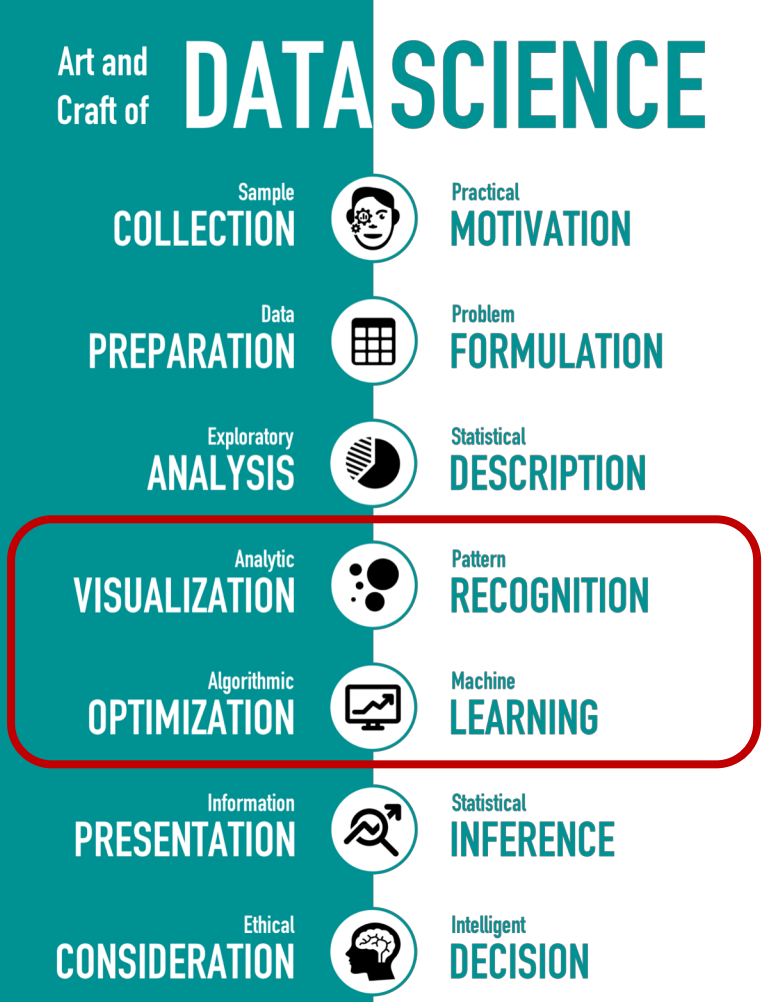
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Clustering Patterns

Sourav SEN GUPTA
Lecturer, SCSE, NTU



Data Science

Clustering Patterns

Pattern Recognition

Is there a pattern in the acquired data?

How to learn the underlying pattern?

How to exploit the pattern in data?

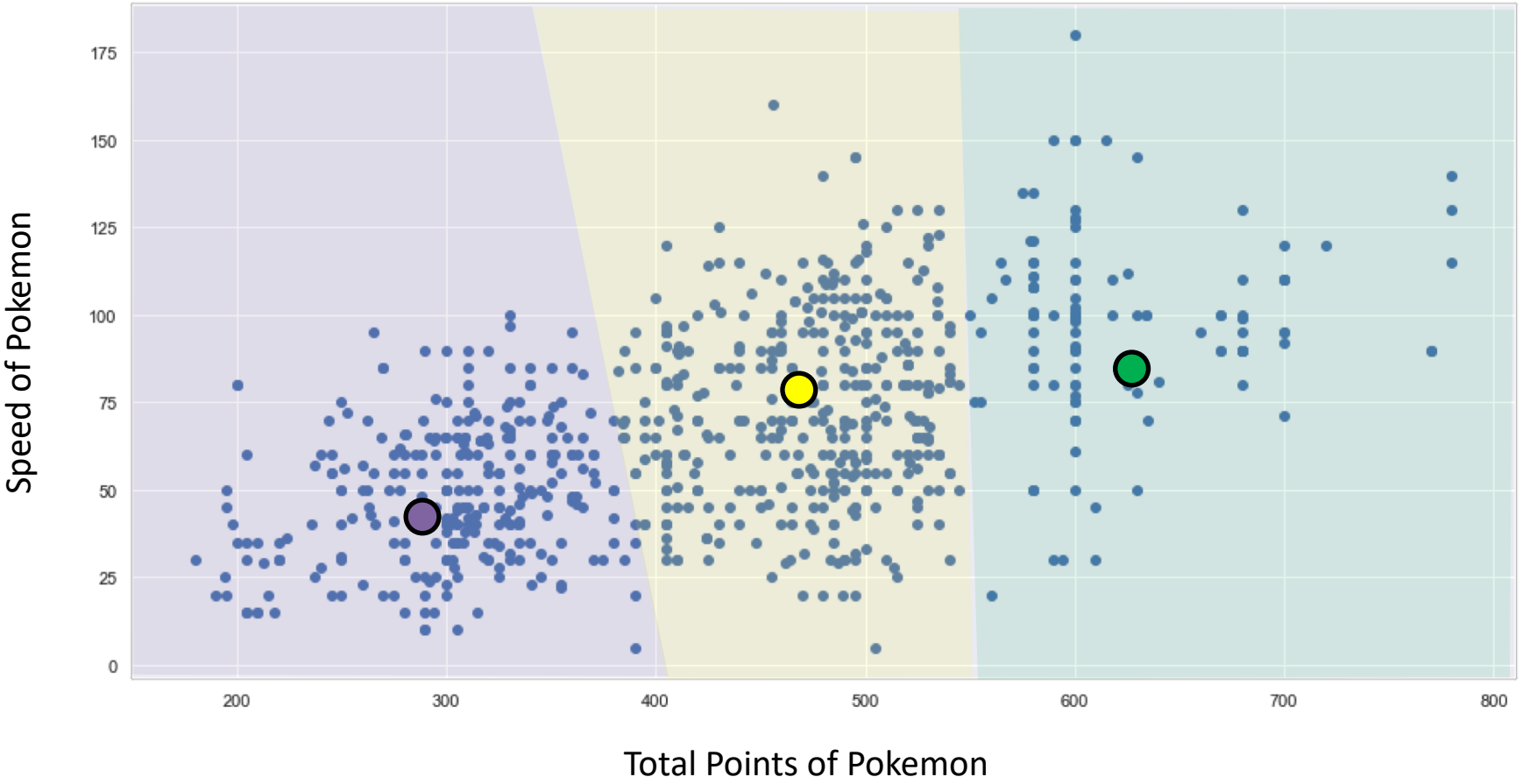
How to optimally learn from the Data?

Data Science

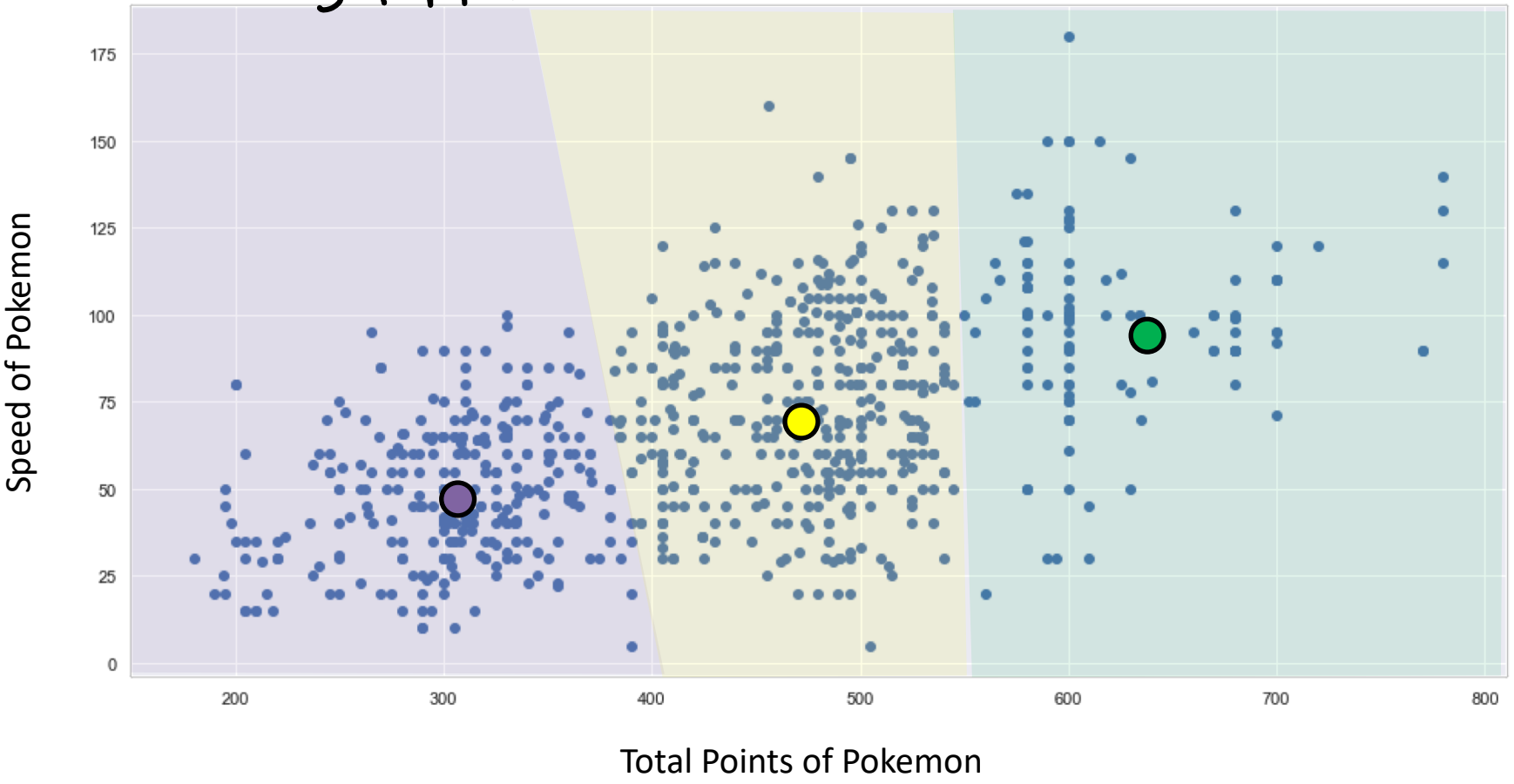
The Pokemon Dataset

#	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

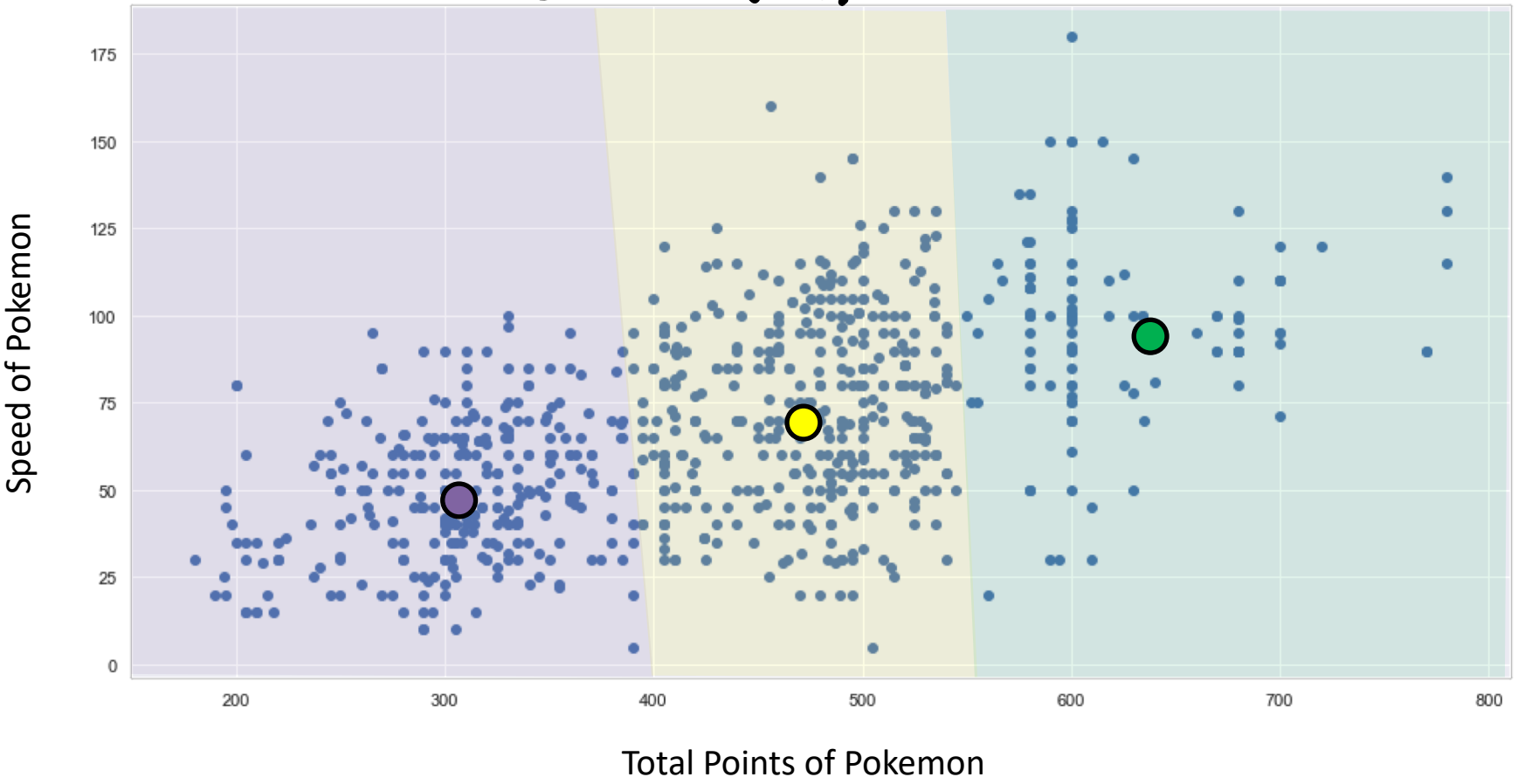
Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

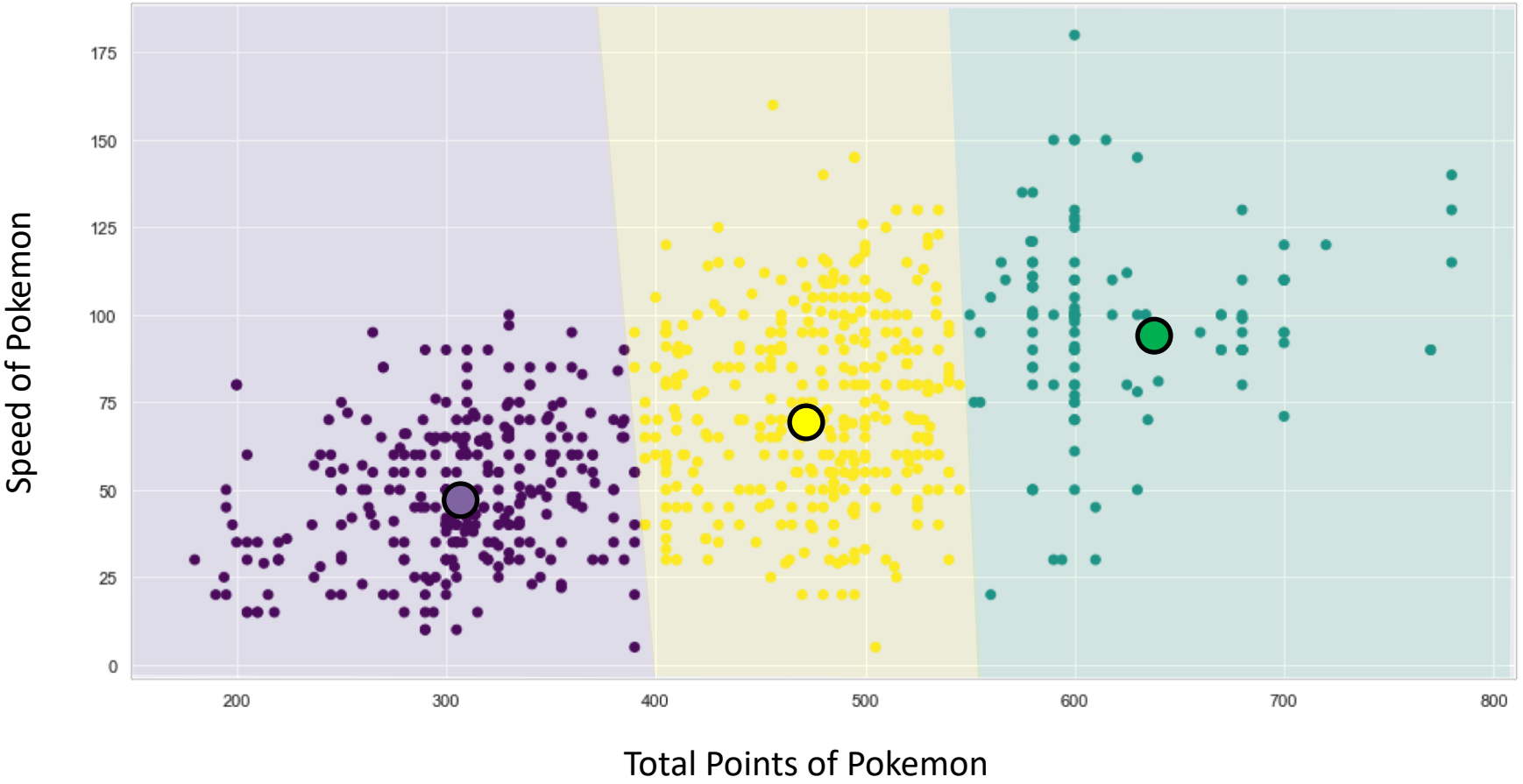

Speed of Pokemon
175
150
125
100
75
50
25
0
200
300
400
500
600
700
800
Total Points of Pokemon

Estimation


Speed of Pokemon
175
150
125
100
75
50
25
0
200
300
400
500
600
700
800
Total Points of Pokemon

with centroid




Speed of Pokemon
175
150
125
100
75
50
25
0
200
300
400
500
600
700
800
Total Points of Pokemon

Data Science

Clustering Patterns

K-Means Clustering

Total	Total Points of Pokemon
Speed	Speed of Pokemon

Choose ***K*** – the potential number of clusters — parameter

Choose ***K*** cluster centroids from the dataset — initialization

for *each point* in the dataset
    Re-Label according to nearest centroid

for *each cluster* of data points
    Re-Compute the centroid of the cluster

iteration

Machine Learning Questions

- How many Clusters are “visible”?
- Can we identify those Clusters?
- What do the Clusters signify?

Data Science

Clustering Patterns

K-Means Clustering

Total Total Points of Pokemon

Speed Speed of Pokemon

Features	Total	Speed
Cluster 0:	305.65	49.36
Cluster 1:	622.57	97.08
Cluster 2:	474.27	73.55

Within Cluster Sum of Squares = 2118651

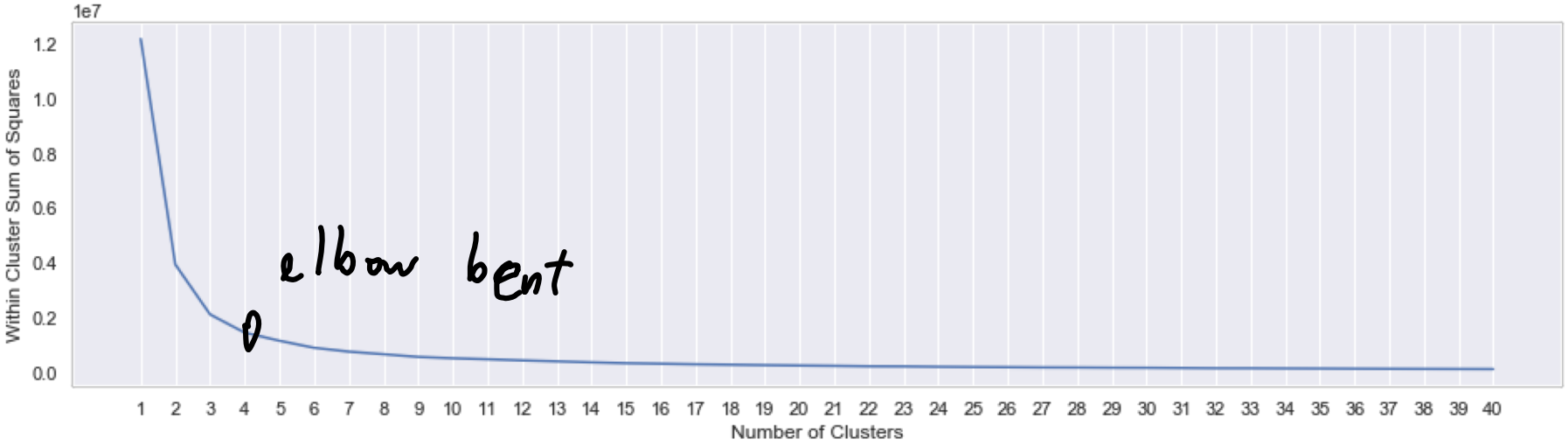
Optimization Questions

- What is ”optimal” Cluster count?
- Can we justify the Cluster count?
- What is a nice Clustering metric?

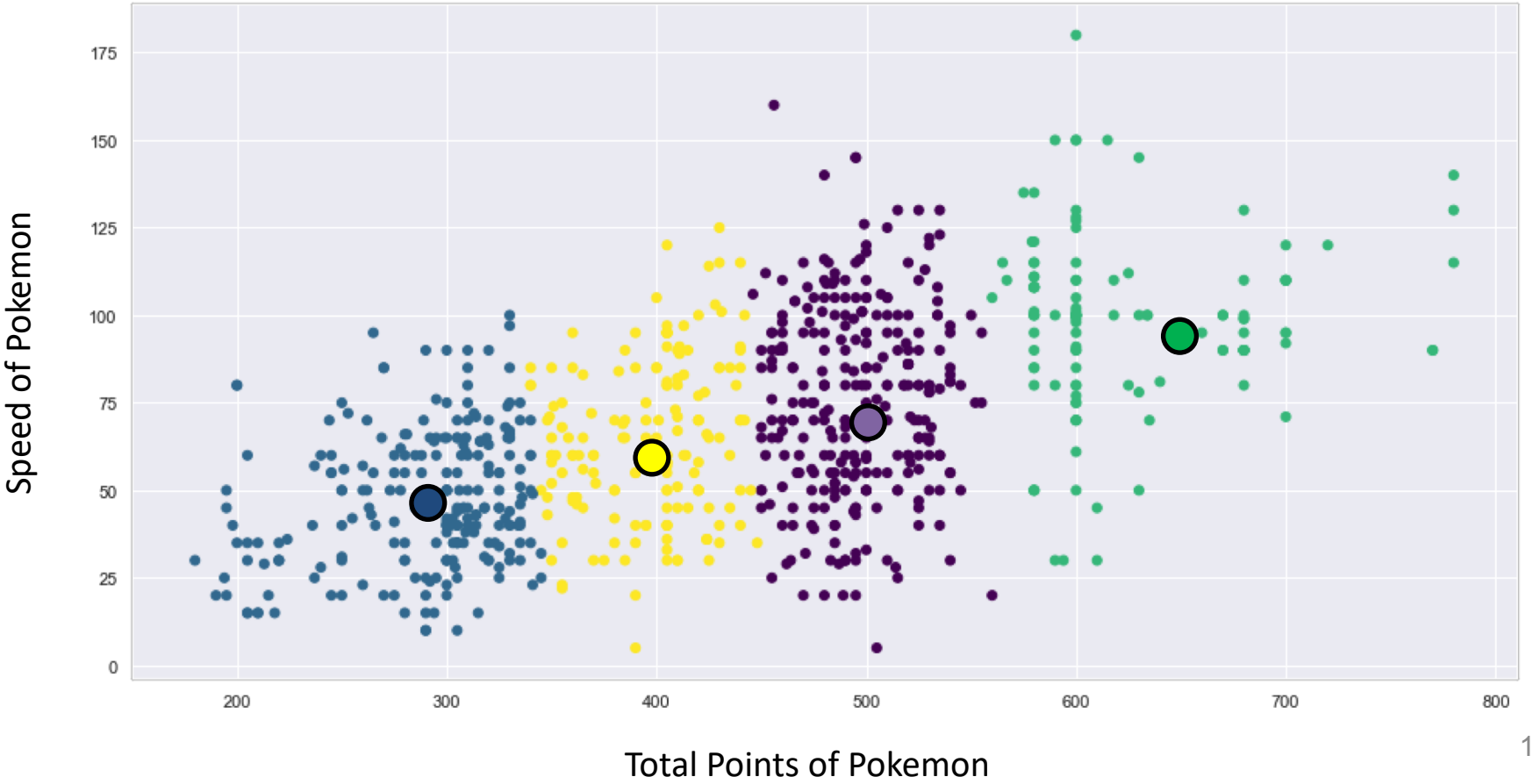
Data Science

Clustering Patterns

angle plot

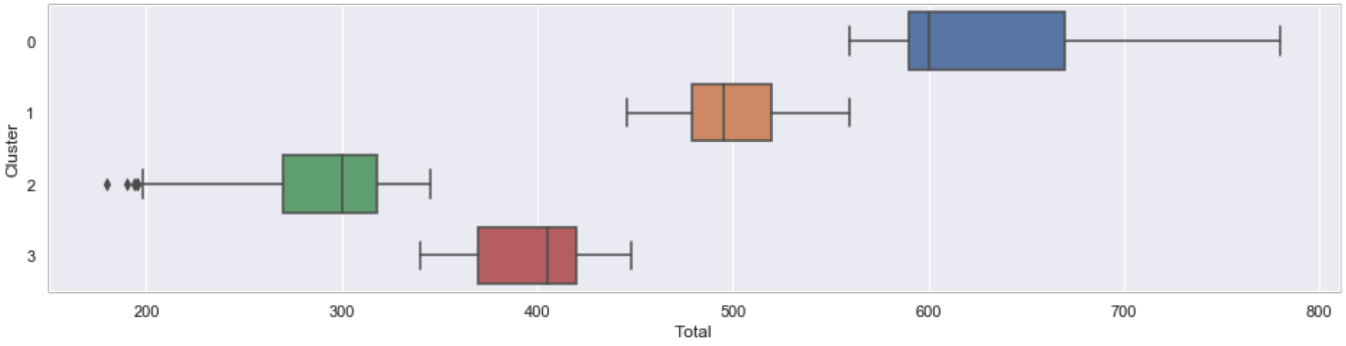


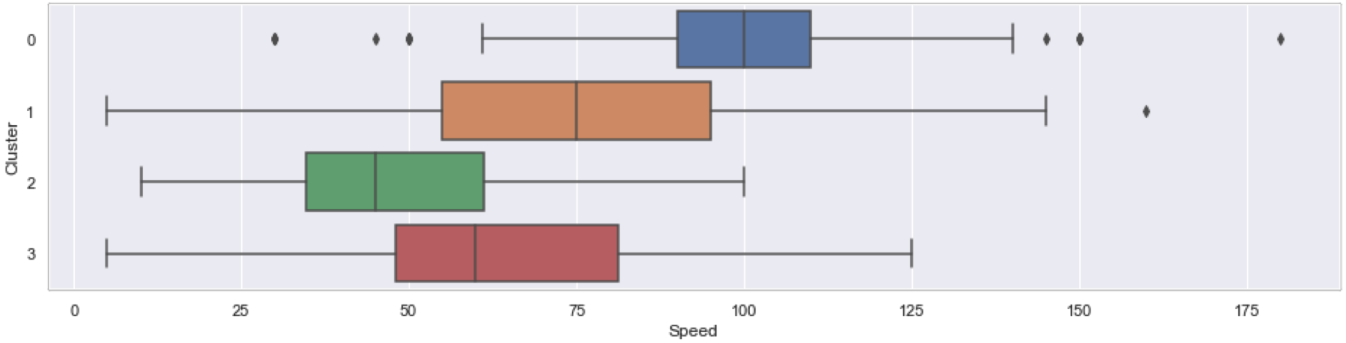
Guess for Optimal Number of Clusters = 4

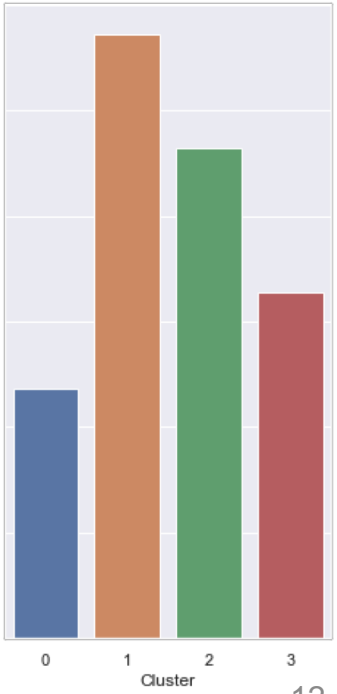

Speed of Pokemon
175
150
125
100
75
50
25
0
200
300
400
500
600
700
800
Total Points of Pokemon

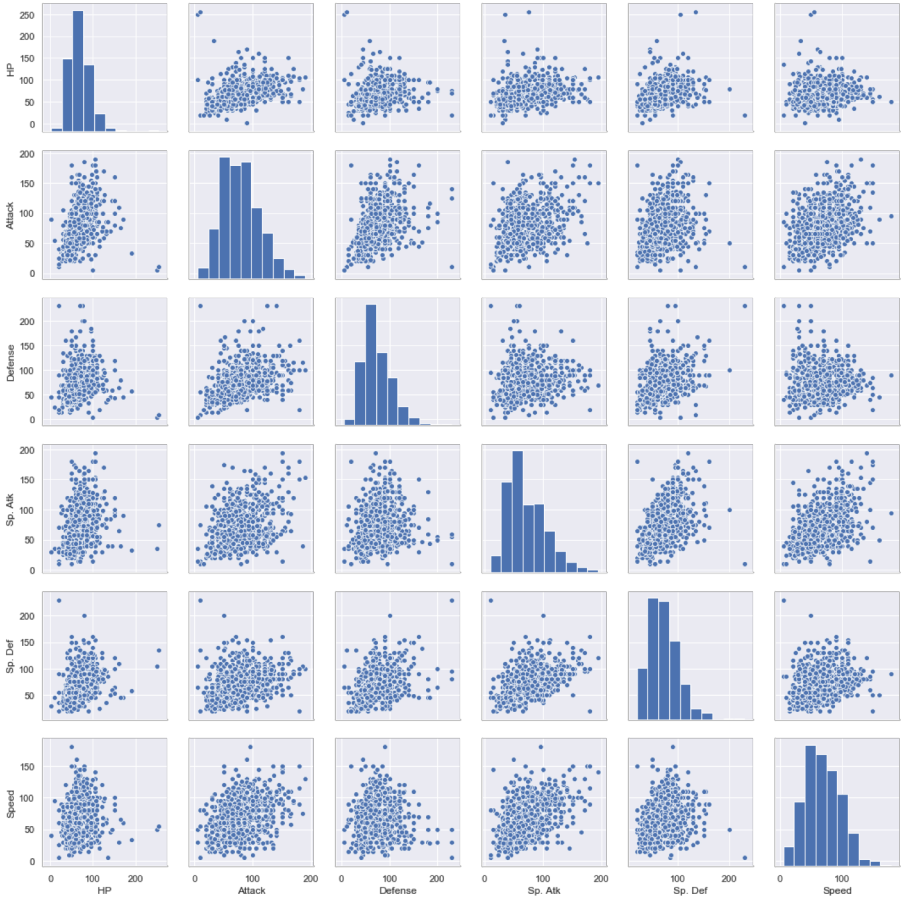
Data Science

Clustering Patterns









Data Science

Clustering Patterns

K-Means Clustering

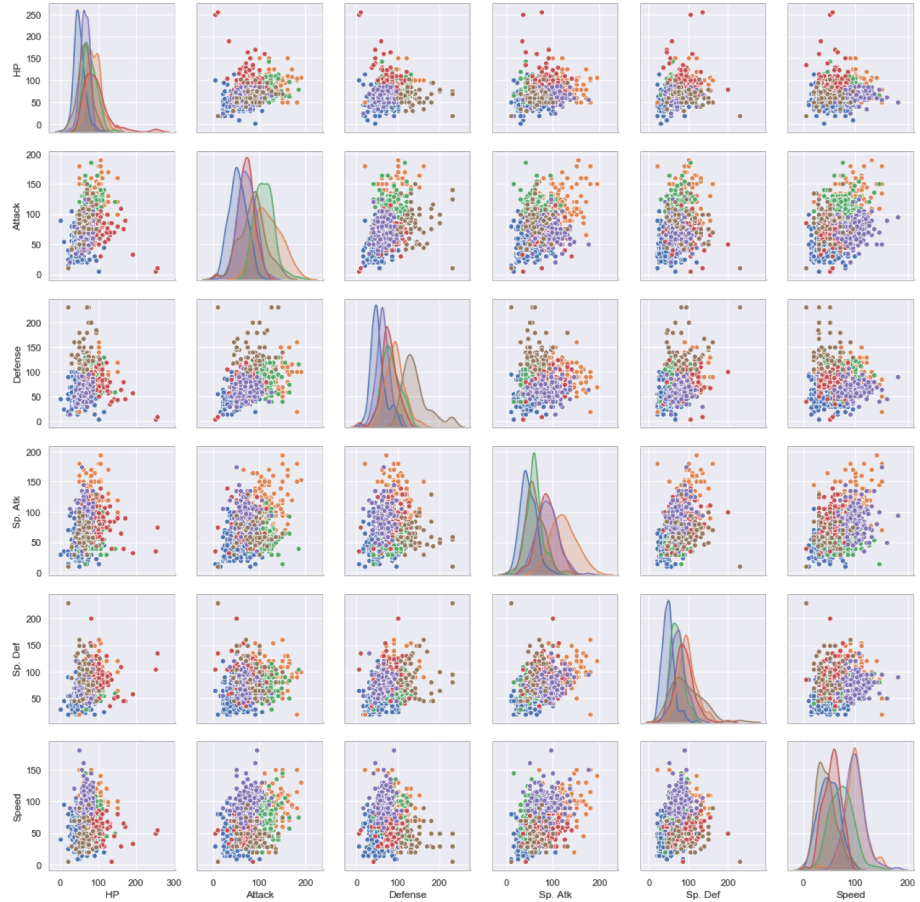
HP	Hit Points
Attack	Attack Points
Defense	Defense Points
Sp. Atk	Special Attack
Sp. Def	Special Defense
Speed	Speed of Pokemon

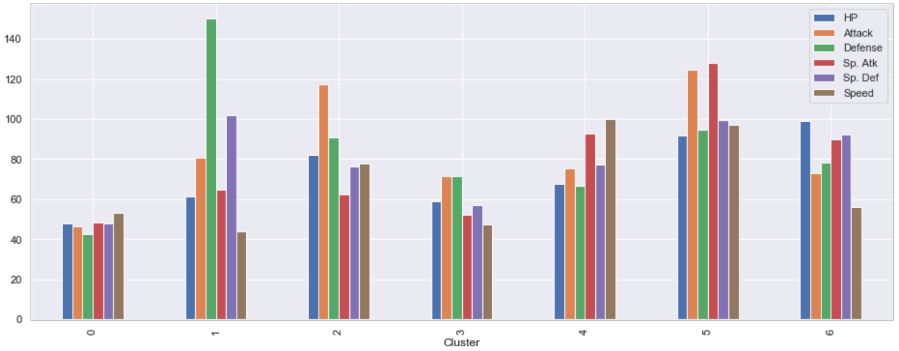
$$\text{Centroid} = (\bar{x}_1, \bar{x}_2 \ldots \bar{x}_6)$$

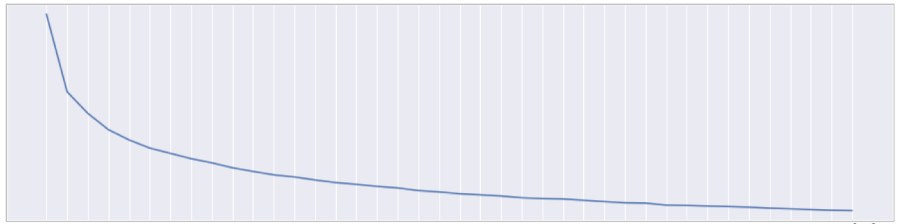
Data Science

Clustering Patterns

K-Means Clustering







Data Science Pipeline

Pattern Recognition

How to learn from the acquired Data?

How to find pattern in acquired Data?

How to utilize the pattern in the Data?

How to optimally learn from the Data?