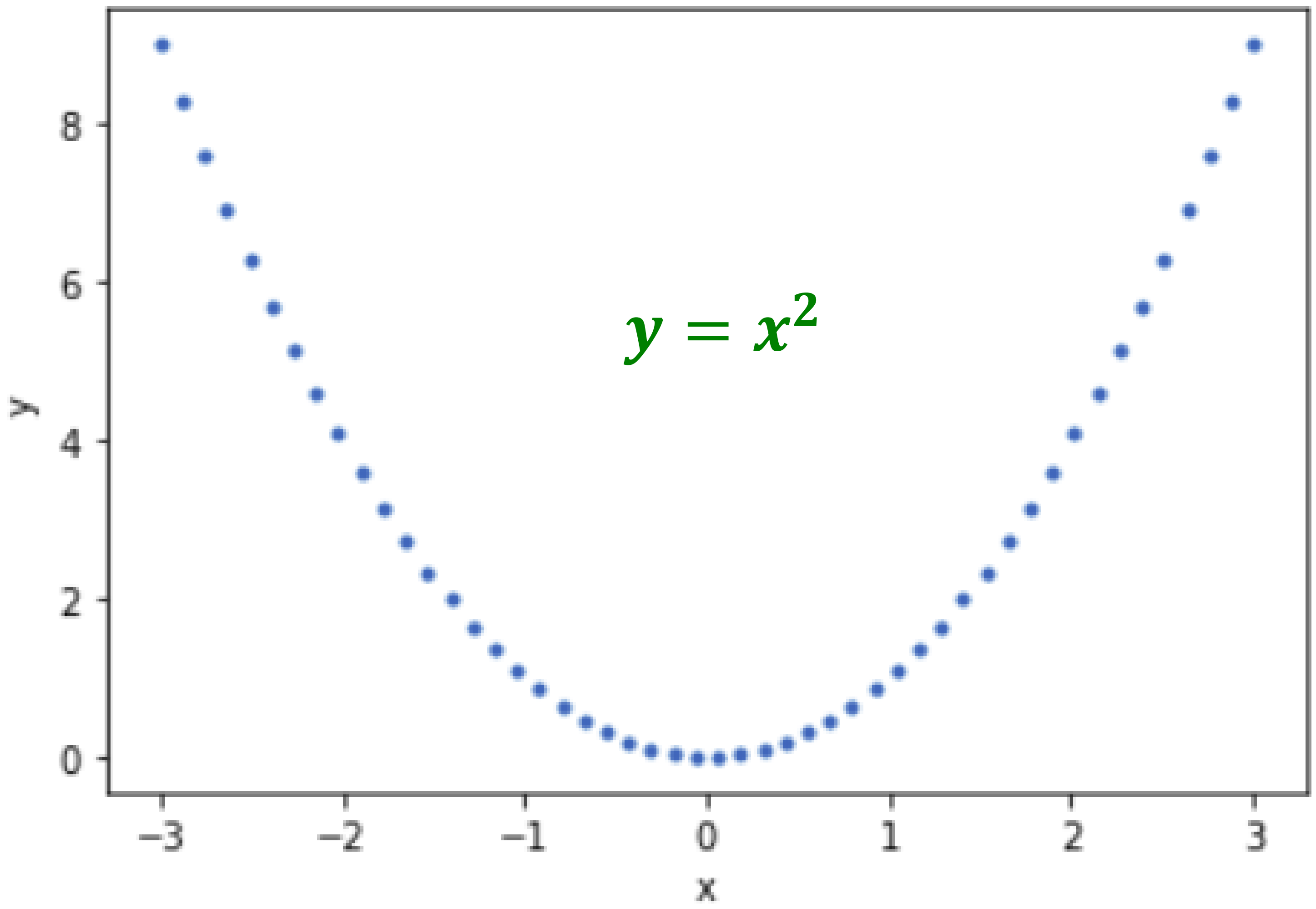
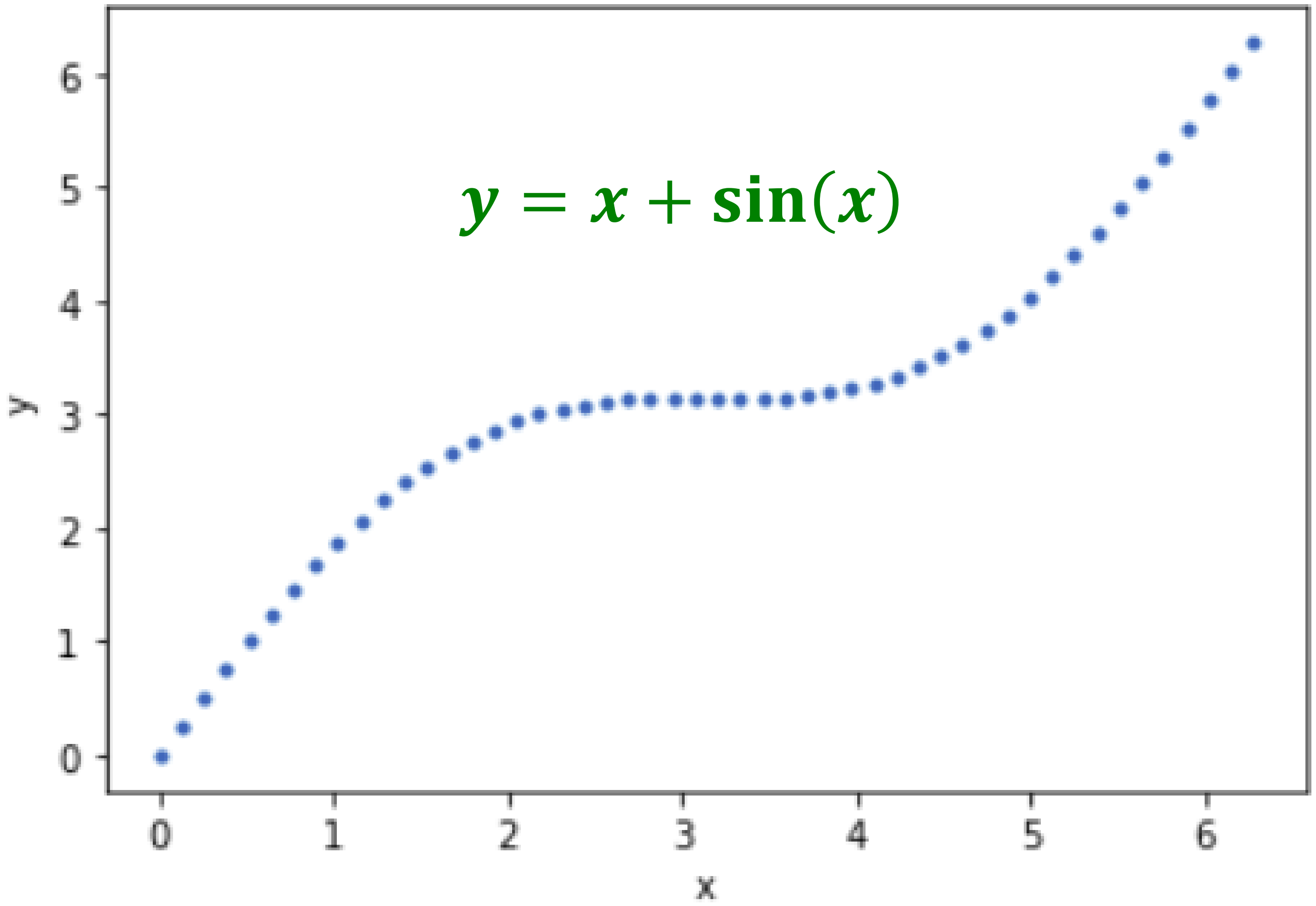
Symbolic Regression with Genetic Programming

Geoffrey Bomarito

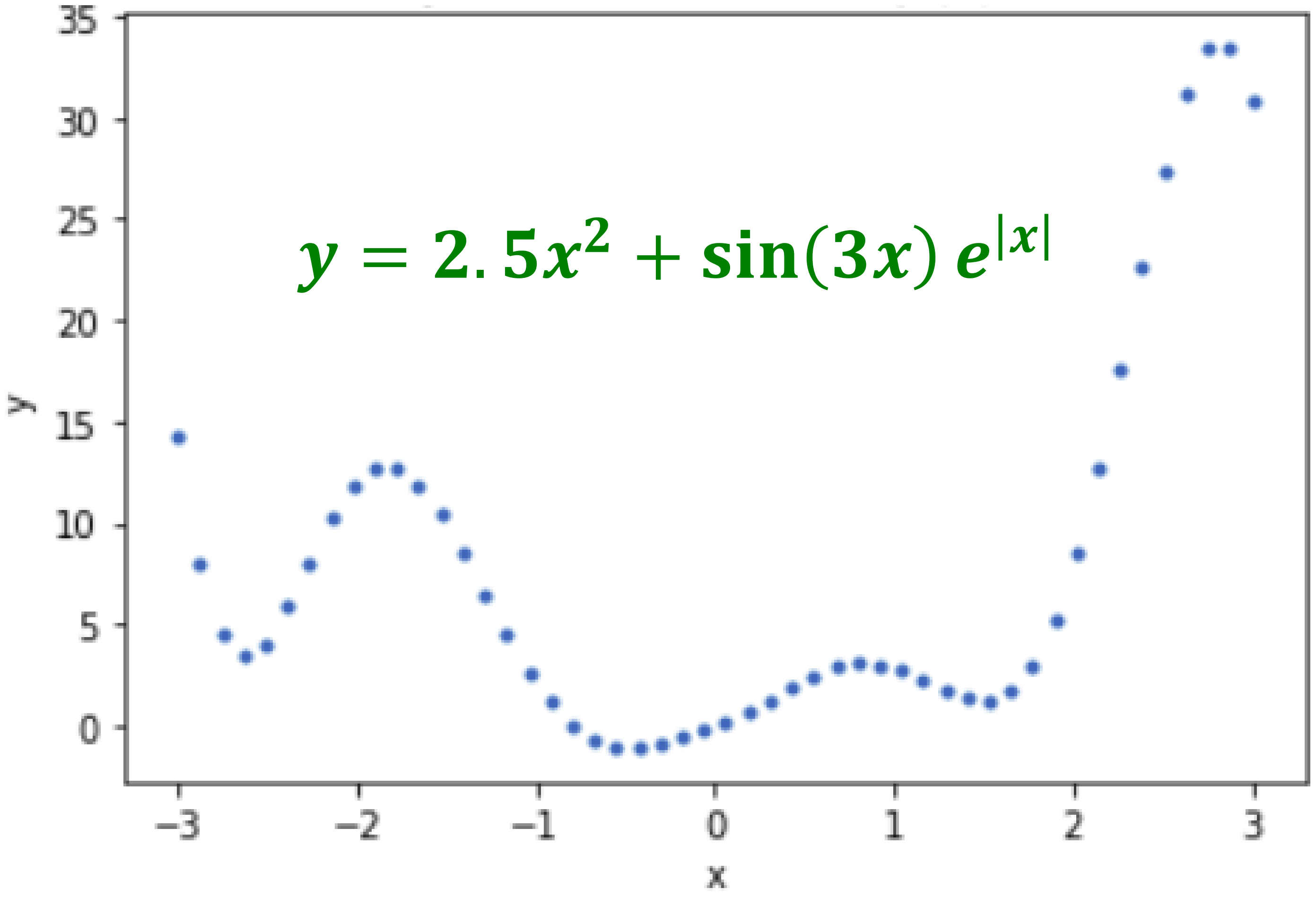
Can you name the equation?



Can you name the equation?



Can you name the equation?



Symbolic regression

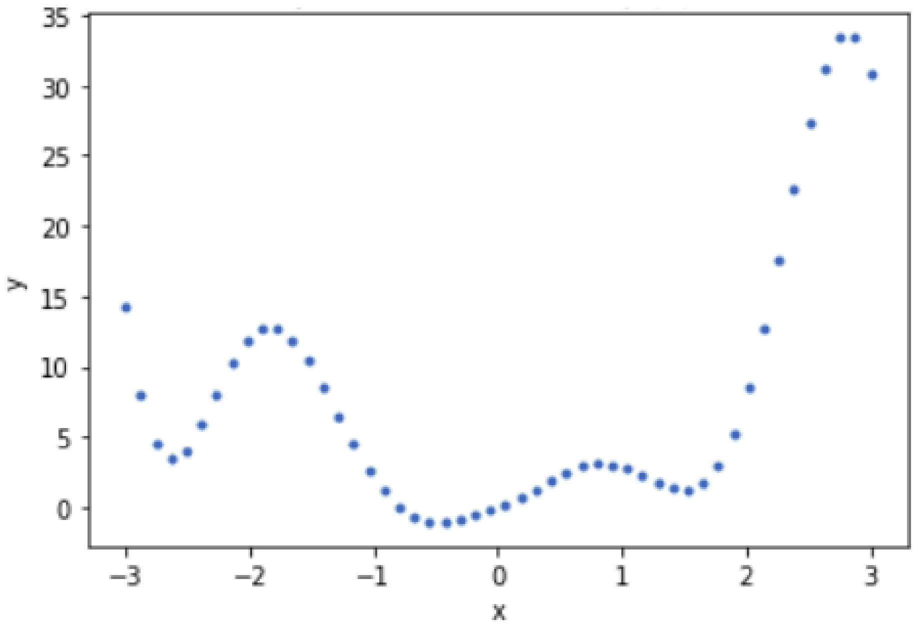
$$y = a_1 x + a_2 x^2 + a_3 x^3 + \ \dots$$

$$y = a_1 \sin(a_2 x) + a_3 x + \ \dots$$

Standard regression

$$y = 2.5x^2 + \sin(3x)\, e^{|x|}$$

Symbolic regression



Genetic Programming (GP)

The evolutionary approach to symbolic regression

- Why GP?
  - Discrete optimization
  - Not the most compute efficient
  - But has seen huge success

- Alternatives:
  - Deep symbolic regression
  - Prioritized grammar enumeration

$$\boldsymbol{y = a_1 \sin(a_2 x) + a_3 x}$$

$$\updownarrow$$

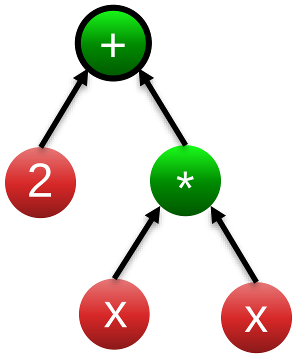
$$\boldsymbol{y = a_1 + a_3 x^2 + \frac{a}{x}}$$

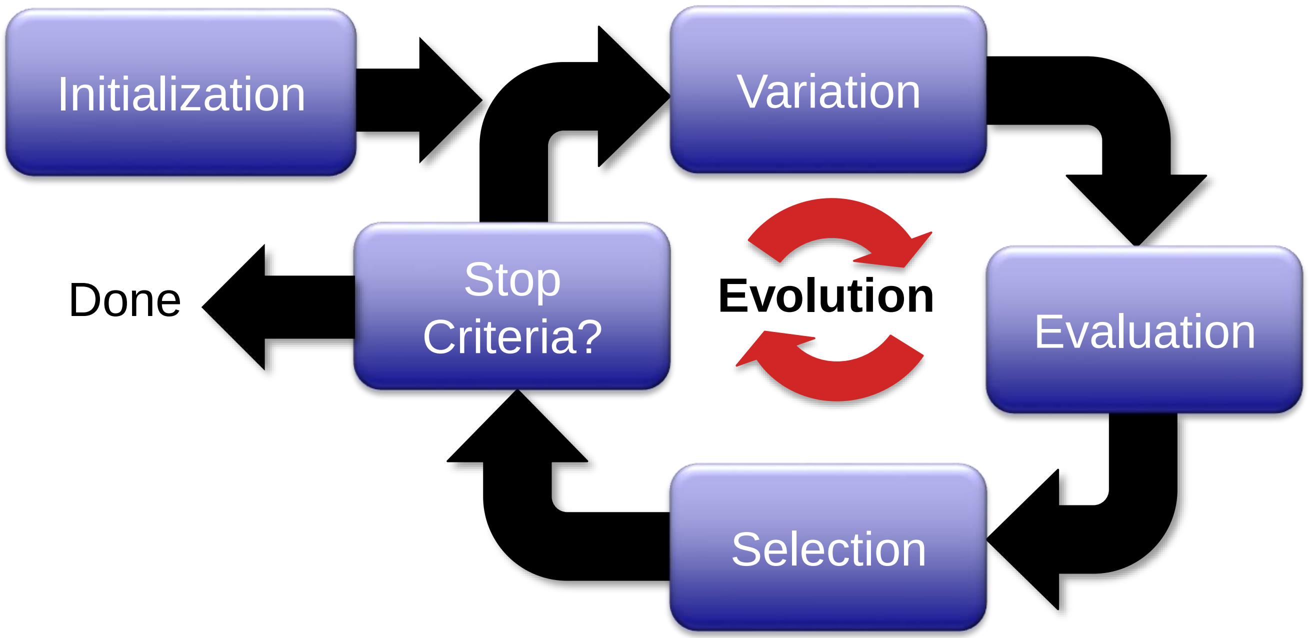
Genetic Programming (GP)

The evolutionary approach to symbolic regression

- Evolving equations to match data
- Equations as graphs

$$2 + x^2$$

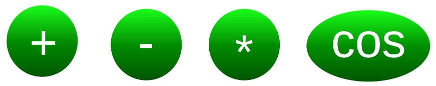




Initialization

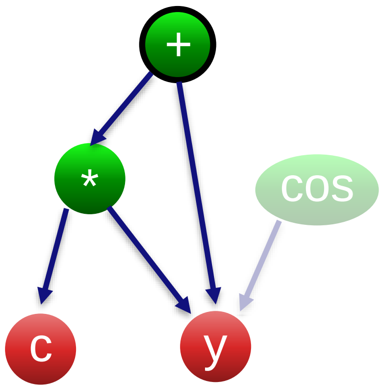
Randomly add nodes (and connections if necessary)

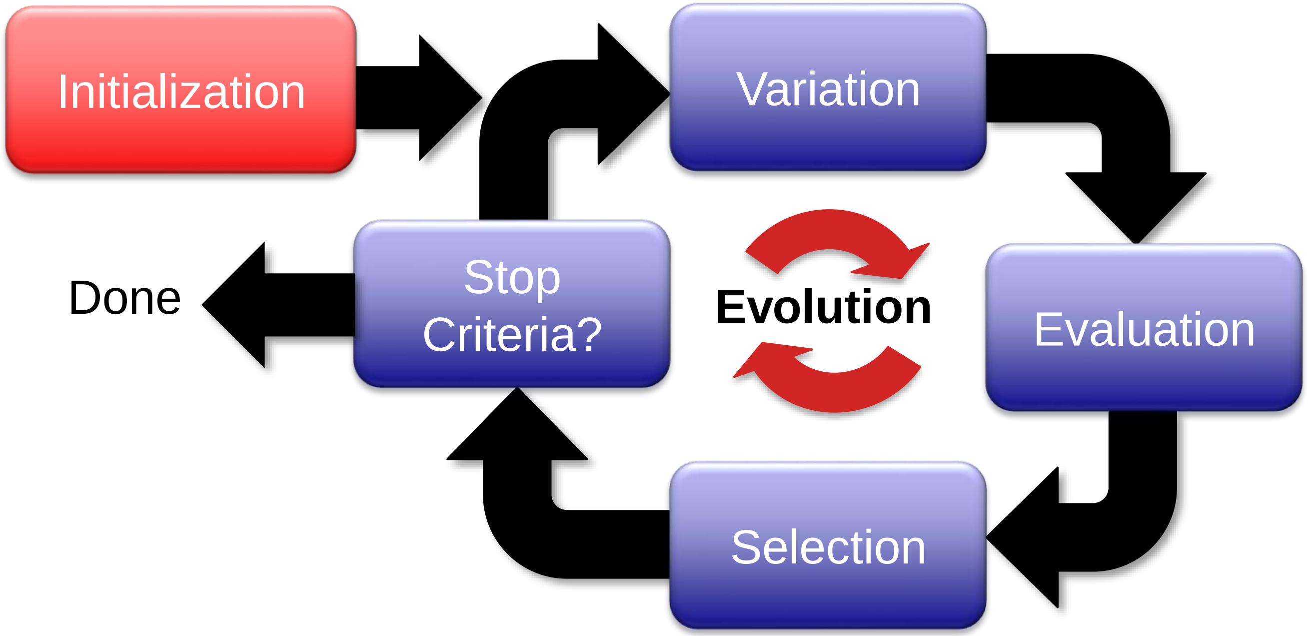
- Set of operators



- Set of terminals

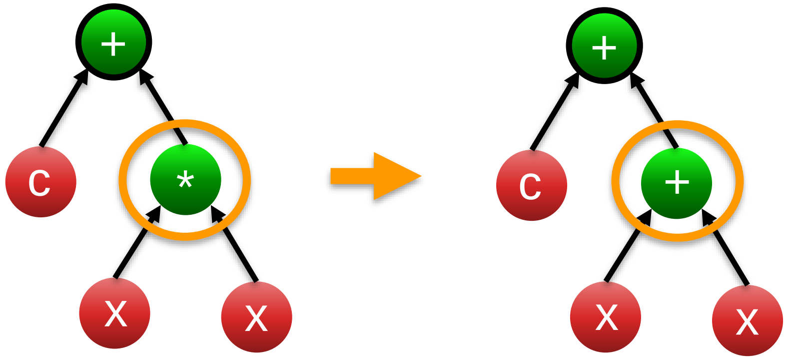
$$cy + y$$



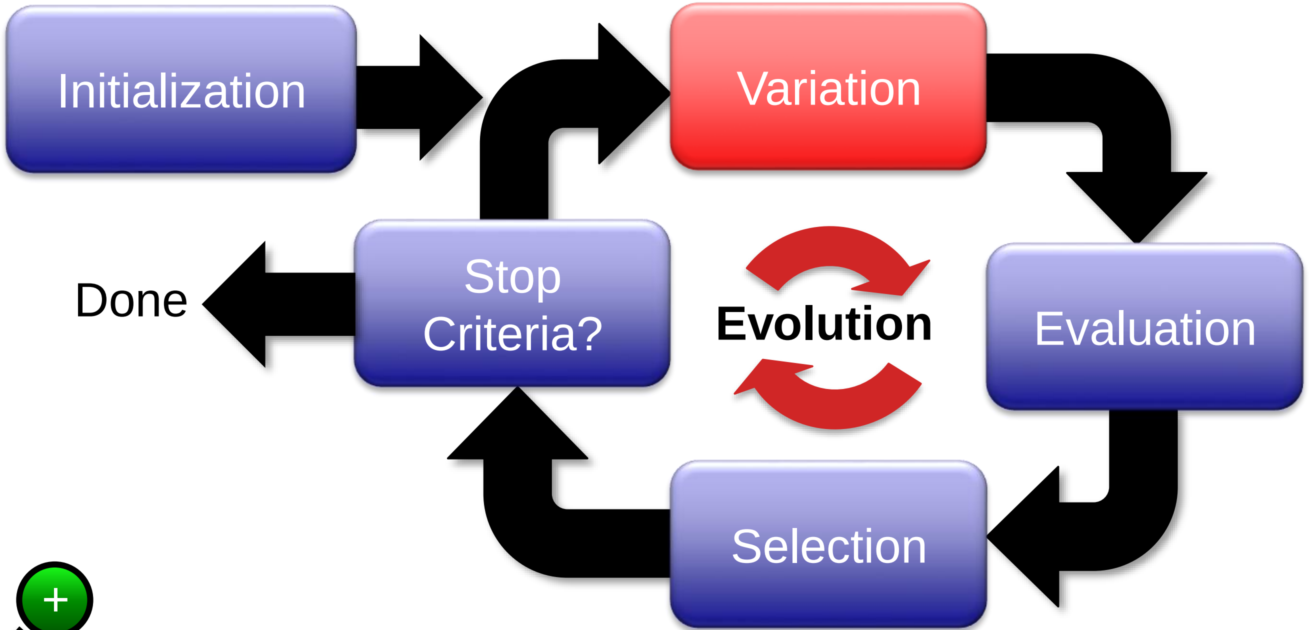


Variation

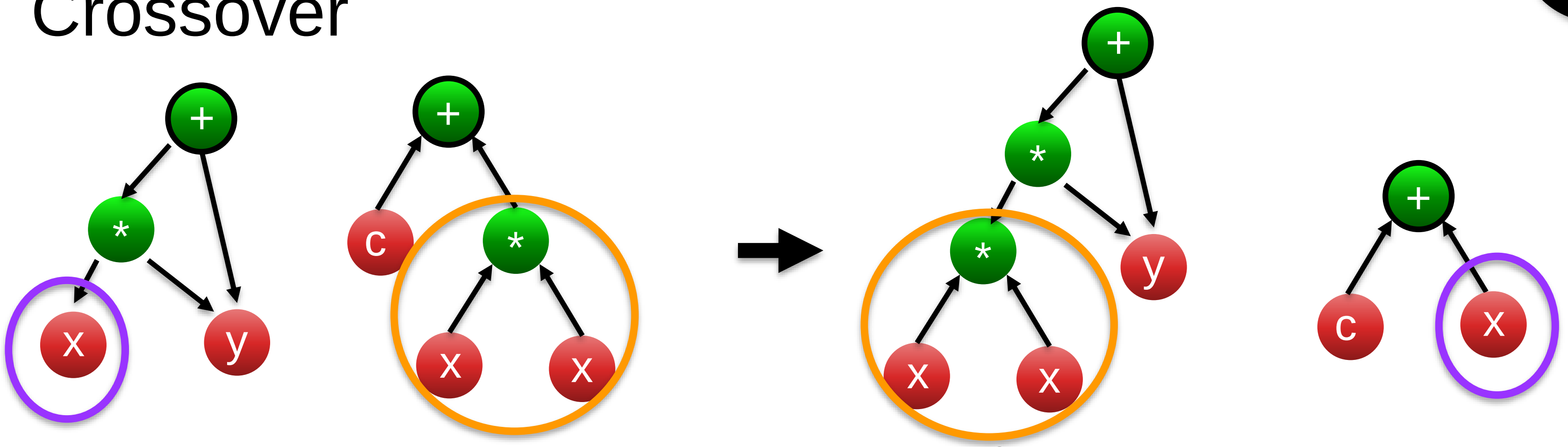
Mutation



$c + x^2$ $\quad$ $c + 2x$



Crossover



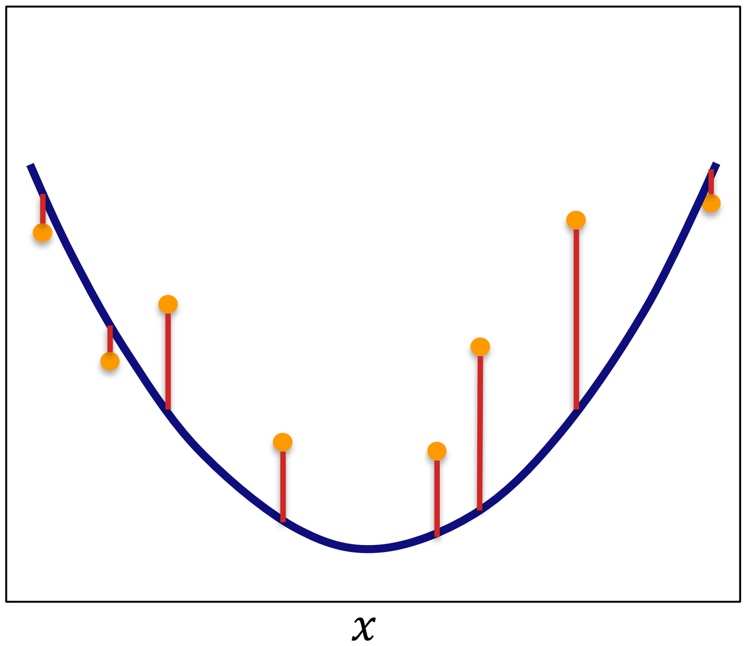
$xy + y$ $\quad$ $c + x^2$ $\quad$ $x^2y + y$ $\quad$ $c + x$

Evaluation

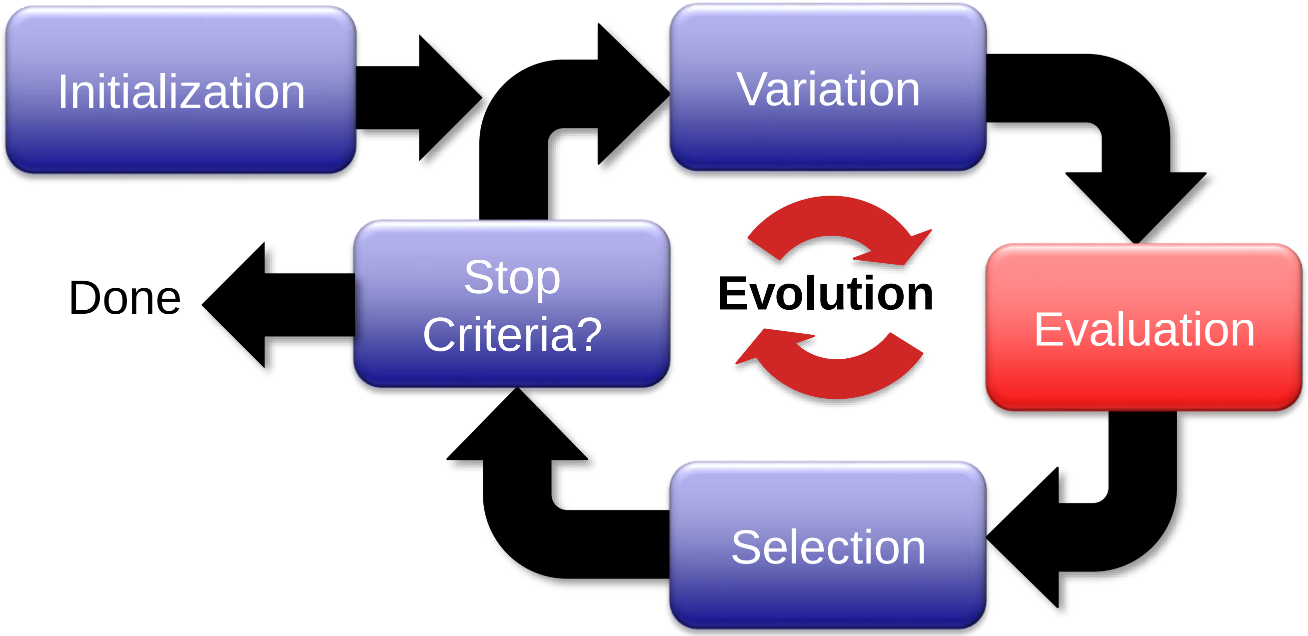
Fitness: How well the equation matches the training data

Explicit Regression

$c + x^2$



Mean absolute error
Mean squared error
Etc.

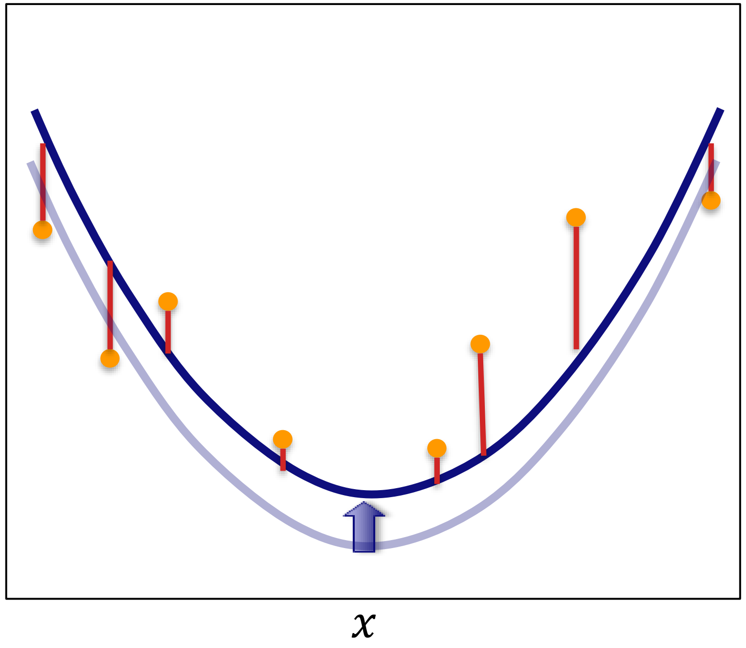


Evaluation

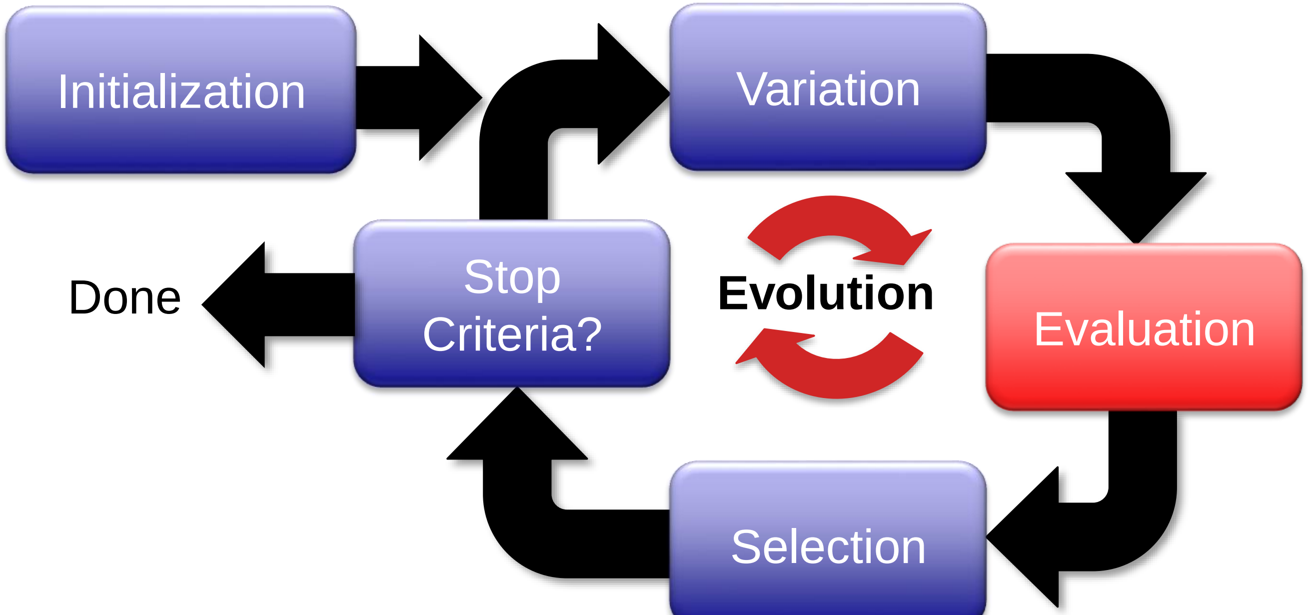
Fitness: How well the equation matches the training data

<u>Explicit Regression</u>

$$c + x^2 \Rightarrow 2.7 + x^2$$



Local optimization performed for constants

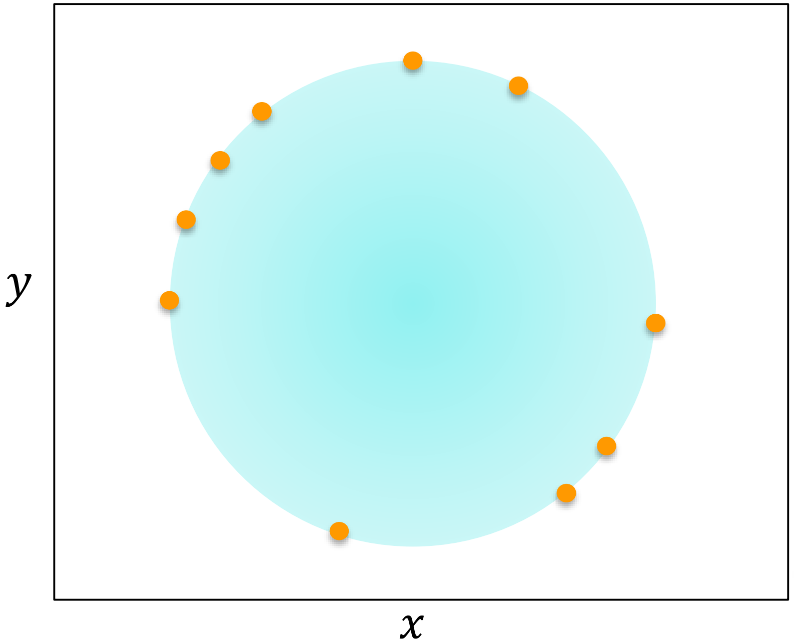


Evaluation

Fitness: How well the equation matches the training data

Implicit Regression

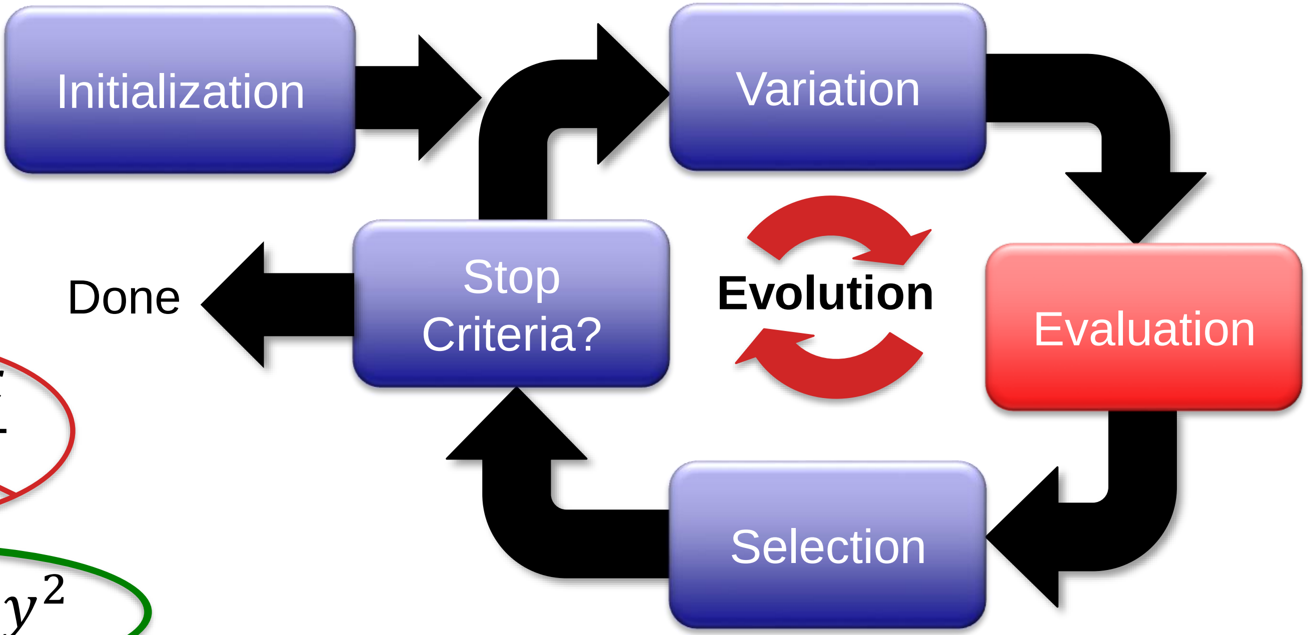
Searching for invariants in the data



$$\cancel{f = 0}$$

$$\cancel{f = \frac{x + x}{x}}$$

$$f = x^2 + y^2$$

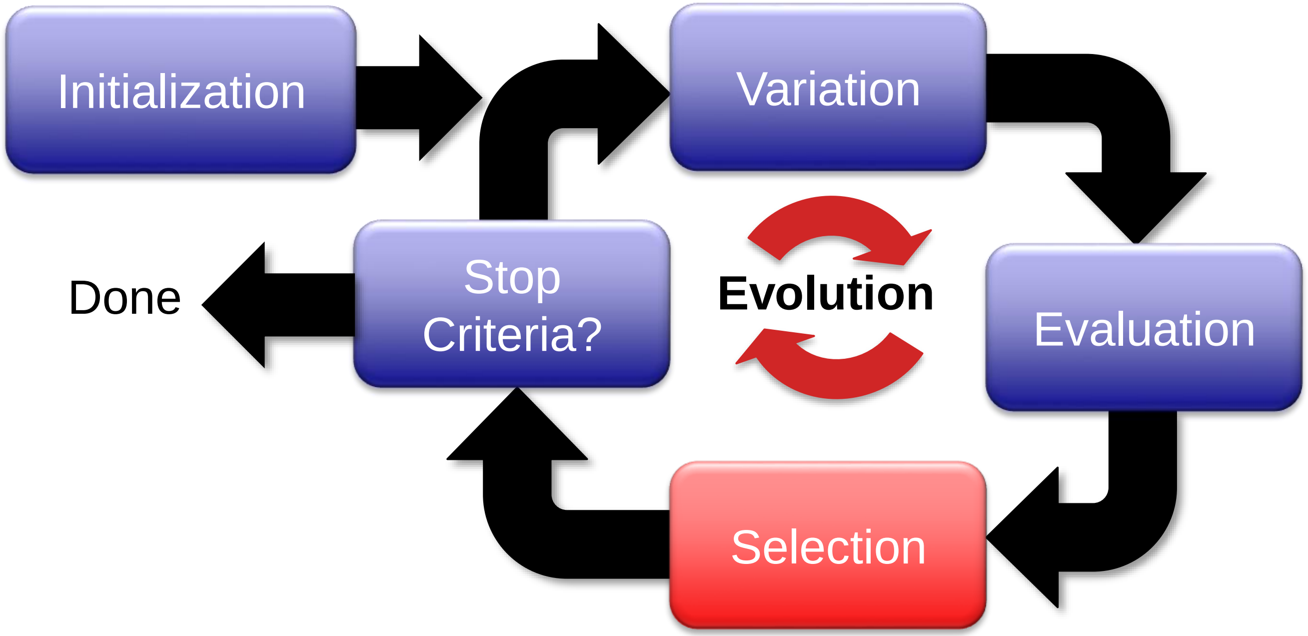


$$f(x, y) = constant$$

Selection

Survival of the fittest

- Tournament
  - Randomly selected groups compete
  - Most common
- Deterministic crowding
  - Children compete against parents
  - Promotes niching
- Age-fitness
  - Equations compete against other similarly aged equations
  - Mitigates premature convergence



Stopping criteria

Is evolution done?

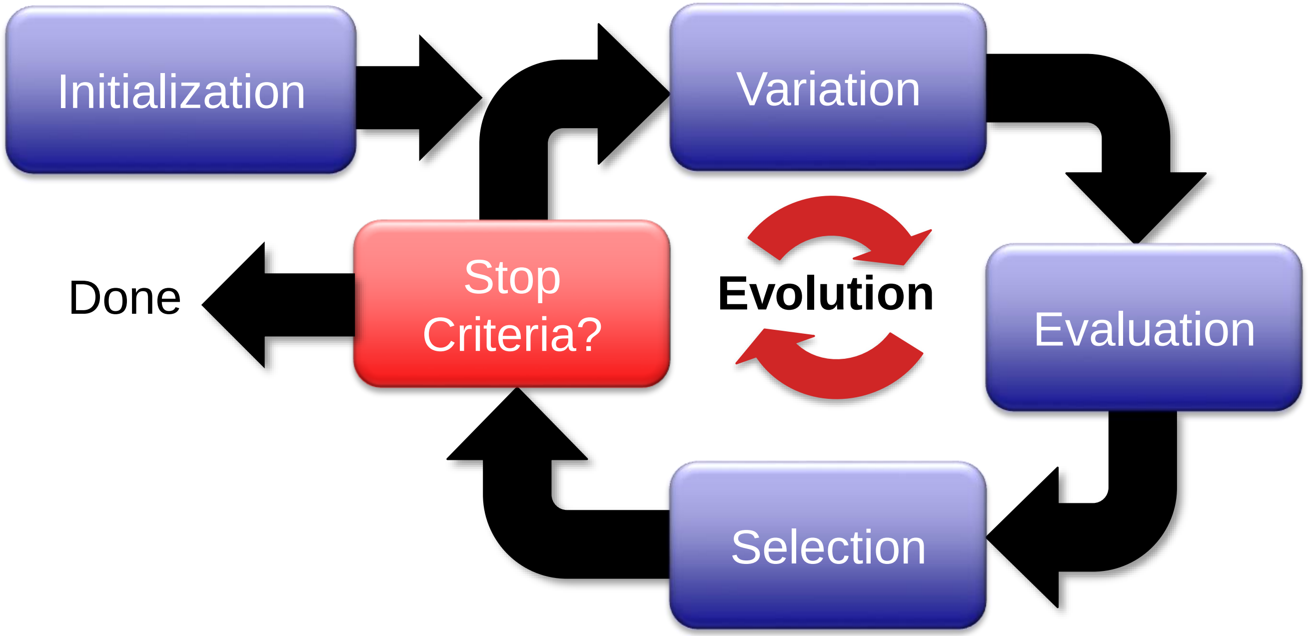
- Convergence
- Stagnation
- Time limit

NO:

repeat

YES:

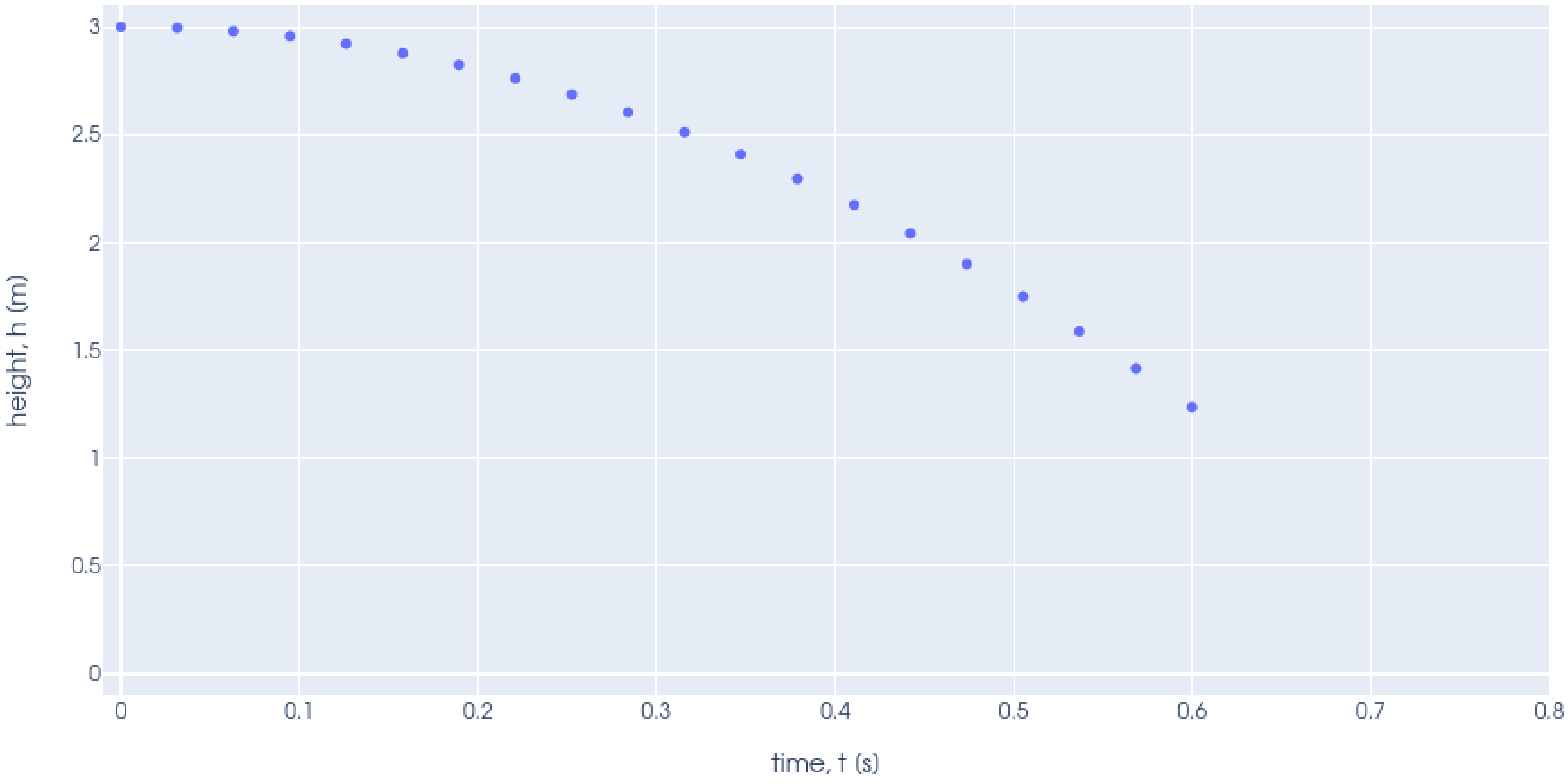
results in equation(s) that best match data



Live Demo

Apple dropping example

Explicit Regression

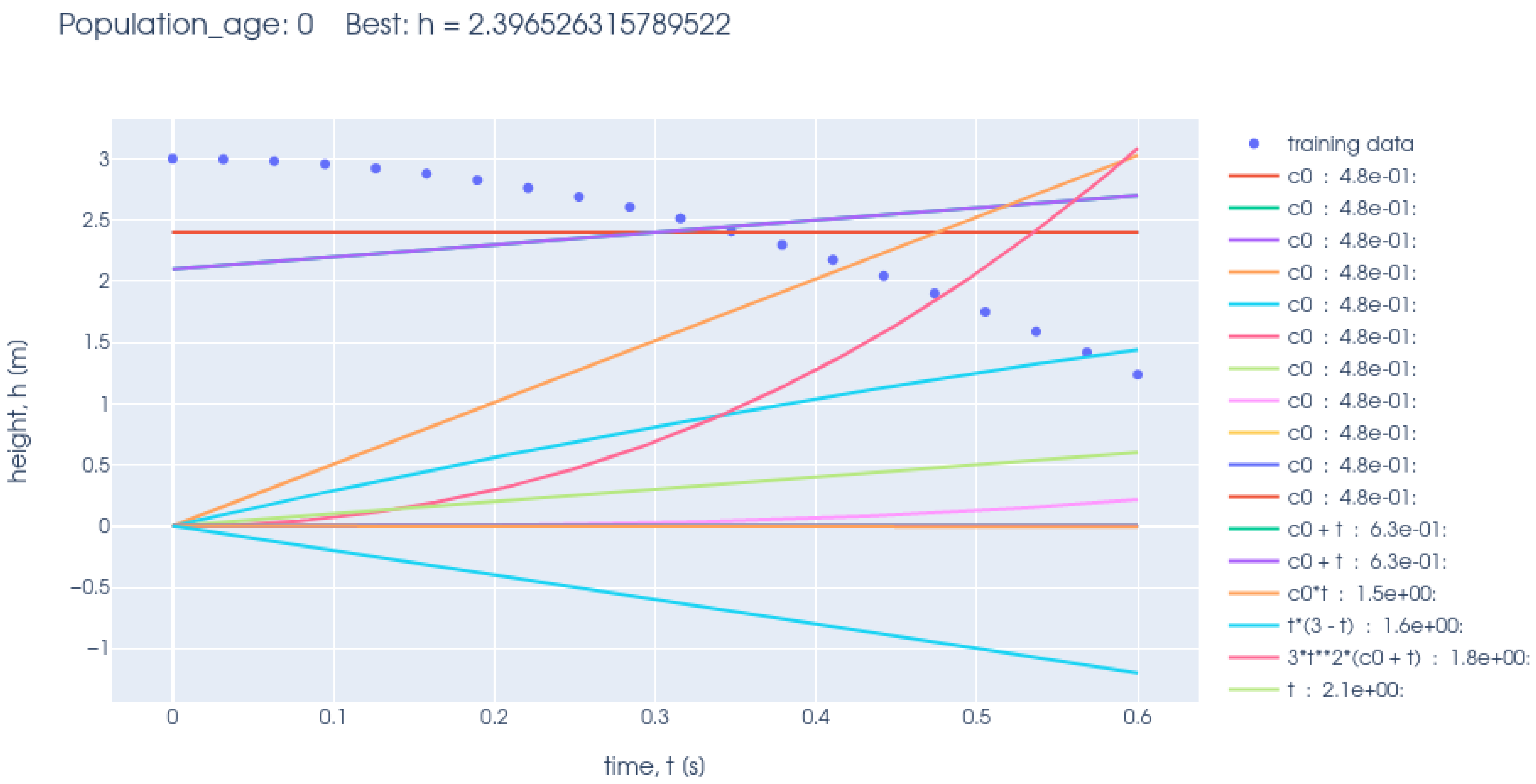


Initial Population

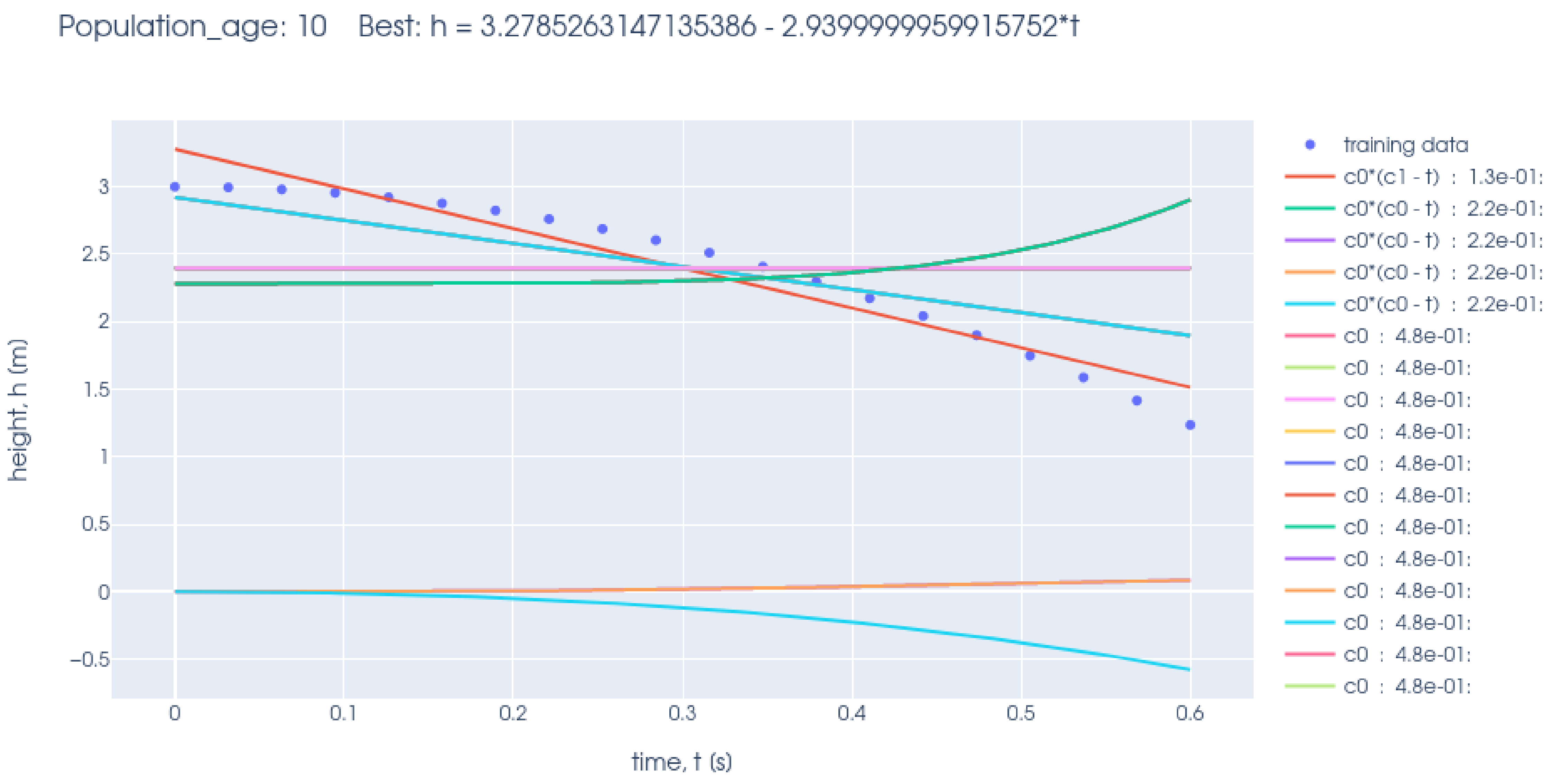
Inputs: Variable: t
Operators: +, -, *

Training data from previous slide

```
0: t**5*(t - 1)*(2*t - 1)
1: c0
2: c0
3: c0
4: c0 + t
5: 0
6: c0
7: c0 + t
8: 0
9: 0
10: t**3
11: c0*t
12: c0
13: t
14: c0
15: 0
16: 0
17: c0
18: c0
19: t*(3 - t)
20: c0
21: 3*t**2*(c0 + t)
22: c0
23: -2*t
24: c0
```



Population Evolution

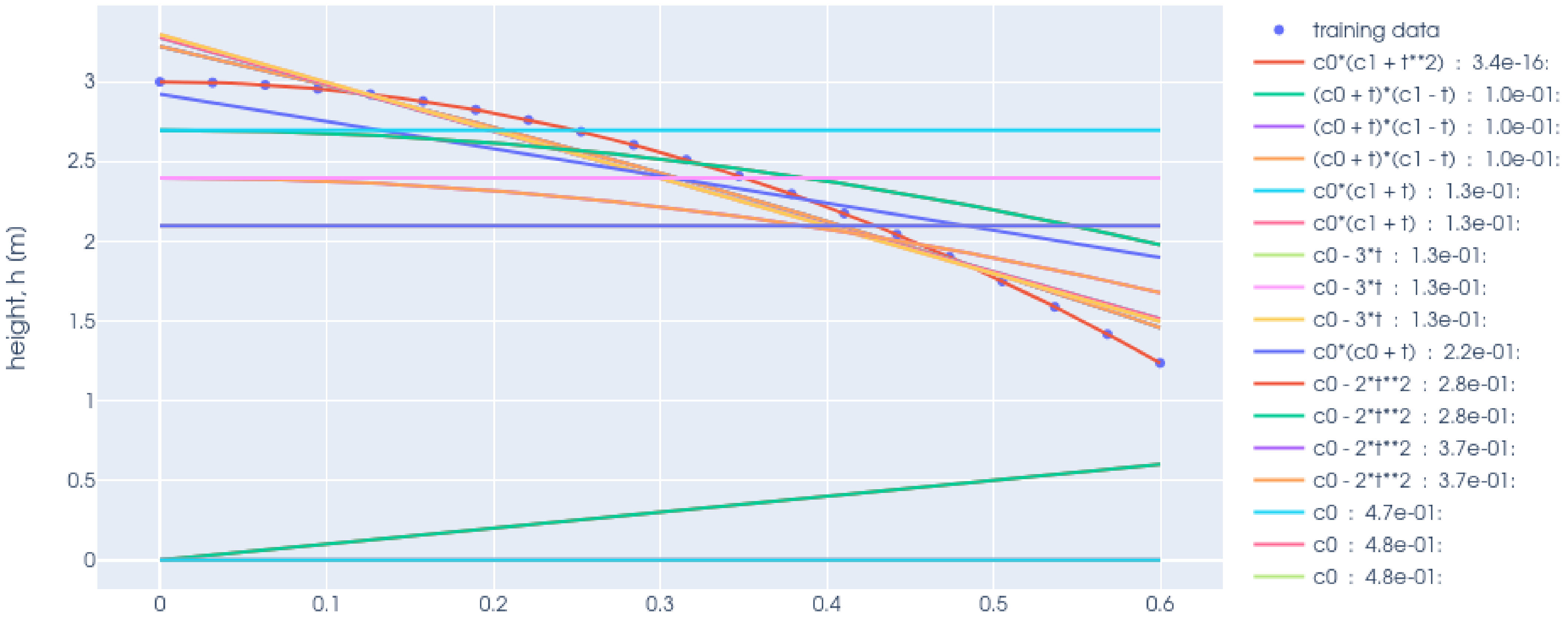


Population Evolution

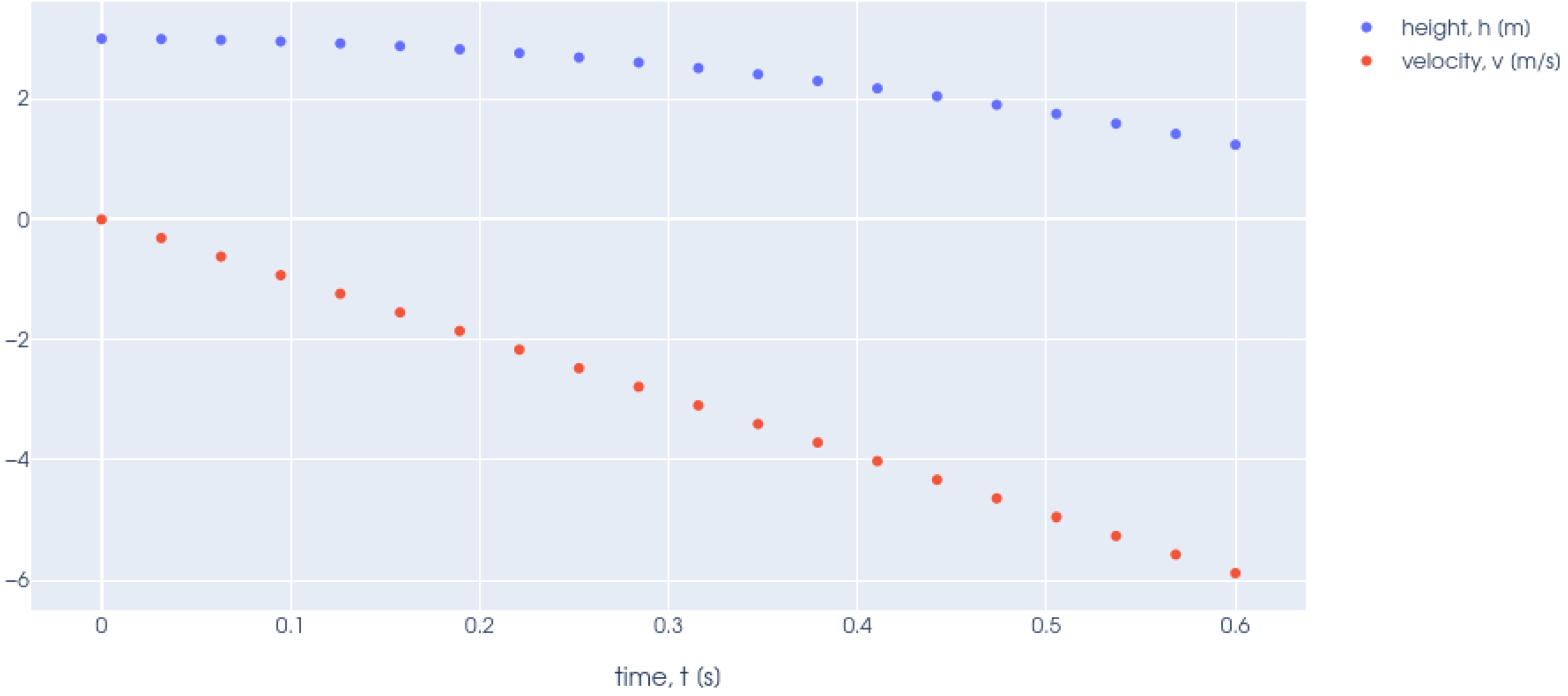
$$h_0 \qquad \frac{1}{2}g$$

Population_age: 187   Best: h = 3.0 - 4.900000000000001*t**2

$$h = h_0 + \frac{1}{2}gt^2$$



Implicit Regression

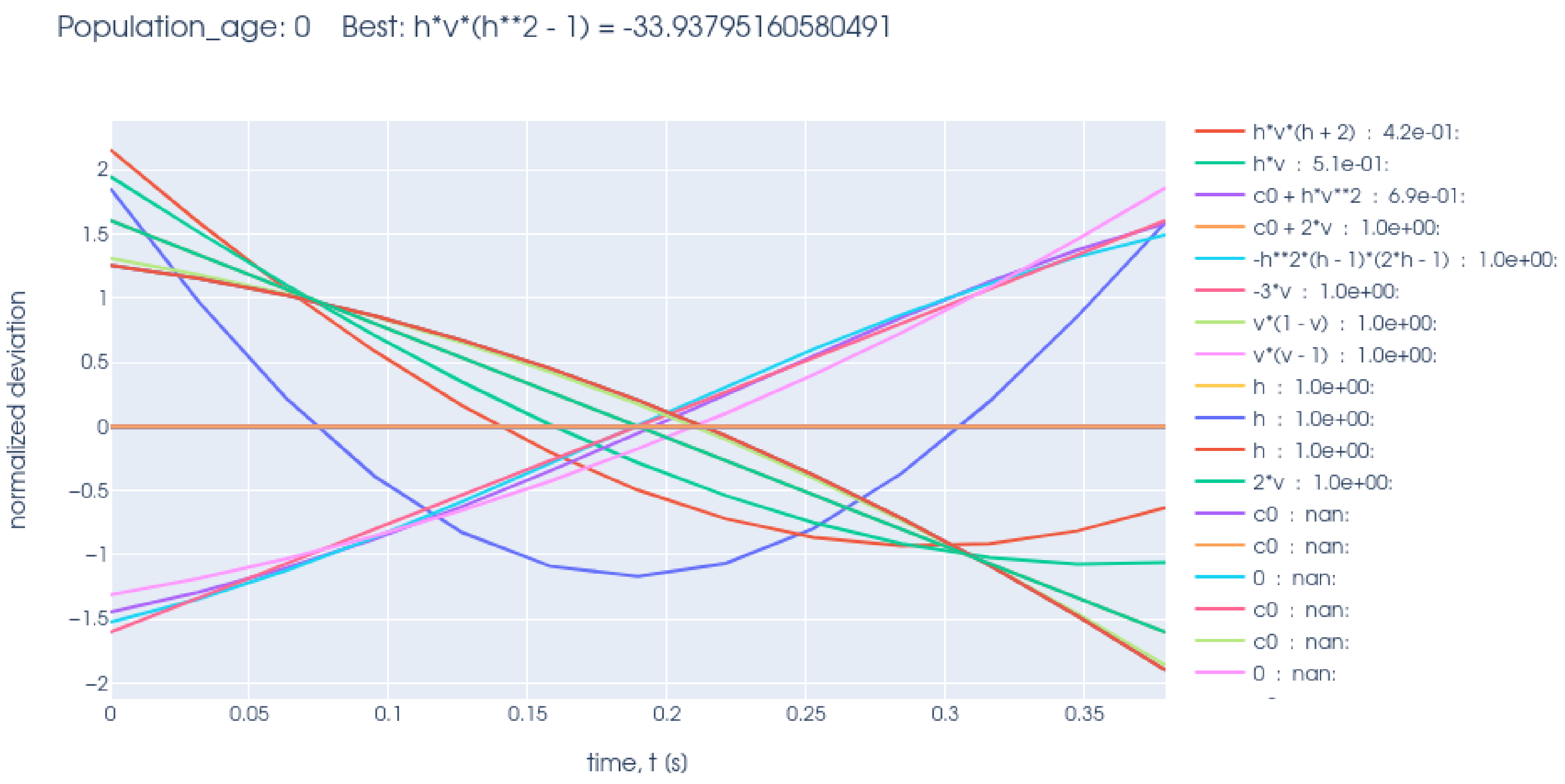


$$f(h, v) = constant$$

Initial Population

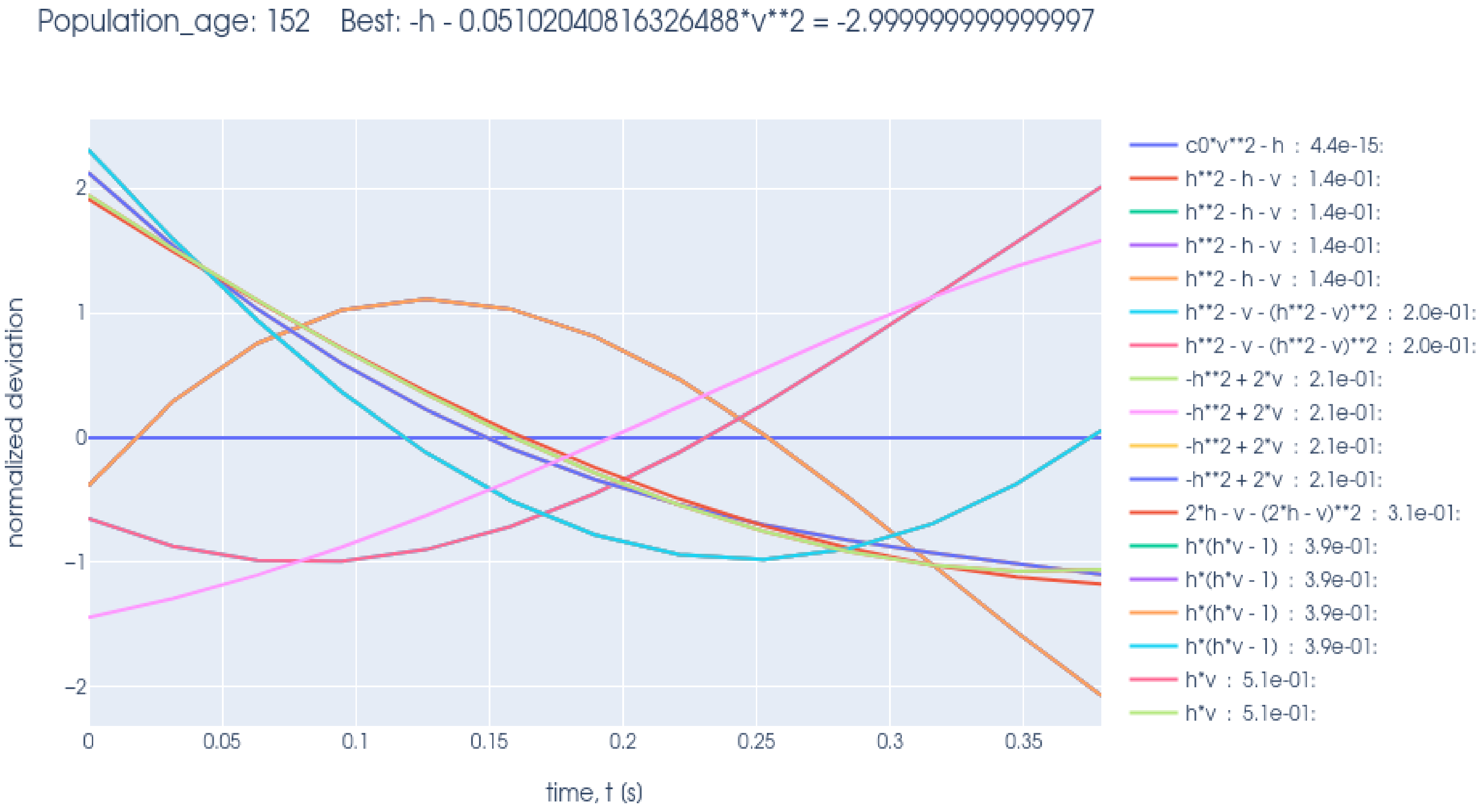
Inputs: Variable: h, v · Operators: +, -, * · Training data from previous slide

```
0: 0
1: c0
2: c0
3: 0
4: c0
5: h
6: h*v*(h**2 - 1)
7: h
8: 2*v
9: h*v
10: v*(v - 1)
11: c0
12: c0 + 2*v
13: 0
14: v*(1 - v)
15: c0
16: c0 + h*v**2
17: -3*v
18: c0
19: h*v*(h + 2)
20: 0
21: 0
22: -h**2*(h - 1)*(2*h - 1)
23: c0
24: h
```



Population Evolution

Best: -h - 0.05102040816326488*v**2 = -2.999999999999997



$$h + 0.05102v^2 = 3$$

$$9.8h + \frac{1}{2}v^2 = 3(9.8)$$

$$mgh + \frac{1}{2}mv^2 = constant = 3mg$$