Analytics Vidhya
Learn everything about analytics

DATAHACK SUMMIT 2019

Using Genetic Algorithms to build Machine Learning Pipellines

Sahil Verma

13 - 16 November 2019   Bengaluru

Agenda

1. Introduction to Genetic Algorithms (GA)
2. Evolutionary Cycle
   a. Initialization
   b. Selection
   c. Crossover
   d. Mutation
3. Toolkit for your GA
4. Live case studies
5. Drawbacks of GA

Introduction to GA

What is a Genetic Algorithm?

An optimization technique(Heuristic) inspired from NATURE that can be used to solve any optimization problem (*YES ANY…..*)

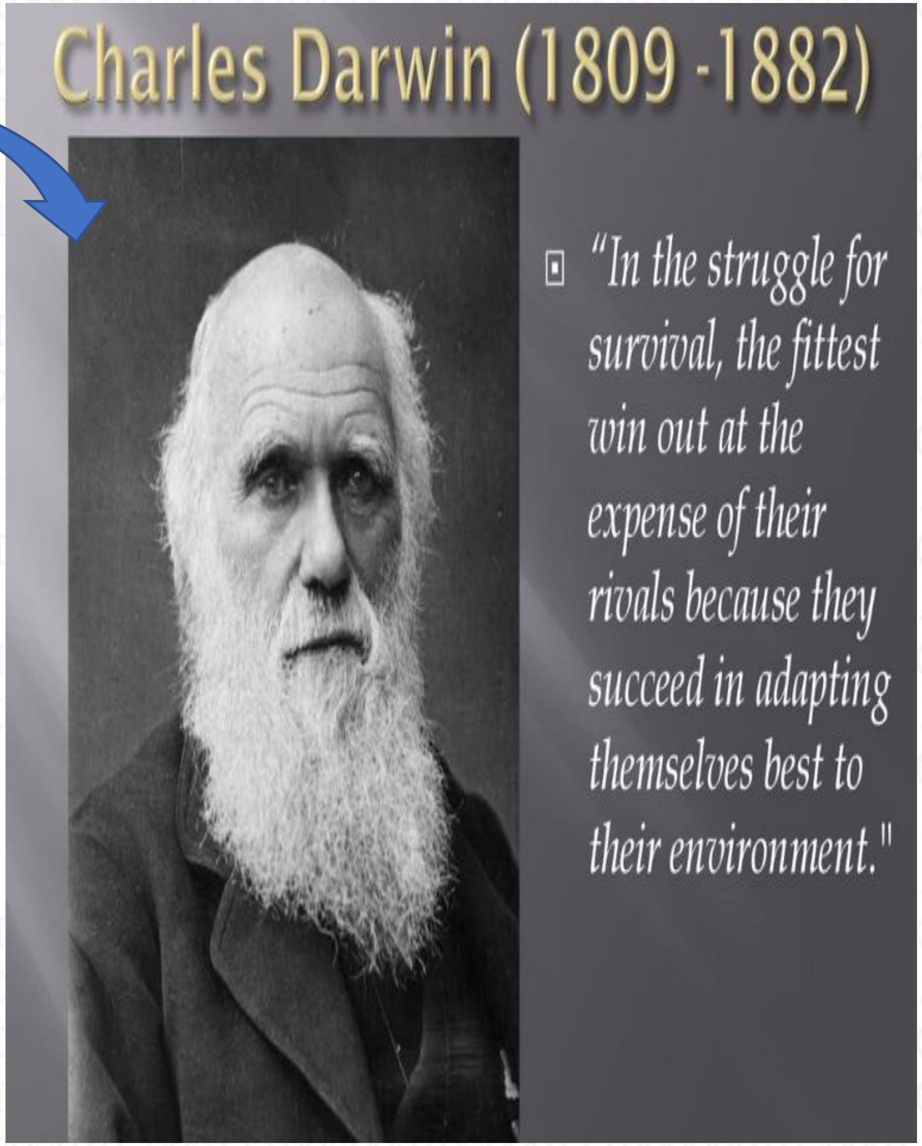
Are they even effective?

Google AI

H_2O.ai

TPOT

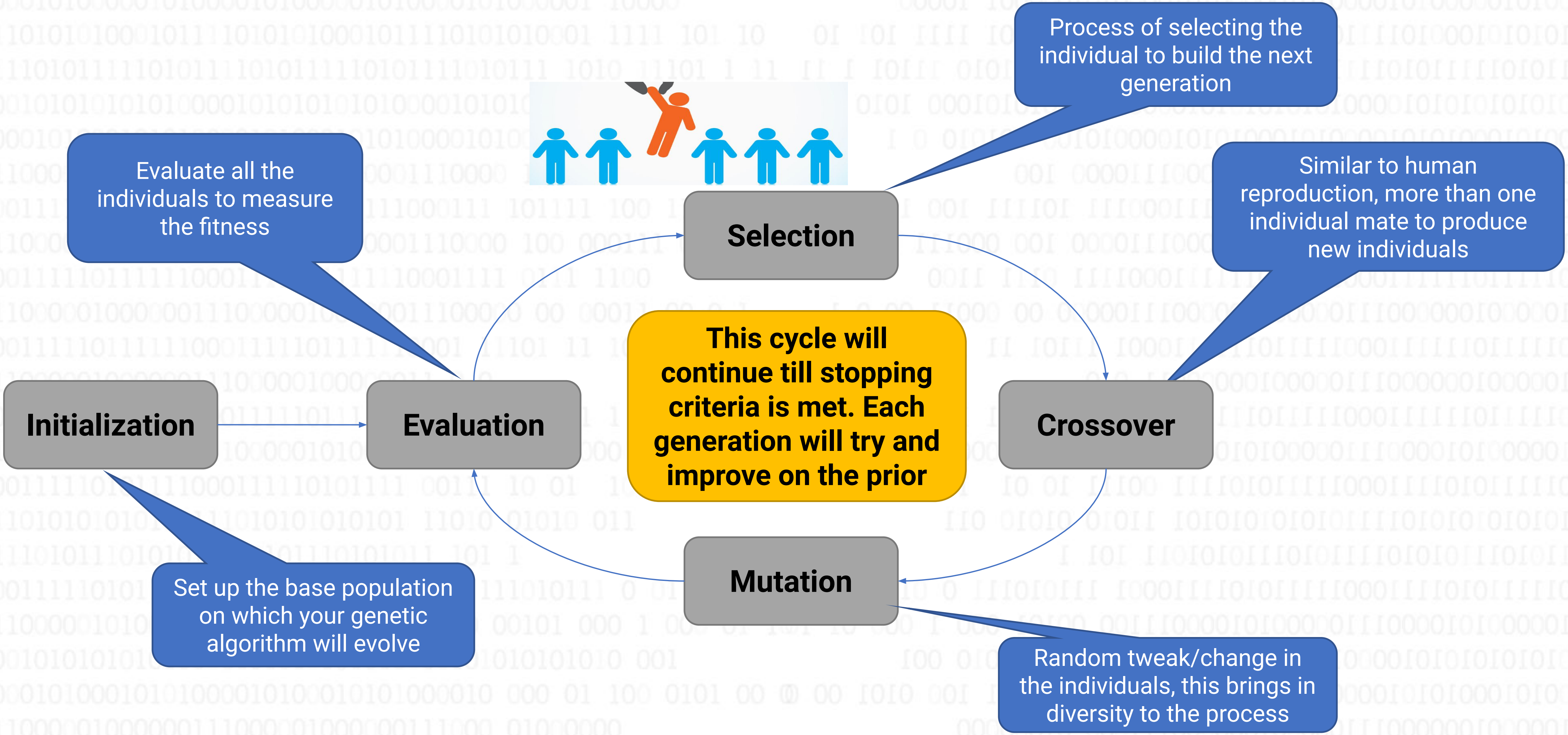
y = a + bX + e



Charles Darwin (1809 -1882)

- "In the struggle for survival, the fittest win out at the expense of their rivals because they succeed in adapting themselves best to their environment."

Evolutionary Cycle



Initialization

Evaluation

Selection

Crossover

Mutation

Set up the base population on which your genetic algorithm will evolve

Evaluate all the individuals to measure the fitness

Process of selecting the individual to build the next generation

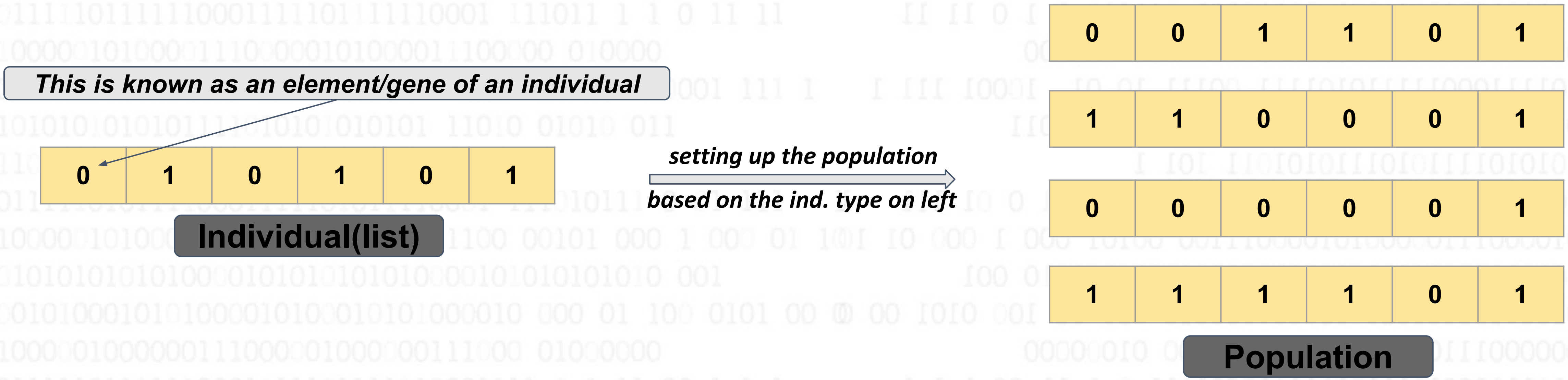
Similar to human reproduction, more than one individual mate to produce new individuals

Random tweak/change in the individuals, this brings in diversity to the process

This cycle will continue till stopping criteria is met. Each generation will try and improve on the prior

Evolutionary Cycle (Initialization)

- We define an individual here, which is used for setting up the population
- Creating an individual is the **most important step**
- Individuals **vary depending on the problem**
- Size of the population is based on hit and trial:
  - *Small population size: suboptimal solution, faster computation*
  - *Large population size: Better Solution, slower computation*
- Types of individuals we will see today:
  - List
  - Dictionary



This is known as an element/gene of an individual

0	1	0	1	0	1

Individual(list)

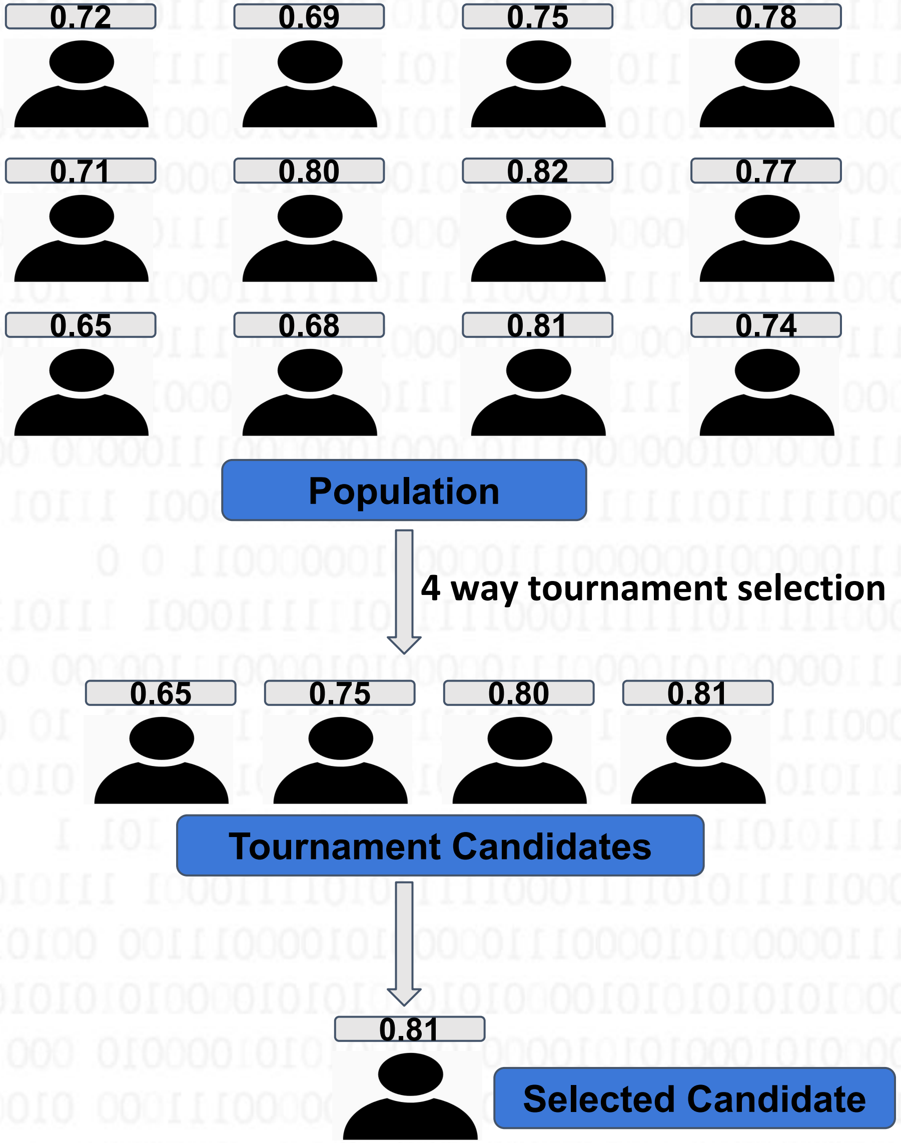
setting up the population
based on the ind. type on left

0	0	1	1	0	1
1	1	0	0	0	1
0	0	0	0	0	1
1	1	1	1	0	1

Population

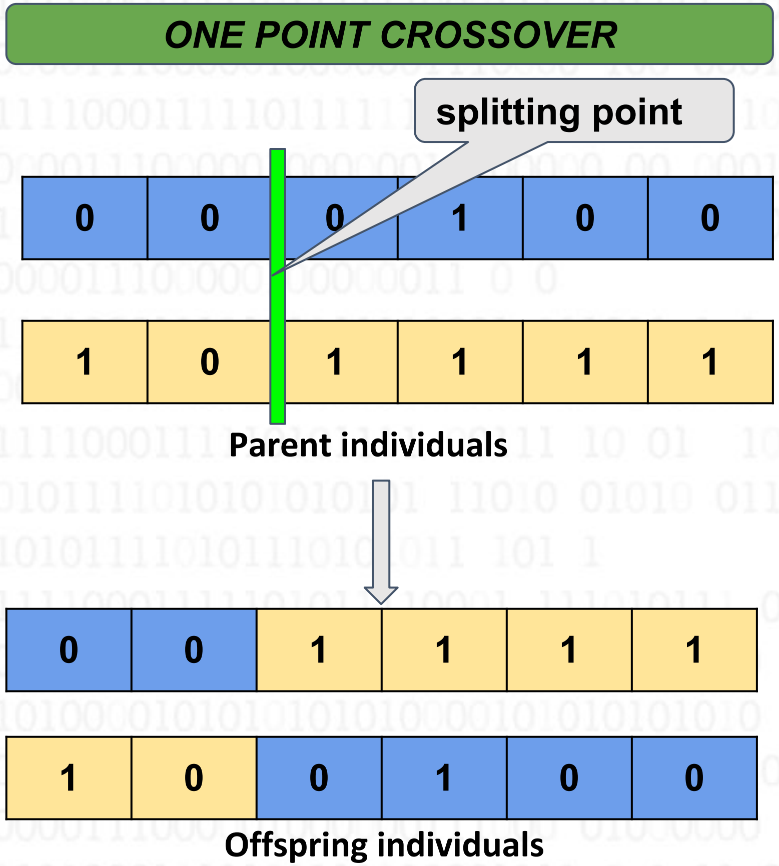
Evolutionary Cycle (Selection)

- Process of selecting the parents from the population
- Parents are responsible for mating/mutating to produce the new generation
- **Higher fitness individual should have more chances of selection**
- Types of selection processes:
  - *Tournament Based Selection* (We will use this for all our use-cases)
  - *Roulette Wheel Selection*
  - *Random Selection*
- **K-way tournament selection,** involves selecting k random individuals from the population and select the best of the K
- We repeat the K-way tournament selection, as many time as the size of the population
- Ideal value of K is based on hit and trials:
  - **K==1**: *meaning we select 1 random individual for the tournament, this is equivalent to random selection*
  - **K== population size**: *meaning we select all the individuals in the population, this will always give the same result*



Evolutionary Cycle (Crossover)

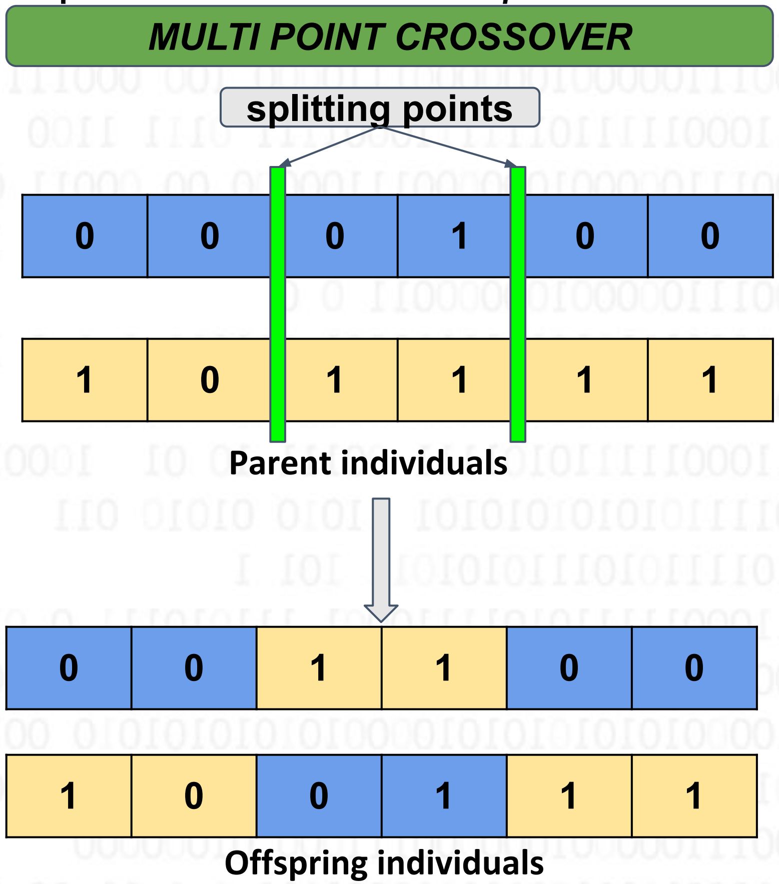
- Analogue of Reproduction
- Two Individuals mate to produce offsprings
- Types of Crossover:
  - *One Point Crossover* (we will be using this in our case studies)
  - *Multi point Crossover*
- Crossover is applied probabilistically (with a probability ***cxpb***), we prefer the value of *cxpb* closer to 1



ONE POINT CROSSOVER

Parent individuals

Offspring individuals



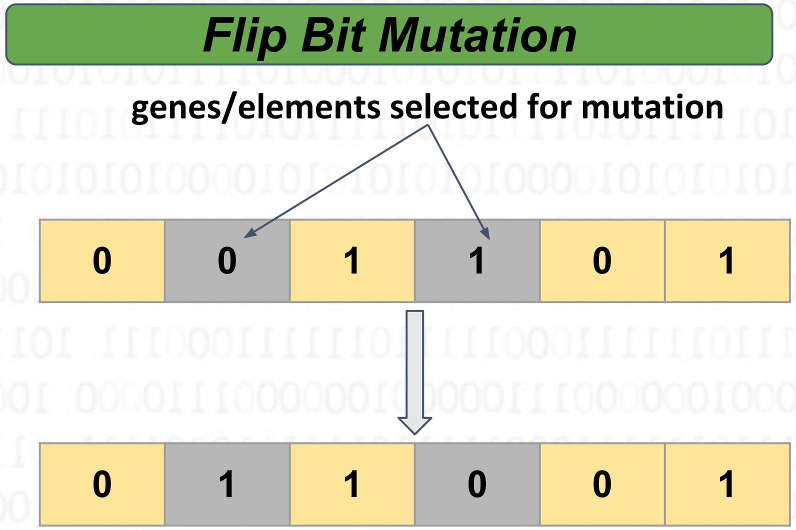
MULTI POINT CROSSOVER

Parent individuals

Offspring individuals

Evolutionary Cycle (Mutation)

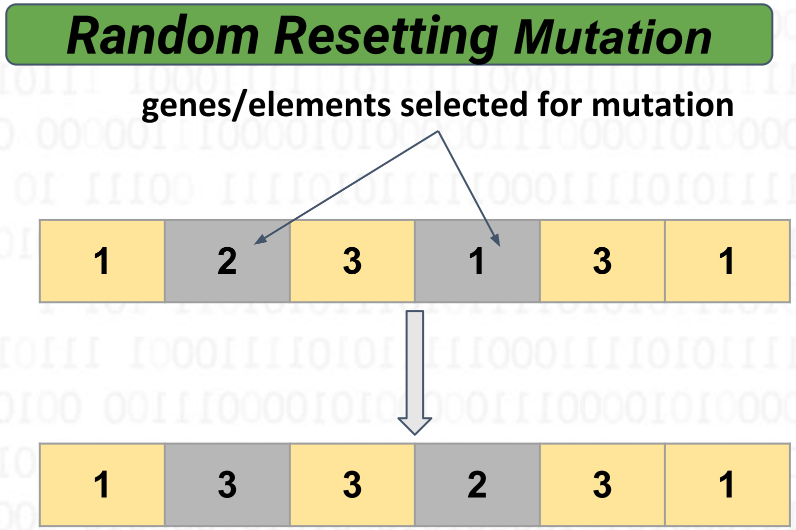
- Process of **random change/tweak** in the element/gene of an individual
- This process helps to **introduce diversity** in the population
- Mutation is also applied probabilistically (with a probability ***mutpb***)
- Because of the random nature of mutation, we ***prefer the value of mutpb to be low***
- Types of mutations we will use:
  - *Flip Bit mutation*
  - *Random resetting*
- In ***Flip Bit mutation***, we select one or more gene/element and flip the values
- **Random resetting**, is an extension of flip bit, instead of flipping the values we reset the value from a set of permissible range of values
- When the permissible range becomes [0,1], we have the flip bit mutation



Flip Bit Mutation

genes/elements selected for mutation

0	0	1	1	0	1
0	1	1	0	0	1



Random Resetting Mutation

genes/elements selected for mutation

1	2	3	1	3	1
1	3	3	2	3	1

*permissible set was [1,2,3]

Toolkit for setting up your GA

To define a GA you need the following 4 things:

- **Individual**: base element
- **Crossover Function:** will be responsible for combining individuals to give new individuals
- **Mutation Process:** this will introduce random changes within the individuals resulting in new individuals
- **Fitness Function:** metric for evaluating the fitness of each individual

→

DISTRIBUTED
EVOLUTIONARY
ALGORITHMS IN
PYTHON

Problem Statements

We will formulate a framework of GA to solve for the following:

- Introduction to DEAP
- Feature Selection
- Feature Creation

Let’s CODE

Be Careful with GA

- GA are useful if used with care
- They suffer from two major drawbacks:

❖ **Tendency to Overfit**: GA are known for overfitting a bit. But this effect can be reduced by adding a suitable regularization in your GA framework

❖ **Computationally heavy**: Because the search space is generally very large, we end up setting very long GA. We can try and setup smarter GAs (Adaptive GAs)

Analytics Vidhya
Learn everything about analytics

DATAHACK
SUMMIT
2019

Thank you!