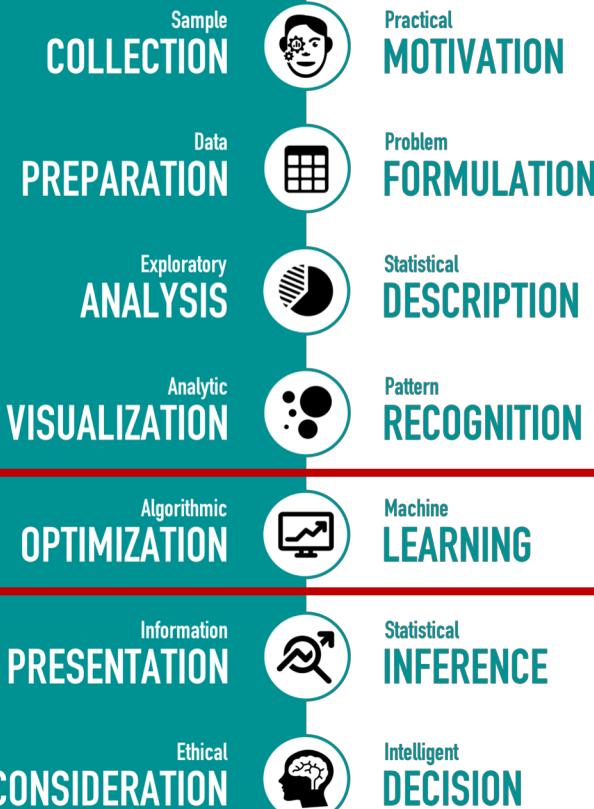


Basics of Machine Learning

Sourav SEN GUPTA
Lecturer, SCSE, NTU





Data Science Machine Learning

What do we know about the Problem?
Does the Data provide us with Hints?

Can we Learn from the Hints in Data?
Can we Test how well we Learned?

How to optimally learn from the Data?



Data Science

The Pokemon Dataset

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
430	Honchkrow	Dark	Flying	505	100	125	52	105	52	71	4	False
338	Solrock	Rock	Psychic	440	70	95	85	55	65	70	3	False
32	Nidoran♂	Poison	NaN	273	46	57	40	40	40	50	1	False
442	Spiritomb	Ghost	Dark	485	50	92	108	92	108	35	4	False
480	Uxie	Psychic	NaN	580	75	75	130	75	130	95	4	True
536	Palpitoad	Water	Ground	384	75	65	55	65	55	69	5	False
360	Wynaut	Psychic	NaN	260	95	23	48	23	48	23	3	False
478	Froslass	Ice	Ghost	480	70	80	70	80	70	110	4	False
76	Golem	Rock	Ground	495	80	120	130	55	65	45	1	False
177	Natu	Psychic	Flying	320	40	50	45	70	45	70	2	False

Source : Kaggle Datasets | [Pokemon with stats](#) by Alberto Barradas | <https://www.kaggle.com/abcsds/pokemon>

Data Science Common Problem



Prediction : Numeric

How much?
How many?

What are the overall Strengths of
the Pokemons shown beside?
Can you *estimate* Total, at least?



Data Science Machine Learning

Prediction : Numeric Regression

Model : Total = f (Variables)

Given	Some Pokémons to Train
Learn	The Model for Total Points
Predict	Estimate Total for others



Data Science Machine Learning

Prediction : Numeric Regression

HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Total
100	125	52	105	52	71	505
70	95	85	55	65	70	440
46	57	40	40	40	50	273
50	92	108	92	108	35	485
75	75	130	75	130	95	580
75	65	55	65	55	69	384
95	23	48	23	48	23	260
70	80	70	80	70	110	480
80	120	130	55	65	45	495
40	50	45	70	45	70	320

Labels

$$\text{Model : Total} = f(\text{Variables})$$

- Given Some Pokemons to Train
Learn The Model for Total Points
Predict Estimate Total for others



Data Science Common Problem

Prediction : Classes

Is it type A
or type B?

What do you think of the Pokemons
shown here; are they Legendary?
Can you *predict* the probability?



Data Science Machine Learning

Prediction : Classes Classification

Model : \mathcal{P} (Legend) = f (Variables)

Given Some Pokemons to Train
Learn The Model for Legendary
Predict Determine Class for others



Data Science Machine Learning

Prediction : Classes

Classification

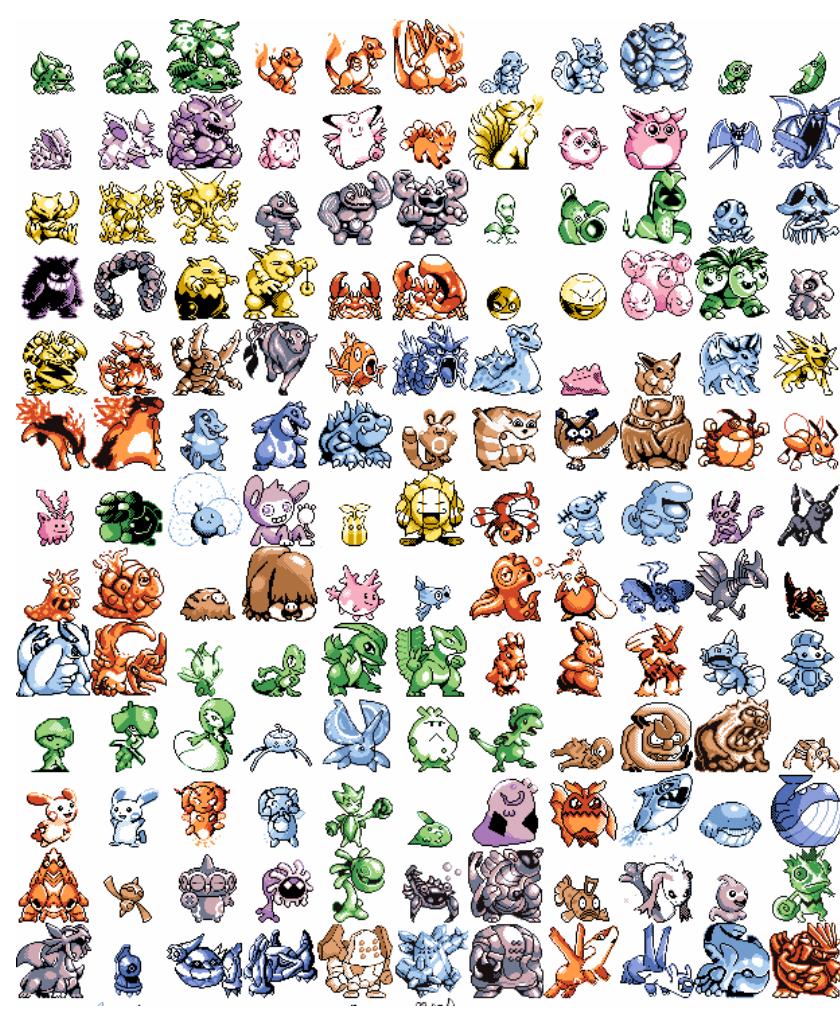
HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	
100	125	52	105	52	71	
70	95	85	55	65	70	
46	57	40	40	40	50	
50	92	108	92	108	35	
75	75	130	75	130	95	
75	65	55	65	55	69	
95	23	48	23	48	23	
70	80	70	80	70	110	
80	120	130	55	65	45	
40	50	45	70	45	70	

Labels →

Legendary
False
False
False
False
True
False

$$\text{Model} : \mathcal{P}(\text{Legend}) = f(\text{Variables})$$

- Given Some Pokemons to Train
Learn The Model for Legendary
Predict Determine Class for others

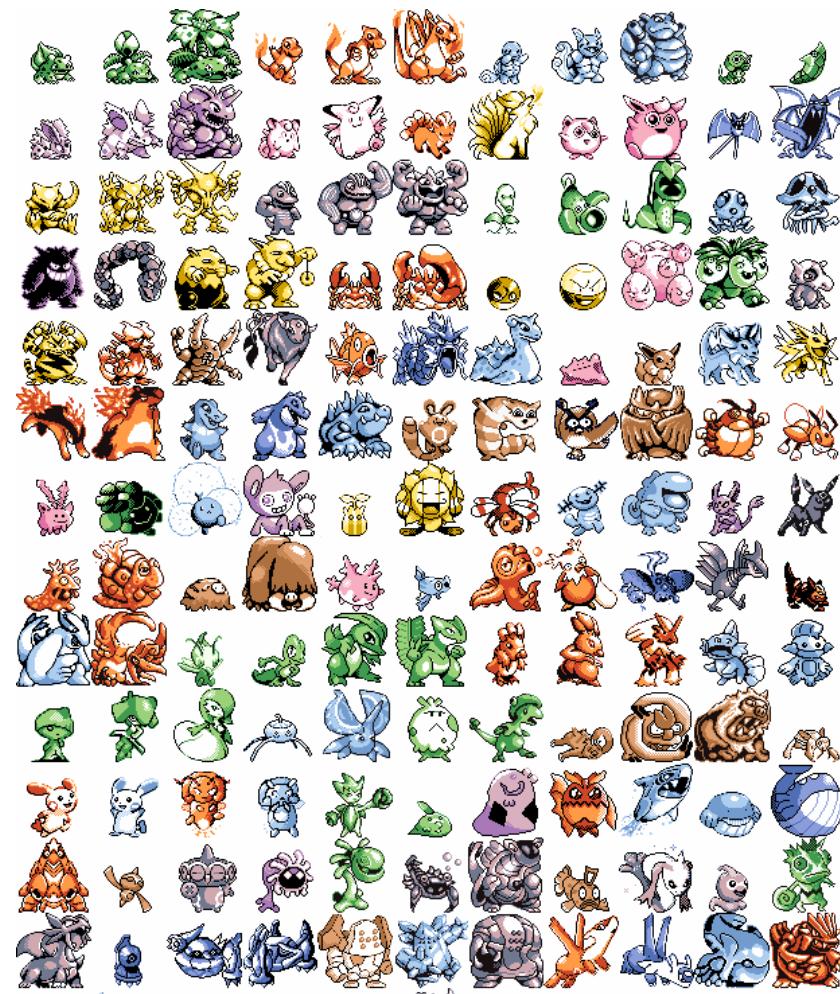


Data Science Common Problem

Detection : Structure

How is this organized?

Is there any structure apparent
within the Pokemons shown here?
Can you *group* them in any way?



Data Science Machine Learning

Detection : Structure Clustering

Grouping depends on “Distance”

- | | |
|---------|------------------------------|
| Given | Distance on all Pokemons |
| Find | Optimal Groups in the Data |
| Justify | Interpretation of the Groups |



Data Science Machine Learning

Detection : Structure Clustering

Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
505	100	125	52	105	52	71
440	70	95	85	55	65	70
273	46	57	40	40	40	50
485	50	92	108	92	108	35
580	75	75	130	75	130	95
384	75	65	55	65	55	69
260	95	23	48	23	48	23
480	70	80	70	80	70	110
495	80	120	130	55	65	45
320	40	50	45	70	45	70

Grouping depends on “Distance”

- | | |
|---------|------------------------------|
| Given | Distance on all Pokemons |
| Find | Optimal Groups in the Data |
| Justify | Interpretation of the Groups |



Data Science Common Problem

Detection : Anomaly

**Is it weird
behavior?**

Are these Pokemons more or less like the others, or are they Weird?
Can you *identify* the difference?

HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
100	125	52	105	52	71
70	95	85	55	65	70
46	57	40	40	40	50
50	92	108	92	108	35
75	75	130	75	130	95
75	65	55	65	55	69
95	23	48	23	48	23
70	80	70	80	70	110
80	120	130	55	65	45
40	50	45	70	45	70



Total	Legendary
505	False
440	False
273	False
485	False
580	True
384	False
260	False
480	False
495	False
320	False



Data Science Machine Learning

Supervised Learning

Regression

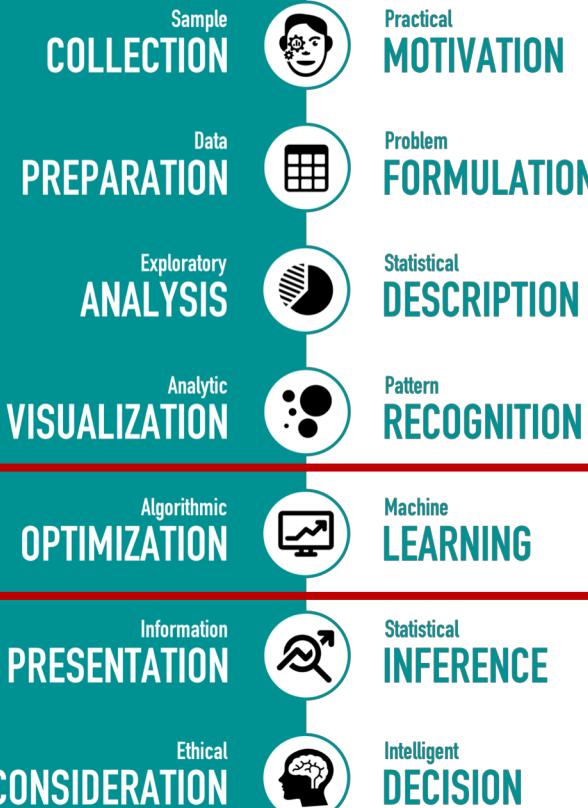
Classification

Un-Supervised Learning

Clustering

Anomaly Detection

Photos : Pokedex Website | <https://www.pokemon.com/us/pokedex/>



Data Science Pipeline Machine Learning

How to learn from the acquired Data?

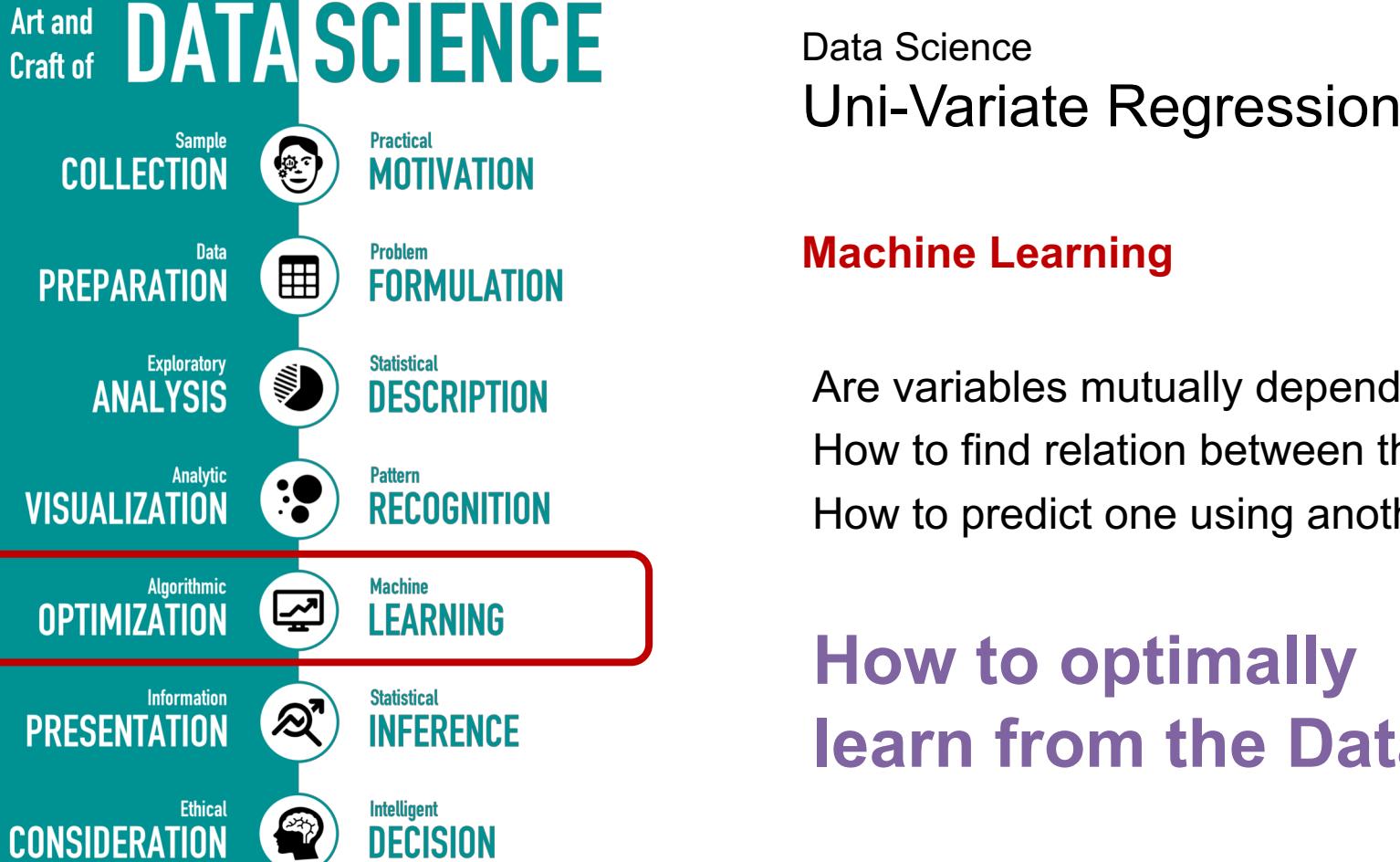
- How to model, train and predict?
- How to identify the structures?

How to optimally learn from the Data?

Uni-Variate Linear Regression

Sourav SEN GUPTA
Lecturer, SCSE, NTU







Data Science

The Pokemon Dataset

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
430	Honchkrow	Dark	Flying	505	100	125	52	105	52	71	4	False
338	Solrock	Rock	Psychic	440	70	95	85	55	65	70	3	False
32	Nidoran♂	Poison	NaN	273	46	57	40	40	40	50	1	False
442	Spiritomb	Ghost	Dark	485	50	92	108	92	108	35	4	False
480	Uxie	Psychic	NaN	580	75	75	130	75	130	95	4	True
536	Palpitoad	Water	Ground	384	75	65	55	65	55	69	5	False
360	Wynaut	Psychic	NaN	260	95	23	48	23	48	23	3	False
478	Froslass	Ice	Ghost	480	70	80	70	80	70	110	4	False
76	Golem	Rock	Ground	495	80	120	130	55	65	45	1	False
177	Natu	Psychic	Flying	320	40	50	45	70	45	70	2	False

Source : Kaggle Datasets | [Pokemon with stats](#) by Alberto Barradas | <https://www.kaggle.com/abcsds/pokemon>

Data Science

Bi-Variate Exploration

Statistical Summary

	HP		Total
count	800.000000	count	800.000000
mean	69.258750	mean	435.10250
std	25.534669	std	119.96304
min	1.000000	min	180.000000
25%	50.000000	25%	330.000000
50%	65.000000	50%	450.000000
75%	80.000000	75%	515.000000
max	255.000000	max	780.000000

HP Hit Points of a Pokemon
Total Total Points of a Pokemon

Machine Learning Questions

- What is the mutual relationship?
- Can we predict Total given HP?

	HP		Total
count	600.000000	count	600.000000
mean	69.026667	mean	432.715000
std	26.611267	std	122.365283
min	1.000000	min	180.000000
25%	50.000000	25%	325.000000
50%	65.000000	50%	440.000000
75%	80.000000	75%	515.000000
max	255.000000	max	780.000000

Train Dataset
75% of the Data

Data Science

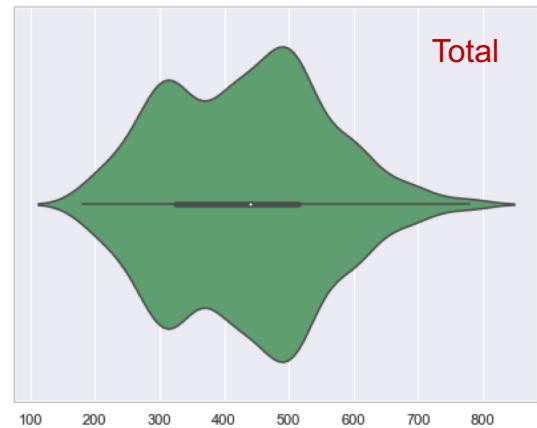
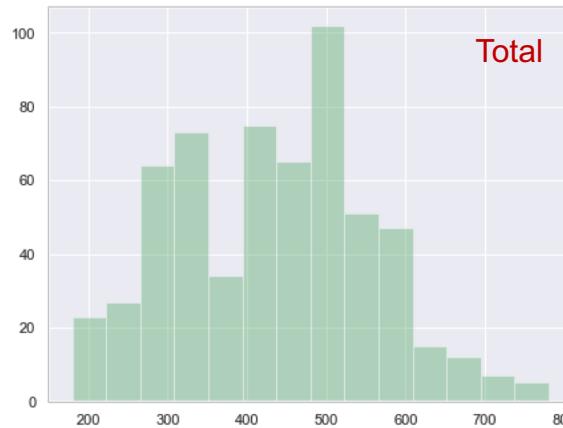
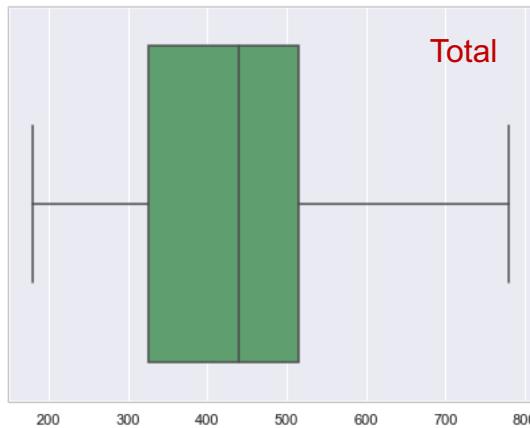
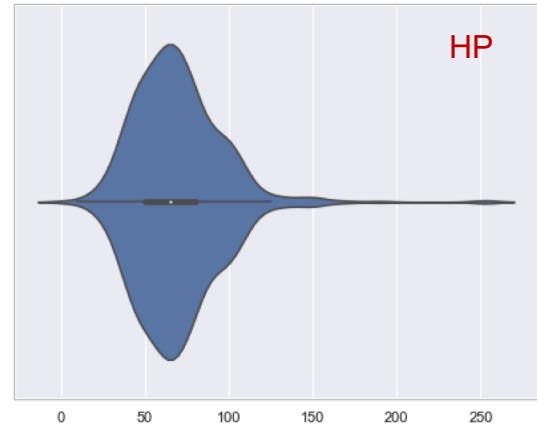
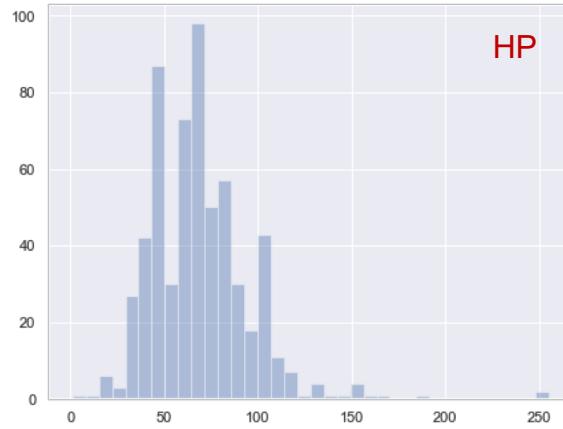
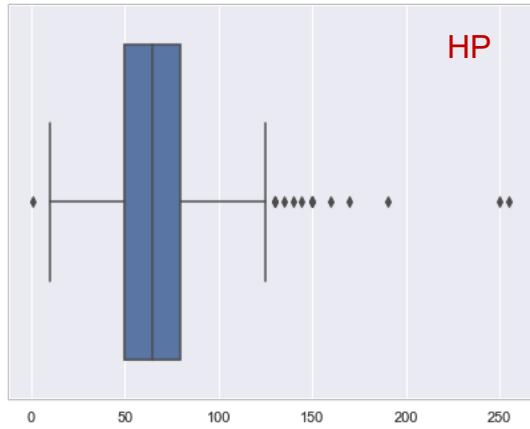
Bi-Variate Exploration

Split the Dataset

- Train** Used to train the model
- Test** Used to test the model

The Objectives

- Learn the relationship from Train
- Try to predict the Total on Test



	Total	HP
Total	1.000000	0.590826
HP	0.590826	1.000000



Data Science Bi-Variate Exploration

Correlation Matrix and Plot

Natural Intuition

Dependence of HP and Attack

Statistical Intuition

No Dependence

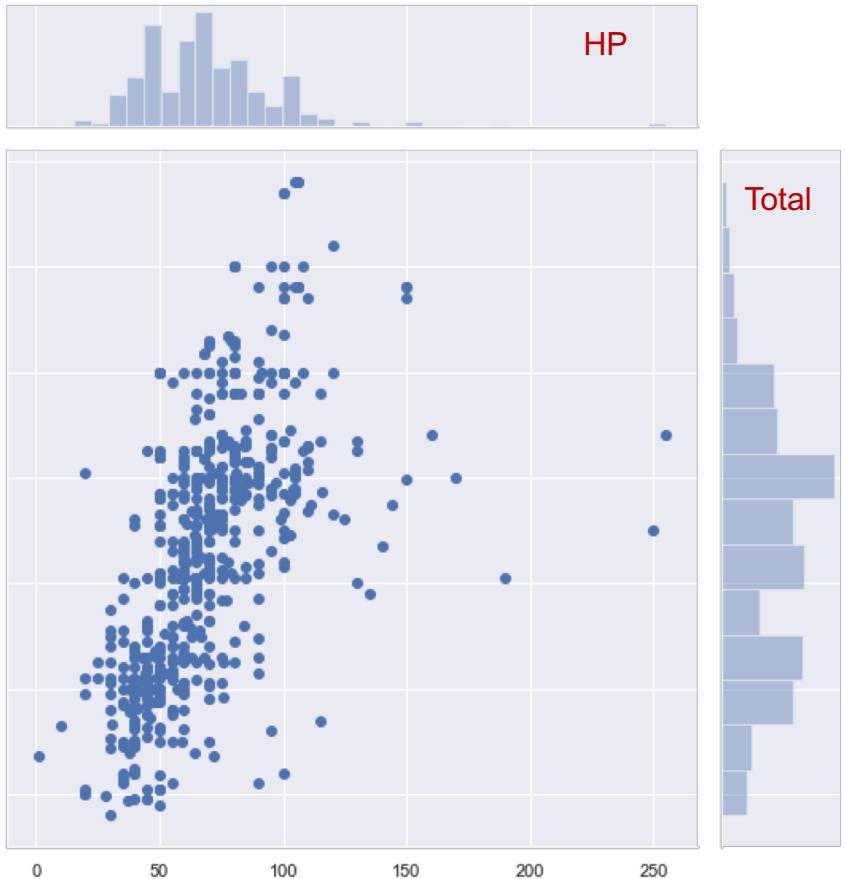
Corr = 0

Perfect Positive

Corr = + 1

Perfect Negative

Corr = - 1



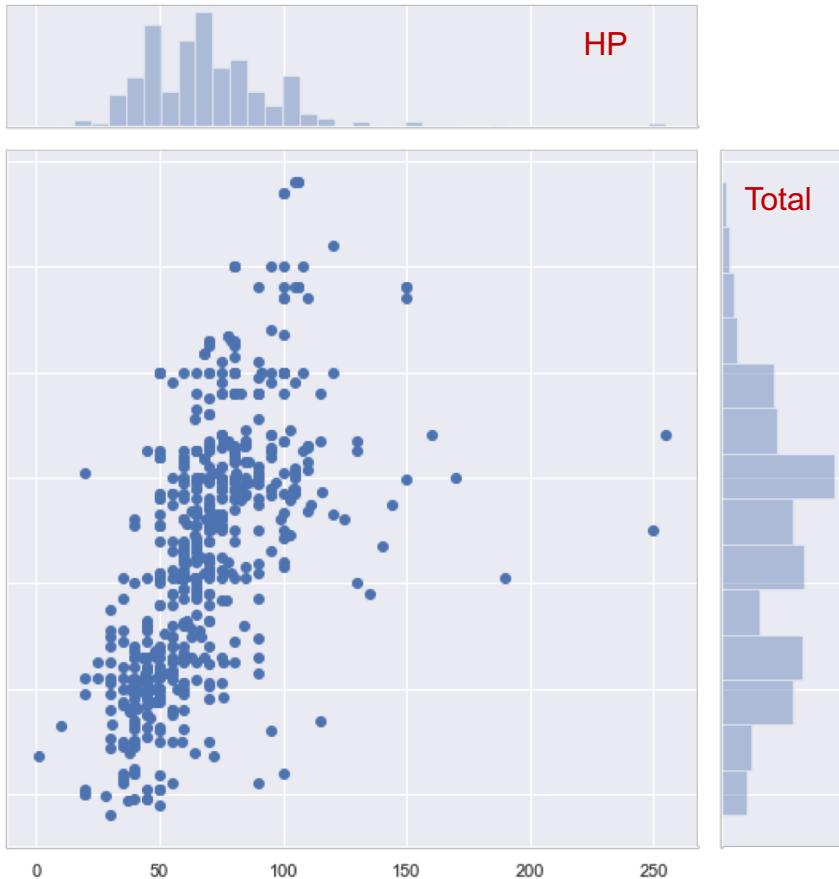
Data Science Bi-Variate Exploration

Statistical Relation

- Total** Plotted along Y axis
- HP** Plotted along X axis

The Objectives

- Learn the relationship from Train
- Try to predict the Total on Test



Data Science

Uni-Variate Regression

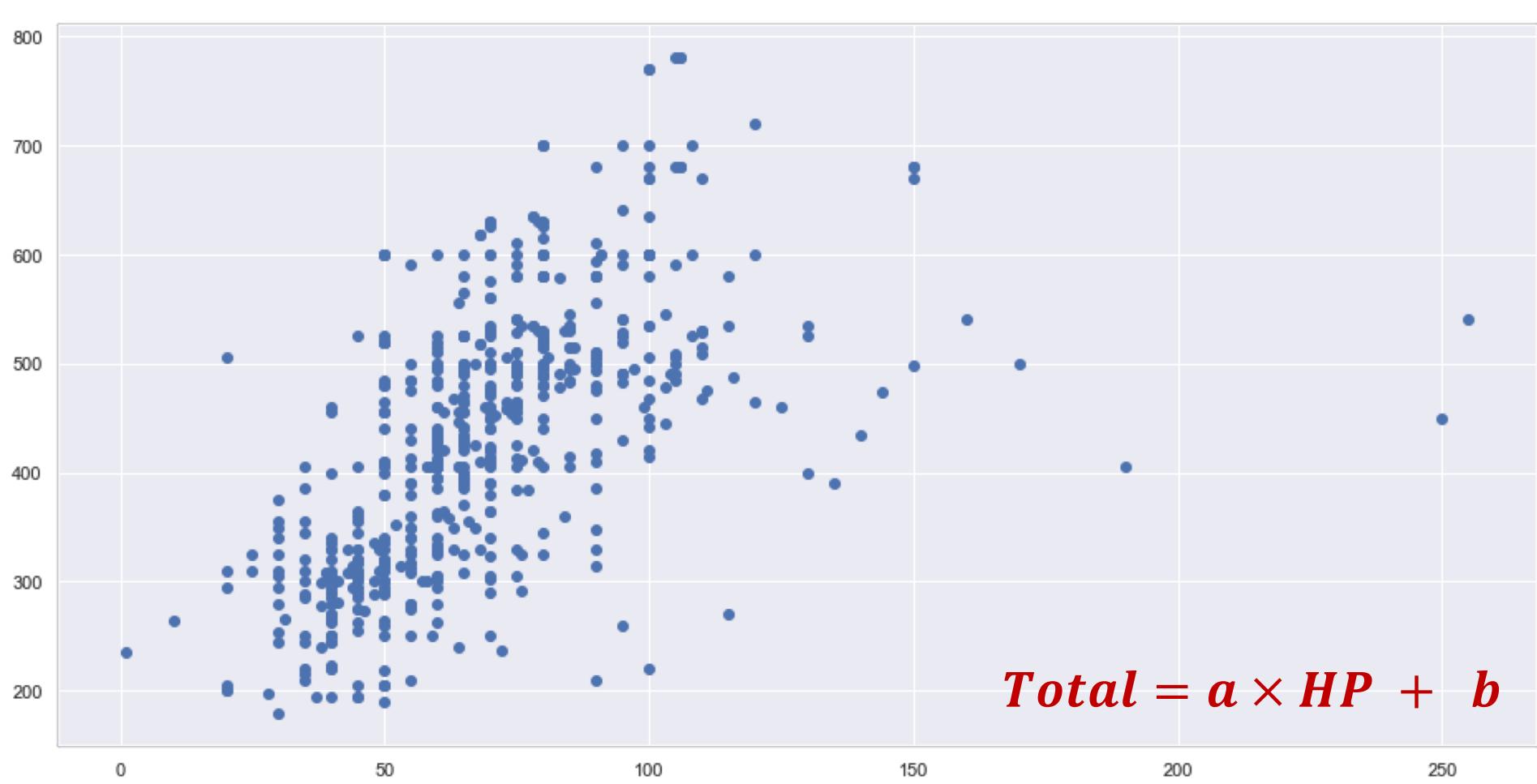
Statistical Modeling

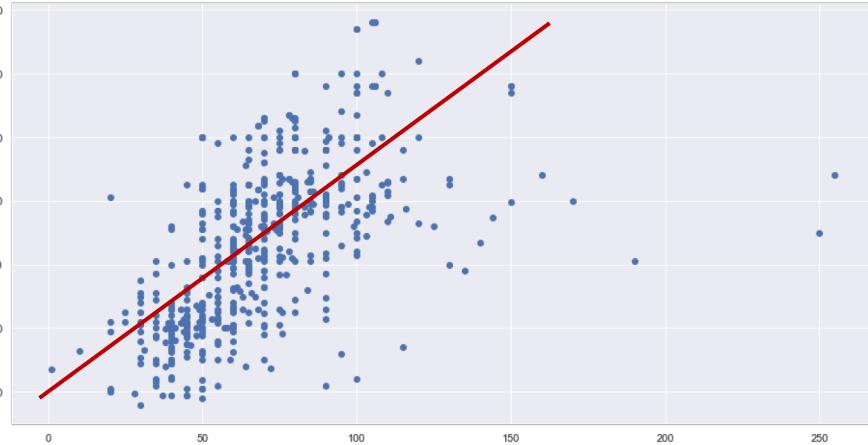
Hypothesize a Linear Model

$$\text{Total} = a \times HP + b + \epsilon$$

The Objectives

- Learn the parameters of the Model
- Try to use the Model for Prediction





Data Science

Uni-Variate Regression

Algorithmic Optimization

Hypothesize a Linear Model

$$\text{Total} = a \times HP + b + \epsilon$$

Steps in Linear Regression

Guess parameters a and b in the model

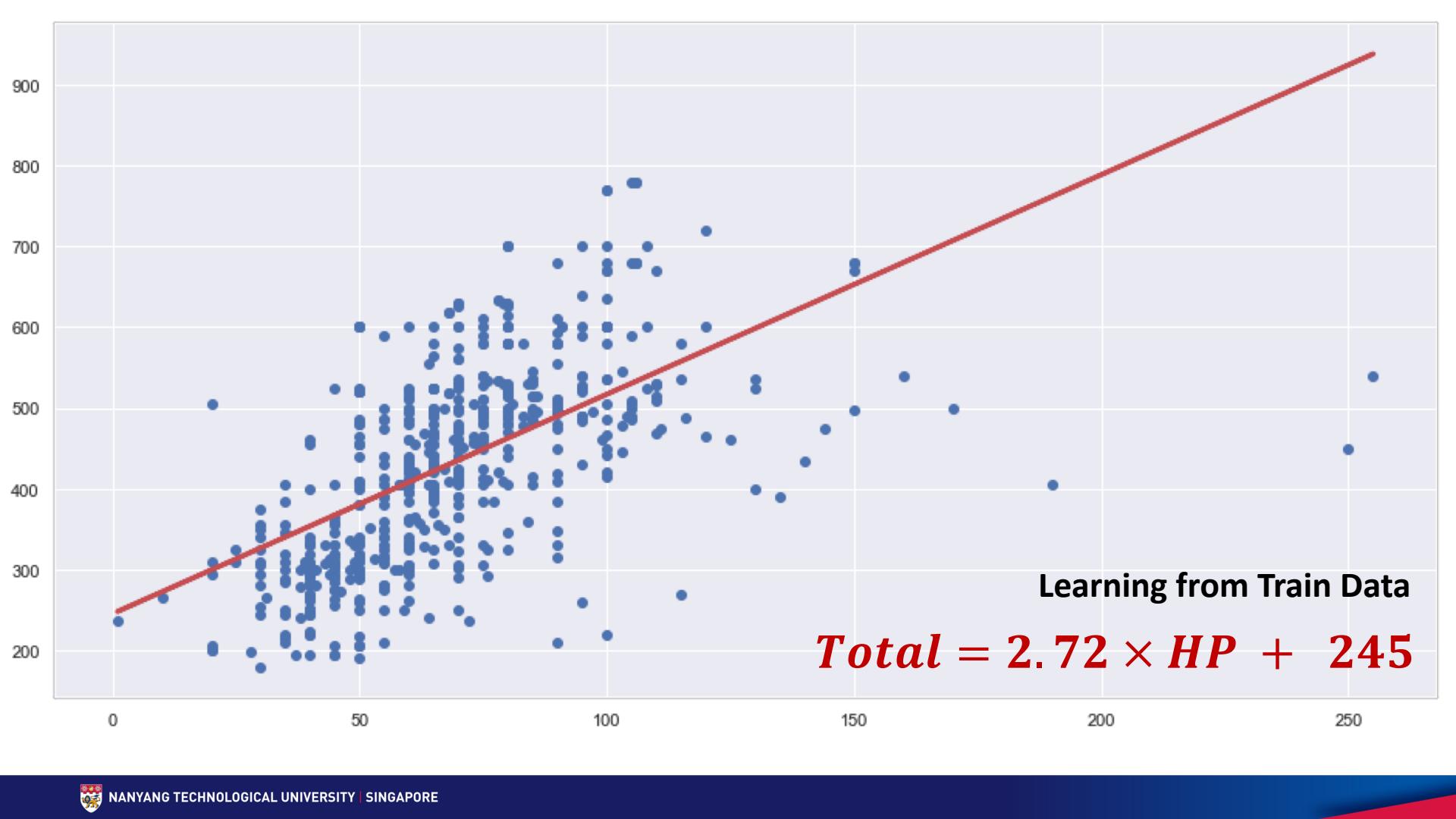
Predict the values of Total in Train Data

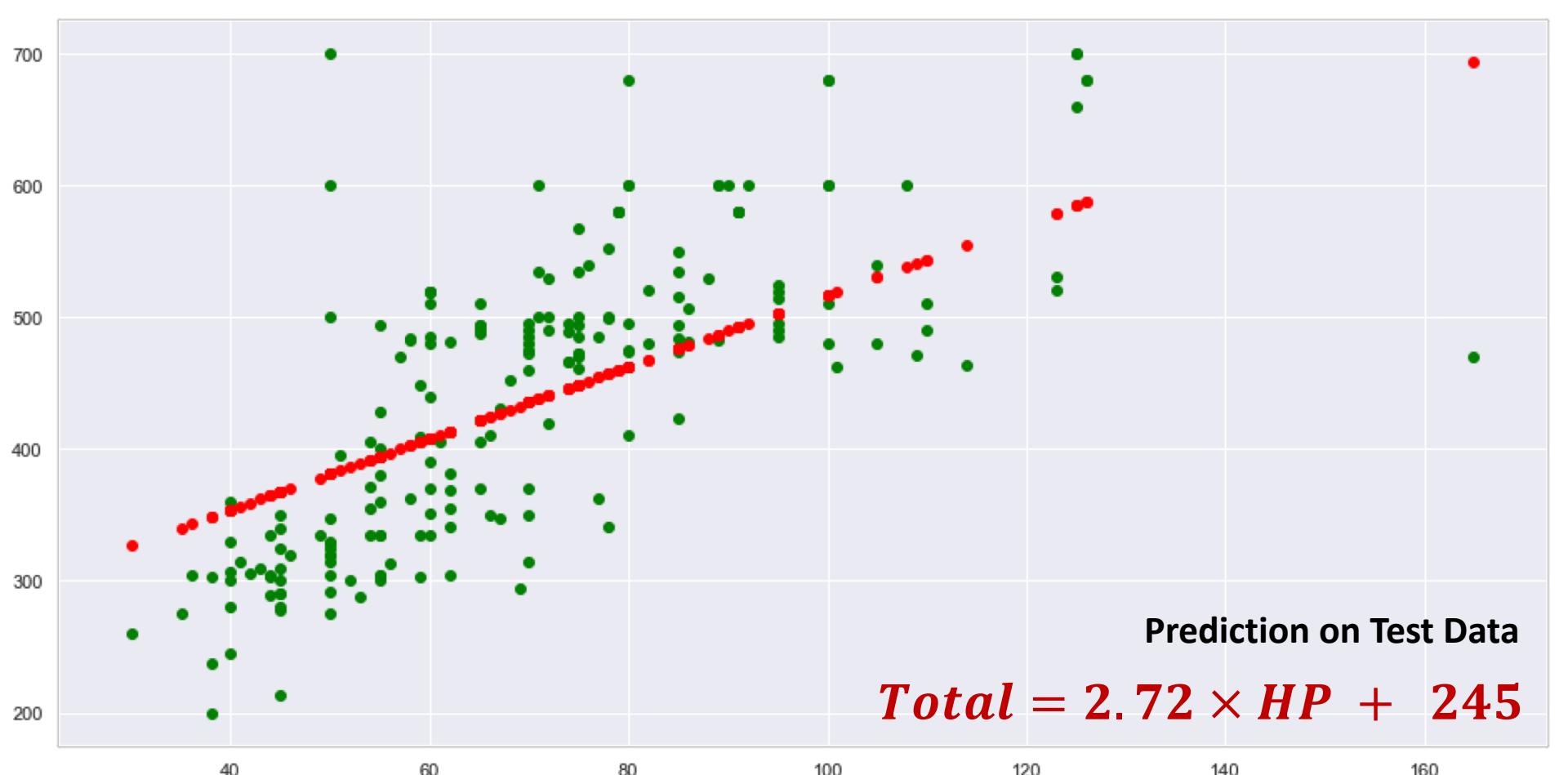
Calculate the Errors compared to actual

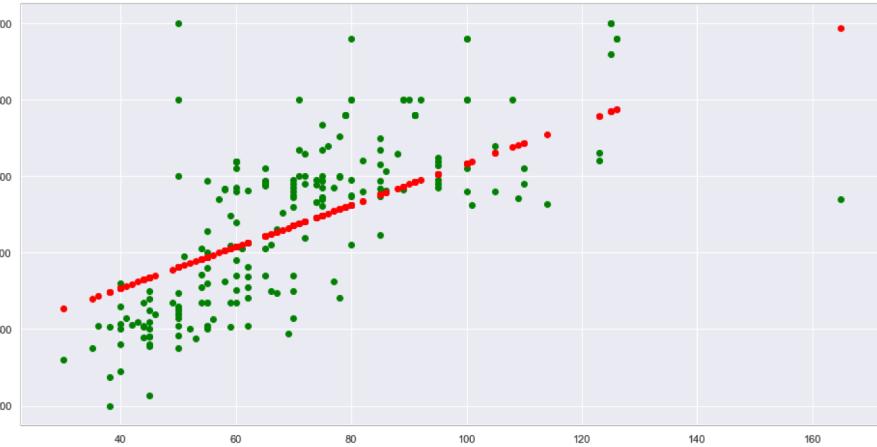
Tune the parameters to minimize Errors

Cost Function to Minimize

$$J(a, b) = \sum (\text{Total} - a \times HP - b)^2$$







Data Science

Uni-Variate Regression

Goodness of Fit of the Model

Hypothesis : The Linear Model

$$Total = a \times HP + b + \epsilon$$

Explained Variance (R^2)

$$R^2 = 1 - \frac{\sum(Total - a \times HP - b)^2}{\sum(Total - \bar{Total})^2}$$

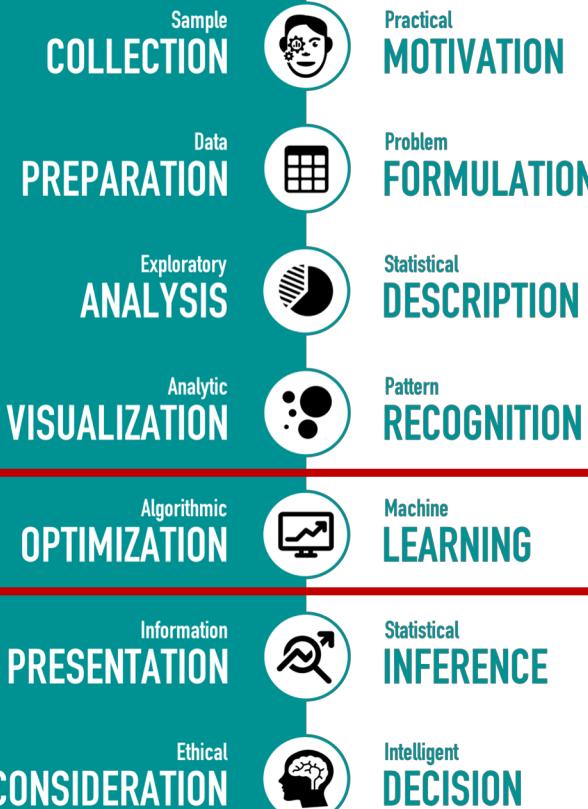
The higher the R^2 the better the Model

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum (Total - a \times HP - b)^2$$

The lower the MSE the better the Model





Data Science Pipeline Machine Learning

How to learn from the acquired Data?
How to model the acquired Data?
How to predict on new Data?

How to optimally
learn from the Data?