

ML03: Symbolic Regression: Survival of the Best Fit

Name: Hector Leitch

Title: Programmer I

Company: Veramed Ltd.

Hector has 1 year experience in the industry.

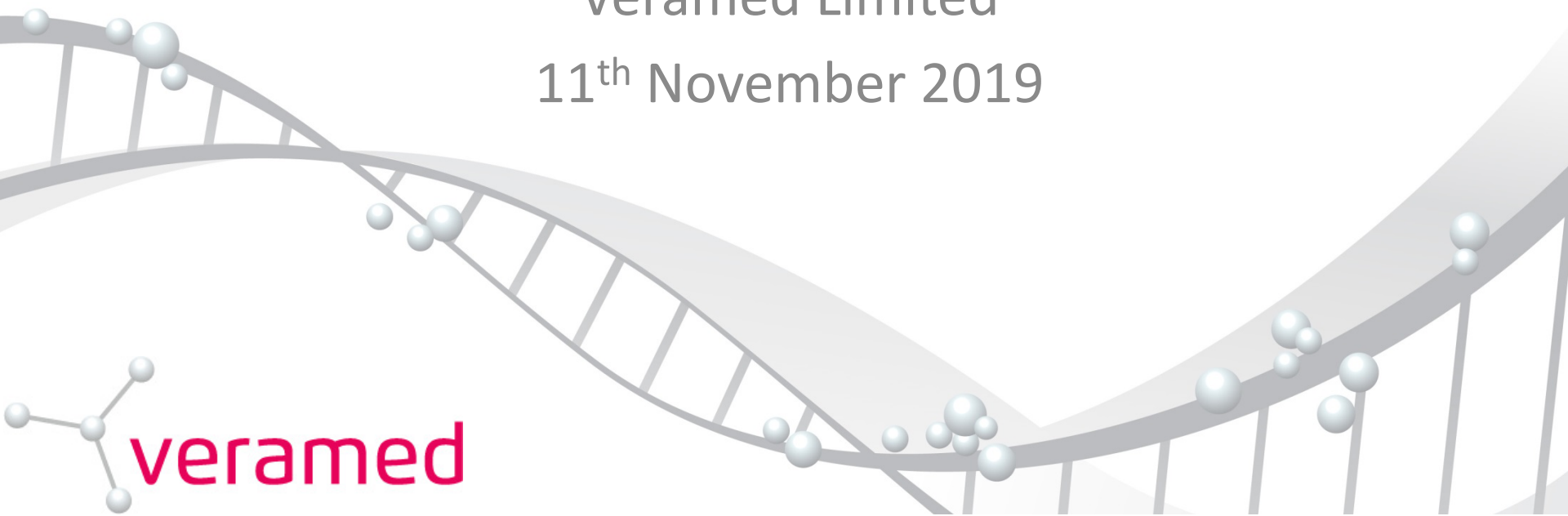
Hector's interests are squash, cricket and music.

Symbolic Regression: Survival of the Best Fit

Hector Leitch

Veramed Limited

11th November 2019

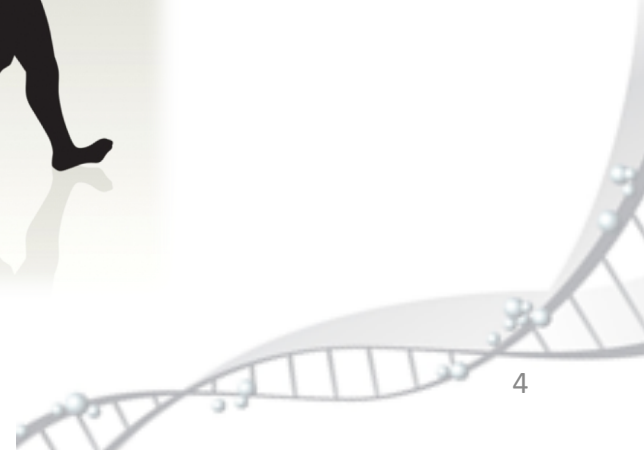
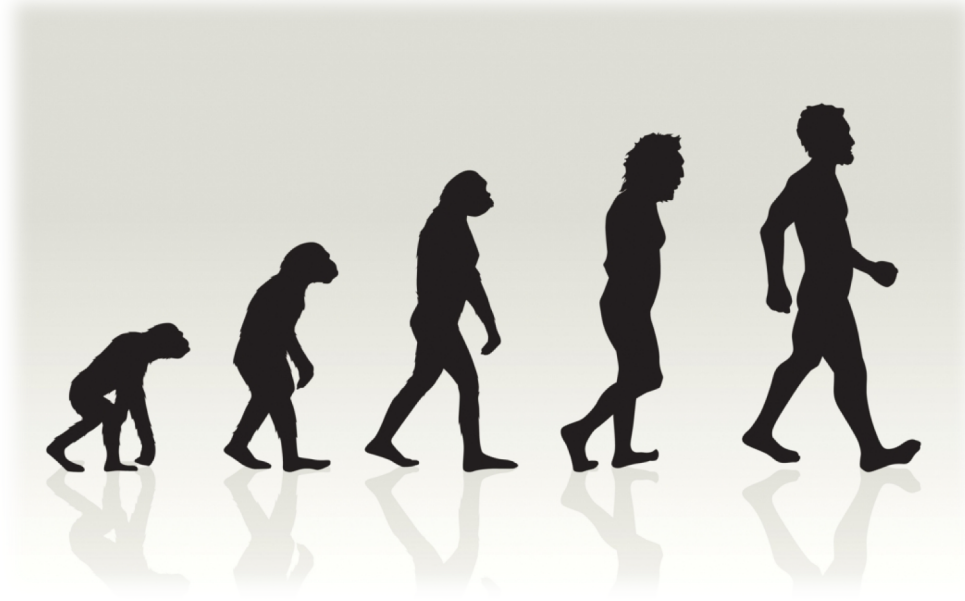


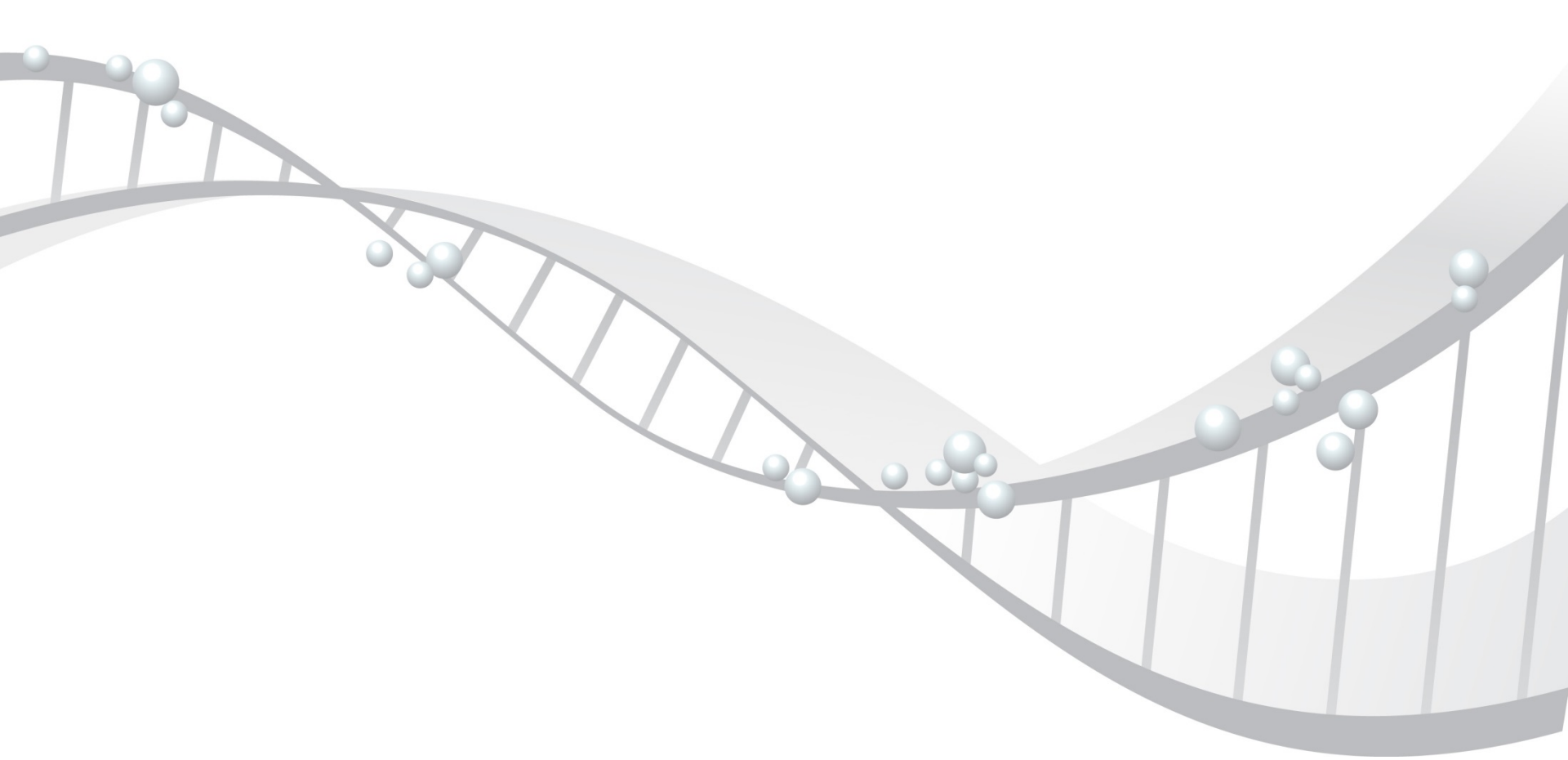
Overview

- What is Symbolic Regression?
- Genetic Programming (GP)
 - Key concepts to be applied to model selection later
- Building a Symbolic Regression (SR) program
 - Attaching familiar ideas to the concepts of GP
- Advantages/Disadvantages
 - How could it be used in the pharmaceutical industry?

What is Symbolic Regression?

- Symbolic Regression is a model selection algorithm that improves a population of models over time using techniques inspired by the theory of natural selection and genetic inheritance

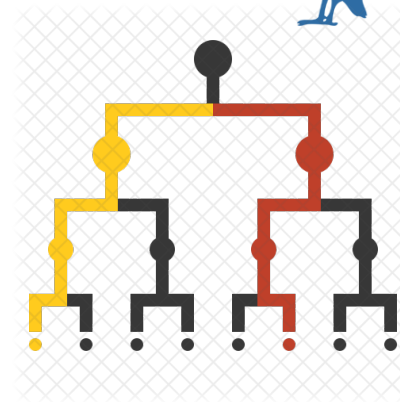
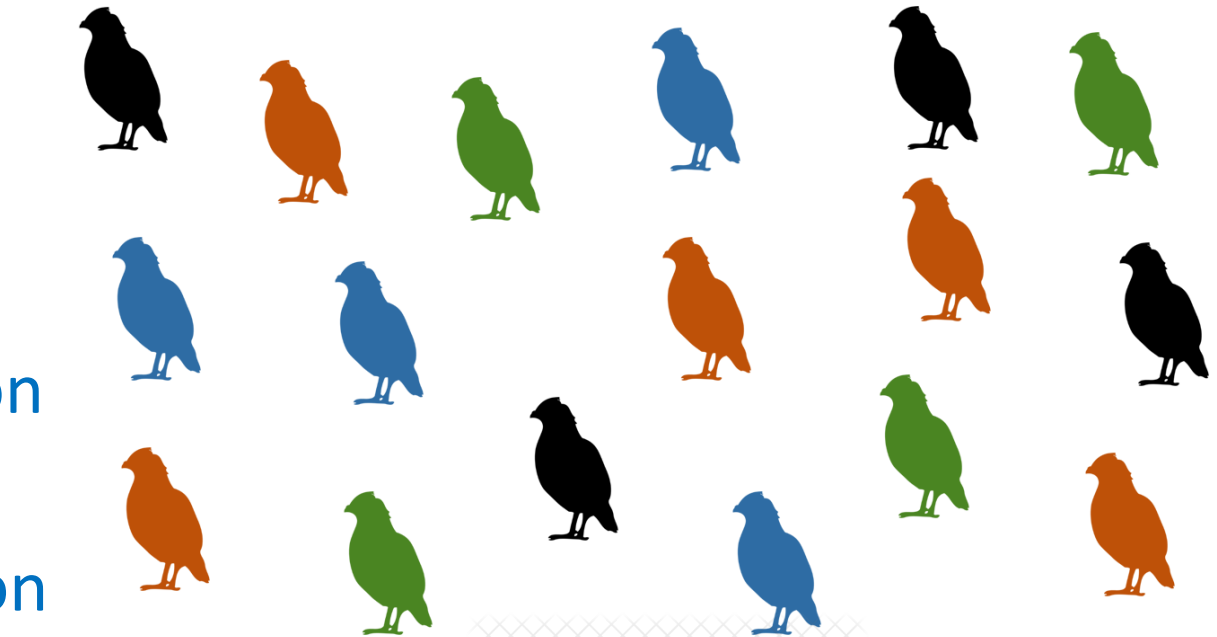




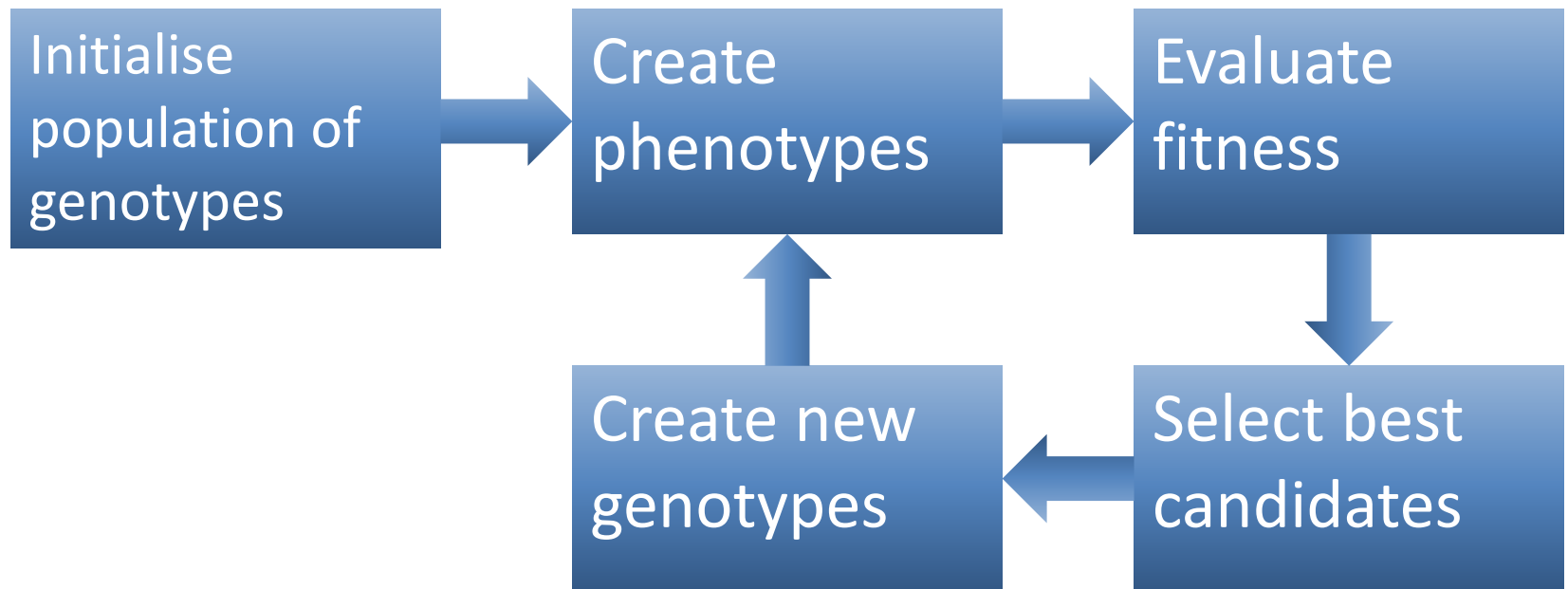
GENETIC PROGRAMMING (GP)

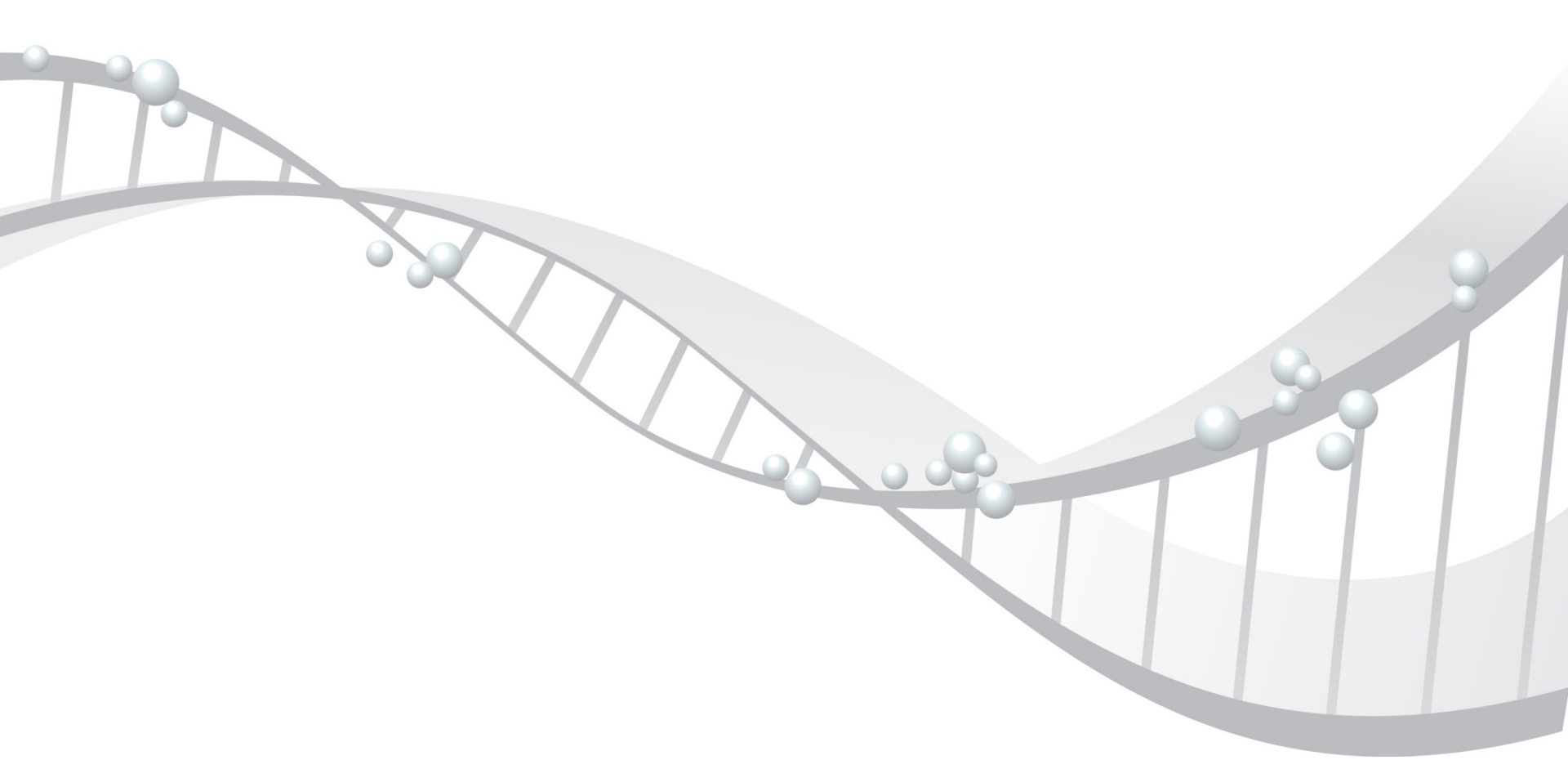
Structure of a Genetic Program

- Population
- Genotype
- Phenotype
- Fitness function
- Selection
- Next generation



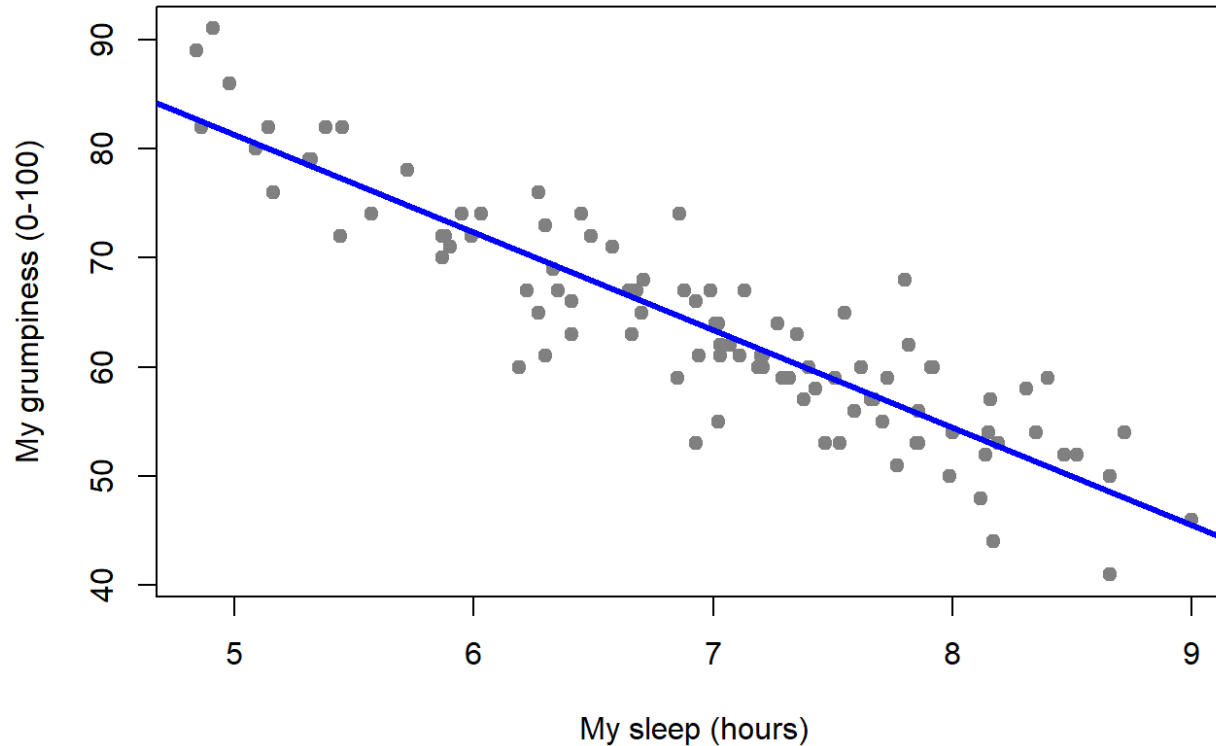
Algorithm





SYMBOLIC REGRESSION (SR)

Linear Regression



$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Stepwise Selection

- Start with full/null model
- Add or remove a single term
- Select best model according to a score function such as RMSE

$$\begin{array}{ccc} & y = \beta_0 + \cdots + \beta_p x_p & \\ \swarrow & & \searrow \\ y = \beta_0 + \cdots + \beta_{p-1} x_{p-1} & & y = \beta_0 + \cdots + \beta_p x_p + \beta_{p+1} x_{p+1} \end{array}$$

- Limited model space

Back to GP

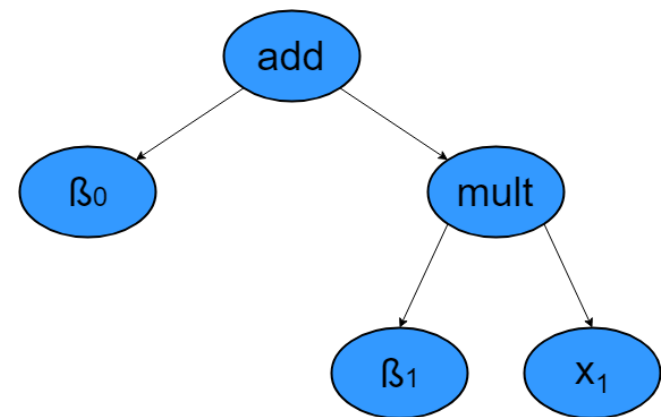
GP Term	Linear Models/Stepwise Selection Term
Genotype	Model formula
Phenotype	Complete model formula with parameter estimates, residuals, etc.
Fitness function	Score function (RMSE) – Smaller is better!

- Creating new genotypes?

Representing a Model Formula

- Choose a set of building blocks including operators (+, -, *, /), analytic functions (exp, sin, arctanh) and parameters/variables
- Represent as a tree or list
- Easy to check for validity
- Easy to evaluate

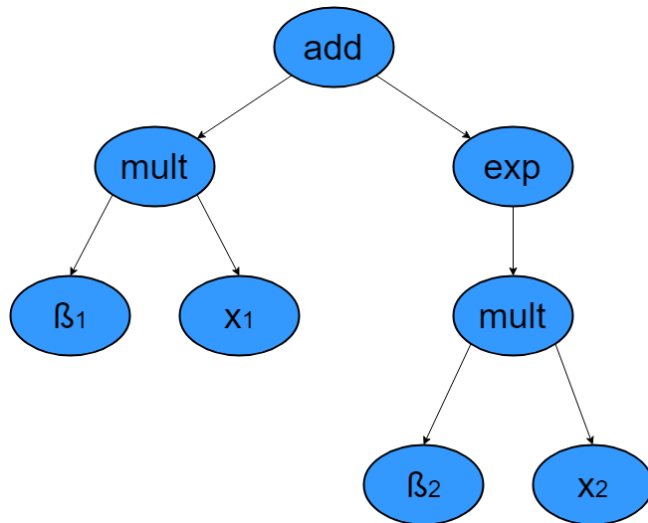
$$y = \beta_0 + \beta_1 x_1$$



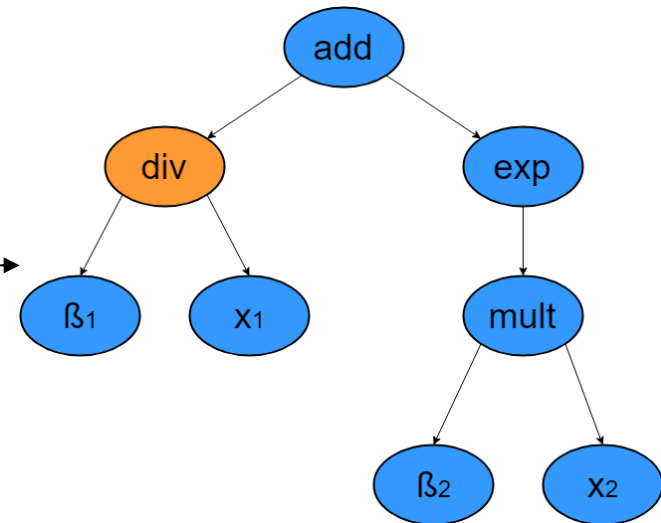
Creating New Genotypes

- Mutation – explore model space

$$y = \beta_1 x_1 + \exp(\beta_2 x_2)$$



$$y = \frac{\beta_1}{x_1} + \exp(\beta_2 x_2)$$

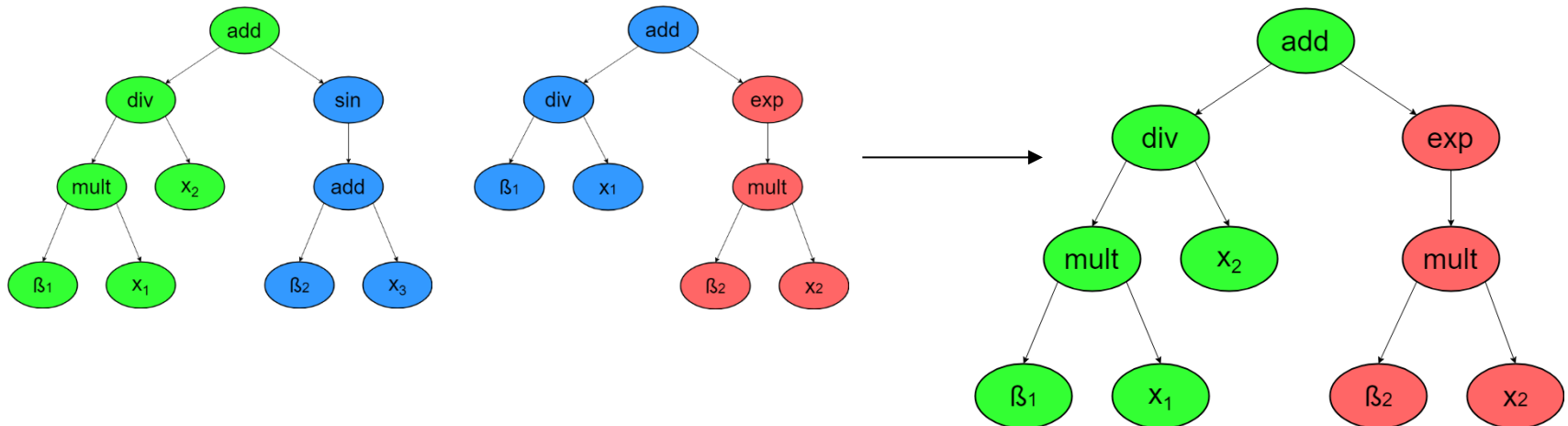


Creating New Genotypes

- Crossover – retain strong features

$$y = \frac{\beta_1 x_1}{x_2} + \sin(\beta_2 + x_3), \quad y = \frac{\beta_1}{x_1} + \exp(\beta_2 x_2)$$

$$y = \frac{\beta_1 x_1}{x_2} + \exp(\beta_2 x_2)$$

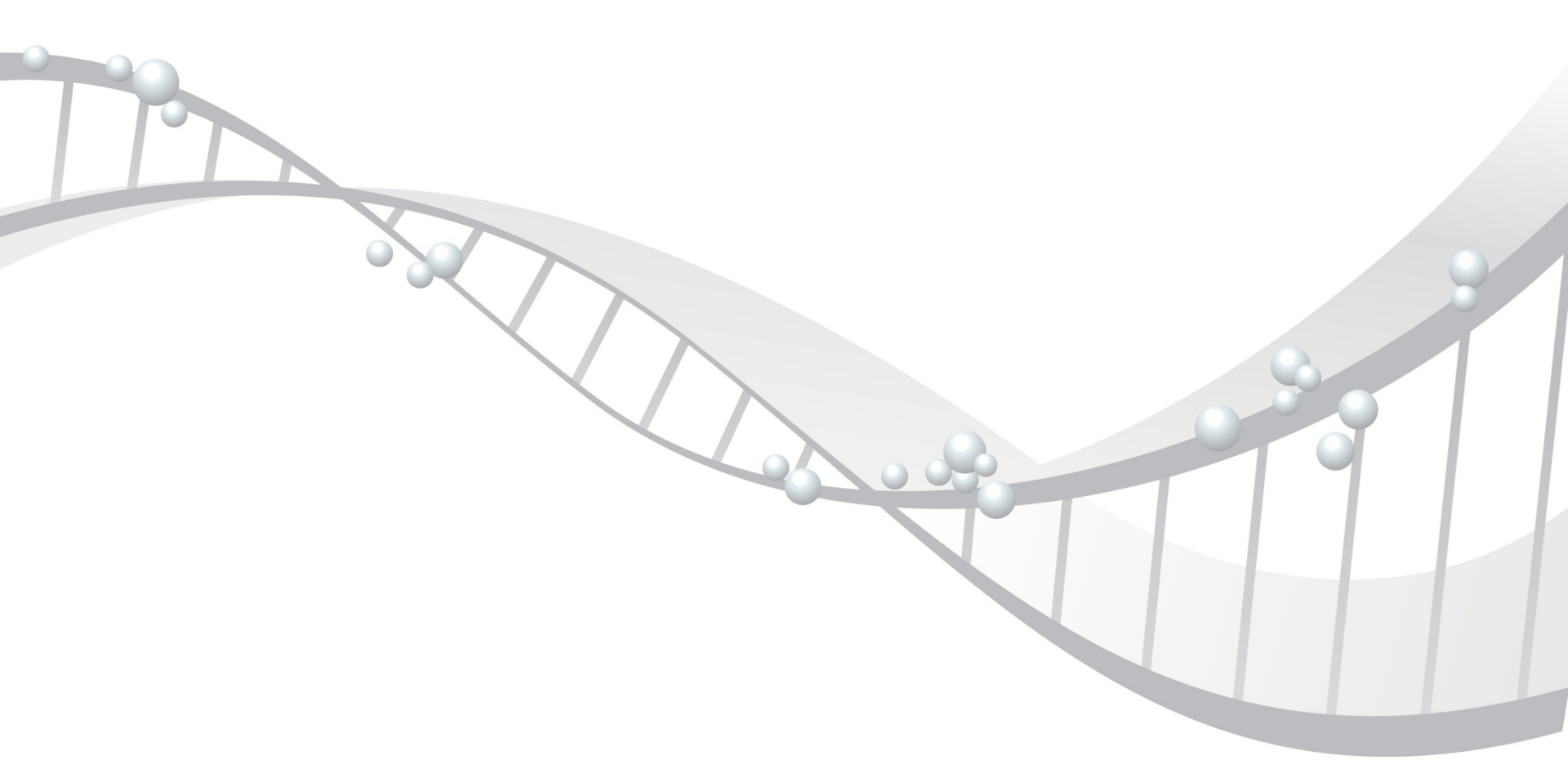


Considerations

- Crossover/mutation balance
 - Retaining features vs. exploration
- Maximum tree depth
 - Limiting the model space improves learning speed and avoids impractical models
- Modifying the fitness function
 - Add a penalty to model complexity

Termination

- SR is an iterative process and therefore requires pre-defined stopping criteria
- Options include:
 - A pre-defined fitness score has been reached
 - A certain number of generations have passed with no further improvement
 - Run for a set number of generations
- Good idea to use a combination of fitness/improvement and generations/run time



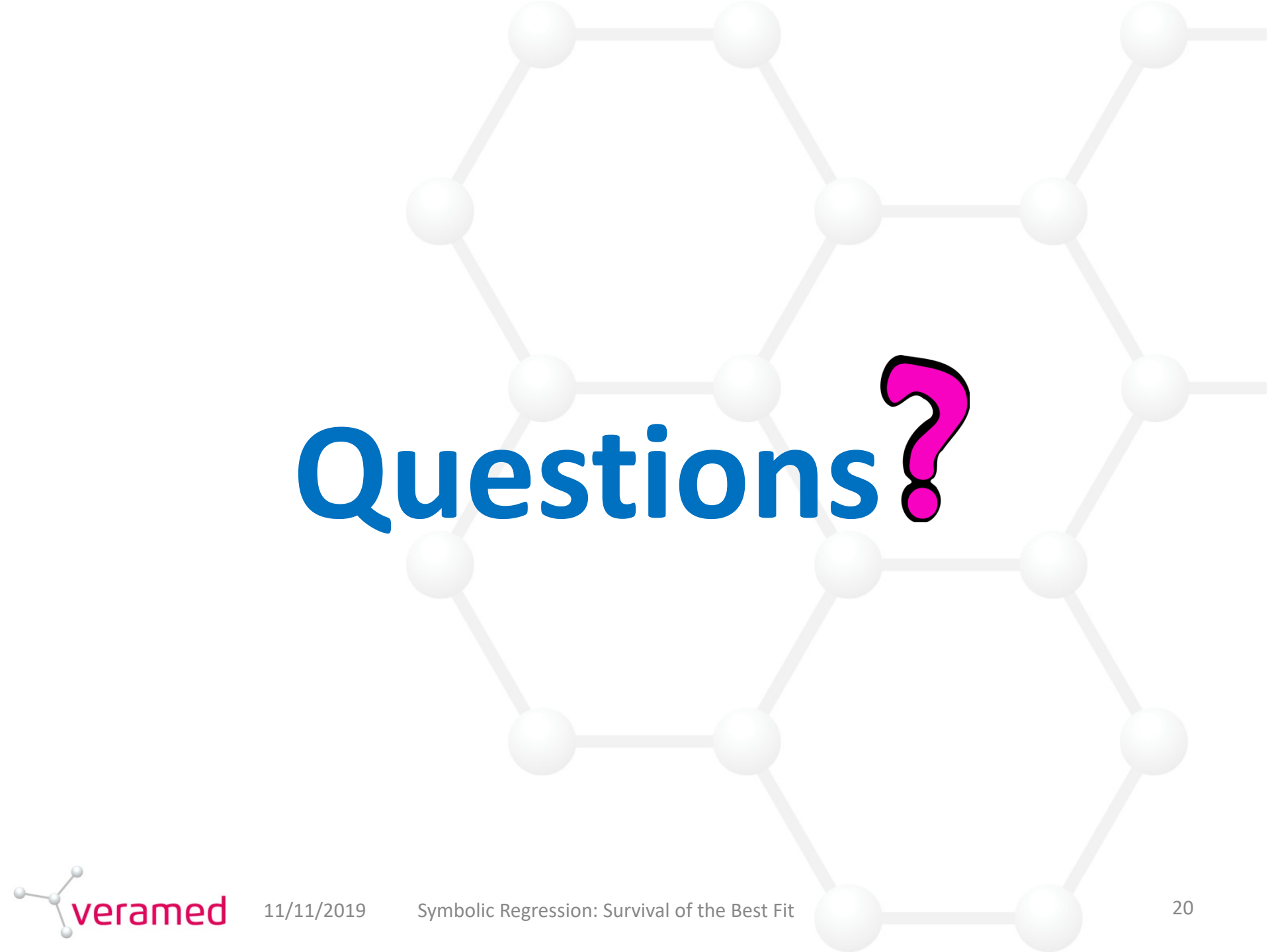
ADVANTAGES, DISADVANTAGES AND USE IN THE PHARMACEUTICAL INDUSTRY

Advantages/Disadvantages

Advantages	Disadvantages
We have the option to examine a whole population of models in addition to the best candidate	Time consuming to program from scratch and learn how to tune
Includes transformations to variables that may explain the data better	Time consuming to run
Offers flexibility through tuning	Not deterministic and output may be difficult to interpret

Use in the Pharmaceutical Industry

- Examples of its use exist already for developing QSAR (Quantitative Structure Activity Relationship) models for toxicity and predictive pharmacokinetic models
- Could be used as a tool for exploratory or adhoc analysis to provide starting points for further investigation
- Software packages are available to try out!



Questions?