UFABC

Symbolic Regression or:

How I Learned to Worry About my Machine Learning Models

Fabrício Olivetti de França
Universidade Federal do ABC
folivetti@ufabc.edu.br

First there was data...

Huge amount of data can be collected nowadays due to:

- Increase of available storage space
- New techniques for data collection (not necessarily computer related)
- New ways to process large quantity of raw data

… and then the models were created

With such humongous quantity of data, we had to mine some answers from them!

So we’ve created mathematical models that can answer things like:

“This sample belongs to this category, but this does not!”

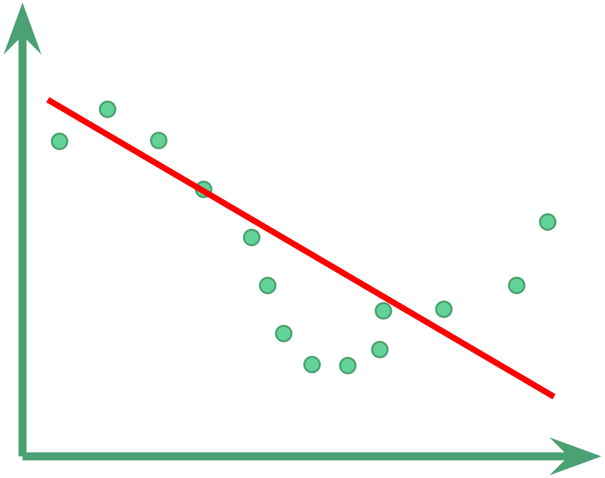
“These set of samples defines a group”

“Given these variables, the expected response is **x**”

The Good, ...

Linear Regression:

$\hat{y} = w * x + b$



Clear interpretation: for every **xi**, we (de)increase the total value by **wi**.

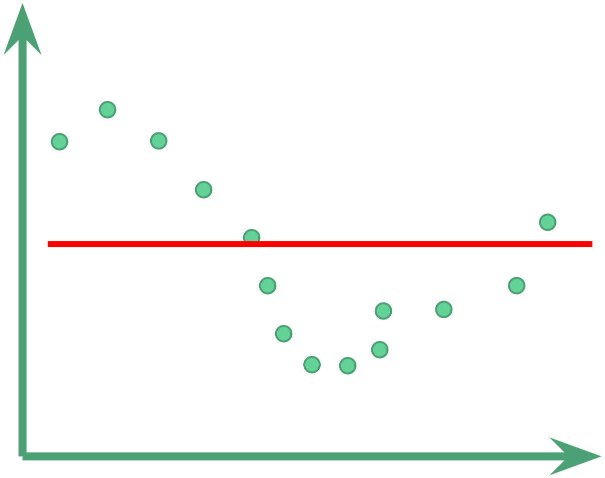
Surprisingly good enough to fit many real world samples

But if the samples present a nonlinear relationship...

… the Bad, …

ŷ = mean(y) -- regression

ŷ = mode(y) -- classification



Incredibly simple and incredibly low accuracy!

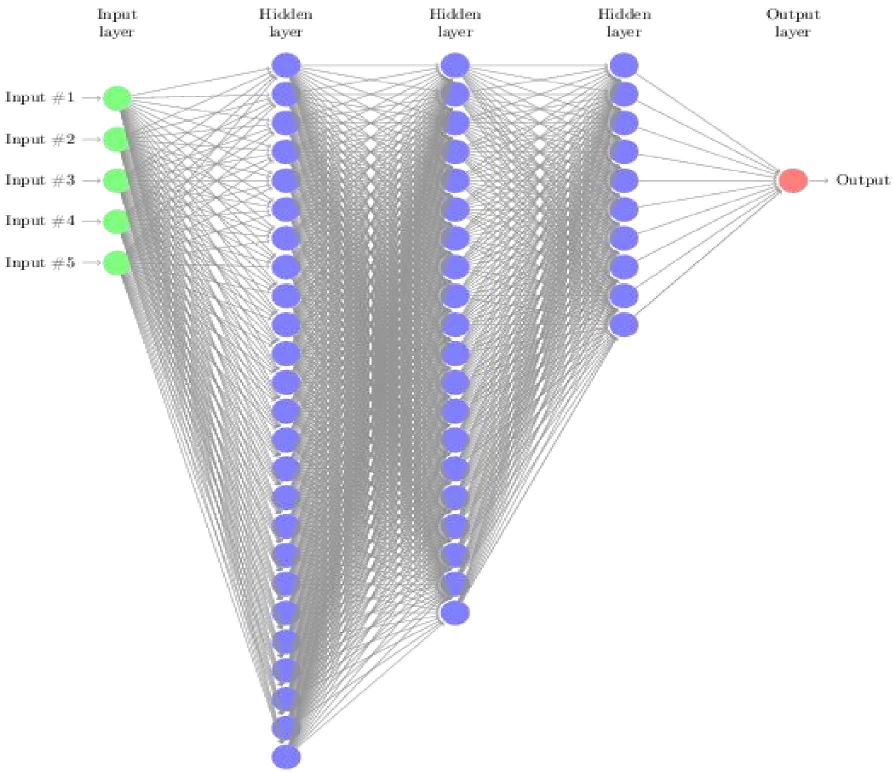
… and the Deep Learning

$$h1 = [\tanh (w1 \,.\, x) \mid w1 \leftarrow W1]$$

$$h2 = [\tanh (w2 \,.\, h1) \mid w2 \leftarrow W2]$$

$$h3 = [\tanh (w3 \,.\, h2) \mid w3 \leftarrow W3]$$

$$\hat{y} = \tanh(w4 \,.\, h3)$$

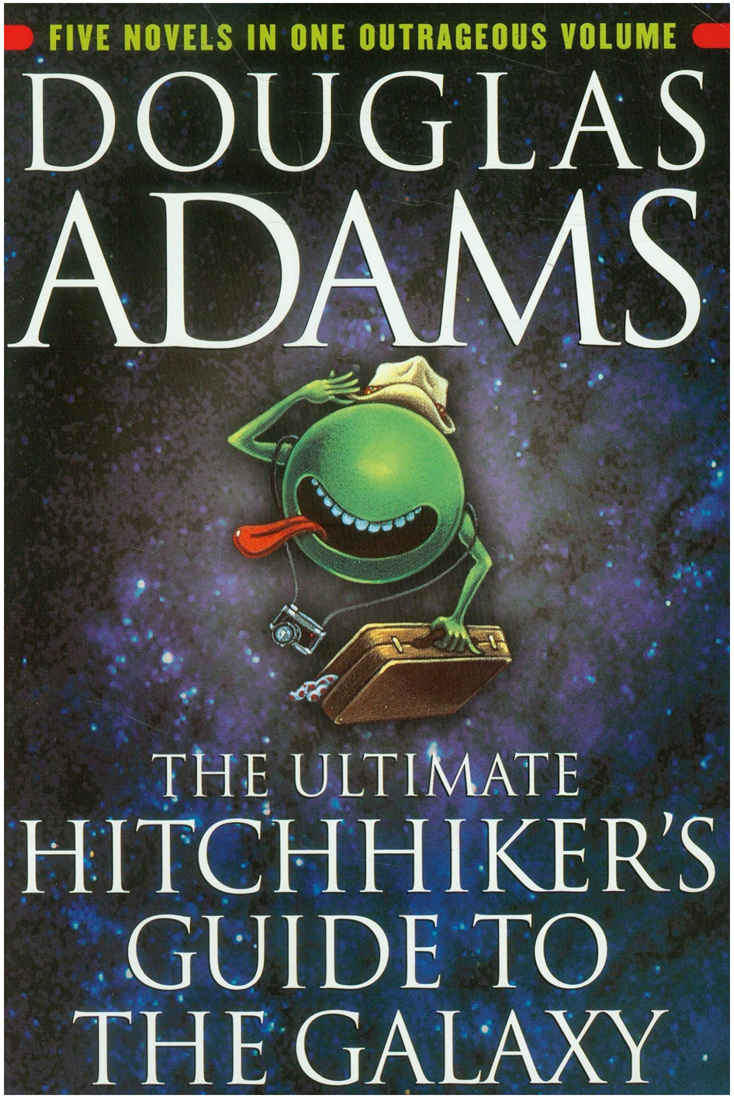


It gets the job done! It can fit anything!

But what does the answer mean? Is the data right?

The answer to the life, the universe and everything

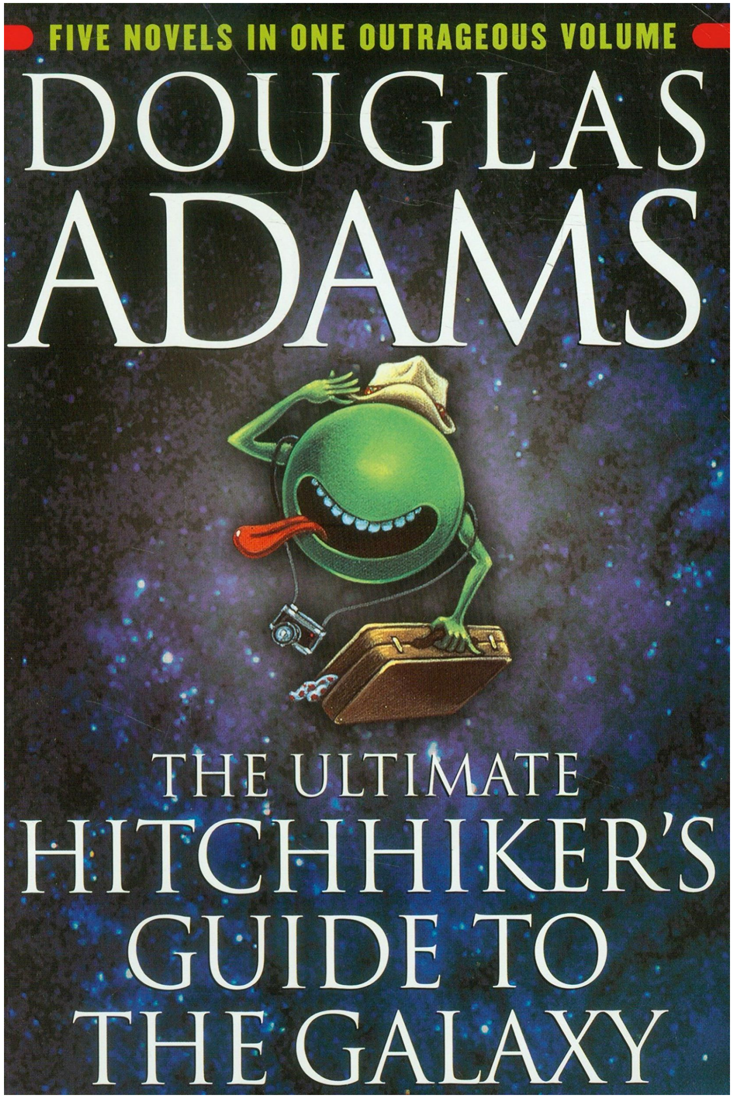
42


FIVE NOVELS IN ONE OUTRAGEOUS VOLUME
DOUGLAS ADAMS
THE ULTIMATE HITCHHIKER'S GUIDE TO THE GALAXY

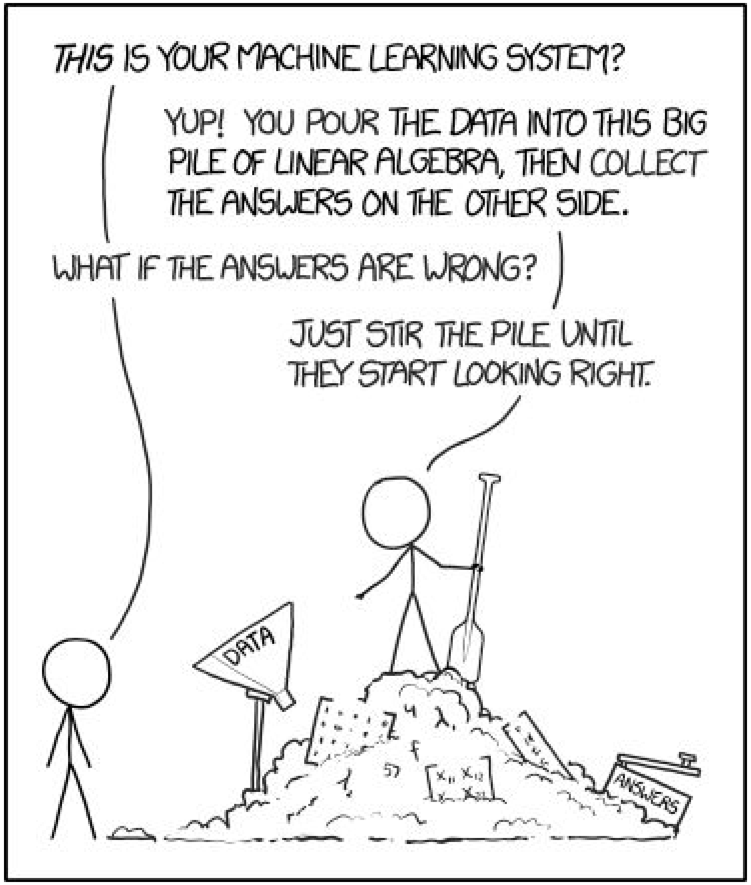
The answer to the life, the universe and everything

42

BUT WHAT IS THE QUESTION?


FIVE NOVELS IN ONE OUTRAGEOUS VOLUME
DOUGLAS ADAMS
THE ULTIMATE HITCHHIKER'S GUIDE TO THE GALAXY

Current state of Data Science in a picture


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://www.xkcd.com/1838/

But isn’t science about learning how things work rather getting a numerical result?

https://www.xkcd.com/1838/

A debt situation

A Machine Learning model can help answer:

- This client will pay his debt?
- How much will he owe?

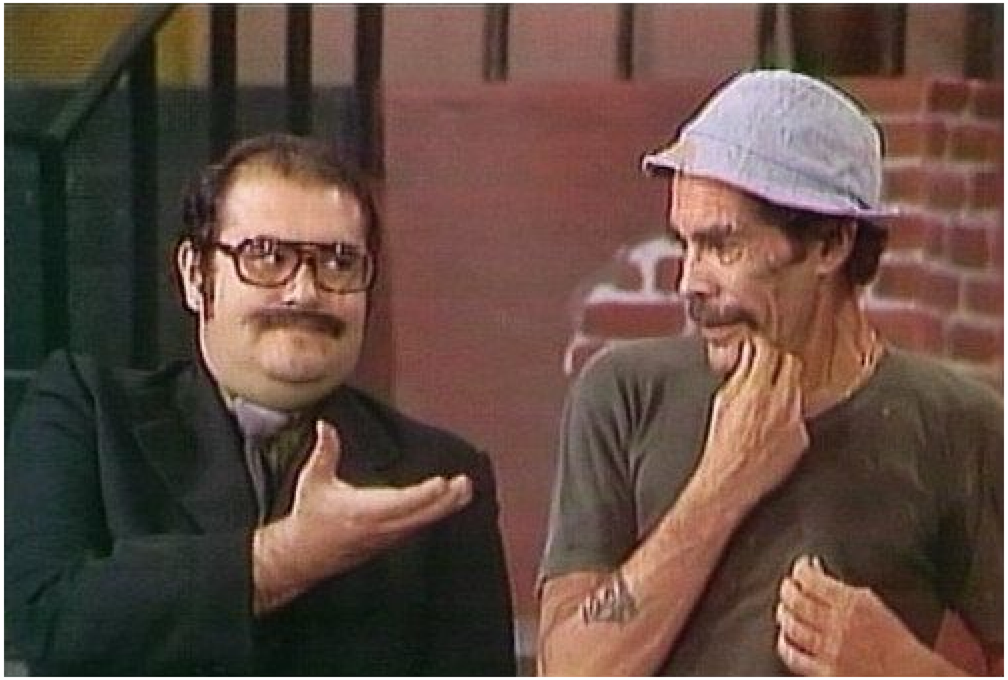
A debt situation

This can support the decision process to whether the bank will lend money or not! And how much!

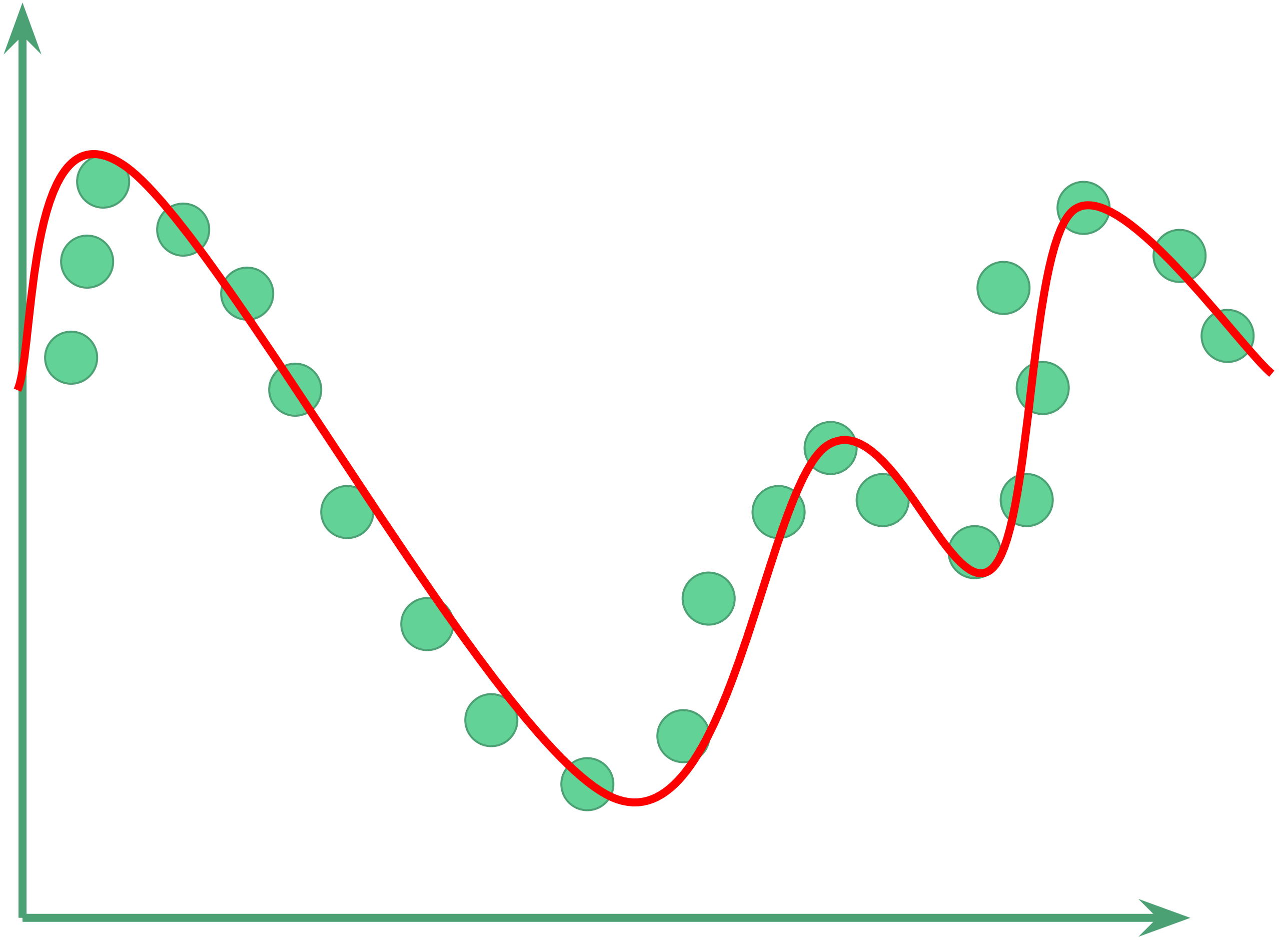
https://pixabay.com/pt/euro-dinheiro-pagar-empr%C3%A9stimos-427528/

A debt situation

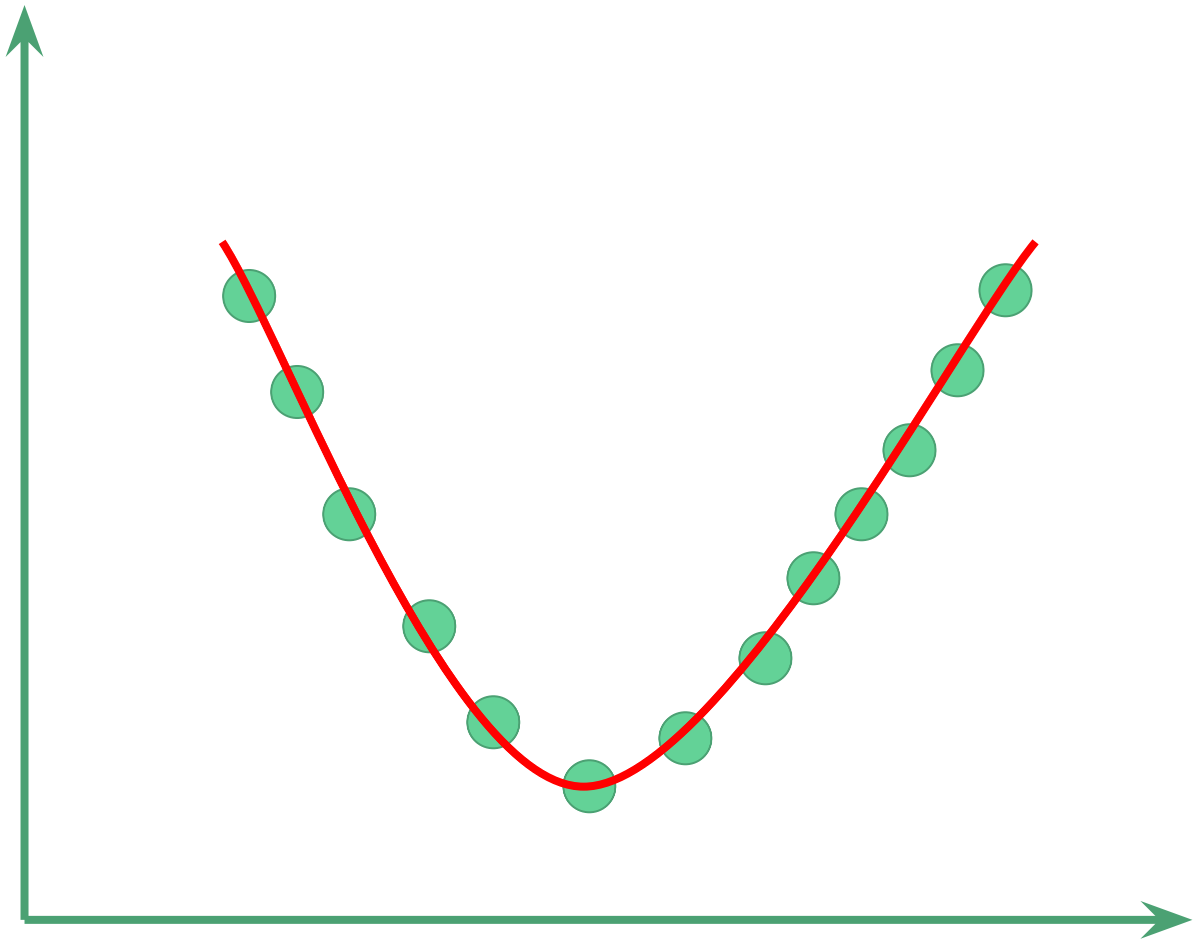
Wouldn’t it be more interesting to understand why a given profile won’t pay back? Is it something the bank can fix? If it can, more money!



And what if the data is wrong?

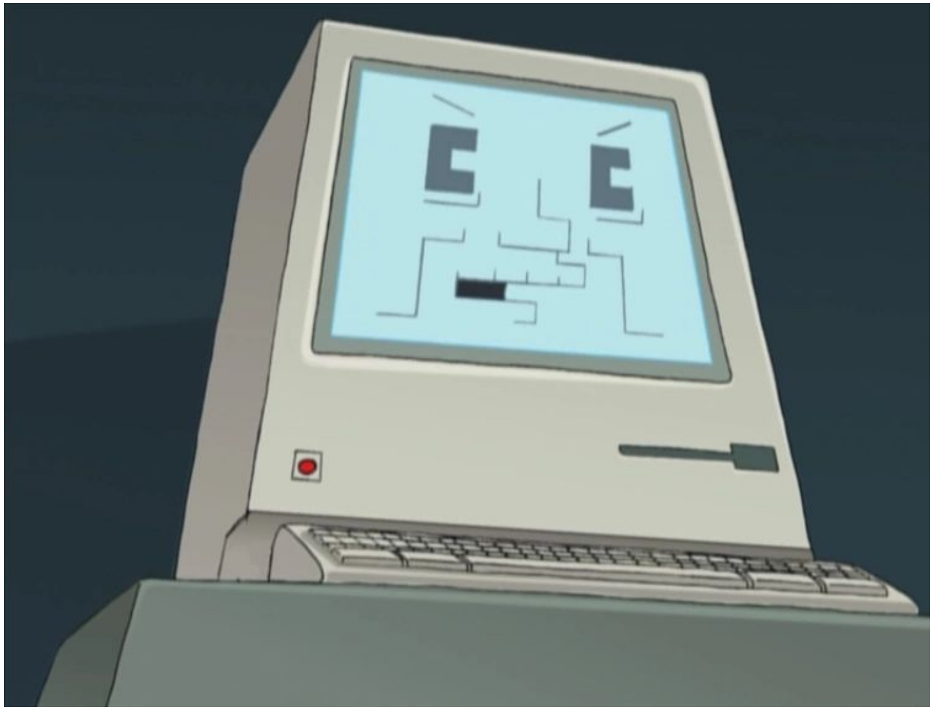


And thus the answer is wrong?



Judge Dredd

Can a Machine Learning model estimate recurrence of criminal? Can also estimate the number of jail time a judge should give?



Futurama: Season 01, Episode 05

Judge Dredd

The model will be fitted to the current data, the current data will include collected statistics regarding past crimes…

… and judges mistakes, social prejudice, and the Machine **will** learn that.

The meaning of life

Sometimes the interesting part of the study is to discover what equations govern a given natural process.

Symbolic Regression

Regression model that searches the space of mathematical expressions for one (or more) that:

- Fits the sampled data
- Is simple!

Symbolic Regression

Regression model that searches the space of mathematical expressions for one (or more) that:

- Fits the sampled data
- **Is simple!**

The Simple Life of a Mathematical Model

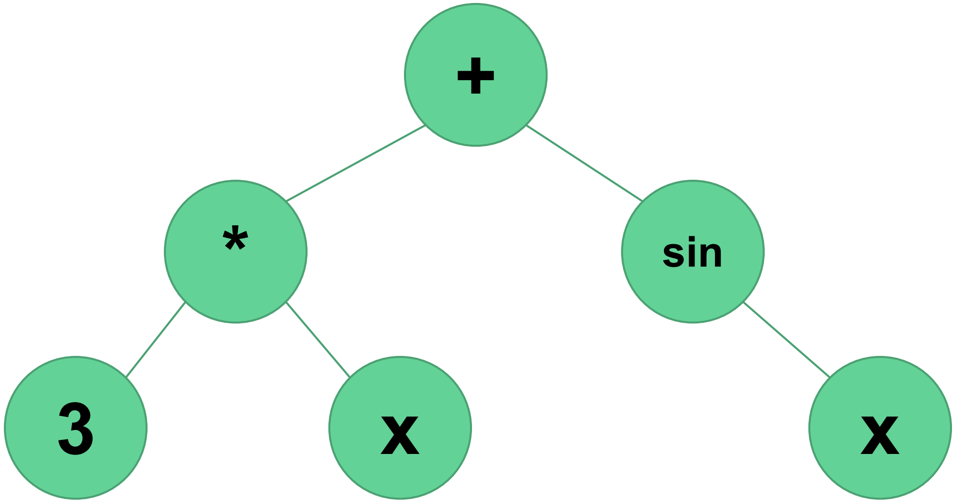
How can we measure the simplicity of a Mathematical Equation?

There is no formalization yet! But we all agree that:

- The smaller the better
- Less nonlinearity is simpler

Genetic Programming

Evolving computer programs described as expressions:



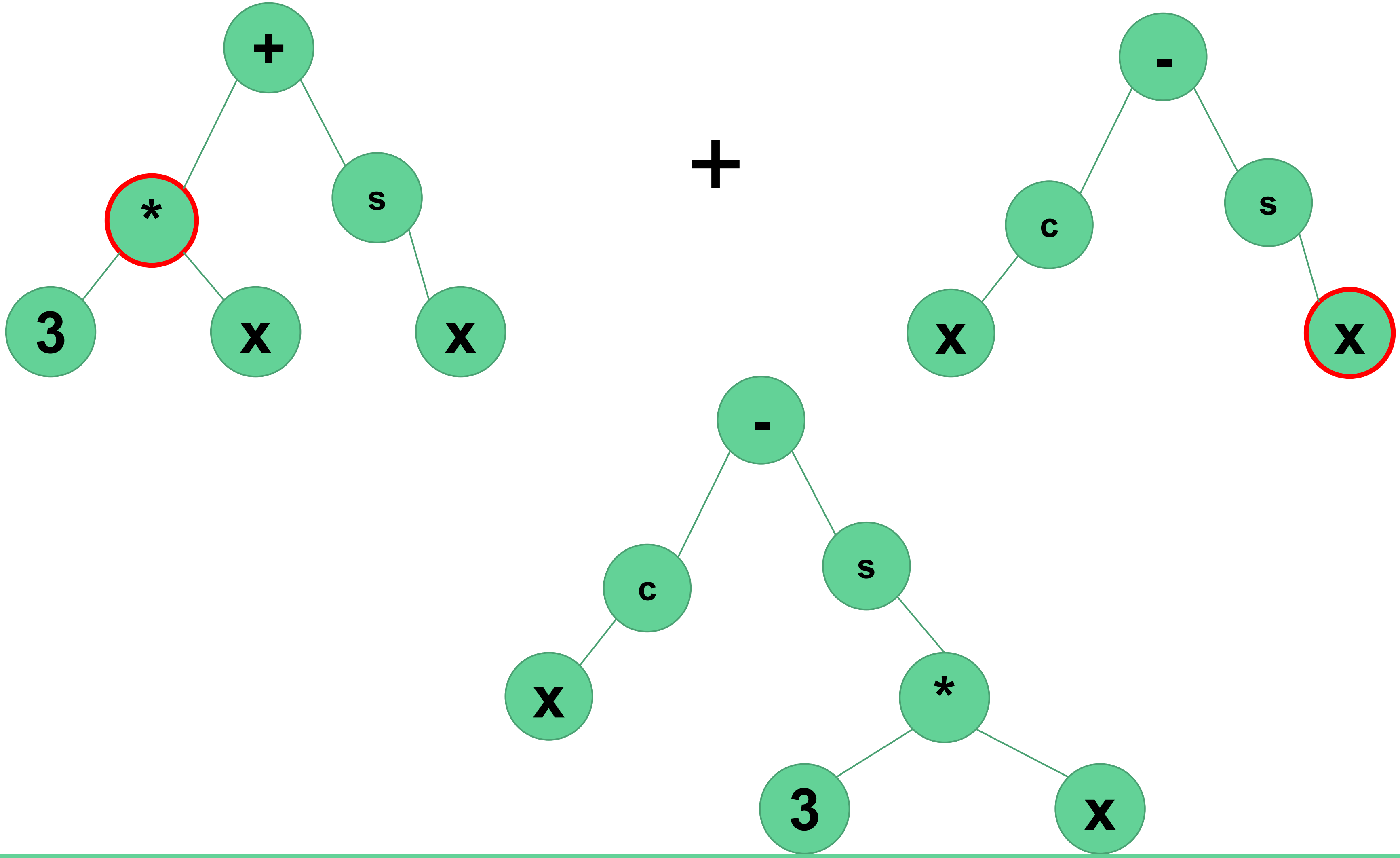
λx ➡ (3*x) + sin(x)

Genetic Programming

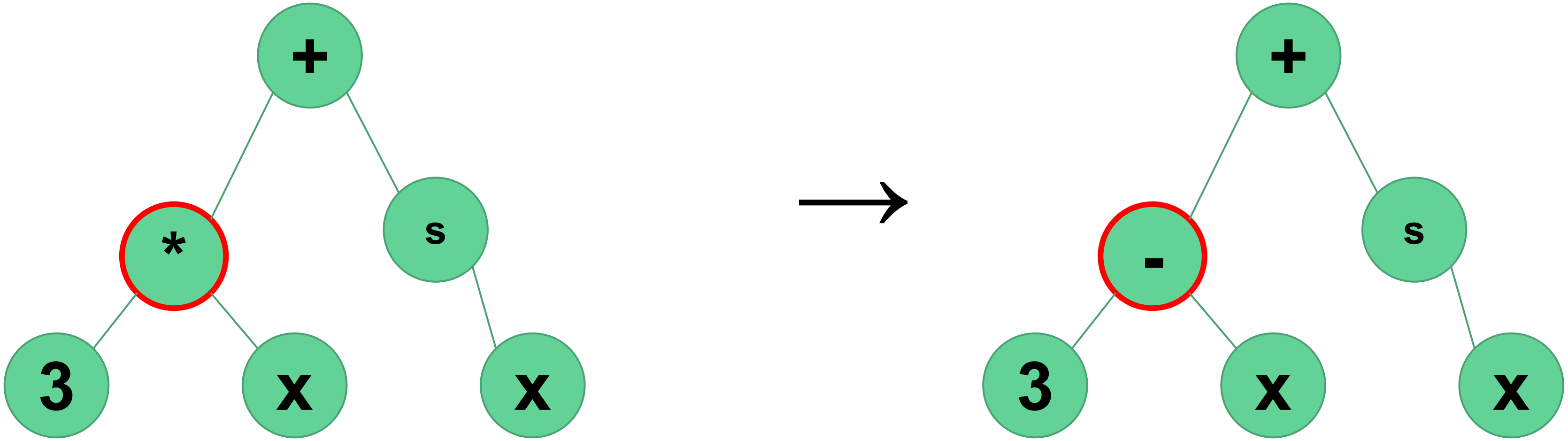
Simple steps:

- Combine expressions
- Change expressions
- Select expressions
- Repeat

Combine Expressions



Change Expressions



Select Expressions

Sample the expressions with probability proportional to the simplicity and inversely proportional to the fitting error.

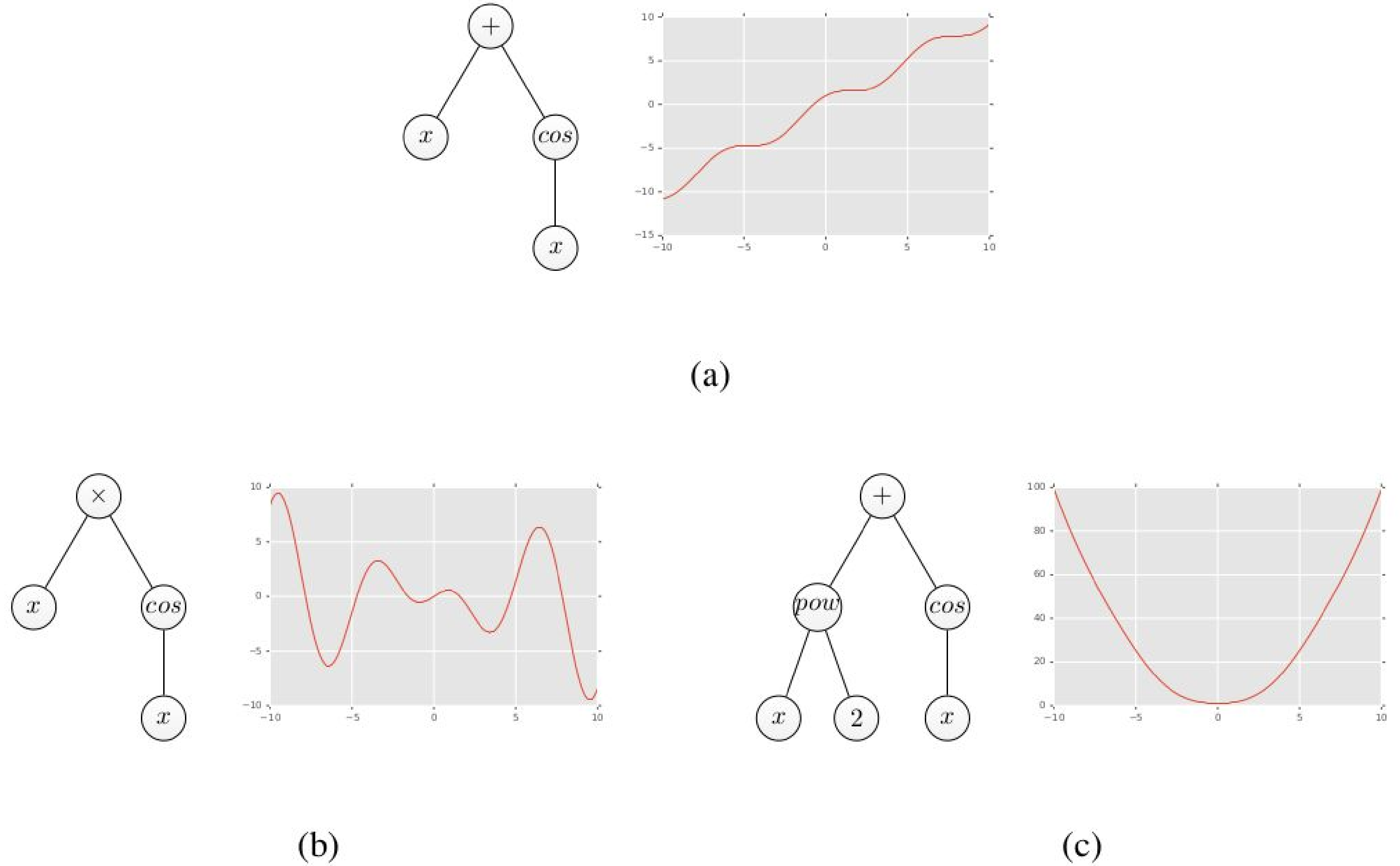
Expected to find a good expression sometime in the near future....

The near future

This algorithm usually take hours to fit simple data with just a few variables...days for complex data...and we usually end up with:

```
6.379515826309025e-3 + -0.00*id(x_1^-4.0 * x_2^3.0 * x_3^1.0) + -0.00*id(x_1^-4.0 * x_2^3.0 * x_3^2.0) +
-0.01*id(x_1^-4.0 * x_2^3.0 * x_3^3.0) + -0.02*id(x_1^-4.0 * x_2^3.0 * x_3^4.0) + 0.01*cos(x_1^-3.0 * x_2^-1.0) +
0.01*cos(x_1^-3.0) + 0.01*cos(x_1^-3.0 * x_3^1.0) + 0.01*cos(x_1^-3.0 * x_2^1.0) + 0.01*cos(x_1^-2.0 * x_2^-2.0)
+ -0.06*log(x_1^-2.0 * x_2^-2.0) + 0.01*cos(x_1^-2.0 * x_2^-1.0) + 0.01*cos(x_1^-2.0 * x_2^-1.0 * x_3^1.0) +
0.01*cos(x_1^-2.0) + 0.01*cos(x_1^-2.0 * x_3^1.0) + 0.01*cos(x_1^-2.0 * x_3^2.0) + 0.01*cos(x_1^-2.0 * x_2^1.0)
+ 0.01*cos(x_1^-2.0 * x_2^1.0 * x_3^1.0) + -0.00*id(x_1^-2.0 * x_2^2.0) + -0.00*sin(x_1^-2.0 * x_2^2.0) +
0.01*cos(x_1^-2.0 * x_2^2.0) + -0.00*tan(x_1^-2.0 * x_2^2.0) + -0.00*log1p(x_1^-2.0 * x_2^2.0) +
0.01*cos(x_1^-2.0 * x_2^2.0 * x_3^1.0) + 0.00*tan(x_1^-2.0 * x_2^2.0 * x_3^1.0) + …
```

Why?


+
x
cos
x
10
5
0
−5
−10
−15
−10
−5
0
5
10
×
x
cos
x
10
5
0
−5
−10
−10
−5
0
5
10
+
pow
cos
x
2
x
100
80
60
40
20
0
−10
−5
0
5
10

(a)

(b)

(c)

How to improve it?

Many solutions were devised:

- Array-based representation
- Semantic operators that minimize the observable change
- Search space restriction

How to improve it?

Many solutions were devised:

- Array-based representation
- Semantic operators that minimize the observable change
- **Search space restriction through representation**

Interaction-Transformation

$$\hat{y} = \sum_i w_i \cdot f_i(P_i(x))$$

$$P(x) = \prod_j x_j^{p_j}$$

Data Structure: Expression

Poly = [Int] -- coefficients of a polynomial transformation

Term = (Poly, λ) -- the polynomial applied to a function

Expression = {Term} -- a set of terms

What it can and what cannot

$$f(x) = log(x^2) + 5\sqrt{|x|}$$ ✓

$$f(x) = log(x^2 + x) + 5\sqrt{|x|}$$ ✗

$$f(x) = \sin(\cos(\tan(x^2)))$$ ✗

Simple algorithm

```
Start with linear expression
Repeat:
    Generate all interactions
    Generate all transformations
    Incrementally add terms as long as they reduce mse
```

Weights **w** are adjusted as a linear regression.

Interaction

Poly1 = [1, 1]   -- x1 . x2

Poly2 = [0, 2] -- x2^2

Poly1 + Poly2 = [1, 3] -- x1 . x2^3

Poly1 - Poly2 = [1, -1] -- x1 / x2

Transformation

(Poly1, id) ➔ [(Poly1, sin), (Poly1, cos), (Poly1, tan), ...]

Marathon Time Prediction

Data: recorded the training data of marathoners from Prague.

Km4weeks: total Kms run in the last 4 weeks

Sp4weeks: avg. speed in the last 4 weeks

Marathon time: target variable

Results

Multi Layer Perceptron:

tanh activation
50 x 50 x 10 hidden neurons
MSE: 0.030

Linear Regression:

$\hat{y} = 5.76 - 0.17 \, . \, \text{speed}$
MSE: 0.051

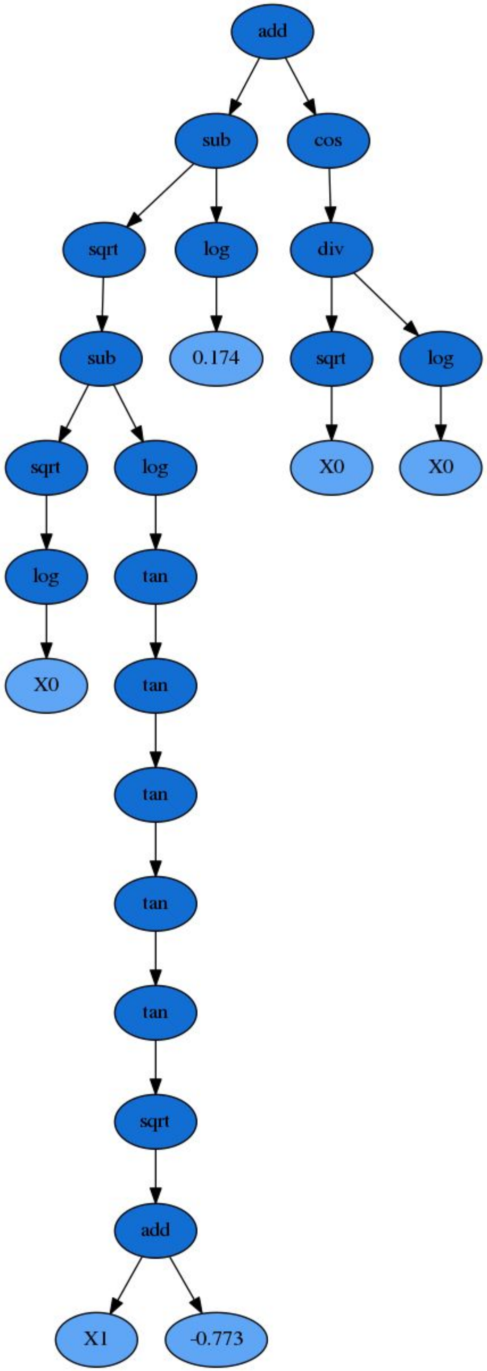
Symbolic Regression:

$\hat{y} = 5.14 - 0.24 \sin(\text{speed}) - 0.12 \, . \, \text{speed} - 0.08 \, . \, \text{km}/\text{speed} - 0.07 \, . \, \cos(\text{km}) - 0.02 \tan(\text{km} \, . \, \text{speed})$
MSE: 0.029

Results

Canonical Genetic Programming:

MSE: 0.038
(after half an hour)



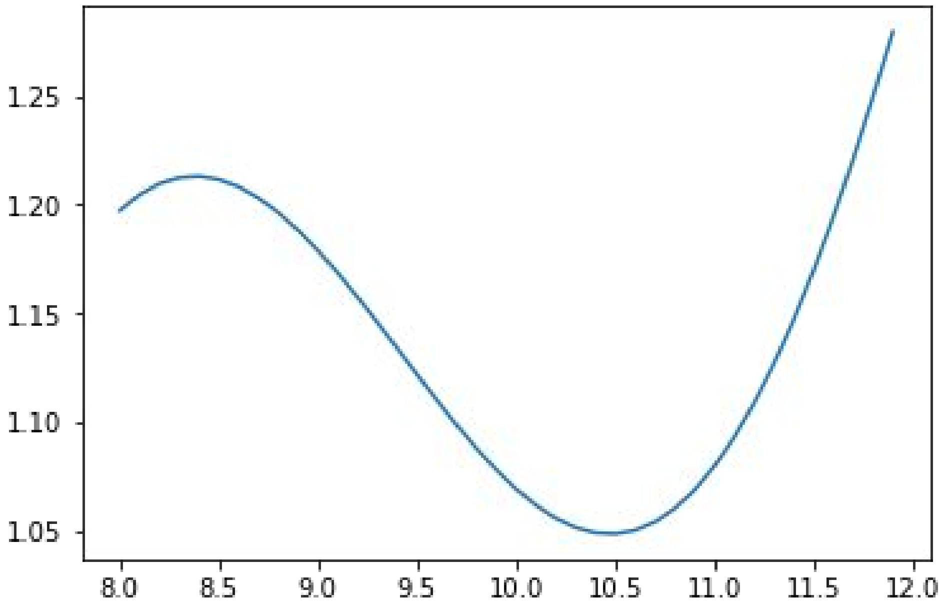
Results

Symbolic Regression also found a solution with MSE = 0.02, but with 20 terms (still much simpler than Neural Network).

Some insights

Evidently, km / speed (fourth term) gives us the expected time!

The first part of the expression. Within the speed range, gives us:



Some insights

This might indicate that:

- The best athletes run faster (duh).
- A “middle class” of athletes keep a lower speed, probably to keep it constant.
- The “lower class” tries to increase their speed at the beginning, raising the average speed, but with a penalty by the final half of the marathon.

Final Remarks

Symbolic Regression can be (in the near future) a very competitive approach to advanced model fitting.

The main advantage is the expressiveness of the model that can help us detect biases in the data and understand what the model is doing.

Final Remarks

So far, there is no SR algorithm capable of dealing with high dimensions while competing with Neural Network regarding minimization of error and speed.