Feature selection for disease classification

Kieren O’Neil

Supervisor: Dr. Mohamed Tawhid

July 25, 2023

SHAD TRU Seminar Series

What do you think of when you hear AI?

THOMPSON RIVERS UNIVERSITY

SHAD CANADA

What is AI?

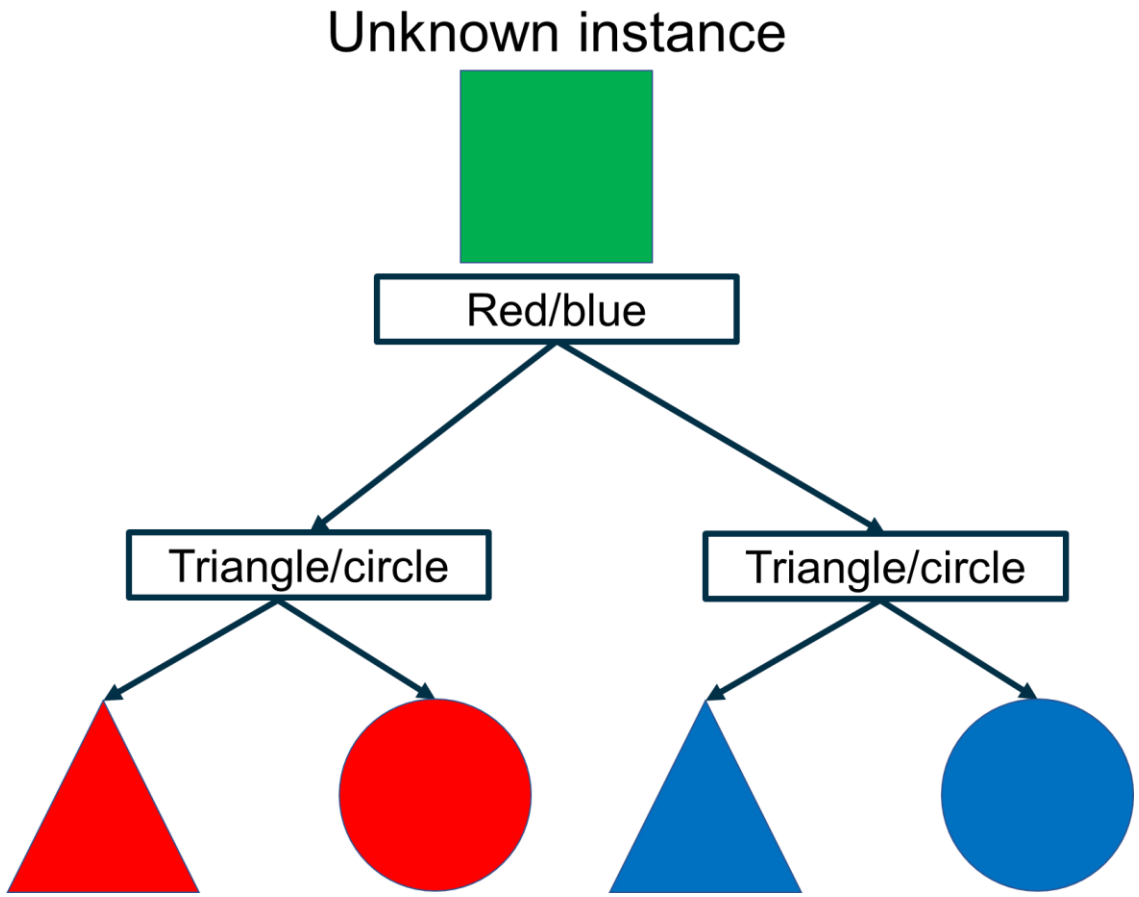
- Very broadly: intelligence of machines or software.
- There are many applications:
  - Search engines
  - Classification
  - Image generation
  - Computer vision
  - Chatbots
  - Recommendation systems
  - Strategic game systems
  - Optimization

Classification

- Automated pattern recognition.
- Grouping of some observations or instances into categories.
- A classifier takes an input with measurable properties and groups it into a class based on those parameters.
- To create a classifier, it is trained on a human categorized dataset; once trained it can classify data other than its training data.

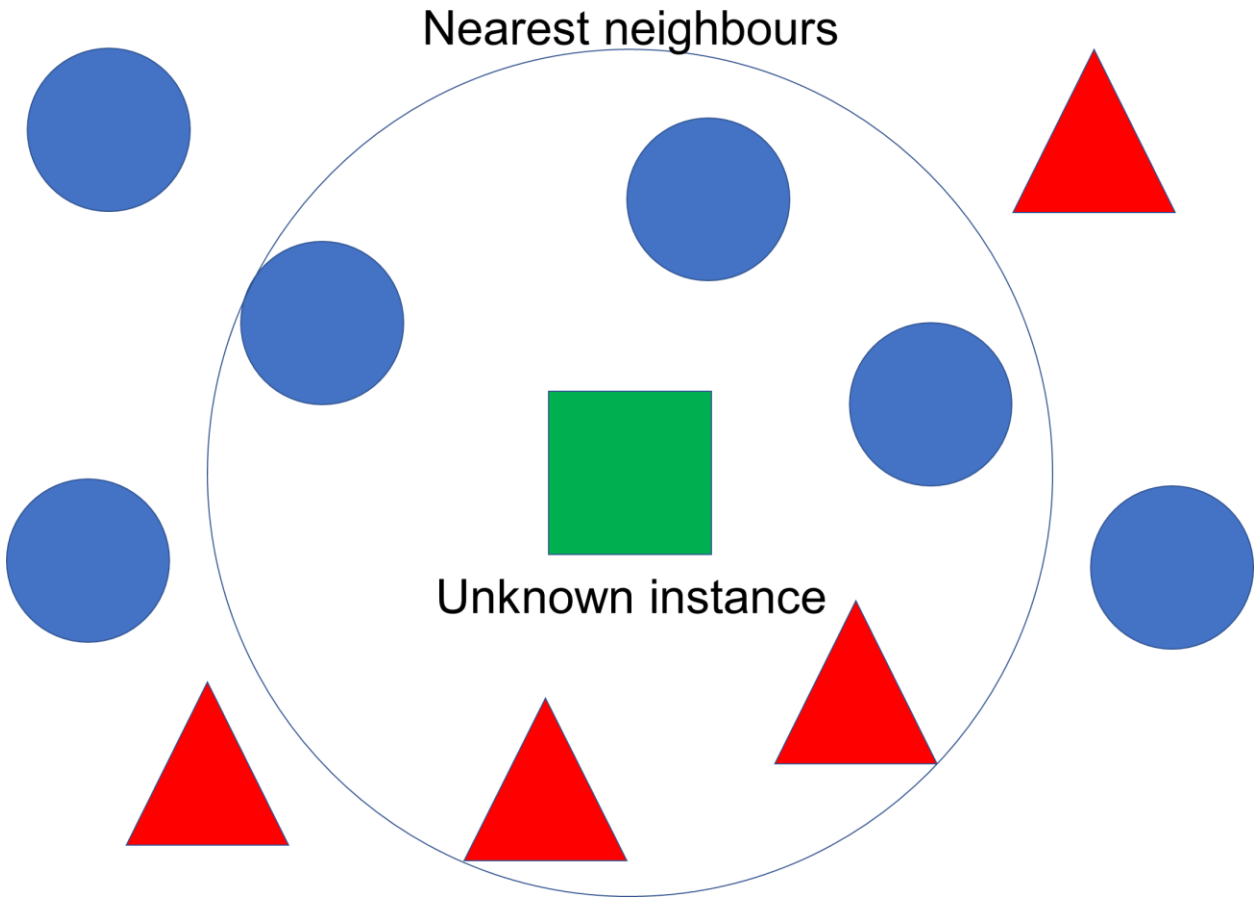
Decision Tree

- Binary search tree.
- A series of binary decisions is created to split instances into classes.
- Each block is a binary choice, through a series of binary choices a class is assigned to the instance.



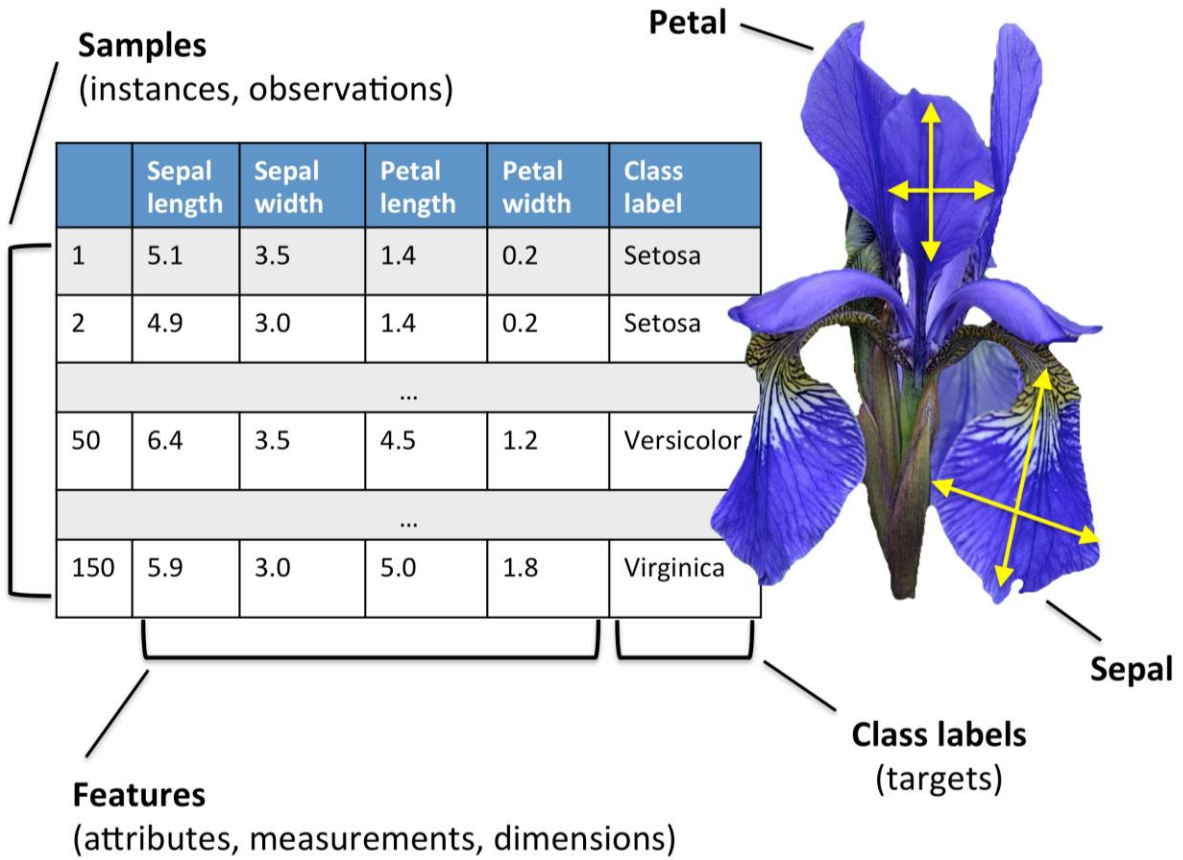
K nearest neighbours

- Graph an object and assign it to the class of the nearest neighbours in the dataset.
- Of the closest K instances assign the unknown one to the class the most neighbors have to an instance.
- We used $K = 5$.



Datasets

- A collection of instances to test classification.
- They are human grouped into classes and can be used to train artificial classifiers.
- They contain measured information and the class it falls into.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

Medical Datasets

- In this research we tested 16 medical datasets.
- Medical datasets contain clinical information such as age, test results, and environmental factors grouped by diagnosis.
- Cleveland heart disease dataset.

https://archive.ics.uci.edu/dataset/45/heart+disease

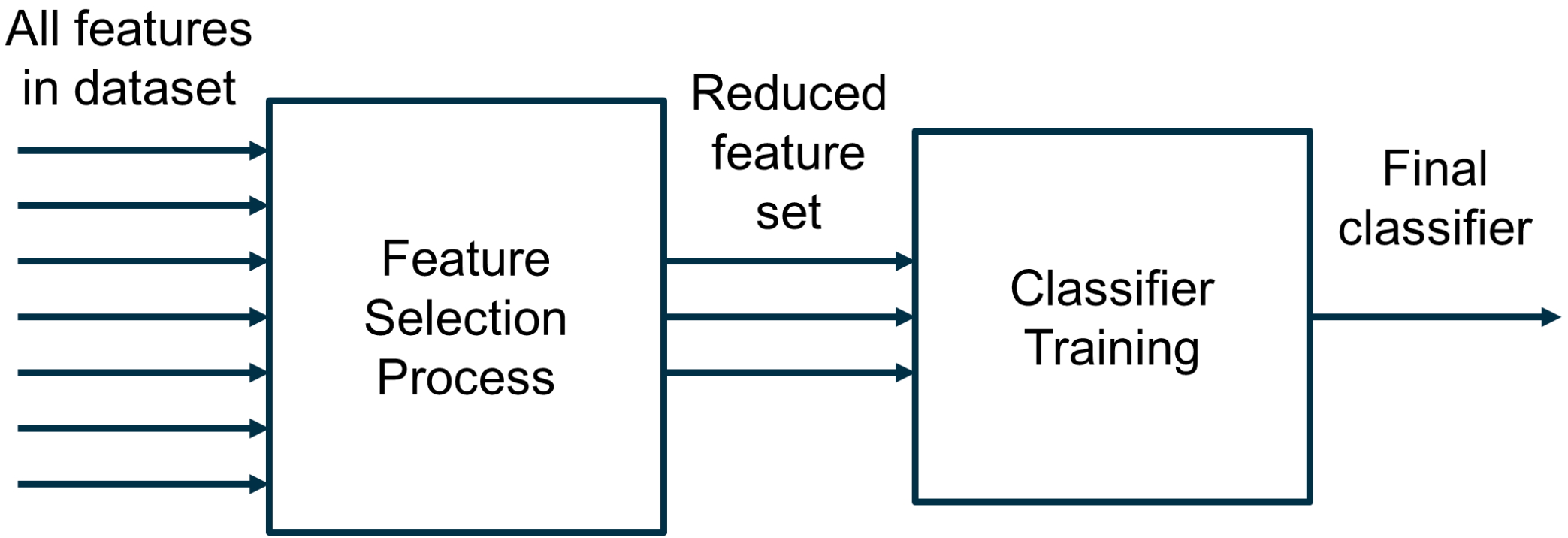
Features

Attribute Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer			years	false
sex	Feature	Categorical				false
cp	Feature	Categorical				false
trestbps	Feature	Integer		resting blood pressure (on admission to the hospital)	mm Hg	false
chol	Feature	Integer		serum cholestoral	mg/dl	false
fbs	Feature	Categorical		fasting blood sugar > 120 mg/dl		false
restecg	Feature	Categorical				false
thalach	Feature	Integer		maximum heart rate achieved		false
exang	Feature	Categorical		exercise induced angina		false
oldpeak	Feature	Integer		ST depression induced by exercise relative to rest		false
slope	Feature	Categorical				false
ca	Feature	Integer		number of major vessels (0-3) colored by flourosopy		false
thal	Feature	Categorical				false
num	Target	Integer		diagnosis of heart disease		false

Feature selection

- Feature selection reduces dataset size by selecting the most important features.
- Feature selection reduces training time, decreases complexity, and improves accuracy of final classifiers.
- In a dataset with $\boldsymbol{n}$ features there are $\boldsymbol{2^{n-1}}$ possible combinations of features.
- Time to check all combinations increases exponentially so it is a not polynomial-hard (NP-hard) problem.

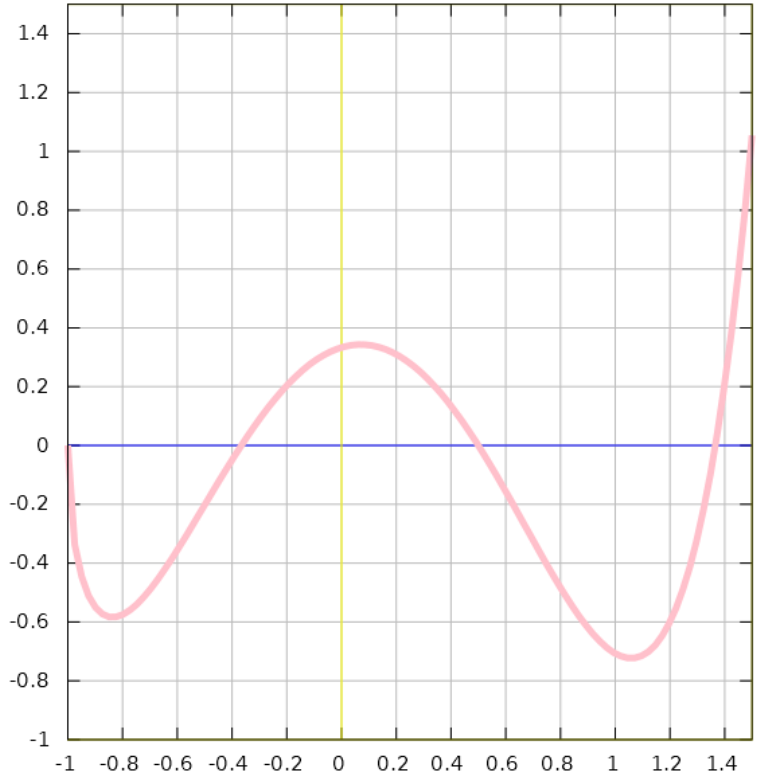
Feature selection



Functions

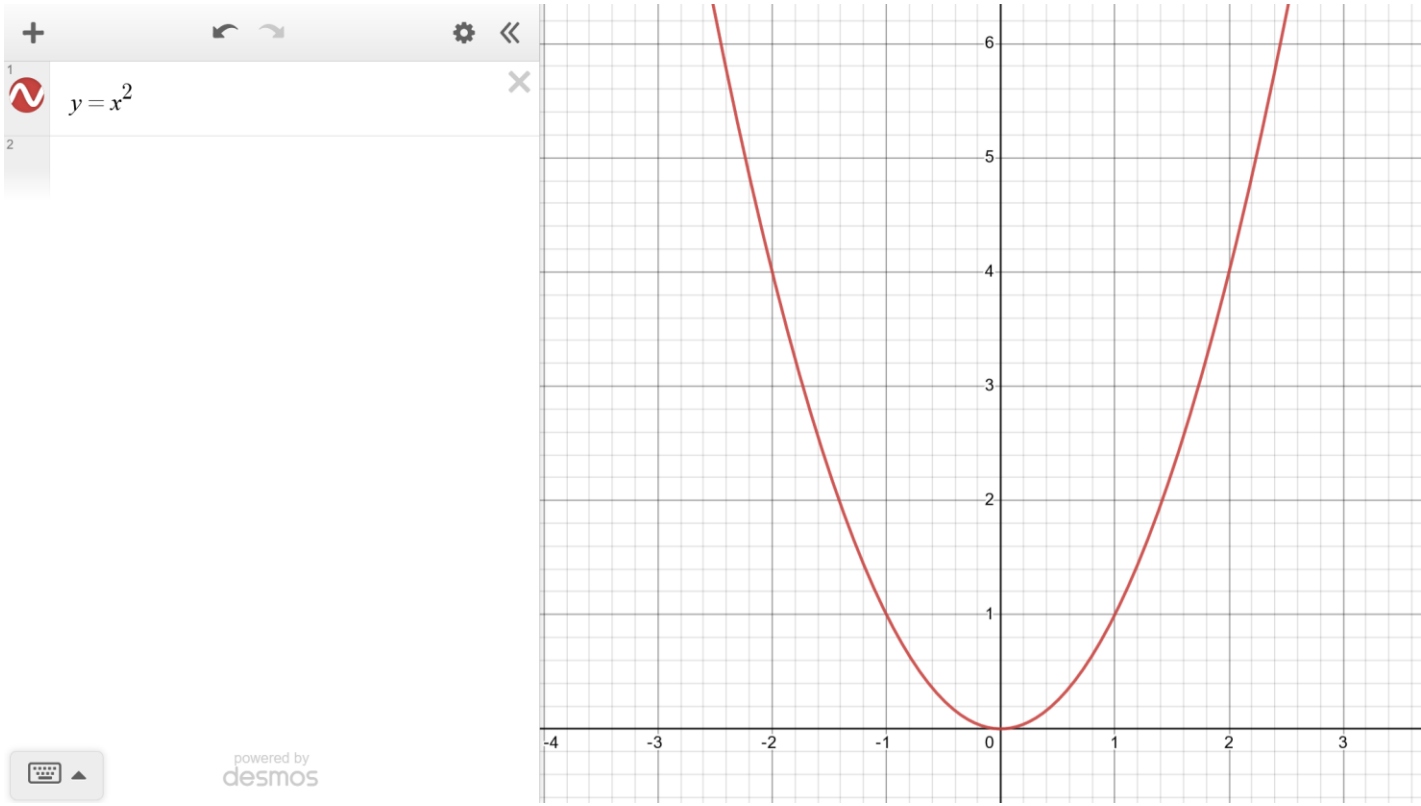
- $y = x$
- $y = 3x^2 - 2x + 14$
- $y = sinx$
- $y = \frac{1}{x}$
- $y = 2x_1 + 5x_2 + 45x_3 - 8x_4$

https://en.wikipedia.org/wiki/Function_(mathematics)#General_properties



Optimization

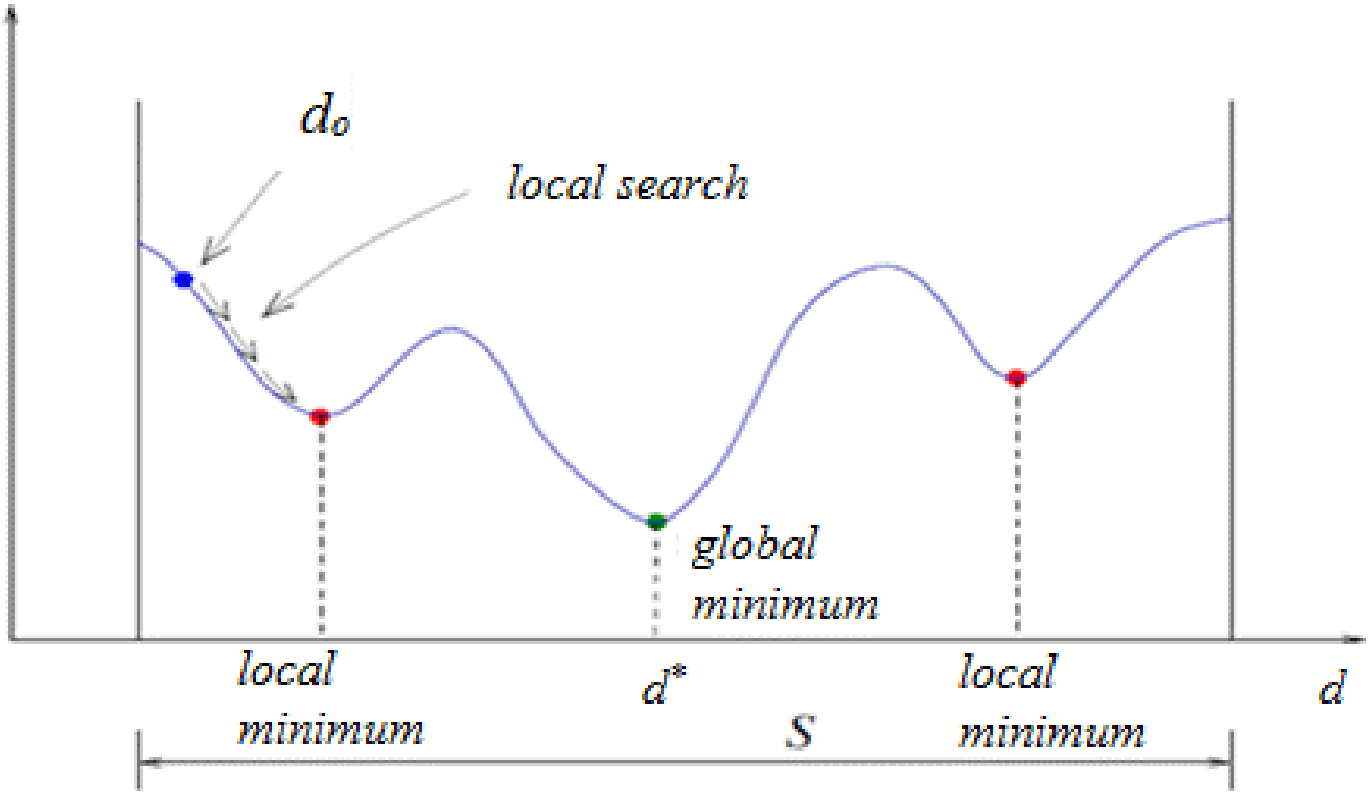
- We want to find the input to a function which minimizes the output.
- The simplest way is going down the hill (gradient descent).



Optimization

- For complex (bumpy) functions there may be many local valleys (minima).
- To find the overall best (global minimum) we need a way to stop from getting stuck in a shallow valley and find the deepest one.

https://www.researchgate.net/figure/Local-and-Global-minima_fig7_278665911

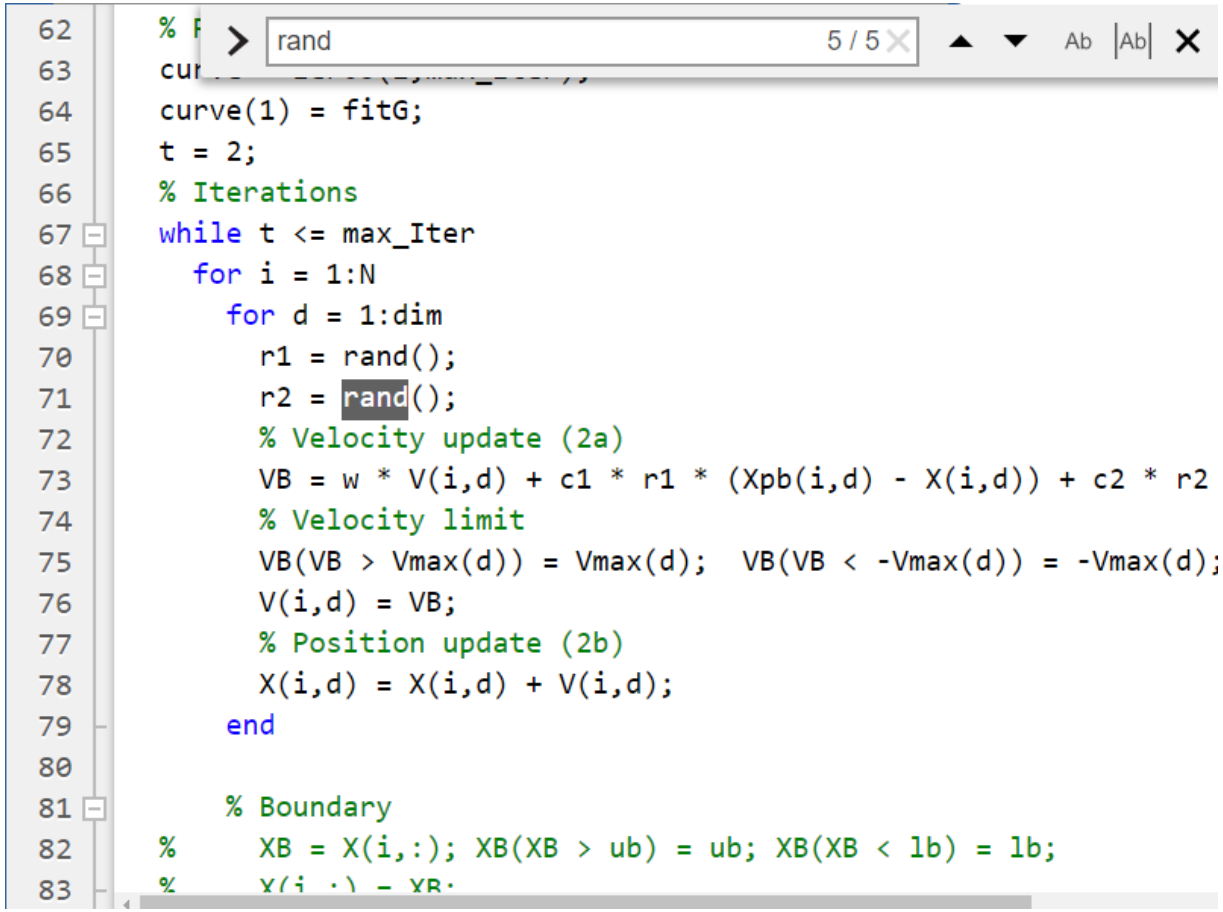


Demo

https://scopeblog.stanford.edu/2018/11/28/tackling-a-sticky-surgical-complication-stanford-researchers-identify-culprit-and-potential-treatment/tangled-ropes/

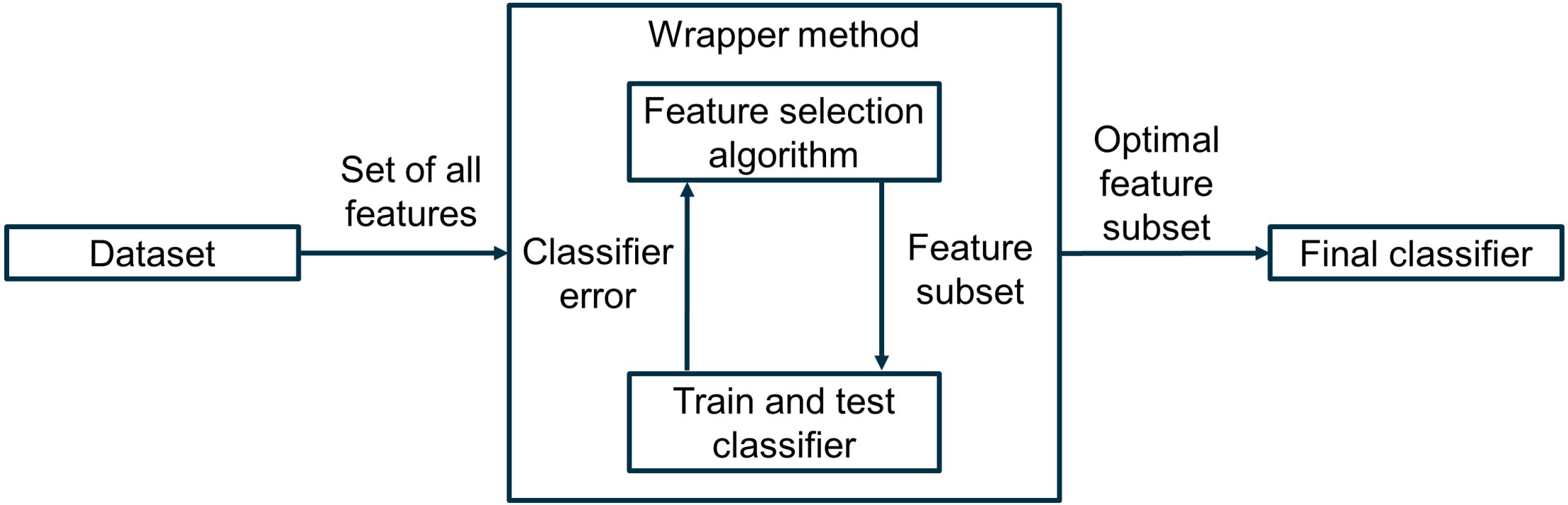
Stochastic Optimization

- Stochastic essentially means random.
- This means the optimization we're doing incorporates some randomness in it.
- The randomness is to excite out of the local minima and find the global minimum.



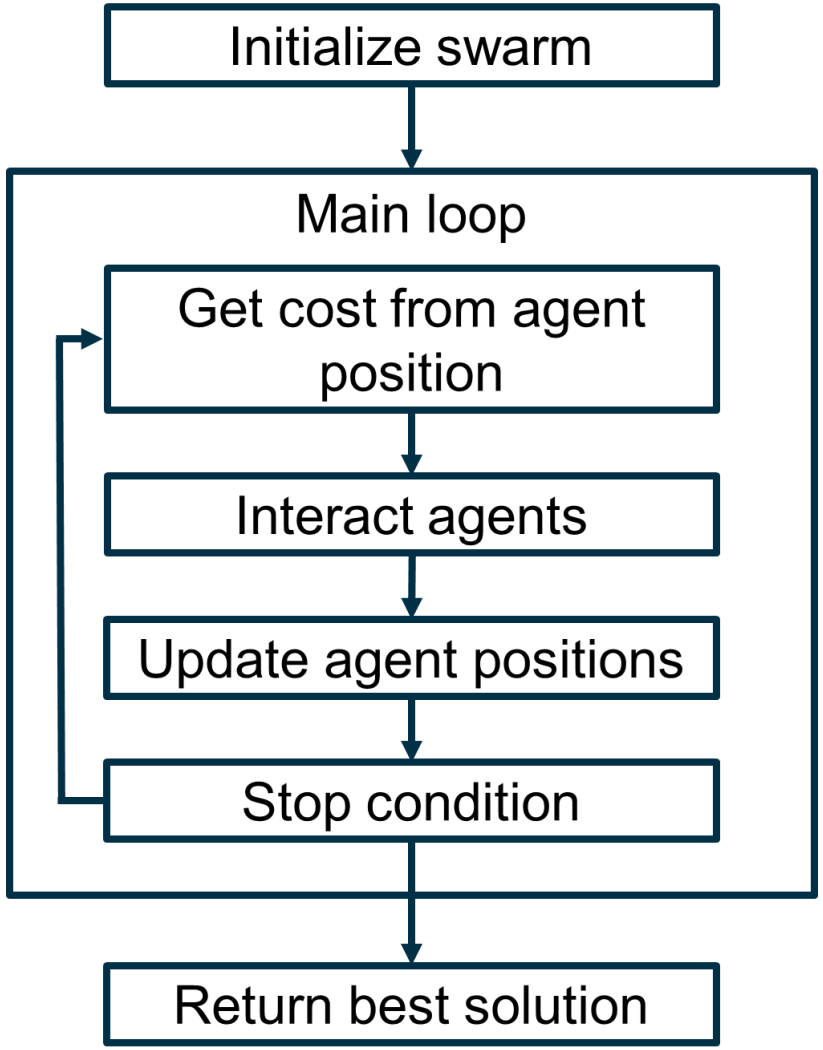
Wrapper method

- A selection algorithm optimizes for the set of features that produce the lowest error (the lowest fraction of data incorrectly categorized) in a trained classifier.



Swarm algorithms

- Swarm algorithms are what we used for feature selection.
- They get their name from a large population of search agents which test many inputs.
- From simple interactions between search agents complex search behavior emerges.



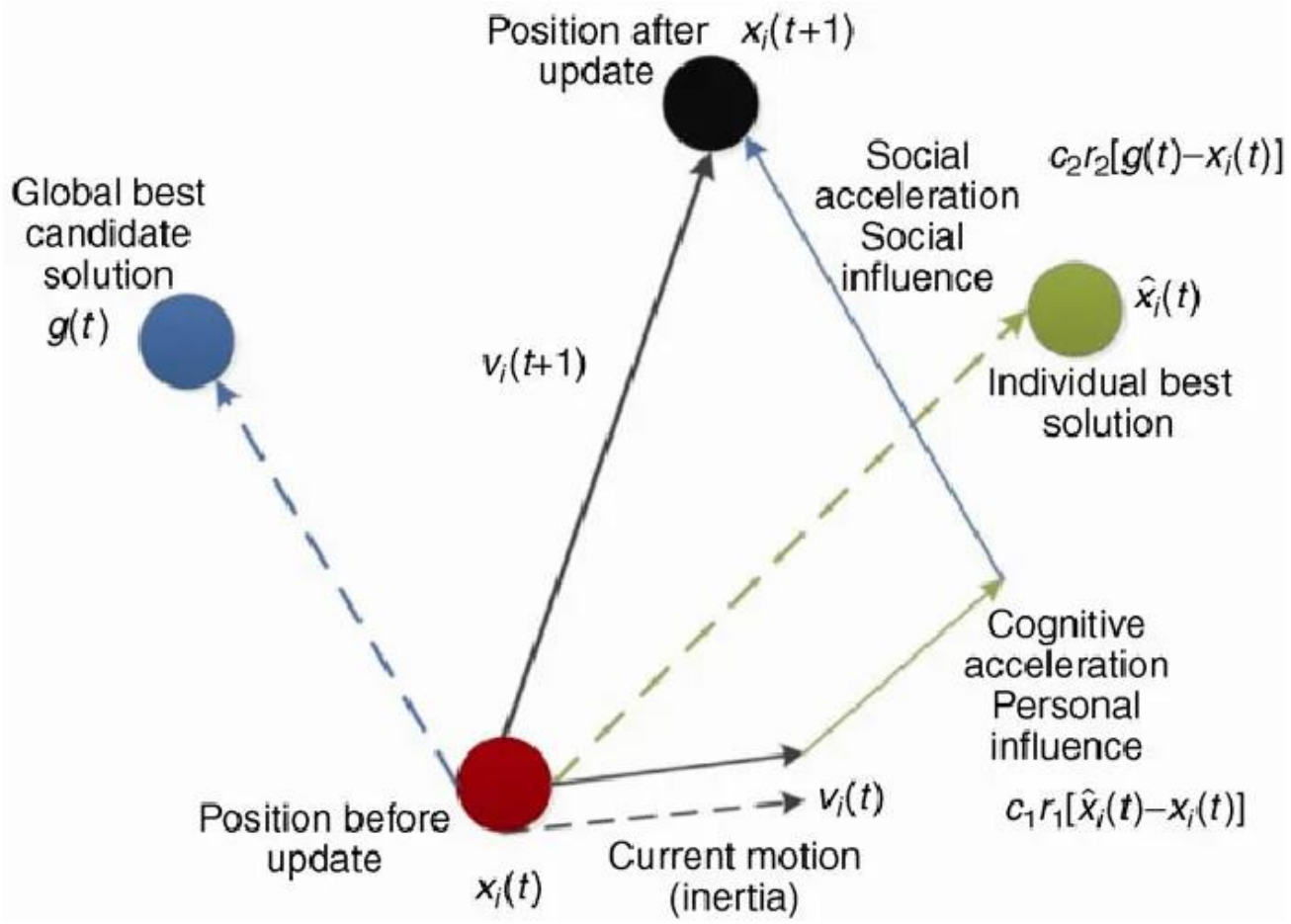
Swarm algorithms

- These are inspired from modeling real-world behavior.
- Often biological systems are used as the inspiration for these:
  - Ants
  - Birds
  - Bees
- Other concepts can also be used.

https://www.morrocoastaudubon.org/2020/09/video-sep-21st-community-program-why.html

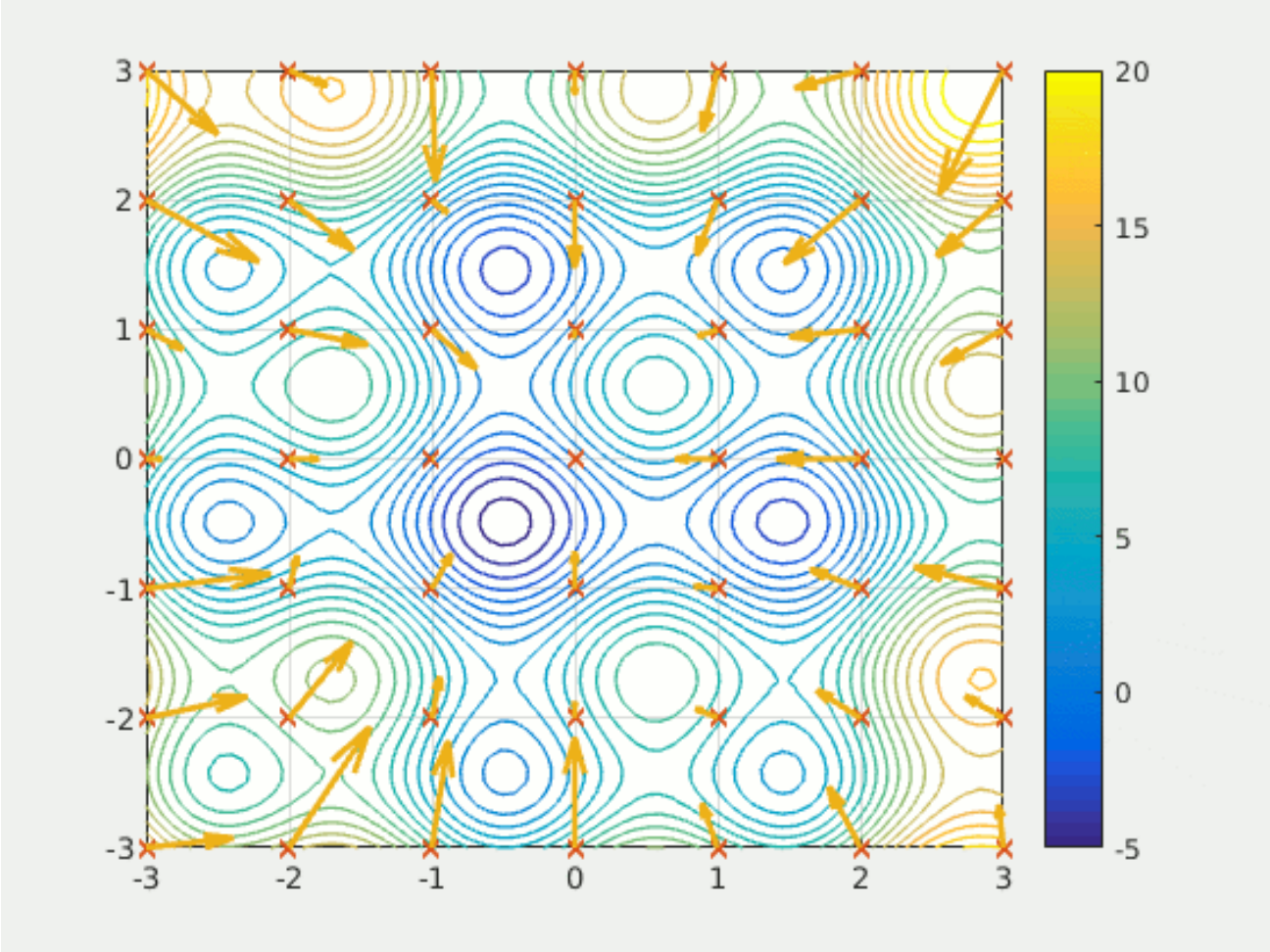
Particle swarm optimization

- Inertia – move in the same direction as the last move.
- Cognitive – move towards the best position found by that individual.
- Social – move toward the best position found by any individual.
- Others are similar.



https://medium.com/analytics-vidhya/implementing-particle-swarm-optimization-pso-algorithm-in-python-9efc2eb179a6

Particle swarm optimization

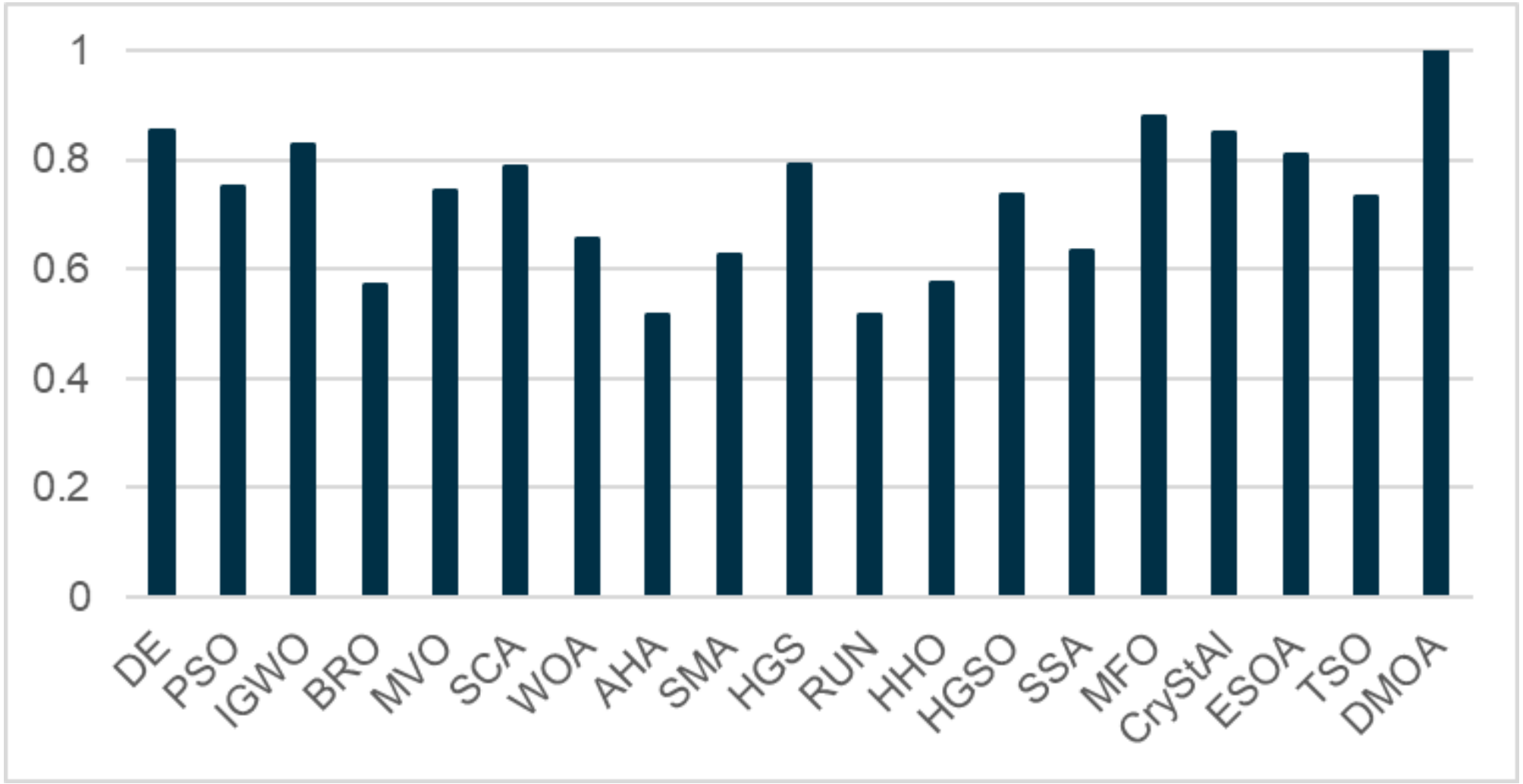


https://en.wikipedia.org/wiki/Particle_swarm_optimization#/media/File:ParticleSwarmArrowsAnimation.gif

What we did

- We compared 19 swarm algorithms on 16 medical datasets.
- The swarm algorithms were run 25 times with a population of 25 for 100 iterations.
- The results were compared with the Friedman test across all classifiers and datasets.
- The output is a rank of the algorithms where the highest value is the best.

Results


1
0.8
0.6
0.4
0.2
0
DE
PSO
IGWO
BRO
MVO
SCA
WOA
AHA
SMA
HGS
RUN
HHO
HGSO
SSA
MFO
CryStAl
ESOA
TSO
DMOA

Results

- We tested 19 algorithms on feature selection of 16 medical datasets for training 2 classification algorithms. Dwarf Mongoose Optimization Algorithm was overall the best for feature selection.
- We are currently developing improvements to DMOA and applying feature selection to gene array micro regime datasets.
- The medical datasets we tested here are 10-30 features, the gene array datasets are thousands of features where feature selection is essential.

Any Questions?