VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

COMP307/AIML420 Genetic Programming for Regression and Classification: Tutorial

Dr Fangfang Zhang
fangfang.zhang@ecs.vuw.ac.nz

COMP307 Week 6 (Tutorial)

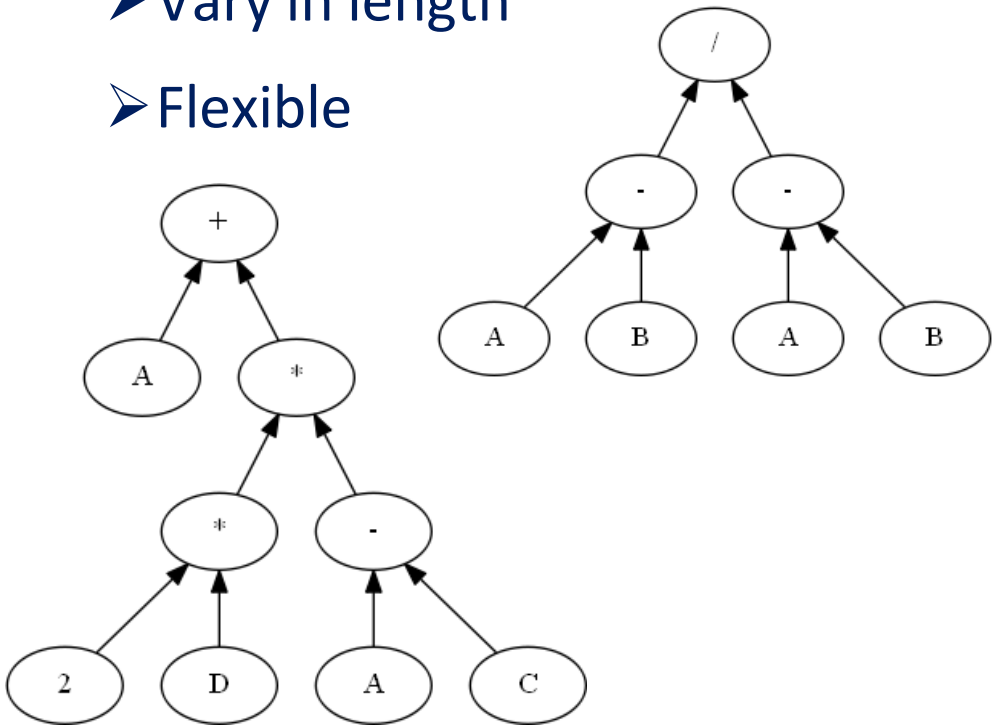
- Announcements
  - Assignment 2
  - Helpdesk (teaching break)
  - Andrew's teaching evaluation (until 12th April)
- GA to GP
  - Representation
  - Evolutionary operators
- Genetic Programming
  - Terminal set
  - Function set
  - Fitness function
  - GP Parameters
  - Stopping criterion
- GP for Regression
  - Statistical Regression VS. Symbolic Regression
  - Fitness function
- GP for Binary Classification
  - Fitness function
  - Classifier
- Tutorial for today
  - Overview
  - Go through part 2 of A2
  - An GP example
  - Report

GA to GP (Representation)

Genetic Algorithm	Genetic Programming
➢ Bit string representation	➢ Tree-like structure
➢ Fixed in length	➢ Vary in length
➢ Inflexible	➢ Flexible

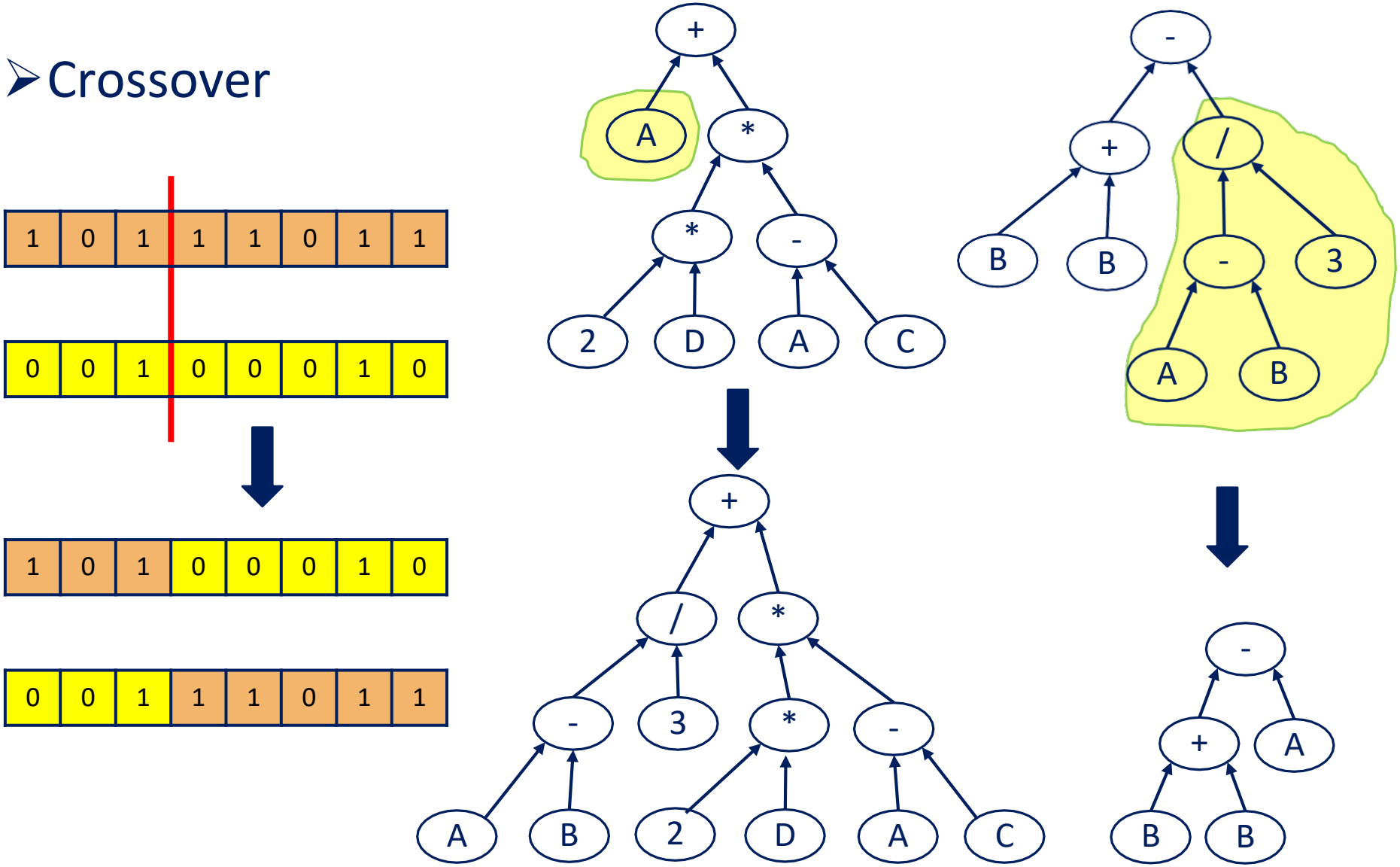
1	0	1	1	0	1	0	1	1	0

0	0	0	1	0	1	1	1	0	1



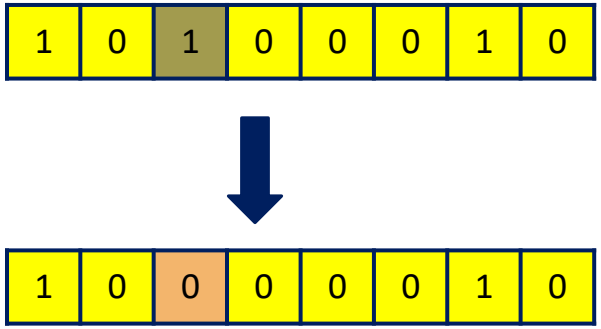
GA to GP (evolutionary operators)

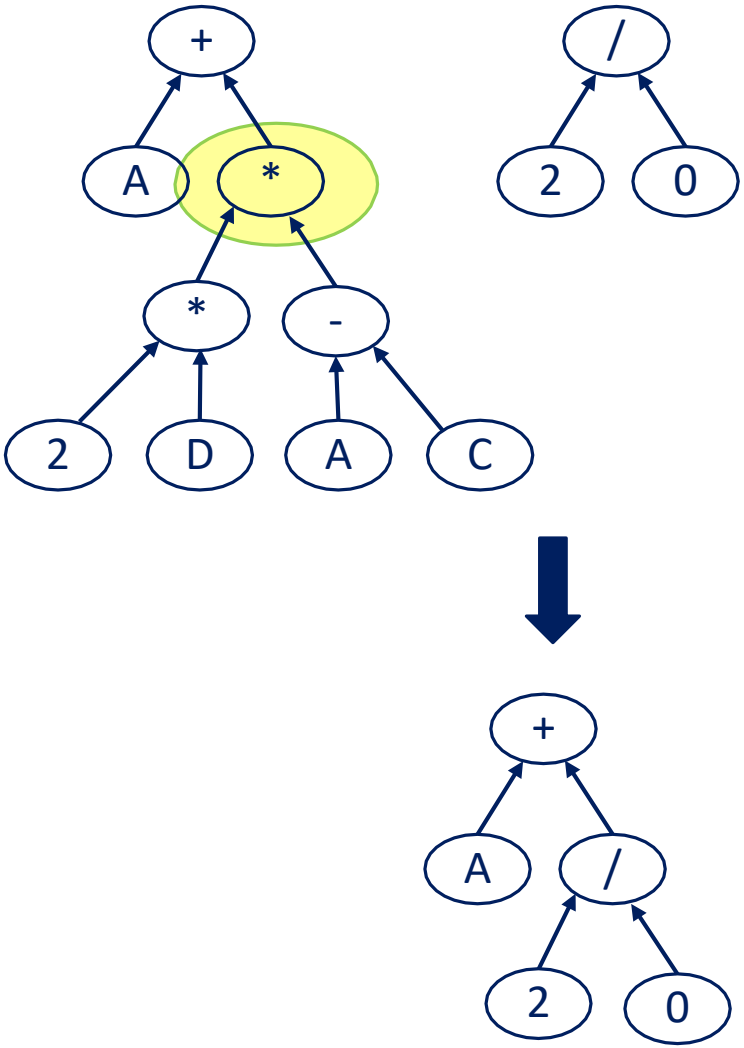
- Crossover



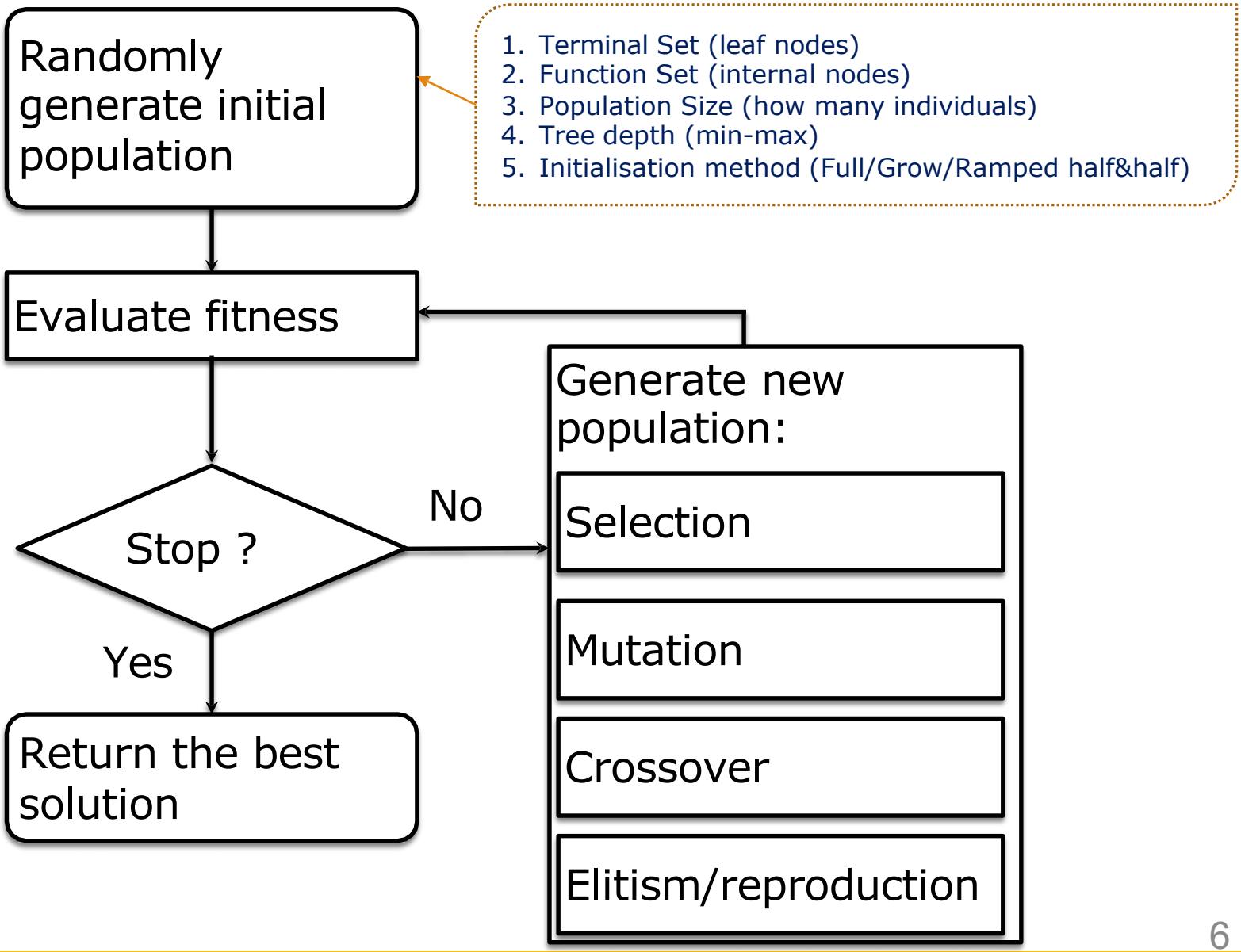
GA to GP (evolutionary operators)

- Mutation



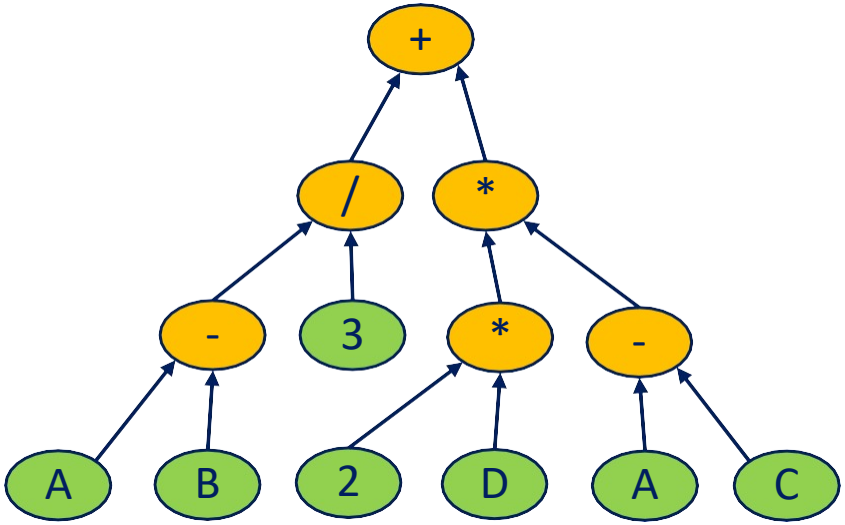


Flowchart of GP

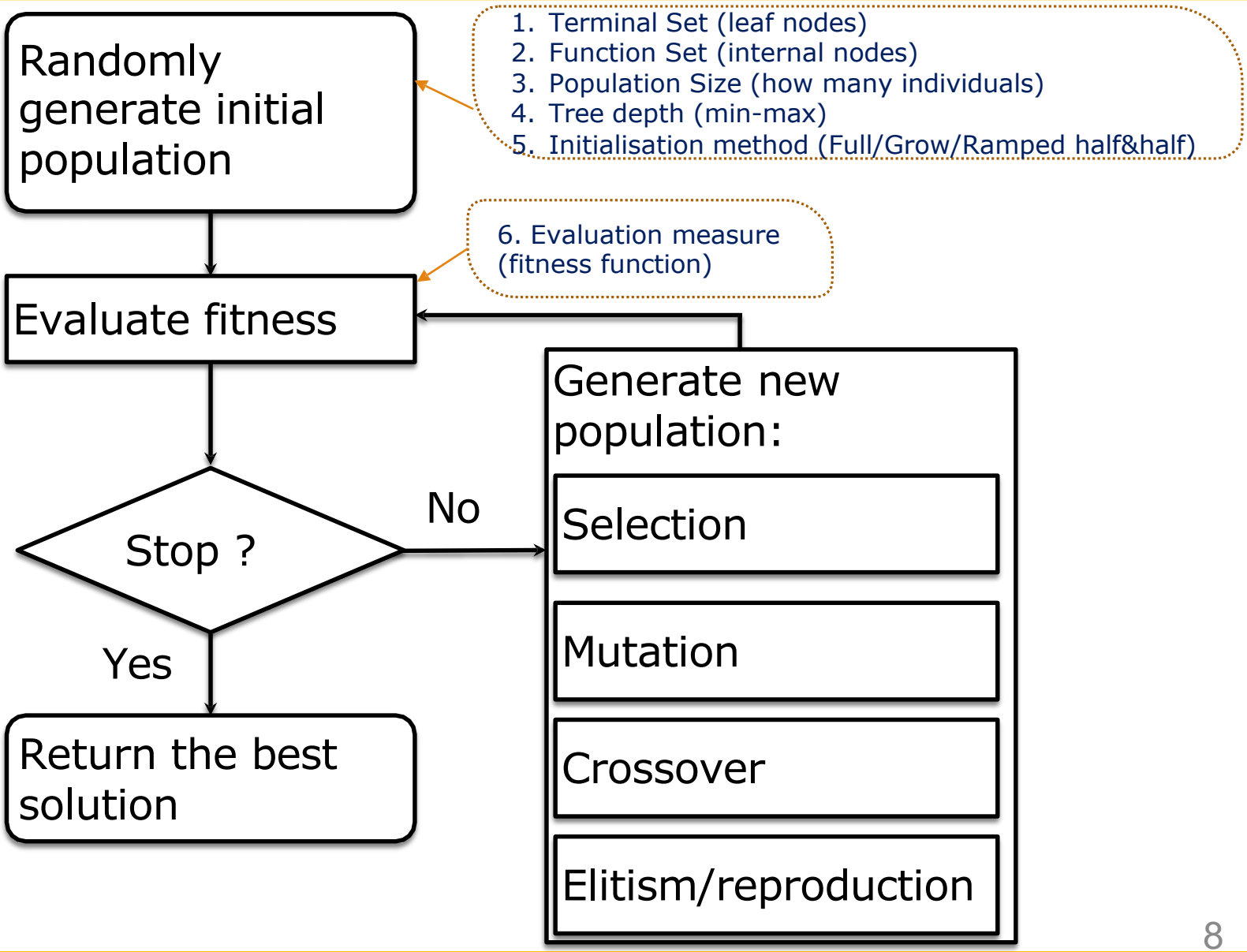


Terminal Set and Function Set

- Terminal Set
  - {A, B, C, D, E, F, *rand*}
- Function Set
  - {+, -, *, / (protected)}



Flowchart of GP



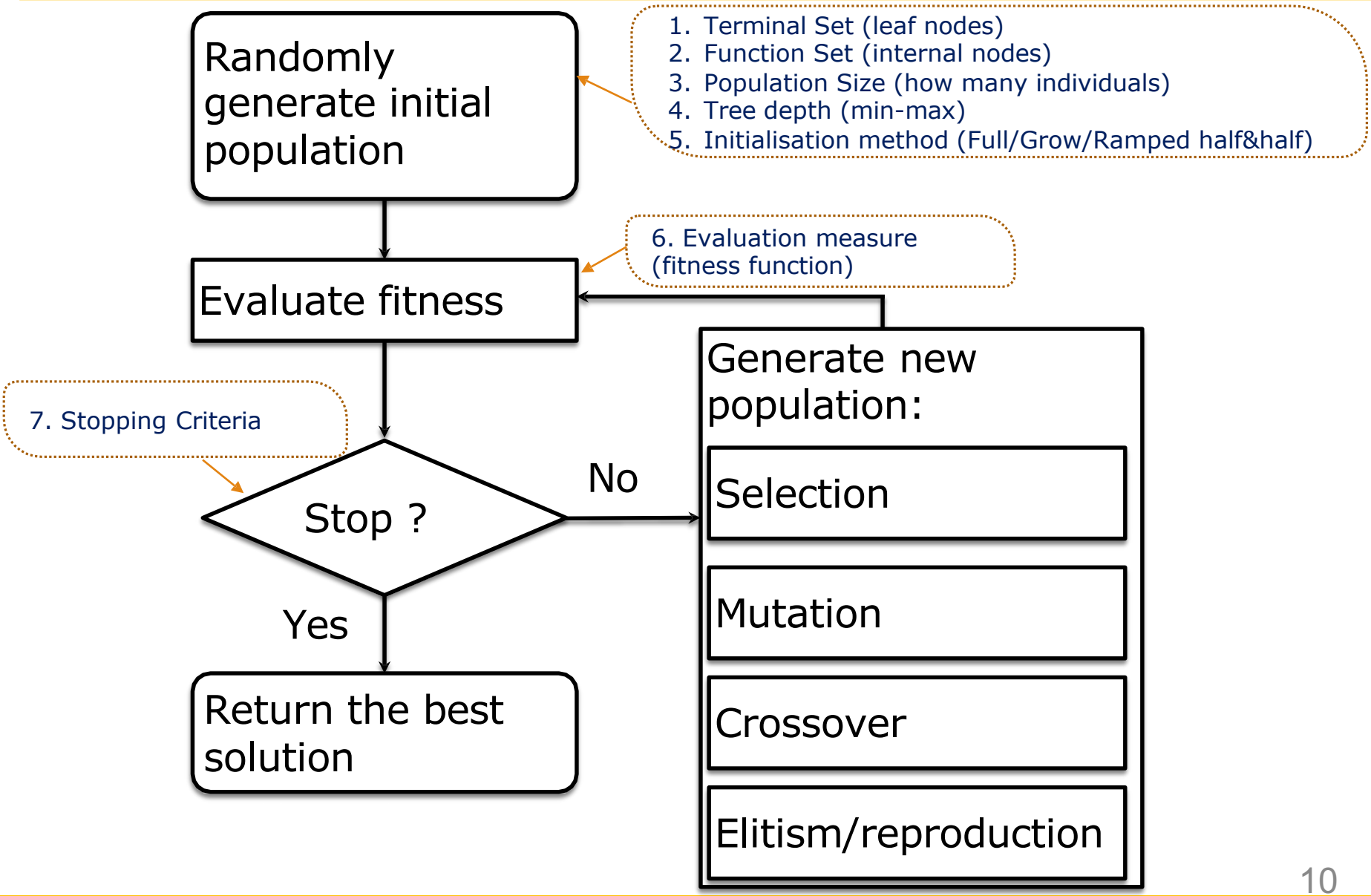
Fitness Function

- Fitness function measures the fitness of a program
  - How well a program performs on the training set
  - Very important
  - Varies from domain to domain

    Regression: MSE (mean squared error), RMSE,

    Classification: Classification Accuracy, Error Rate

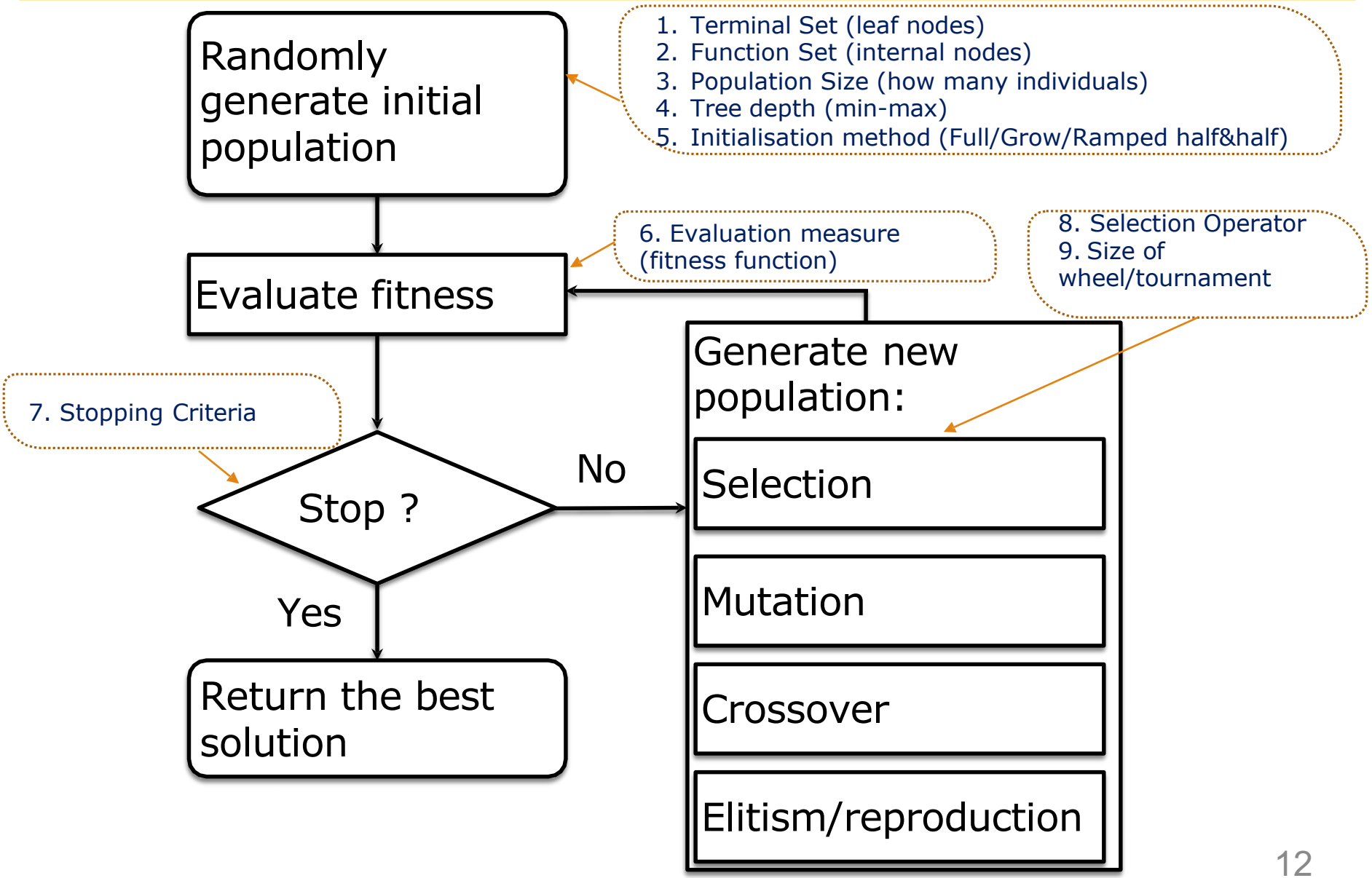
Flowchart of GP



Stopping Criteria

- When to stop the evolutionary
  - Satisfactory solutions found (e.g., error < 0.01)
  - Reach the maximum number of generations (e.g., 100 generations)
  - Some other criteria

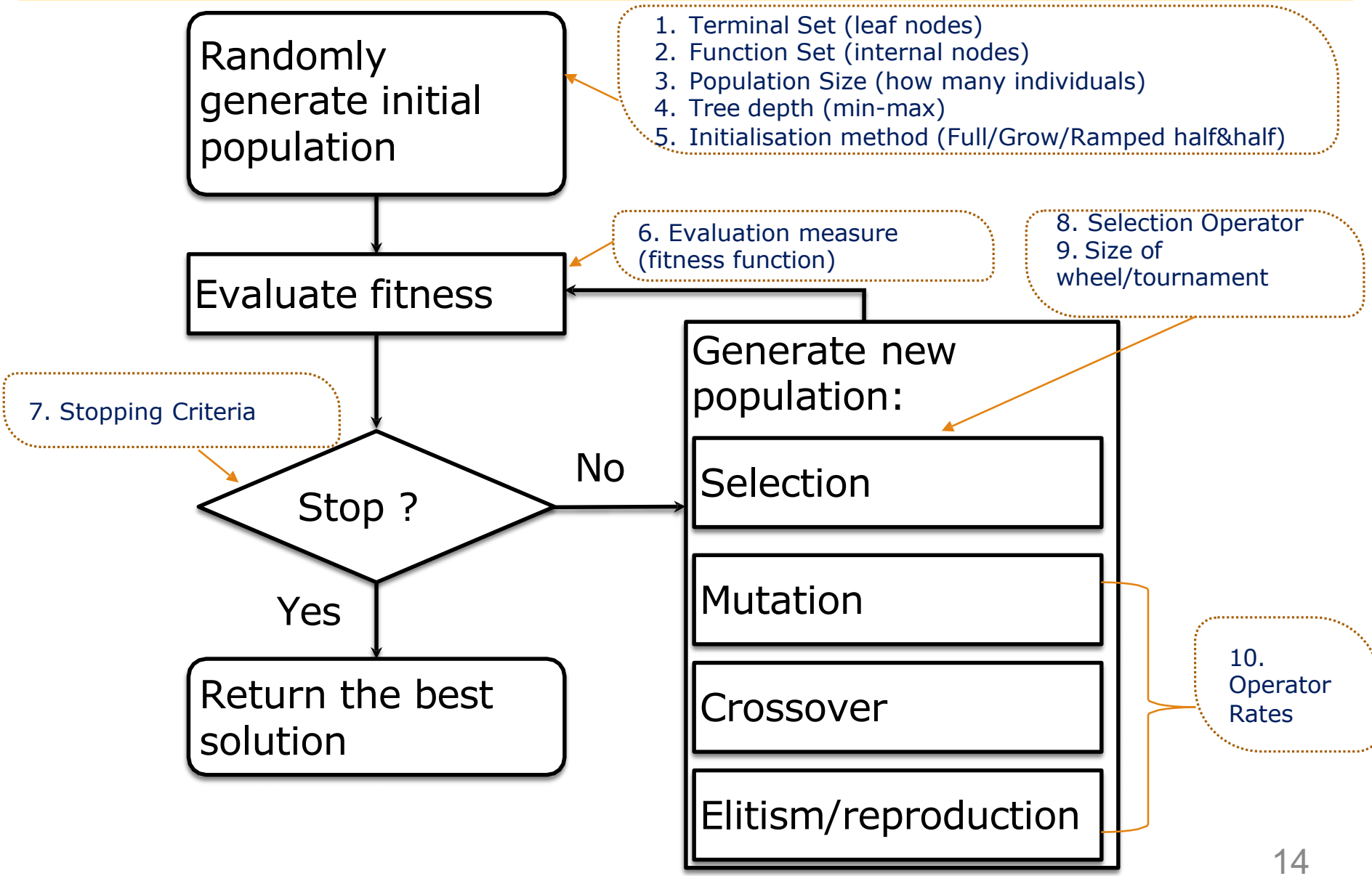
Flowchart of GP



Selection

- Selection operators
  - Roulette wheel selection (i.e., Proportional selection)
  - Tournament selection
- Size of tournament selection
  - The number of programs selected from the population

Flowchart of GP



Genetic Operators

- Genetic Operators
  - Reproduction/Elitism
  - Crossover
  - Mutation
- Operator Rates
  - Reproduction/Elitism rate + Crossover rate + Mutation rate =1
  - Small number of elitism rate
  - Higher Crossover rate and lower mutation rate

Tackling a Problem with GP

- What is the set of **terminals** used in the program trees?
- What kind of **functions** can be used to form the function set to represent the program tree?
- What is the **fitness measure**?
- What values can be given for the **parameters** and variables for controlling the evolutionary process, for example, population size and number of generations?
- When to **terminate** a run?
- How do we know the result is good enough?
- What genetic operators, at what frequencies, are going to be applied?

Tackling a Problem with GP (Summary)

- The preparatory steps:
  - The set of terminals
  - The set of functions
  - The fitness function
  - The parameters

    (for controlling the GP run)

    - Impossible to make general recommendations
    - Some typical settings
  - The criterion for terminating a run

1. Population size [>=100]
2. Initial tree depth (min-max) [2-6]
3. Maximum tree depth [<=17]
4. Number of generations [51]
5. Size of tournament selection [7]
6. Operator rates (elitism, crossover and mutation rate) [5%, 90%, 5%]

GP for Regression

- (Statistical) Regression

$$y = \alpha + \beta x + \epsilon$$

Symbolic Regression

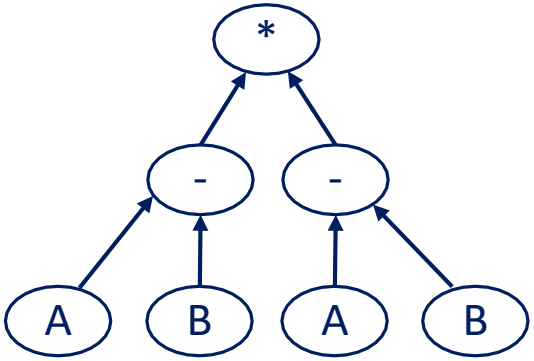
$$y = ??$$

- GP for symbolic regression
  - Fitness function

$$MSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n$$

$$RMSE = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n}$$

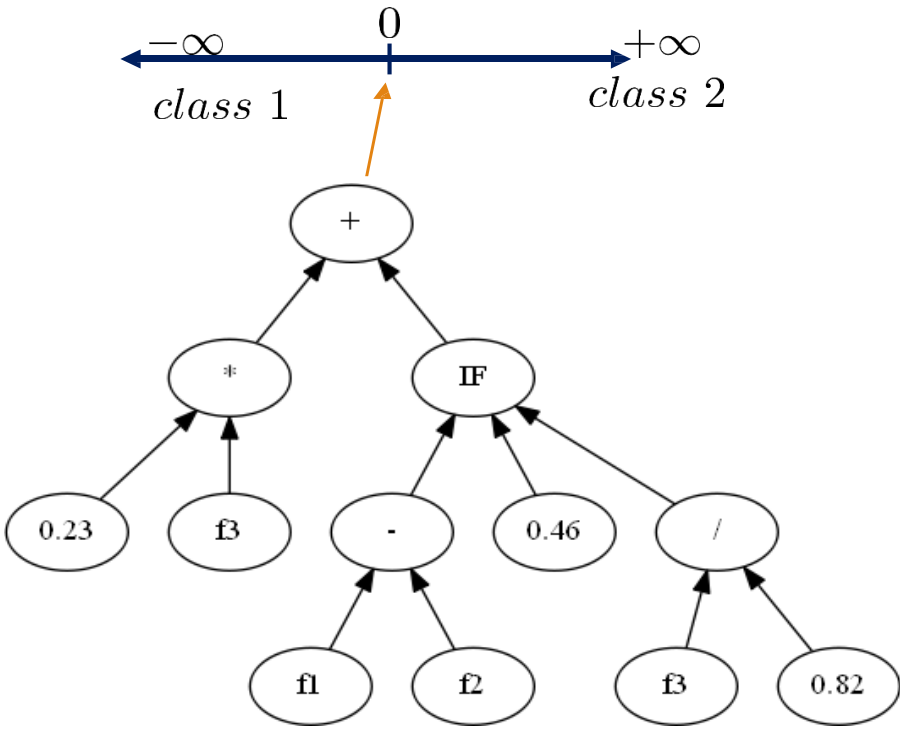
many others



$$y = (A - B)^2$$

GP for Classification

- Fitness function
  - Classification accuracy
  - Error rate
- Classifier
  - Each program is a classifier



$$\texttt{if } \mathrm{ProgOut} < 0 \texttt{ then } class\ 1 \texttt{ else } class\ 2$$

Go through part 2 of A2

What is a good Report?

1. Answer the question **directly** (i.e., marking is based on key points)

   --- good to start with a summary sentence and then explain details

2. Answer/give **what are asked** ("useless information" is not helpful)

   --- source file

3. **Show your thinking** rather than "copying"!!! (e.g., depends…)

4. Proper font size (e.g., clear, especially for A3 and A4)

5. Not necessary to be very long (keep concise and accurate)

What is a good Report?

6. Show your **working process**

7. Do not bury **your effort** (e.g., improvements of assignments)

8. **Visualisation** (e.g., Figures, tables)

9. **Bullet points** to make your report easier to read

10. No ideas about the assignments at all (e.g., show your thinking)

    --- do not hand in a blank

Summary

- ✓ Overview --- GP for regression and classification
- ✓ Go through part 2 of A2
- ✓ An GP example
- ✓ Report

Good luck for your assignments.

See you in the second half of the course.