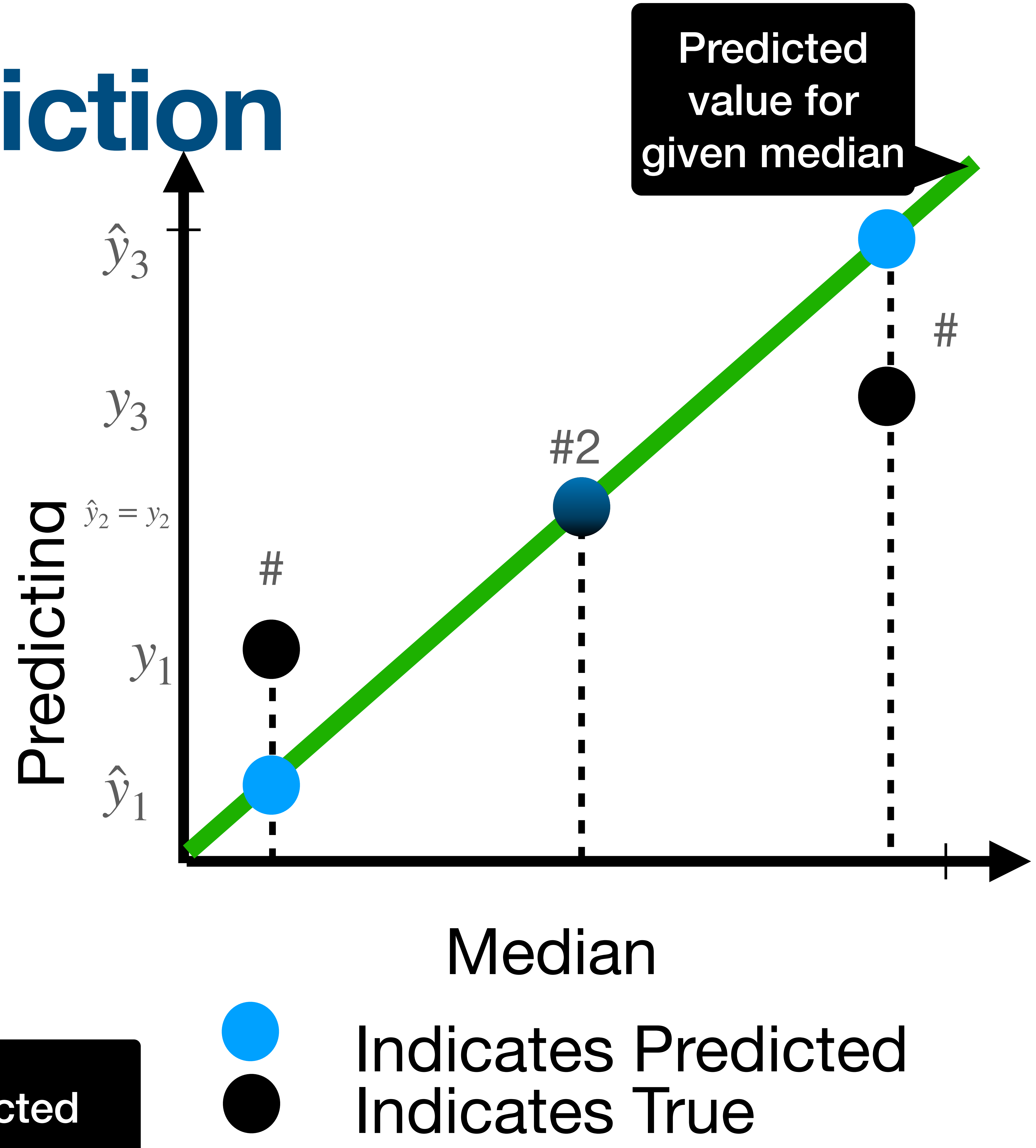
A simplified Linear Regression Model

- Compute ***mean absolute error (MAE)*** by computing the error in the prediction for each sample, and averaging this error over all samples

$$MAE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

True

Predicted



Median House Value Prediction

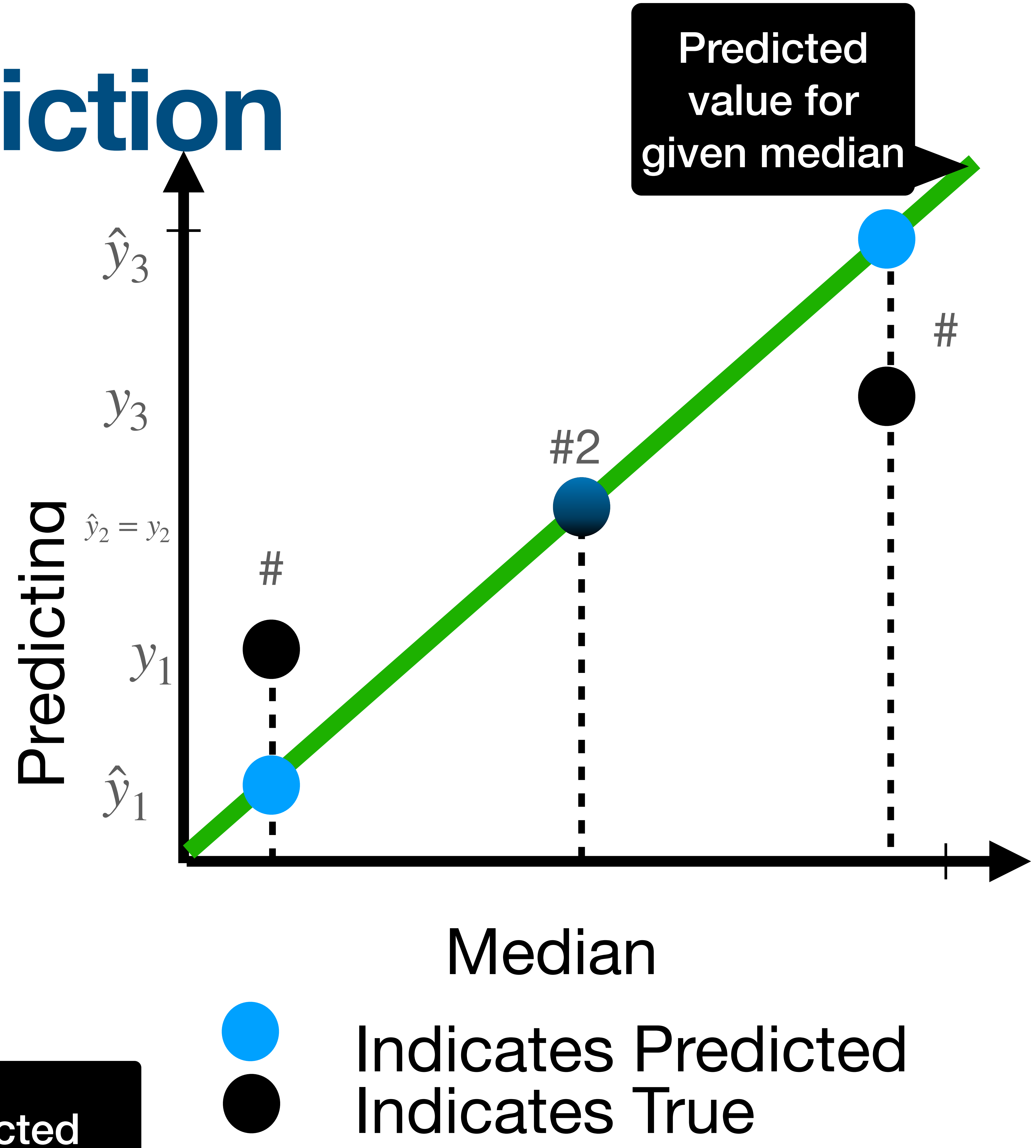
A simplified Linear Regression Model

- Compute ***mean square error (MSE)*** by computing the error in the prediction for each sample, squaring each error, and averaging this result over all samples
  - May also take root of MSE (e.g. RMSE)

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

True

Predicted



- **Other metrics exist (R^2, F*, t-test,…), but we'll cover these on an as-needed basis**

Probability Review

Meaning of Probability

What is the meaning of probability?

- ***Probability*** is: (1) A means of representing and reasoning about uncertainty and (2) Describing the repeatability of an event

- **Definition 1: Frequentist probability**
  - An outcome has a probability p of occurring
  - If we repeated, then proportion p of the repetitions would result in that outcome
  - Example: Drawing a 'king of spades' from a deck of cards

- **Definition 2: Bayesian (or Degree of Belief) Probability**
  - Qualitative levels of certainty (Not based on repeated trials)
  - Value of 1 indicates absolute certainty, 0 means no possibility
  - Example: There is a 40% chance that this patient has the flu

We Treat Frequentist and Bayesian Probabilities the Same

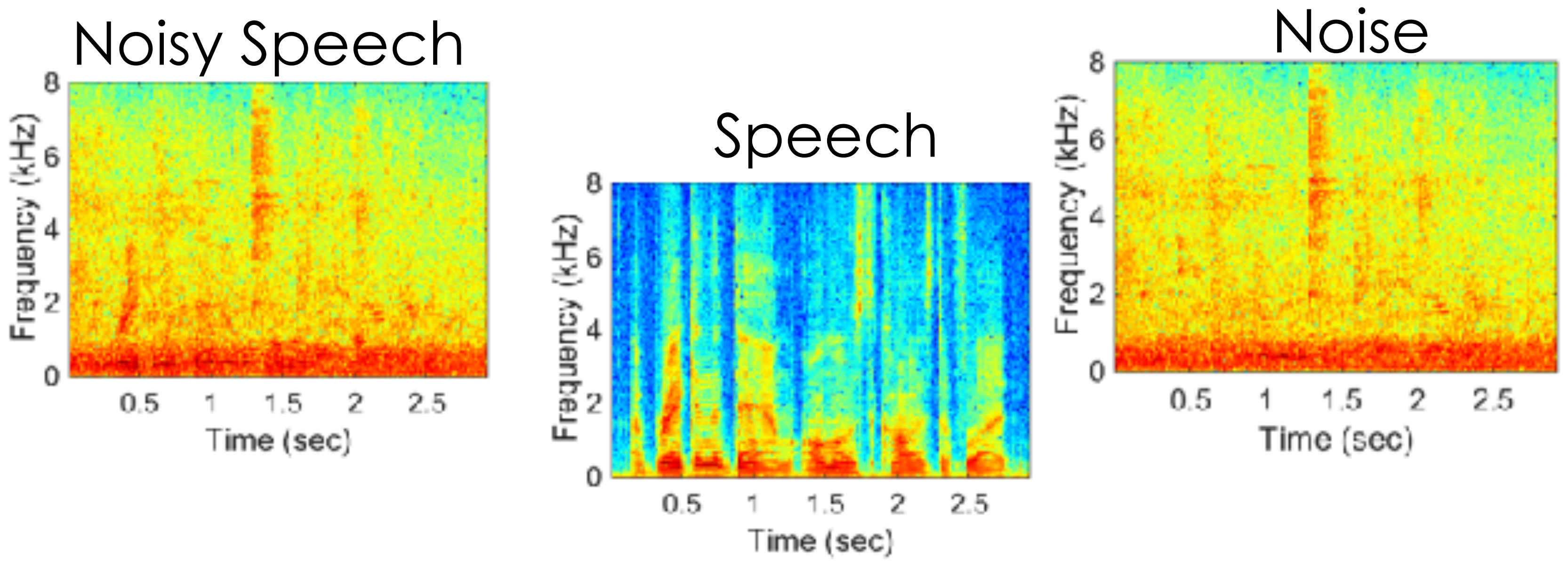
Example

- A large bin of electrical resistors (R). 100 R marked 1 ohm, 500 R marked 10 ohm, 150 R marked 100 ohm, 250 R marked 1 kilo-ohm.

- **If one resistor is pulled out at random, what are the possible outcomes**?
  - There are four possible outcomes
    - P(1 ohm) = 100/1000 = 0.1
    - P(10 ohm) = 500/1000 = 0.5
    - P(100 ohm) = 150/1000 = 0.15
    - P(1 kilo-ohm) = 250/1000 = 0.25

Random Variables

- We often use ***random variables*** to model events (or inputs that vary)
- A ***random variable***, X, is a function that maps a sample space {S} onto the real number line
    - $X : S \to \mathbb{R}$
- The following inputs can each be represented with random variables:



Cats

Bears

Random Variables

- **Example**: Three consecutive (fair) coin tosses
  - X = the number of heads in the first toss
  - Y = the number of heads in all three tosses
- **What is the outcome sample space?**
  - S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- **What are X and Y?**
  - $X: S \rightarrow \{0,1\}$
  - $Y: S \rightarrow \{0,1,2,3\}$

ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(\omega)$	1	1	1	1	0	0	0	0
$Y(\omega)$	3	2	2	1	2	1	1	0

Types of Random Variables

- **Discrete Random Variables (DRV)**
  - If the RV can take integer (discrete) values $\{x_1, x_2, \ldots, x_m\}$ only, no values in between
  - **Example**:
    - X = the number of car accidents in Bloomington next weekend
    - *Possible values:* 0, 1, 2, ... -> X is a discrete RV

- **Continuous Random Variables (CRV)**
  - If the RV can take any value between two limits, where the number of possible values is uncountable
  - **Example**
    - X = a random person's height (in inches), measured to an infinite degree of accuracy.
    - *Possible values*: Any number in the interval [20, 100]

Discrete Random Variables

- For any real number $x \in X$, what do we mean by $P(X = x)$?
- Formal Definition:
    - $P_X(X=x) = P\{s \in S : X(s) = x\}$
    - P_X is called the induced probability function on X, defined in terms of the original probability function P on S
- From Fair Coin Example
    - Y = the number of heads in all three tosses

$$
\begin{aligned}
P_Y(Y=0) &= P\{s \in S : Y(s) = 0\} \\
&= P\{TTT\}
\end{aligned}
$$

$$
\begin{aligned}
P_Y(Y=1) &= P\{s \in S : Y(s) = 1\} \\
&= P\{HTT, THT, TTH\}
\end{aligned}
$$

Probability Mass Function

Distribution for Discrete Random Variables

- For a DRV, X, the probability model, $P_X(X=x)$, is called a ***probability mass function (PMF).*** The PMF gives values for all x
- Example: Two traffic lights
  - $S = \{R_1R_2, R_1G_2, G_1R_2, G_1G_2\}$
  - T -> the RV for the number of red lights
  - T = {0,1,2}

$$P_T(T=t) = ?$$
$$P_T(T=0) = P\{G_1G_2\} = 1/4$$
$$P_T(T=1) = P\{R_1G_2, G_1R_2\} = 1/2$$
$$P_T(T=2) = P\{R_1R_2\} = 1/4$$

- Properties of PMF
  - $1 \geq P_X(X = x) \geq 0$ for all x
  - $\sum_{x \in X} P_X(X = x) = 1$
  - $P(X \in A) = \sum_{x \in A} P(X = x)$ for all $A \subset \mathbb{R}$

Important PMF Families

- **<u>Bernoulli distribution</u>**
  - Have an experiment with only two possible outcomes (i.e. S = {success, failure}), where P{success} = *p*
  - Let, X(success) = 1 and X(failure) = 0

$$P_X(x) = \begin{cases} 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- **<u>Binomial distribution</u>**
  - Perform *n* independent pass/fail trials with P{success} = *p*
  - Let X be the # of passes in *n* tests, so X ∈{0,1,…. n}

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$$\underbrace{\binom{n}{x}}_{\text{Number of outcomes}}\ \underbrace{p}_{\text{Probability of success}}{}^{\overbrace{x}^{\text{Number of successes}}}\ \underbrace{(1-p)}_{\text{Probability of fail}}{}^{\overbrace{n-x}^{\text{Number of fails}}}$$

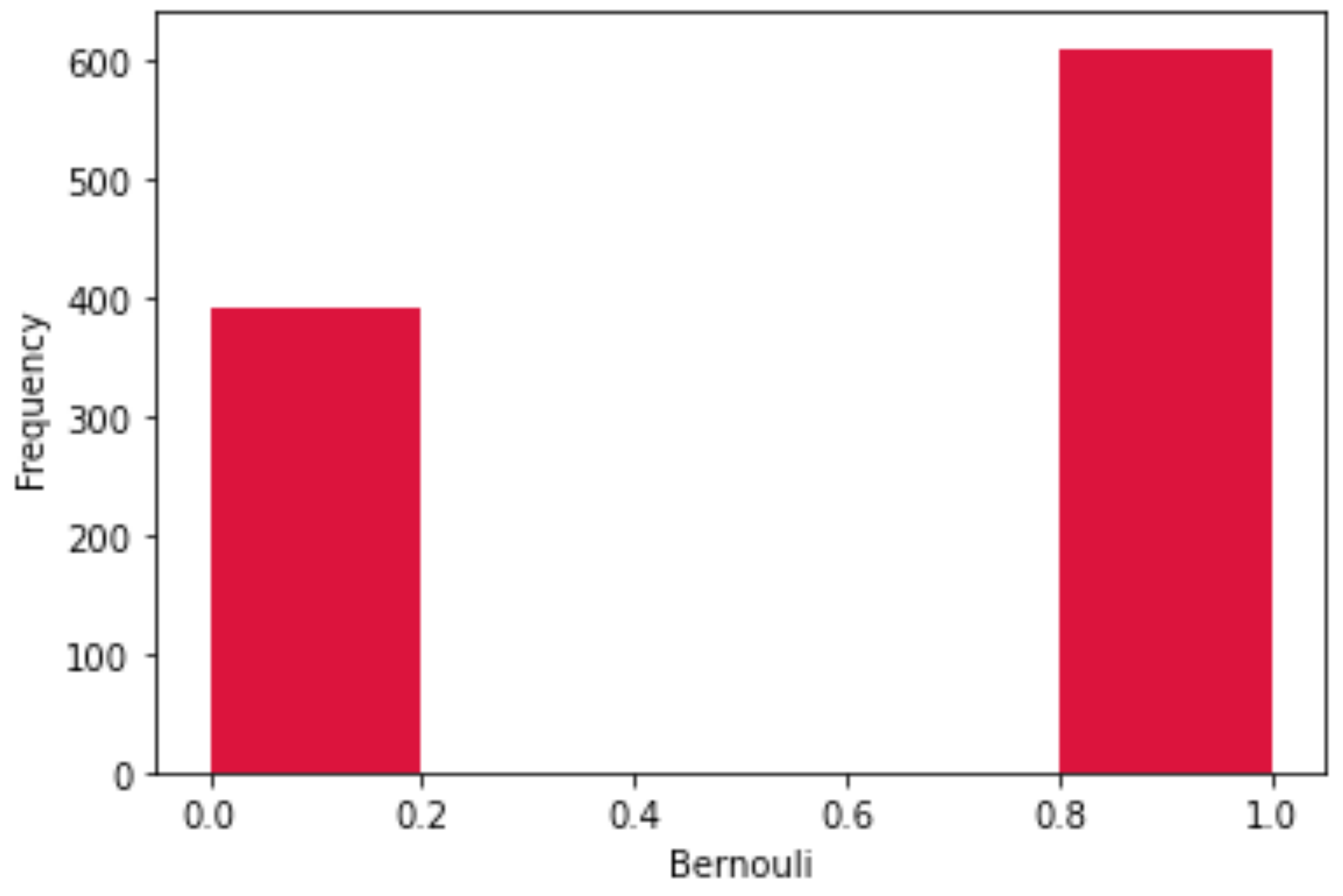
Bernoulli and Binomial Distributions in Python

Generate random samples from distribution

- **<u>Bernoulli distribution</u>**

```
from scipy.stats import bernoulli
import seaborn as sb

data_bern = bernoulli.rvs(size=1000,p=0.6)
ax = sb.distplot(data_bern,
                 kde=False,
                 color='crimson',
                 hist_kws={"linewidth": 25,'alpha':1})
ax.set(xlabel='Bernouli', ylabel='Frequency')
```

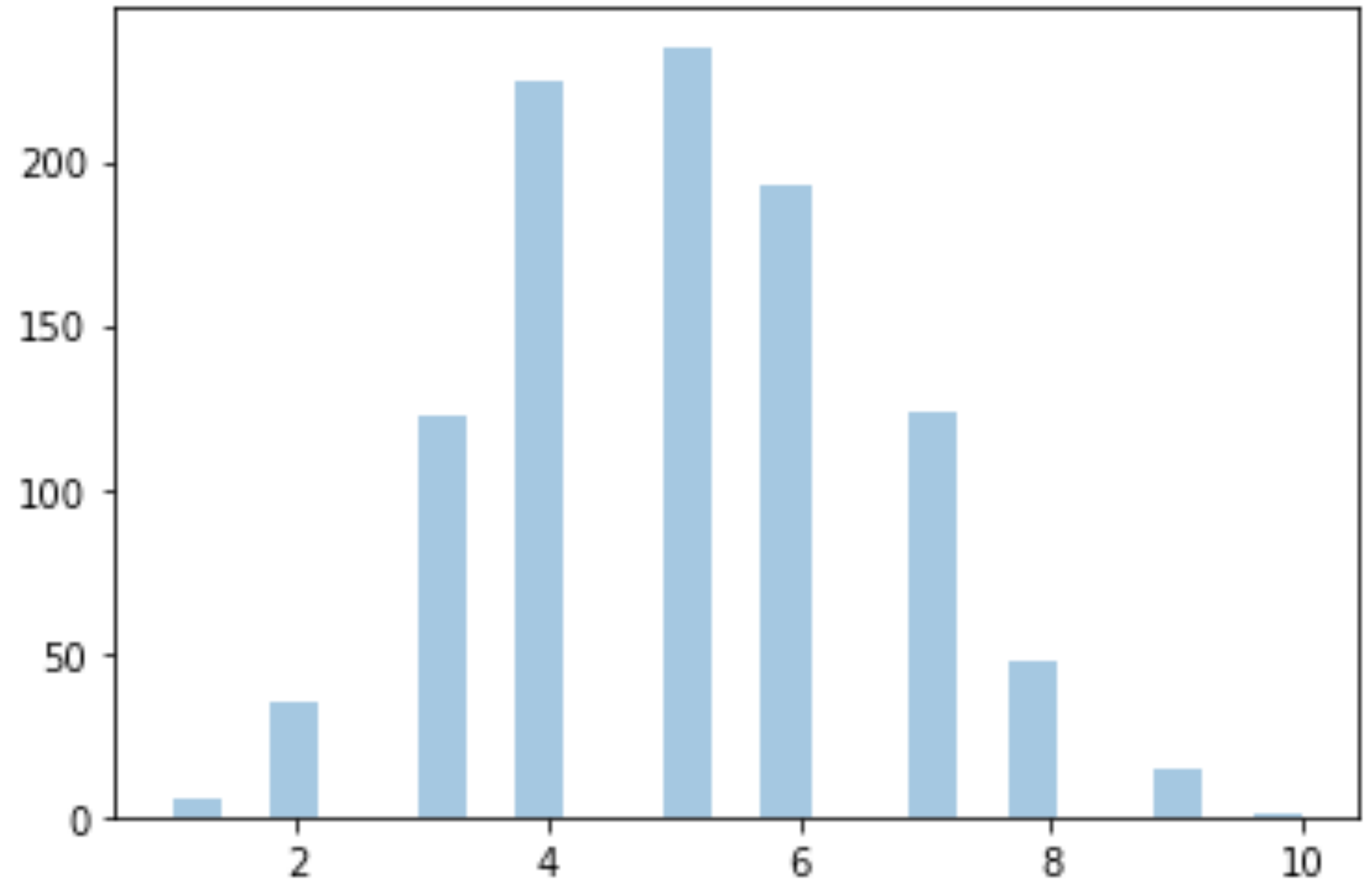


- **<u>Binomial distribution</u>**

```
from numpy import random
import matplotlib.pyplot as plt
import seaborn as sns

sns.distplot(random.binomial(n=10, p=0.5, size=1000), hist=True, kde=False)

plt.show()
```



Cumulative Distribution Function

- The ***cumulative distribution function*** (CDF) is another way of describing a random variable

$$F_X(x) = P_X(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

- Properties of CDF
  - $0 \leq F_X(x) \leq 1$
  - $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$
  - F(x) is right-continuous (i.e. F(2.5) = F(2), if x = 2 is in X, but x = 2.5 is not)
  - $P(a < x <= b) = F(b) - F(a)$
  - $P(X = x) = F(x) - \lim_{y \to x} F(y)$ = "size of jump in F at x"

CDF Example (cont.)

$$F_X(x) = P_X(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

- **<u>Example:</u> Tossing a fair coin 3 times**
  - S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}
  - Let X be the RV for the # of heads, $S_X = \{0,1,2,3\}$
    - $P(X = 0) = 1/8$, $P(X = 1) = 3/8$
    - $P(X = 2) = 3/8$, $P(X=3) = 1/8$
  - What is $F_X(x)$? (e.g. $F_X(0)$, $F_X(1)$, $F_X(2)$, $F_X(3)$)

- <u>Solution</u>
  - $F_X(X \leq 0) = P(X = 0) = 1/8$
  - $F_X(X \leq 1) = P(X=1) + P(X = 0) = ½$

  $F_X(X \leq 2) = F_X(X \leq 1) + P(X = 2) = 7/8$

  $F_X(X \leq 3) = F_X(X \leq 2) + P(X = 3) = 1$

Continuous Random Variables

- ***Continuous RVs*** take any value between two limits.
- Probability model, $f_X(X=x)$, is called a ***probability density function (***PDF)
- PDF satisfies

$$P(a \leq x \leq b) = \int_a^b f(x)dx \quad \text{whenever a} \leq \text{b}$$

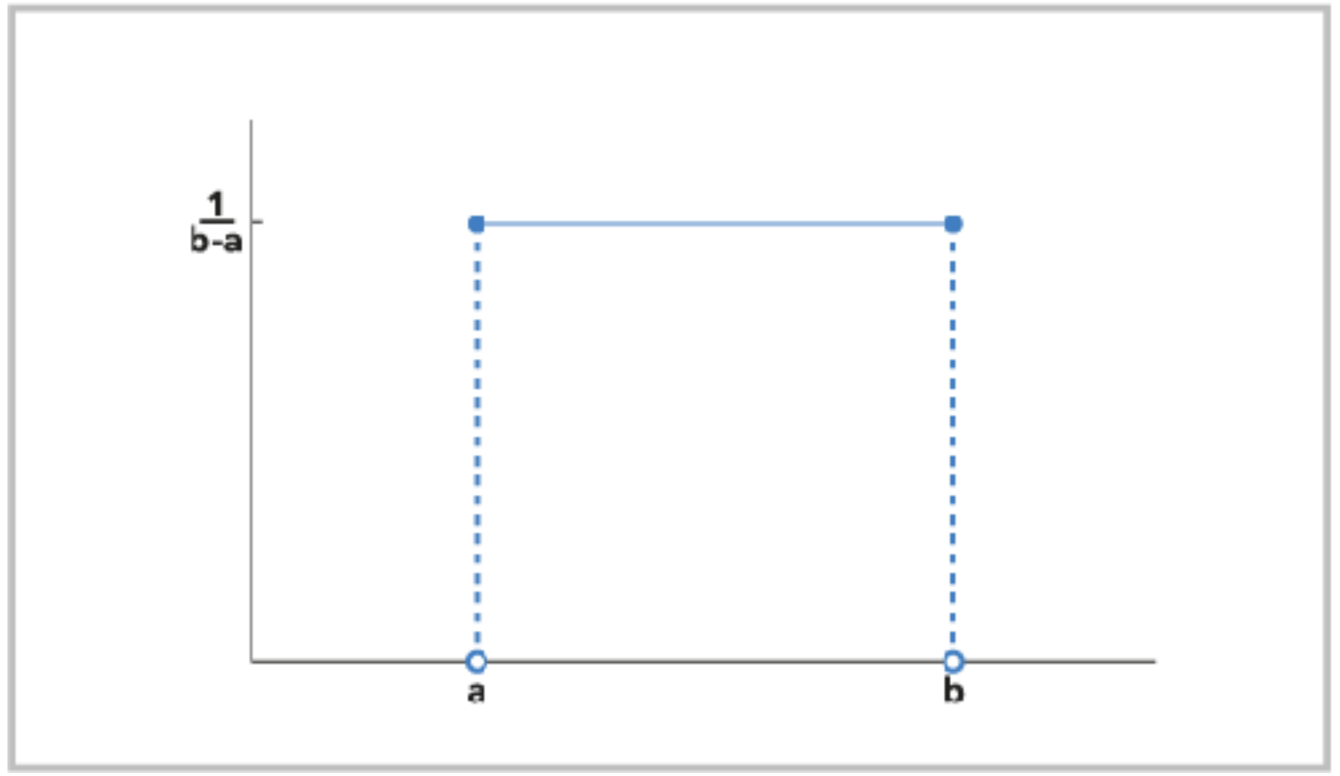
- **Other properties of f(x):**
  - $f(x) \geq 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $P(X = x) = 0$, for all x

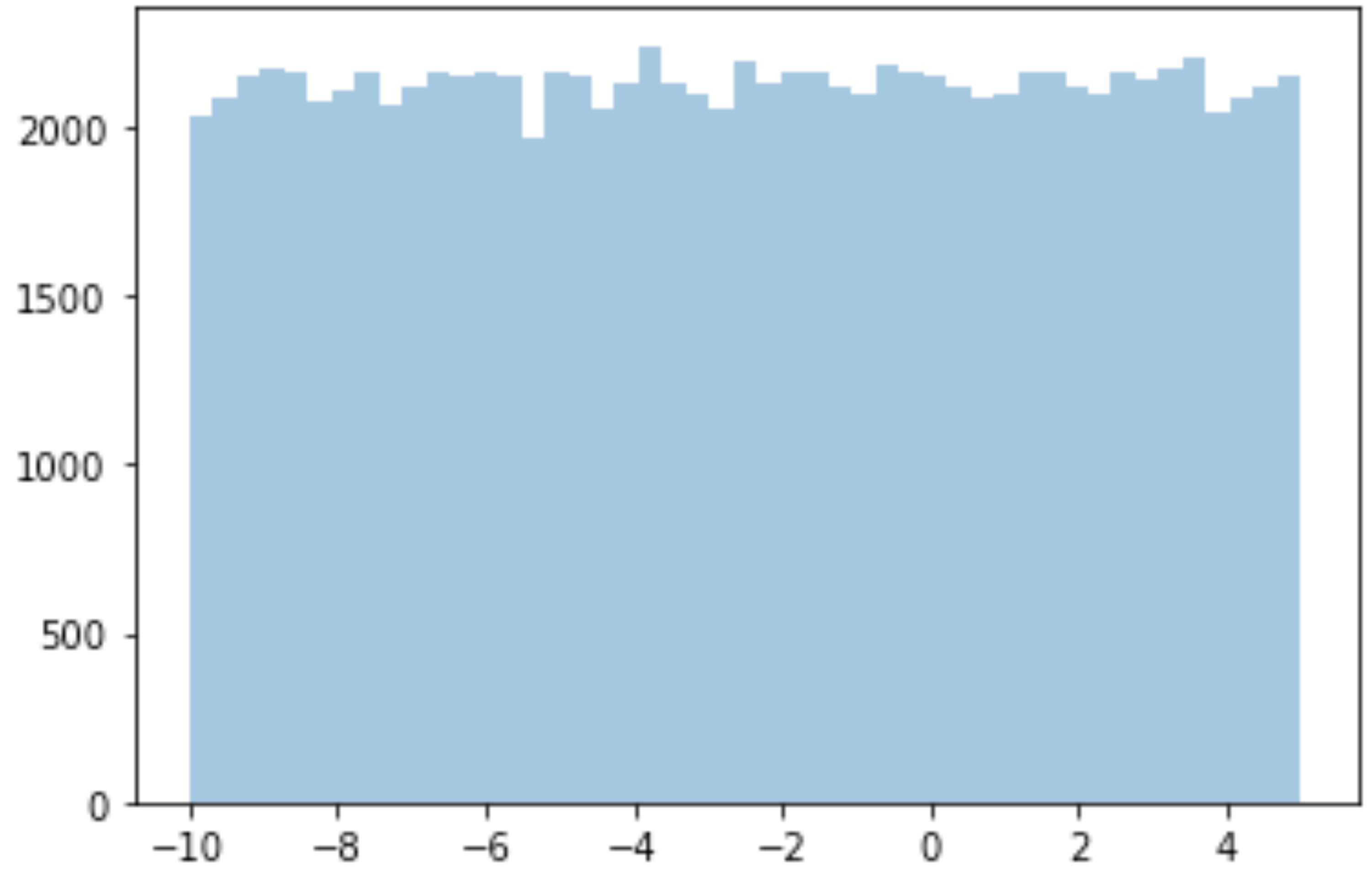
Important PDFs

- **Continuous Uniform Distribution**
  - Pretty much the same as the PMF for DRVs
  - Denoted as X ~ Uniform(a,b)

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$$

```
import numpy as np
s = np.random.uniform(-10,5,100000)
sns.distplot(s, hist=True, kde=False)
plt.show()
```





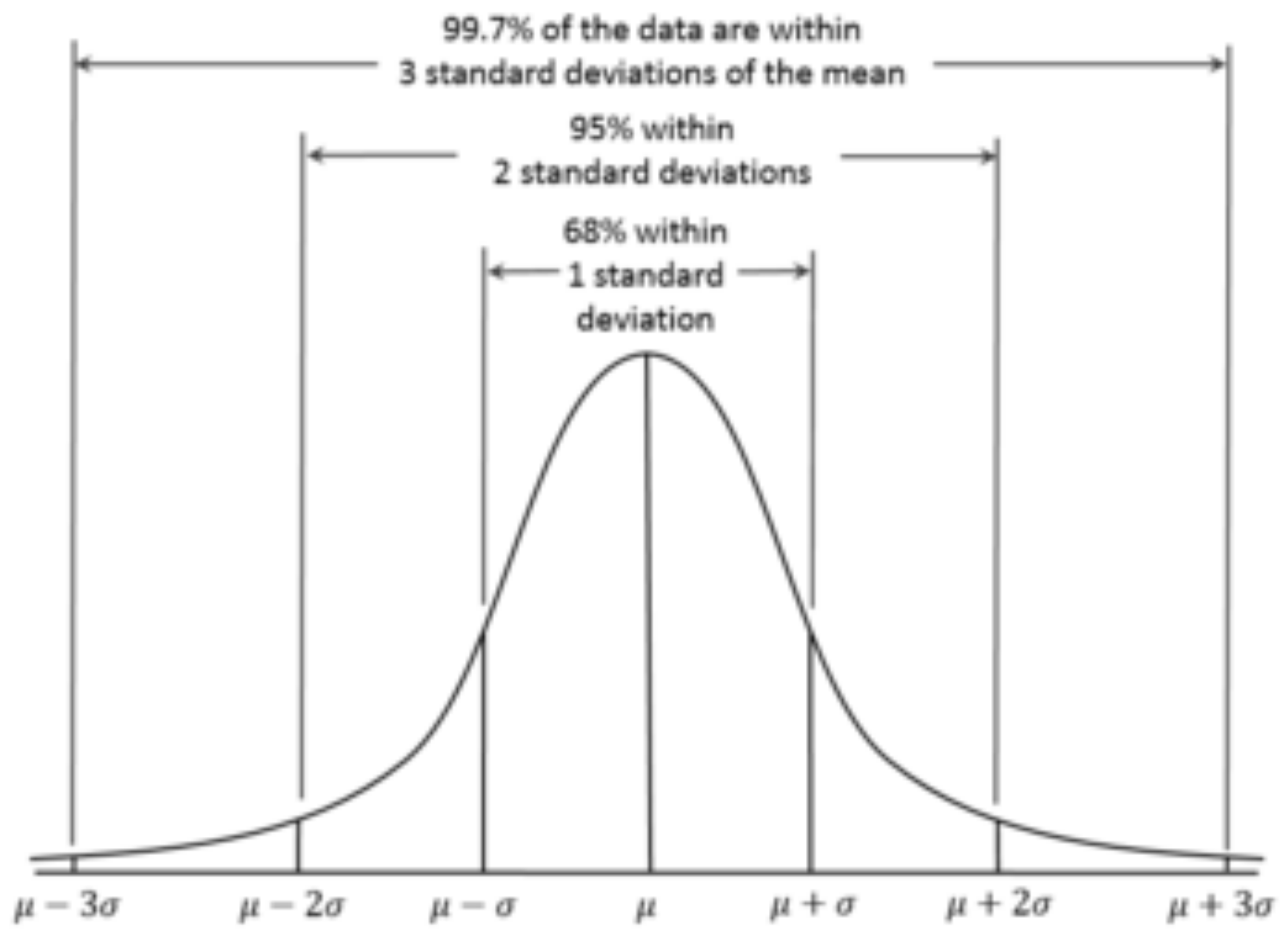
Important PDFs

Gaussian Distribution

- **<u>Normal/Gaussian Distribution</u>**
  - X ~ $N(\mu, \sigma)$
  - X has normal distribution with parameters μ and σ
  - X has PDF f(x) with

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

  - μ is the mean of the distribution
  - σ is the standard deviation of the distribution (more on these later)



- **Important property**
  - If X is $N\left(\mu, \sigma\right)$, then Y = aX + b is $N(a\mu + b, a\sigma)$
  - Any linear transformation of a Gaussian RV produces another Gaussian RV

Important PDFs

Gaussian Distribution

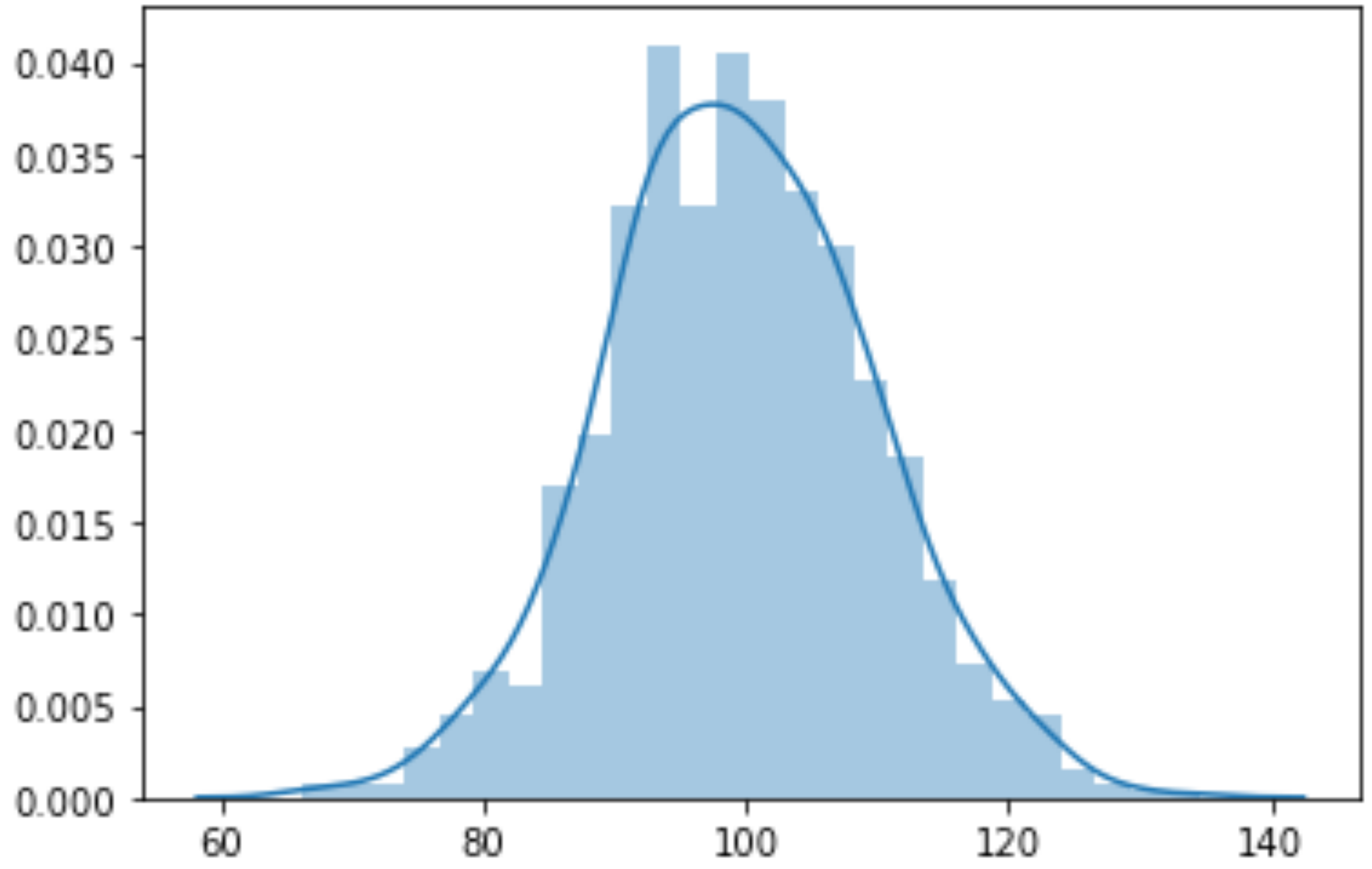
- **<u>Normal/Gaussian Distribution</u>**
  - X ~ $N(\mu, \sigma)$
  - X has normal distribution with parameters μ and σ
  - X has PDF f(x) with

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

  - μ is the mean of the distribution
  - σ is the standard deviation of the distribution (more on these later)

- **Important property**
  - If X is $N\left(\mu, \sigma\right)$, then Y = aX + b is $N(a\mu + b, a\sigma)$

```
mu, sigma = 100, 10 # mean and standard deviation
s = np.random.normal(mu, sigma, 1000)
sns.distplot(s, hist=True, kde=True)
plt.show()
```



CDF for Continuous RVs

- The CDF for continuous RVs is directly related to the PDF

$$F(x) = P(X \leq x) = P(-\infty < X \leq x)$$
$$= \int_{-\infty}^{x} f(t)dt$$

- Additionally

$$F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx$$

- Therefore, given F(x) can compute f(x) $f(x) = \frac{d}{dx}F(x)$

Standard Normal RV

- **The standard normal random variable is *Z* ~ $N(0,1)$.** Zero mean and unit variance
- **Standard Normal CDF:**
  - We can use this CDF to find probabilities of non-Standard Normal CRVs
  - Usually tables exist for CDF values for different values of z

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

- If *X* ~ $N(\mu, \sigma)$, then its CDF is: $$F_X(x) = \Phi(\frac{x-\mu}{\sigma})$$

- The probability that X is in the interval (a,b] is: $$P[a < X \leq b] = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

- **Next Class**

No Class on Tuesday (IU Wellness Day)

Next Thursday: Continue Probability Review

Start Homework #1