# Regresión Simbólica

- Búsqueda de funciones que se ajusten a una serie de puntos dado
- Búsqueda en el espacio de todas las fórmulas matemáticas posibles las que mejor predicen la variable de salida tomando como entrada las variables de entrada
  - Usando un conjunto de funciones base como las funciones aritméticas, trigonométricas y/o exponenciales
  - Espacio de búsqueda enorme
- Al contrario que las técnicas de regresión clásicas como RR.NN.AA. o SVR (SVM para Regresión), el resultado es una ecuación matemática explícita

# Regresión Simbólica

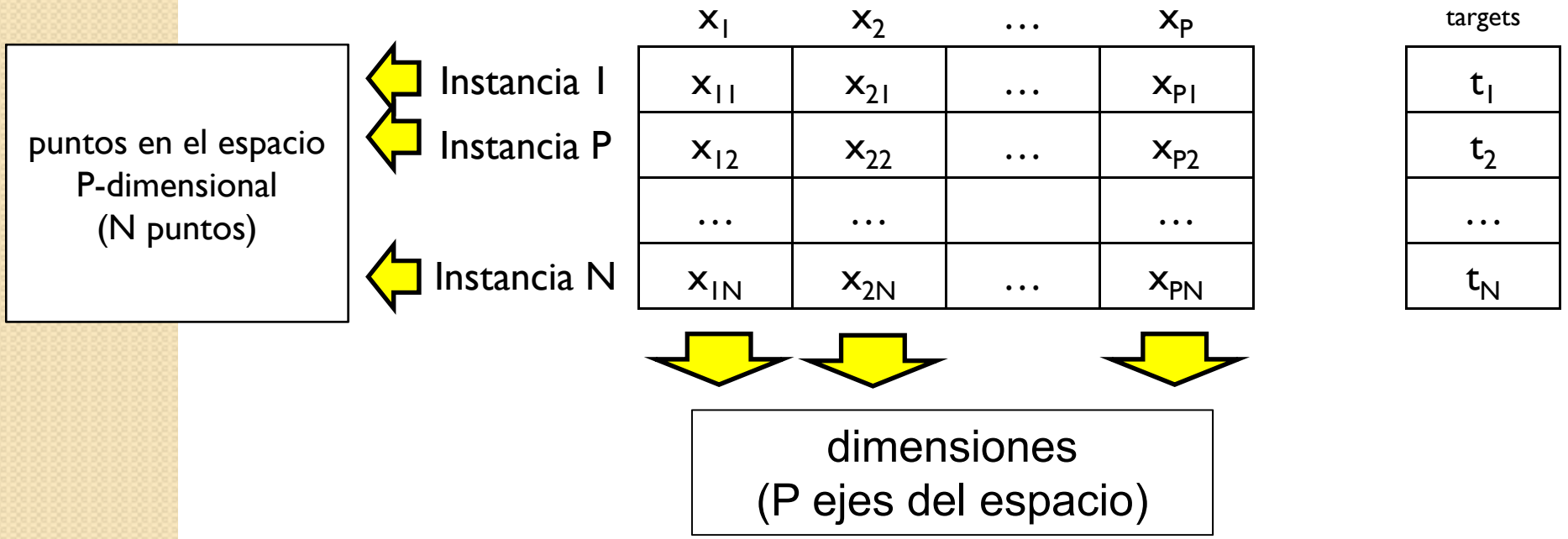
- Principales técnicas, en dos grandes grupos:
  - Basadas en Programación Genética
    - Genetic Programming (Koza, 1992)
    - Age-fitness Pareto Optimization (Schmidt and Lipson, 2011)
    - ∈-Lexicase selection (La Cava et al., 2016)
    - Geometric Semantic Genetic Programming (Moraglio et al., 2012)
    - Multiple Regression Genetic Programming (Arnaldo et al., 2014)

# Regresión Simbólica

- Principales técnicas, en dos grandes grupos:
  - Técnicas de Aprendizaje Automático:
    - Adaptive Boosting (AdaBoost) Regression (Drucker, 1997)
    - Gradient Boosting Regression (Friedman, 2000)
    - Kernel Ridge (Murphy, 2012)
    - Least-Angle Regression with Lasso (Tibshirani, 1994)
    - Linear Regression (Efron et al., 2004)
    - Linear Support Vector Regression (Smola and Schölkopf, 2004)
    - Multilayer Perceptrons (MLPs) Regressor (Kingma and Ba, 2014)
    - Random Forests Regression (Breiman, 2001)
    - Stochastic Gradient Descent Regression (Pedregosa et al., 2011)
    - Extreme Gradient Boosting (Chen and Guestrin, 2016)
    - **Development of Mathematical Expressions**

# Regresión Simbólica

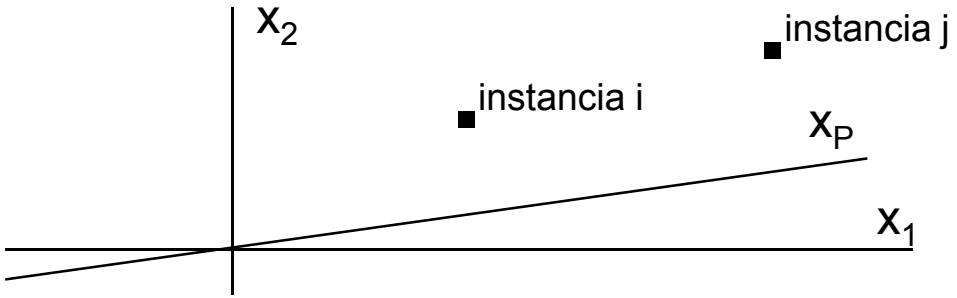
- Development of Mathematical Expressions (DoME)
  - Cambio en la forma de interpretar los datos
    - Forma «clásica» en Aprendizaje Automático
      - N patrones, P variables:



# Regresión Simbólica

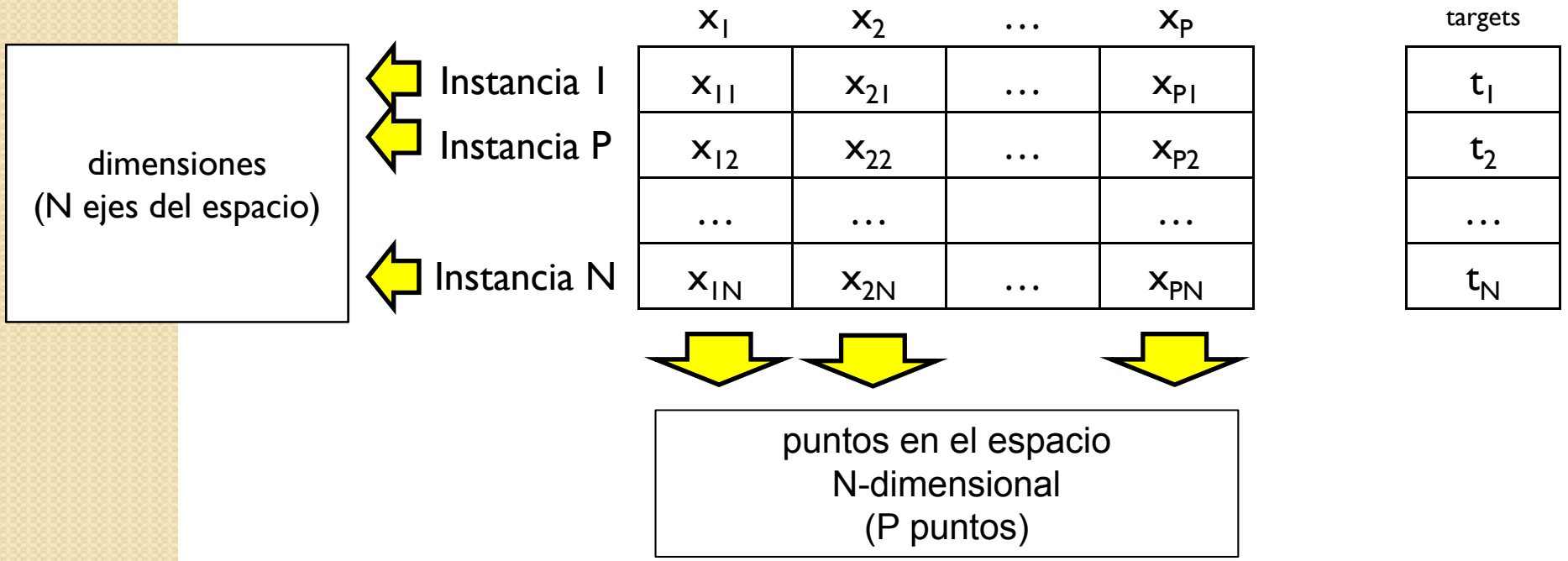
- Development of Mathematical Expressions (DoME)
  - Cambio en la forma de interpretar los datos
    - Forma «clásica» en Aprendizaje Automático
      - N patrones, P variables:

| | $x_1$ | $x_2$ | … | $x_P$ | targets |
|---|---|---|---|---|---|
| Instancia 1 | $x_{11}$ | $x_{21}$ | … | $x_{P1}$ | $t_1$ |
| Instancia P | $x_{12}$ | $x_{22}$ | … | $x_{P2}$ | $t_2$ |
| | … | … | | … | … |
| Instancia N | $x_{1N}$ | $x_{2N}$ | … | $x_{PN}$ | $t_N$ |



# Regresión Simbólica

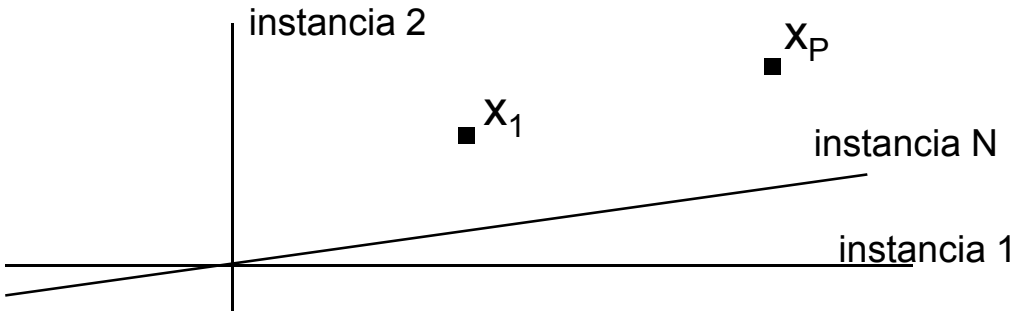
- Development of Mathematical Expressions (DoME)
  - Cambio en la forma de interpretar los datos
    - Espacio semántico
      - N patrones, P variables:



# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Cambio en la forma de interpretar los datos
    - Espacio semántico
      - N patrones, P variables:

| | $x_1$ | $x_2$ | … | $x_P$ | targets |
|---|---|---|---|---|---|
| Instancia 1 | $x_{11}$ | $x_{21}$ | … | $x_{P1}$ | $t_1$ |
| Instancia P | $x_{12}$ | $x_{22}$ | … | $x_{P2}$ | $t_2$ |
| | … | … | | … | … |
| Instancia N | $x_{1N}$ | $x_{2N}$ | … | $x_{PN}$ | $t_N$ |

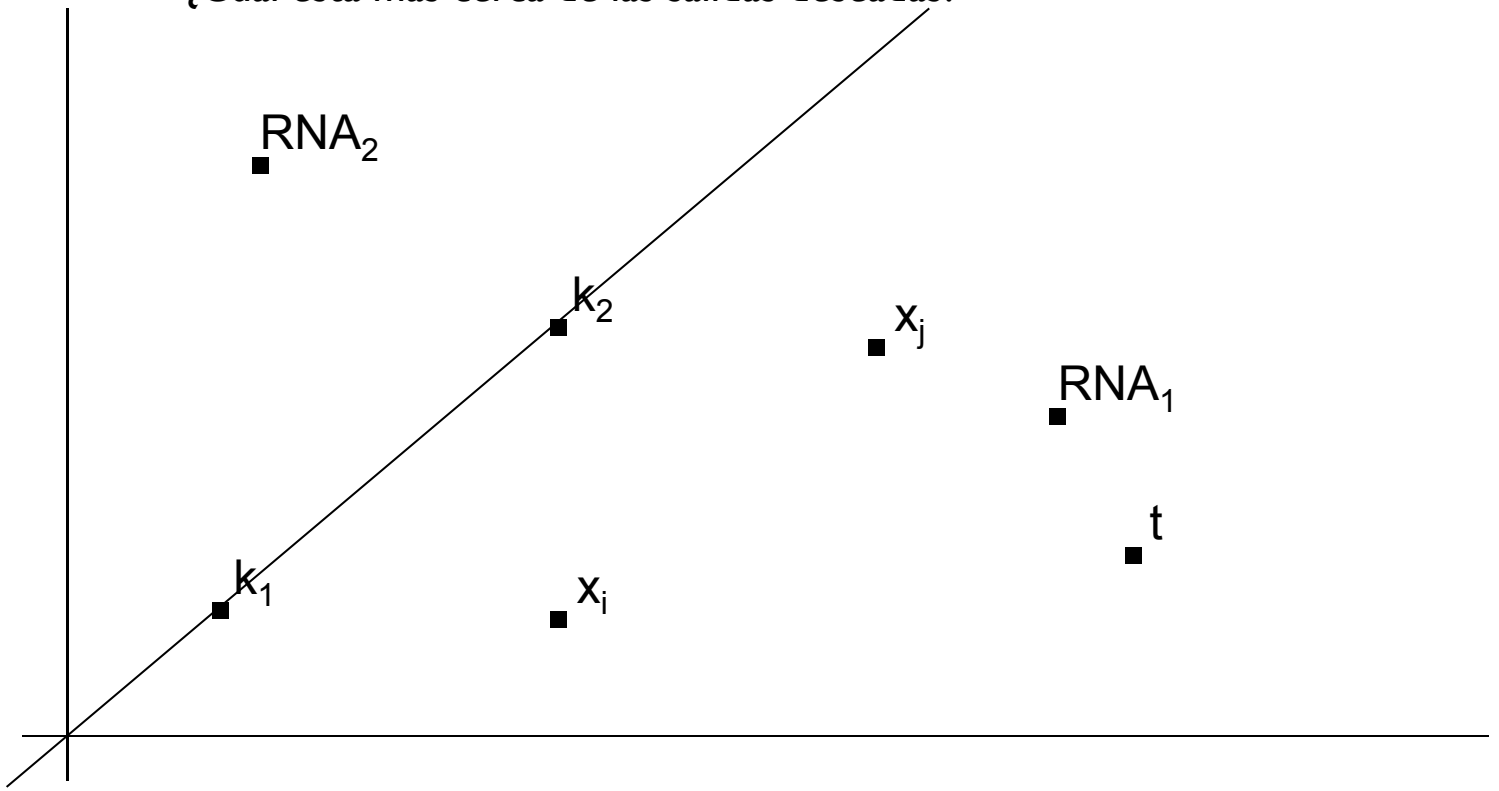


# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - Cada punto del espacio N-dimensional está definido por un valor para cada patrón del conjunto de entrenamiento
    - Por lo tanto, cada punto puede ser:
      - Las salidas de un modelo, para cada patrón
        - Sea el modelo que sea (RNA, SVR, etc.)
      - Una constante
        - Las constantes se pueden interpretar como un modelo:
          - $f(x_1,x_2,\ldots,x_P) = k$
          - Modelo que devuelve el mismo valor para cada patrón
            - Las constantes se sitúan en la línea $(k,k,\ldots,k) = k*(1,1,\ldots,1)$
      - Los distintos valores de cada variable para todos los patrones
        - Las variables se pueden interpretar como un modelo:
          - $f(x_1,x_2,\ldots,x_P) = x_i$
      - Las salidas deseadas (targets)
        - El modelo que se busca

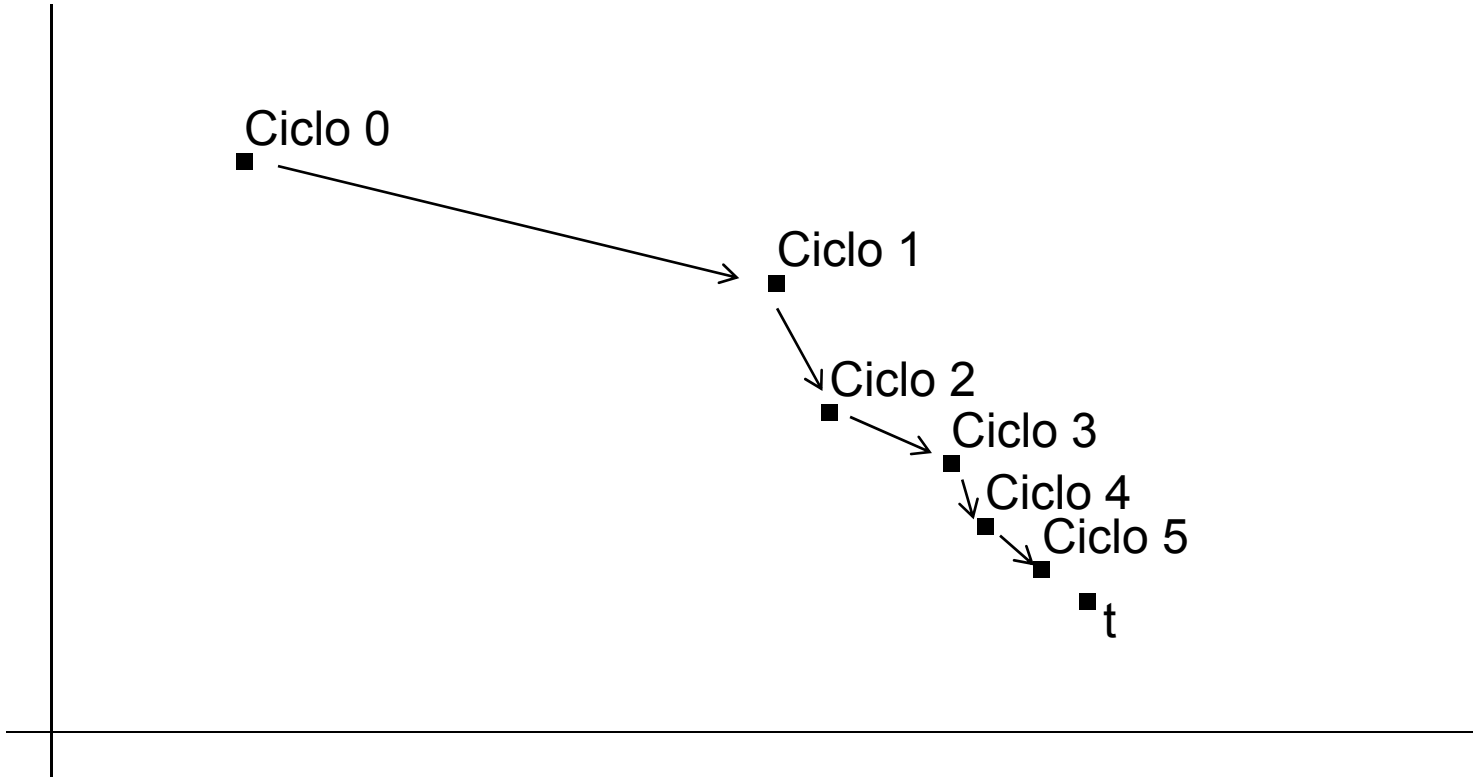
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - ¿Qué modelo de los siguientes es mejor?
      - ¿Cuál está más cerca de las salidas deseadas?



# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - Entrenamiento de una RNA



# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - Medida del error: distancia a las salidas deseadas:

$$\sqrt{\sum_{i=0}^{N}(o_i - t_i)^2} = \sqrt{SSE} = RSSE$$

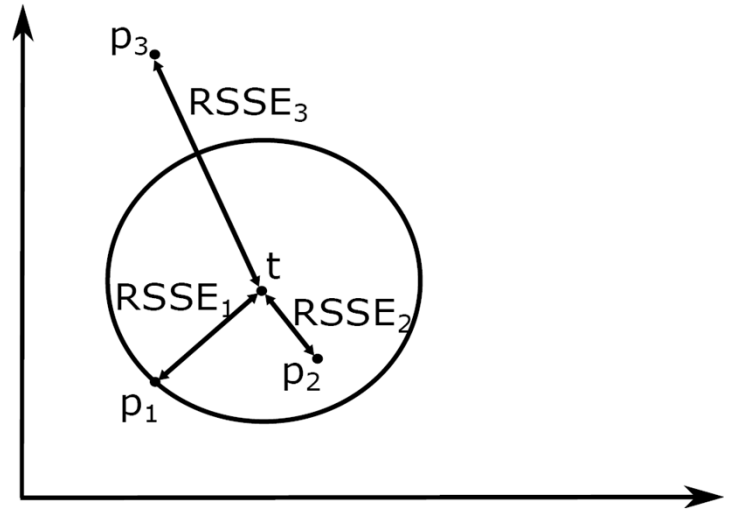
      - $o_i$: salidas del modelo
      - SSE: *Sum Squared Error*
      - RSSE: *Root Sum Squared Error*

    - Ecuación de una esfera N-dimensional centrada en t de radio RSSE
      - En dos dimensiones, la ecuación de un círculo centrado en $(x_0,y_0)$ es

$$(x - x_0)^2 + (y - y_0)^2 = R^2$$

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - Las salidas deseadas t junto con el modelo $p_1$ definen una esfera N-dimensional
      - Radio: $RSSE_1$




    - Un modelo que esté en el interior de la esfera tendrá menor distancia a t, y por tanto menor RSSE
      - Modelo $p_2$, con $RSSE_2$
    - Un modelo en el exterior tendrá mayor RSSE
      - Modelo $p_3$, con $RSSE_3$

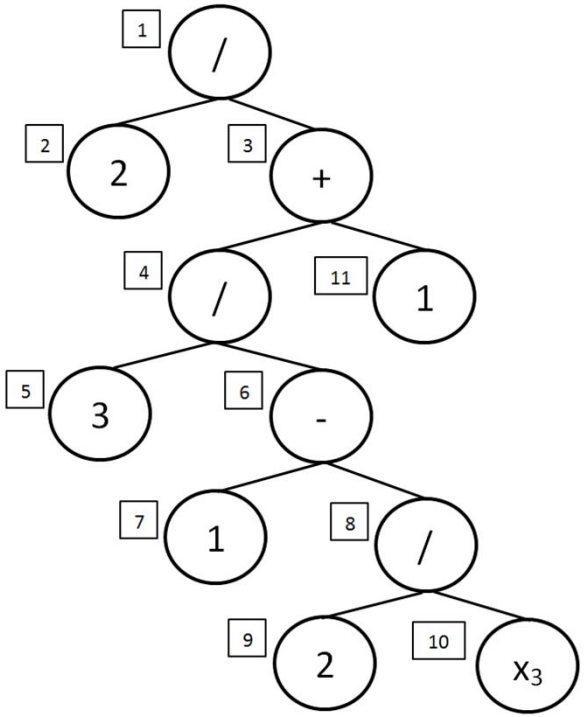
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Espacio semántico
    - Para reducir el impacto del número de patrones, se introduce este en la ecuación
      - La medida del error es el MSE (*Mean Squared Error*):

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(o_i - t_i)^2$$

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Representación en forma de árbol
    - Similar a la Programación Genética clásica
    - Ejemplo:



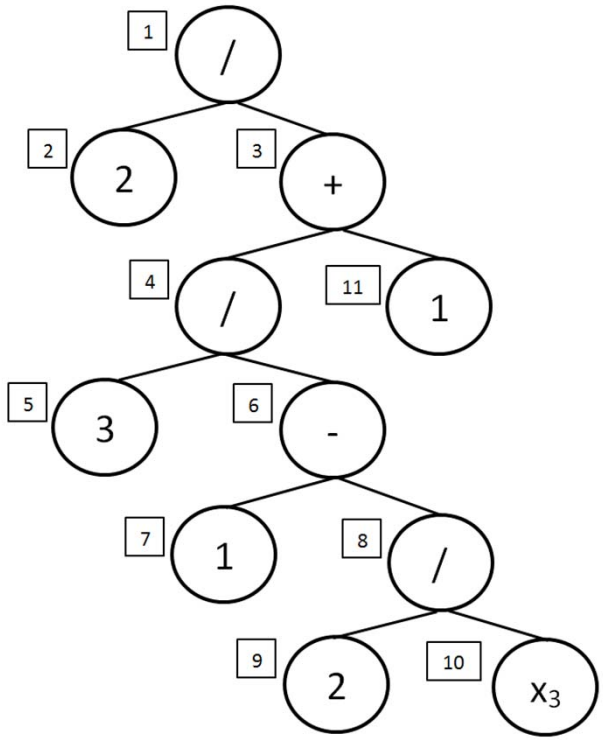

$$f(x_1, x_2, x_3, ...) = \frac{2}{\frac{3}{1-\frac{2}{x_3}} + 1}$$

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Representación en forma de árbol
    - Dos tipos de nodos:
      - Nodos terminales:
        - Variables
        - Constantes
      - Nodos no terminales:
        - Funciones aritméticas: +,-,*,/
    - La semántica del árbol se define como las salidas del árbol para cada patrón del conjunto de entrenamiento
      - Cada nodo del árbol define un subárbol, que se puede interpretar como un nuevo modelo
        - Una nueva ecuación
          - Un «trozo» de la ecuación final
        - Por tanto, **cada nodo tiene su propia semántica**
          - Resultado de evaluar ese subárbol con el conjunto de entrenamiento
          - **Cada nodo es un punto en el espacio N-dimensional**

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Representación en forma de árbol
    - Ejemplo:
      - Valores de $x_3$: (1, 4, -2)




| Node | Semantic |
|---|---|
| 1 | (-1, 0.286, 0.8) |
| 2 | (2, 2, 2) |
| 3 | (-2, 7, 2.5) |
| 4 | (-3, 6, 1.5) |
| 5 | (3, 3, 3) |
| 6 | (-1, 0.5, 2) |
| 7 | (1,1,1) |
| 8 | (2, 0.5, -1) |
| 9 | (2, 2, 2) |
| 10 | (1, 4, -2) |
| 11 | (1, 1, 1) |

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Idea clave: calcular el error cometido **a partir de la semántica de cada nodo del árbol**
    - Cada nodo tiene asociados 4 vectores: a, b, c, d
    - La ecuación para calcular el MSE para un nodo es:

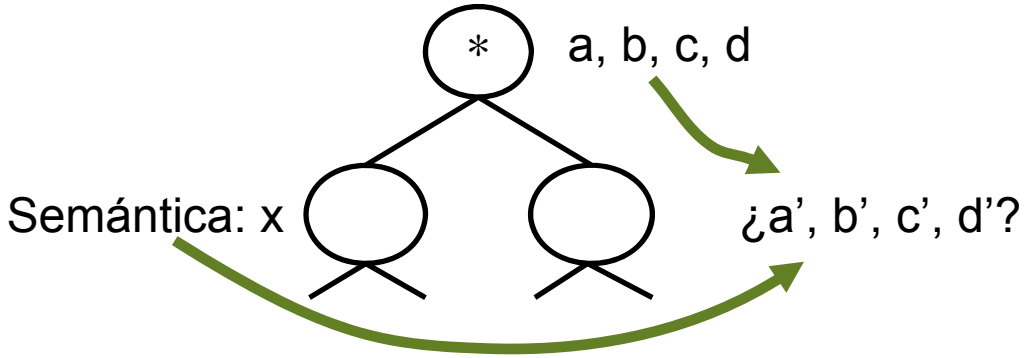
$$MSE = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot o_i - b_i}{c_i \cdot o_i - d_i}\right)^2$$

      - $o_i$: salidas de ese nodo (semántica)
      - Para el nodo raíz: $a_i=1$, $b_i=t_i$, $c_i=0$, $d_i=-1$
        - Lleva a

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(o_i - t_i)^2$$

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - ¿Cómo se calculan los vectores a,b,c,d para cada nodo?
    - Cada nodo no terminal calcula los vectores a',b',c',d' de cada hijo en función de sus propios vectores a,b,c,d y la salida del otro hijo
      - Si x define la semántica del hijo izq., e y define la semántica del hijo dcho
        - Hijo izq: a',b',c',d' se definen en función de a,b,c,d,y
        - Hijo dcho: a',b',c',d' se definen en función de a,b,c,d,x
    - Por ejemplo, nodo *, para el hijo 2:



# REGRESIÓN SIMBÓLICA

- Development of Mathematical Expressions (DoME)
  - ¿Cómo se calculan los vectores a,b,c,d para cada nodo?
    - Cada nodo no terminal calcula los vectores a',b',c',d' de cada hijo en función de sus propios vectores a,b,c,d y la salida del otro hijo
      - Si x define la semántica del hijo izq., e y define la semántica del hijo dcho
        - Hijo izq: a',b',c',d' se definen en función de a,b,c,d,y
        - Hijo dcho: a',b',c',d' se definen en función de a,b,c,d,x
    - Por ejemplo, nodo *, para el hijo 2:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot o_i - b_i}{c_i \cdot o_i - d_i}\right)^2 = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot (x_i \cdot y_i) - b_i}{c_i \cdot (x_i \cdot y_i) - d_i}\right)^2 = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{(a_i \cdot x_i) \cdot y_i - b_i}{(c_i \cdot x_i) \cdot y_i - d_i}\right)^2 = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a'_i \cdot y_i - b'_i}{c'_i \cdot y_i - d'_i}\right)^2$$

que es la ecuación $MSE = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a'_i \cdot o_i - b'_i}{c'_i \cdot o_i - d'_i}\right)^2$ para el segundo hijo con

$$a'_i = a_i \cdot x_i \qquad b'_i = b_i \qquad c'_i = c_i \cdot x_i \qquad d'_i = d_i$$

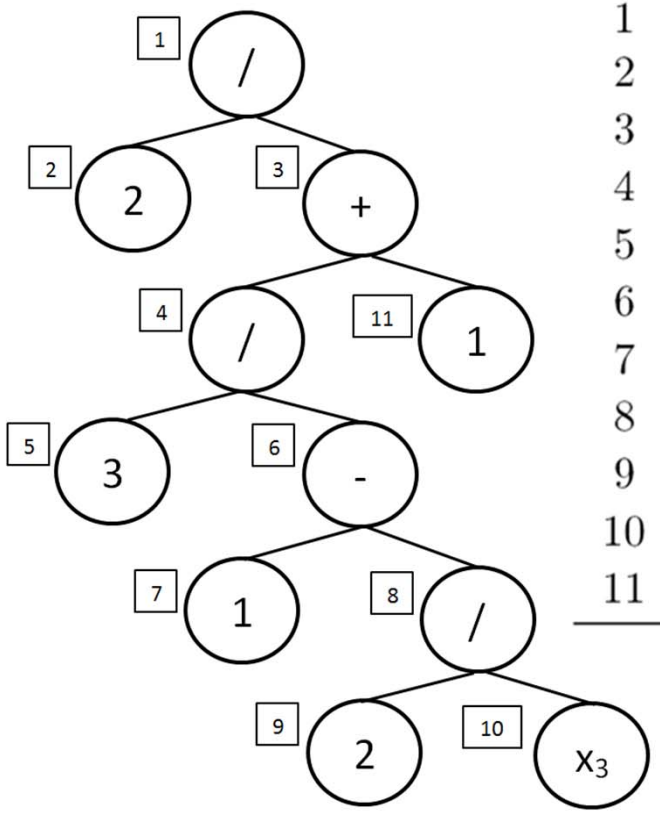
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - ¿Cómo se calculan los vectores a,b,c,d para cada nodo?
    - Haciendo un desarrollo similar con el resto de operaciones e hijos:

| Operation | Left Child | | | | Right Child | | | |
|---|---|---|---|---|---|---|---|---|
| | $a'_i$ | $b'_i$ | $c'_i$ | $d'_i$ | $a'_i$ | $b'_i$ | $c'_i$ | $d'_i$ |
| $+$ | $a_i$ | $b_i - a_i \cdot y_i$ | $c_i$ | $d_i - c_i \cdot y_i$ | $a_i$ | $b_i - a_i \cdot x_i$ | $c_i$ | $d_i - c_i \cdot x_i$ |
| - | $a_i$ | $b_i + a_i \cdot y_i$ | $c_i$ | $d_i + c_i \cdot y_i$ | $a_i$ | $a_i \cdot x_i - b_i$ | $c_i$ | $c_i \cdot x_i - d_i$ |
| $*$ | $a_i \cdot y_i$ | $b_i$ | $c_i \cdot y_i$ | $d_i$ | $a_i \cdot x_i$ | $b_i$ | $c_i \cdot x_i$ | $d_i$ |
| $/$ | $a_i$ | $b_i \cdot y_i$ | $c_i$ | $d_i \cdot y_i$ | $b_i$ | $a_i \cdot x_i$ | $d_i$ | $c_i \cdot x_i$ |

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Ejemplo:




| Node | Semantic | Equation |
|---|---|---|
| 1 | (-1, 0.286, 0.8) | $\frac{1}{3}((o_i-5)^2+(o_i-4)^2+(o_i-1)^2)$ |
| 2 | (2, 2, 2) | $\frac{1}{3}((\frac{o_i+10}{2})^2+(\frac{o_i-28}{-7})^2+(\frac{o_i-2.5}{-2.5})^2)$ |
| 3 | (-2, 7, 2.5) | $\frac{1}{3}((\frac{-5\cdot o_i+2}{o_i})^2+(\frac{-4\cdot o_i+2}{o_i})^2+(\frac{-o_i+2}{o_i})^2)$ |
| 4 | (-3, 6, 1.5) | $\frac{1}{3}((\frac{-5\cdot o_i-3}{o_i+1})^2+(\frac{-4\cdot o_i-2}{o_i+1})^2+(\frac{-o_i+1}{o_i+1})^2)$ |
| 5 | (3, 3, 3) | $\frac{1}{3}((\frac{-5\cdot o_i+3}{o_i-1})^2+(\frac{-4\cdot o_i-1}{o_i+0.5})^2+(\frac{-o_i+2}{o_i+2})^2)$ |
| 6 | (-1, 0.5, 2) | $\frac{1}{3}((\frac{-3\cdot o_i-15}{o_i+3})^2+(\frac{-2\cdot o_i-12}{o_i+3})^2+(\frac{o_i-3}{o_i+3})^2)$ |
| 7 | (1,1,1) | $\frac{1}{3}((\frac{-3\cdot o_i-9}{o_i+1})^2+(\frac{-2\cdot o_i-11}{o_i+2.5})^2+(\frac{o_i-2}{o_i+4})^2)$ |
| 8 | (2, 0.5, -1) | $\frac{1}{3}((\frac{3\cdot o_i-18}{-o_i+4})^2+(\frac{2\cdot o_i-14}{-o_i+4})^2+(\frac{-o_i-2}{-o_i+4})^2)$ |
| 9 | (2, 2, 2) | $\frac{1}{3}((\frac{3\cdot o_i-18}{-o_i+4})^2+(\frac{2\cdot o_i-56}{-o_i+16})^2+(\frac{-o_i+4}{-o_i-8})^2)$ |
| 10 | (1, 4, -2) | $\frac{1}{3}((\frac{-18\cdot o_i+6}{4\cdot o_i-2})^2+(\frac{-14\cdot o_i+4}{4\cdot o_i-2})^2+(\frac{-2\cdot o_i-2}{4\cdot o_i-2})^2)$ |
| 11 | (1, 1, 1) | $\frac{1}{3}((\frac{-5\cdot o_i+17}{o_i-3})^2+(\frac{-4\cdot o_i-22}{o_i+6})^2+(\frac{-o_i+0.5}{o_i+1.5})^2)$ |

# Regresión Simbólica

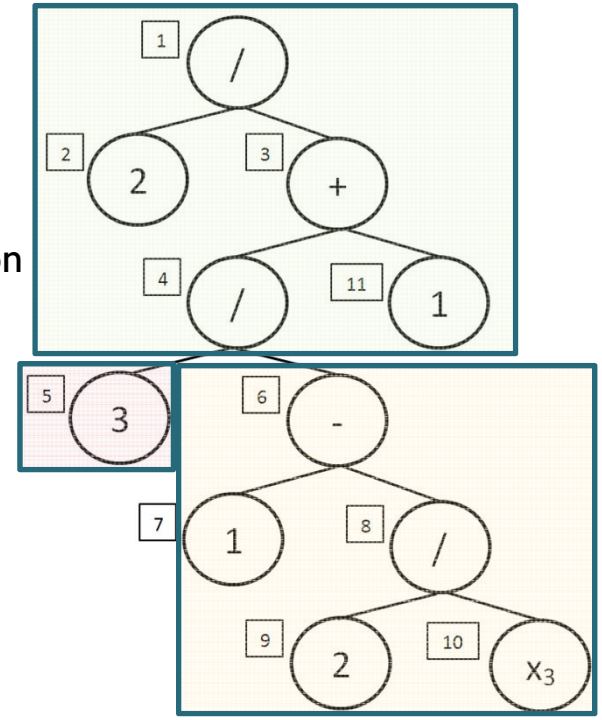
- Development of Mathematical Expressions (DoME)
  - Observaciones (1/4):
    - Cada nodo tiene una ecuación que permite calcular el MSE global
      - Para cada nodo va a dar el mismo valor
    - Esta ecuación depende de:
      - Los vectores a,b,c,d de ese nodo
      - La semántica de ese nodo
        - Las salidas de ese nodo para cada patrón del conjunto de entrenamiento

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Observaciones (2/4):
    - Para calcular los vectores a,b,c,d de ese nodo no es necesario conocer la semántica de ese nodo
      - Es decir, estos vectores son independientes de ese subárbol
      - Dependen del resto del árbol que no cuelga de ese nodo
    - Para calcular la semántica de ese nodo no es necesario conocer el resto del árbol
      - Sólo es necesario conocer ese subárbol
      - Es independiente del resto del árbol que no cuelga de ese nodo

# REGRESIÓN SIMBÓLICA

- Development of Mathematical Expressions (DoME)
  - Observaciones (3/4):
    - Ejemplo: Nodo 6:
      - Para calcular la semántica se necesitan las semánticas de los nodos 6, 7, 8, 9, 10
        - De todo el subárbol que cuelga de ese nodo, expresión «1-(2/$x_3$)»
      - Para calcular los vectores a,b,c,d se necesita:
        - La semántica de su nodo «hermano» 3
          - Todo el subárbol que cuelga de ese nodo
        - Los vectores a,b,c,d del nodo 4
          - Para calcularlos, es necesaria la información de toda la parte superior del árbol

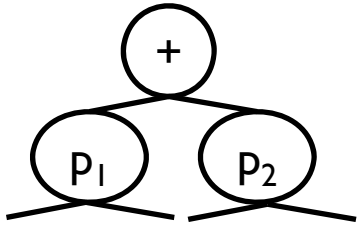


# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Observaciones (4/4):
    - Los vectores a,b,c,d de un nodo son independientes de este y de todo el subárbol que cuelga de él
      - No dependen de la semántica de ese nodo
    - Por tanto, si se cambia ese subárbol por otro, los vectores a,b,c,d no cambian
    - Esto permite calcular cuál sería el valor del MSE global como resultado de cambiar ese subárbol por otro
      - Se calcula la semántica de ese nuevo subárbol
      - Se aplica la ecuación de cálculo de MSE con los vectores a,b,c,d de ese nodo y esta nueva semántica
      - Si el valor del nuevo MSE es inferior al anterior, se puede realizar el cambio del nodo por el nuevo subárbol

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Interpretación geométrica
    - Operadores suma y resta
      - Ejemplo: Árbol p, con hijos $p_1$ y $p_2$:
        - Semántica de $p_1$: x = (1,2)  Semántica de $p_2$: y = (2,1)
        - Ecuaciones de ambos hijos:
          - Hijo1:




$$MSE = \frac{1}{N}\sum_{i=0}^{N}(o_i - t_i)^2 = \frac{1}{N}\sum_{i=0}^{N}((x_i + y_i) - t_i)^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i - (t_i - y_i))^2$$

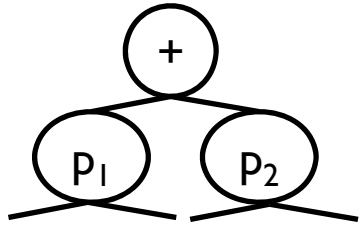
          - Hijo 2:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(o_i - t_i)^2 = \frac{1}{N}\sum_{i=0}^{N}((x_i + y_i) - t_i)^2 = \frac{1}{N}\sum_{i=0}^{N}(y_i - (t_i - x_i))^2$$

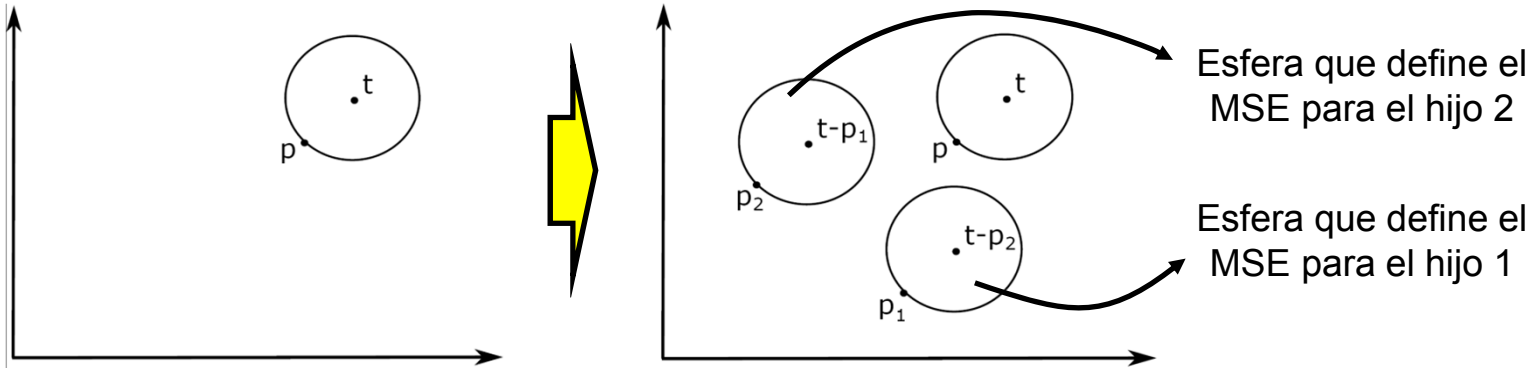
      - Se crean dos nuevas esferas centradas en t-$p_2$ y t-$p_1$, con el mismo radio

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Interpretación geométrica
    - Operadores suma y resta
      - Ejemplo: Árbol p, con hijos $p_1$ y $p_2$:
        - Semántica de $p_1$: x = (1,2)  Semántica de $p_2$: y = (2,1)



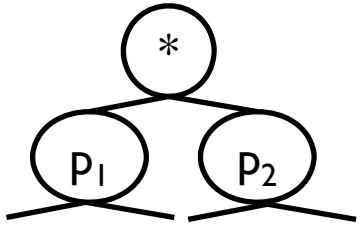

      - Se crean dos nuevas esferas centradas en $t-p_2$ y $t-p_1$, con el mismo radio




      - En lugar de buscar un árbol nuevo dentro de la esfera original, se puede buscar un árbol dentro de las dos nuevas esferas
        - Búsqueda local en cada nodo (nuevos sitios de búsqueda)
        - Si se encuentra, sustituir al hijo correspondiente por ese árbol
          - Como consecuencia, el árbol global p se acercará al punto t

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Interpretación geométrica
    - Operador multiplicación
      - Ejemplo: Árbol p, con hijos $p_1$ y $p_2$:
        - Semántica de $p_1$: x = (1,2)  Semántica de $p_2$: y = (2,1)




        - Ecuaciones de ambos hijos:
          - Hijo 1:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(x_i \cdot y_i - t_i)^2 \Rightarrow \sum_{i=0}^{N}\frac{\left(x_i - \frac{t_i}{y_i}\right)^2}{\left(\frac{SSE}{y_i}\right)^2} = 1$$

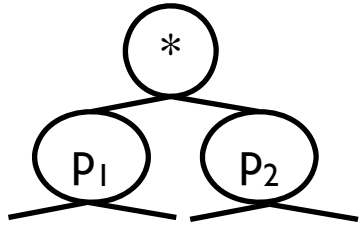
          - Hijo 2:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i \cdot x_i - t_i)^2 \Rightarrow \sum_{i=0}^{N}\frac{\left(y_i - \frac{t_i}{x_i}\right)^2}{\left(\frac{SSE}{x_i}\right)^2} = 1$$

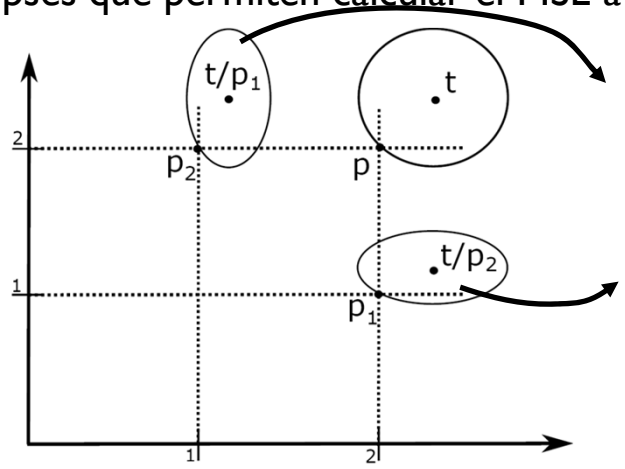
        - 2 nuevas esferas, centradas en $t/p_2$ y $t/p_1$, con los radios modificados cada eje
          - Elipses

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Interpretación geométrica
    - Operador multiplicación
      - Ejemplo: Árbol p, con hijos $p_1$ y $p_2$:
        - Semántica de $p_1$: x = (1,2)  Semántica de $p_2$: y = (2,1)




      - Se crean dos nuevas elipses que permiten calcular el MSE a partir de las semánticas de los hijos




Elipse que define el MSE para el hijo 1

Elipse que define el MSE para el hijo 2

      - Al igual que antes, se puede realizar búsqueda local en cada nodo
        - Buscar un árbol dentro de cada elipse
        - Si se encuentra, se sustituye ese hijo por ese árbol y el árbol global p se habrá acercado a t

# Regresión Simbólica

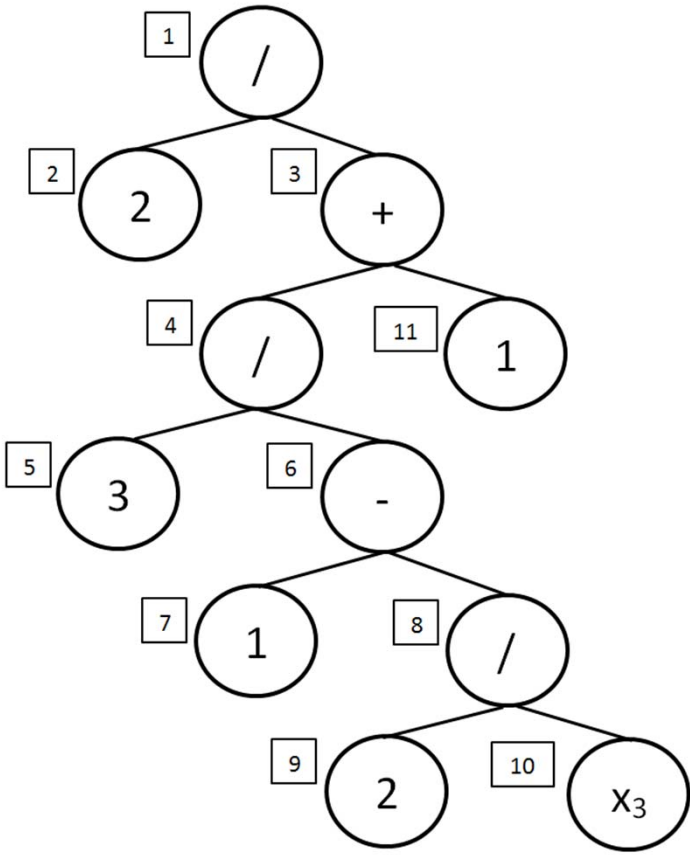
- Development of Mathematical Expressions (DoME)
  - Interpretación geométrica
    - Operador división
      - El lugar de crear esferas o elipses (elipsoides), crea formas totalmente distintas
      - Al igual que antes, se puede realizar búsqueda local en cada nodo
        - Buscar un árbol dentro de cada forma
        - Si se encuentra, se sustituye ese hijo por ese árbol y el árbol global p se habrá acercado a t

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Además, a la vez que se calculan los vectores a,b,c,d para cada nodo, es necesario calcular un conjunto S para cada nodo
    - Contiene qué valores están «prohibidos» en las semánticas de ese nodo
      - Porque llevarían a realizar una división por 0 en un nodo no terminal en alguna parte superior del árbol
        - El conjunto S del segundo hijo de una división (el divisor) siempre contiene la semántica (0,0,…,0)
    - Nodo raíz: $S = \emptyset$

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Ejemplo:

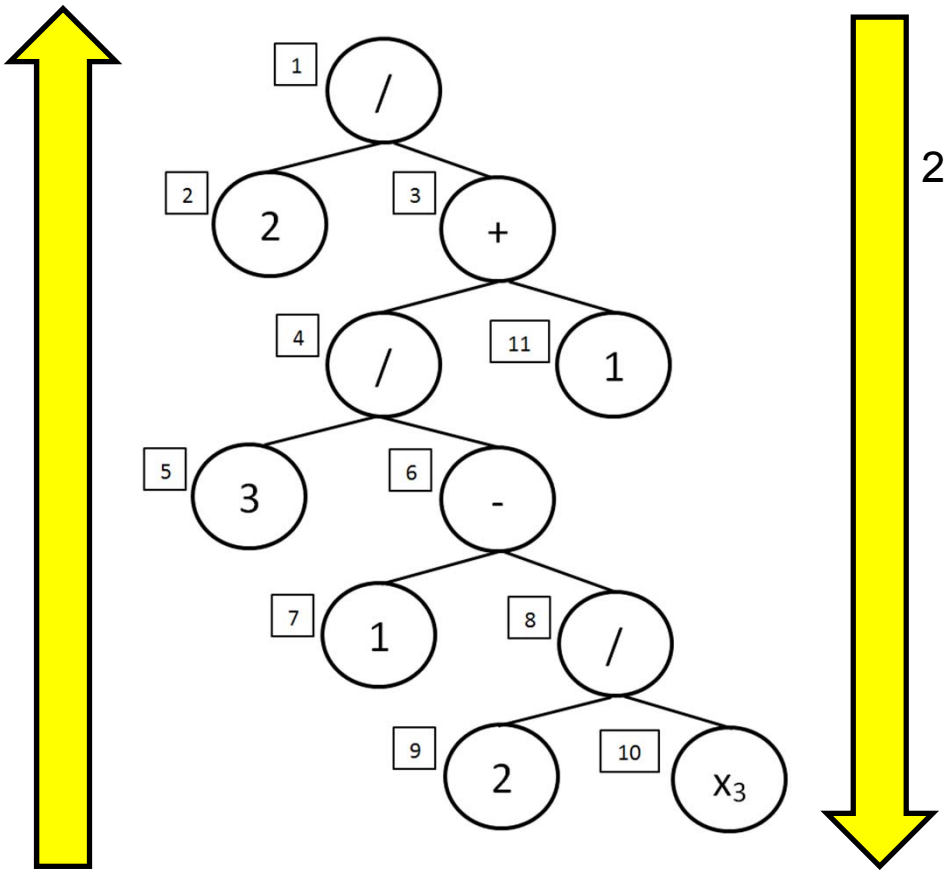



| Node | S |
|---|---|
| 1 | $\emptyset$ |
| 2 | $\emptyset$ |
| 3 | $\{(0,0,0)\}$ |
| 4 | $\{(-1,-1,-1)\}$ |
| 5 | $\{(1,-0.5,-2)\}$ |
| 6 | $\{(0,0,0),(-3,-3,-3)\}$ |
| 7 | $\{(2,0.5,-1),(-1,-2.5,-4)\}$ |
| 8 | $\{(1,1,1),(4,4,4)\}$ |
| 9 | $\{(1,4,-2),(4,16,-8)\}$ |
| 10 | $\{(0,0,0),(2,2,2),(0.5,0.5,0.5)\}$ |
| 11 | $\{(3,-6,-1.5)\}$ |

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Por lo tanto, la evaluación de un árbol se hace en 2 pases:

1. Evaluación del árbol desde los nodos terminales hasta la raíz (cálculo de las semánticas de los nodos)




2. Cálculo de los vectores a,b,c,d y conjunto S de cada nodo desde la raíz hasta los nodos terminales

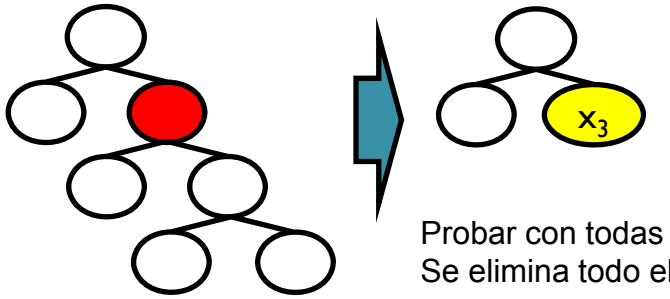
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo
    - Ante un nuevo subárbol con una semántica
      - Se evalúa la ecuación de MSE de ese nodo con la nueva semántica
        - Ecuación definida por los vectores a,b,c,d
      - Si se disminuye el MSE, se puede sustituir el nodo por ese subárbol
        - Eliminar el nodo y todo el subárbol que cuelga de él
    - ¿Cómo hallar un subárbol para un nodo concreto?
      - *Constant search*
        - Sustituir un subárbol por un nodo terminal con una constante
      - *Variable search*
        - Sustituir un subárbol por un nodo terminal con una variable
      - *Constant-variable search*
        - Sustituir un subárbol por un nuevo subárbol con 3 nodos:
          - <constante> <operación> <variable>
      - *Constant-expression search*
        - Desplazar un subárbol «hacia abajo» añadiendo dos nodos
          - <constante> <operación> <subárbol anterior>

# Regresión Simbólica

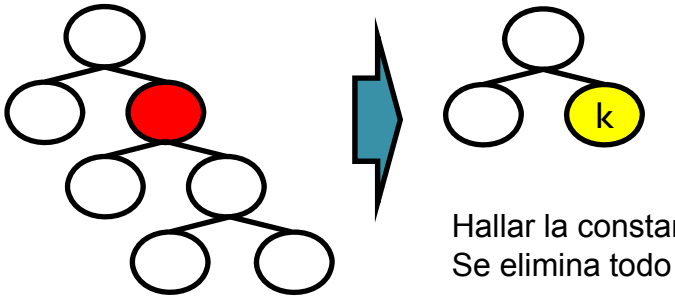
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo
    - Dado un nodo (marcado en rojo)

*Variable search*



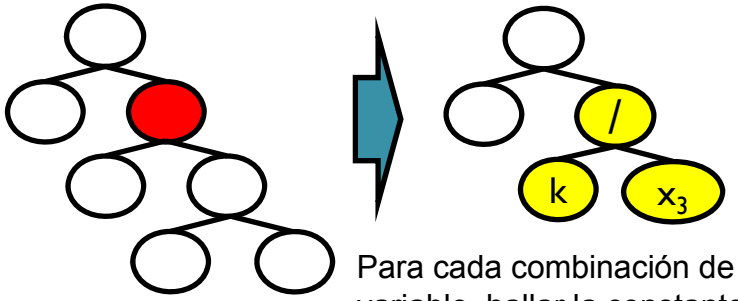

Probar con todas las variables
Se elimina todo el subárbol

*Constant search*



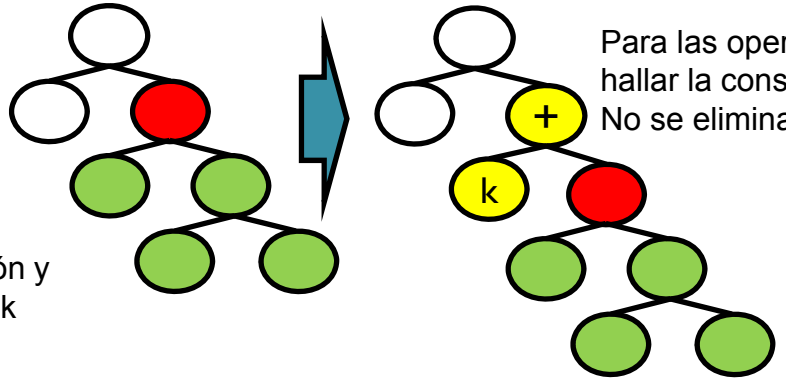

Hallar la constante óptima k
Se elimina todo el subárbol

*Constant-variable search*




Para cada combinación de operación y variable, hallar la constante óptima k
Se elimina todo el subárbol

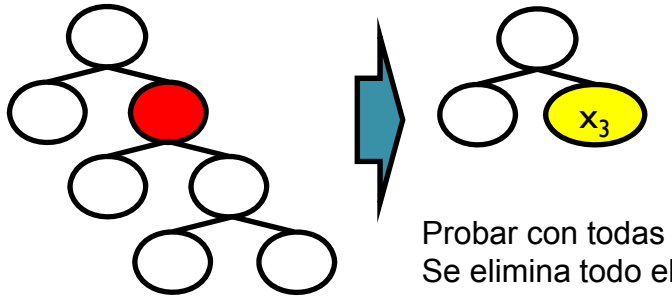
*Constant-expression search*




Para las operaciones + y *, hallar la constante óptima k
No se elimina el subárbol

# Regresión Simbólica

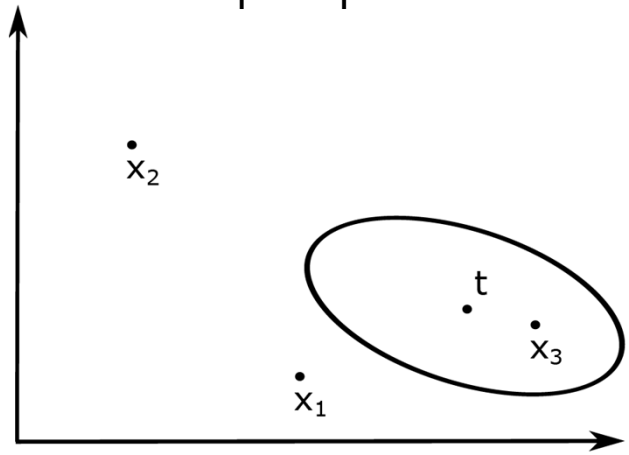
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *variable search*
    - Se busca sustituir el nodo por un nodo terminal con una variable




Probar con todas las variables
Se elimina todo el subárbol

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *variable search*
    - Se busca sustituir el nodo por un nodo terminal con una variable
      - Cada variable tiene una semántica que define un punto en el espacio
      - Si una está en el interior de la forma, se puede crear un nodo terminal con esa variable para sustituir el nodo de búsqueda por ese nodo terminal
        - Ejemplo: variable $x_3$




      - Para saber esto, se usa la ecuación de MSE de ese nodo

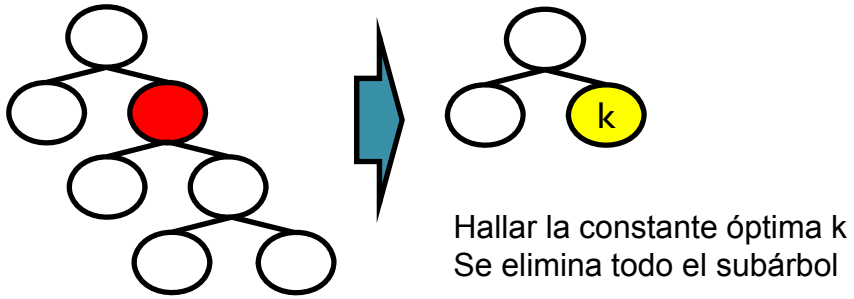
$$MSE = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot o_i - b_i}{c_i \cdot o_i - d_i}\right)^2$$

$o_i$: salida de esa variable para el patrón i

        - Si el MSE es menor, se puede sustituir

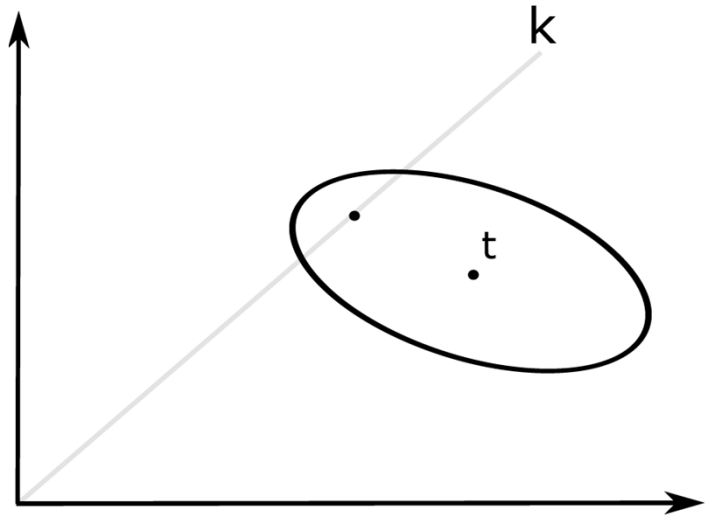
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant search*
    - Se busca un subárbol formado por un nodo terminal que contenga una constante k



# REGRESIÓN SIMBÓLICA

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant search*
    - Se busca un subárbol formado por un nodo terminal que contenga una constante k
      - La semántica será (k,k,…,k)
      - Situada en la línea definida por el origen y el vector (1,1,…,1)
      - **El valor de k será el punto de esa línea más cercano a t**
      - Si este punto está en el interior de la forma correspondiente, el nodo terminal k puede sustituir al nodo donde se está produciendo la búsqueda



# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant search*
    - Para hallar el mejor valor de k para ese nodo:
      - Ecuación para calcular el MSE en ese nodo:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot o_i - b_i}{c_i \cdot o_i - d_i}\right)^2 = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{a_i \cdot k - b_i}{c_i \cdot k - d_i}\right)^2$$

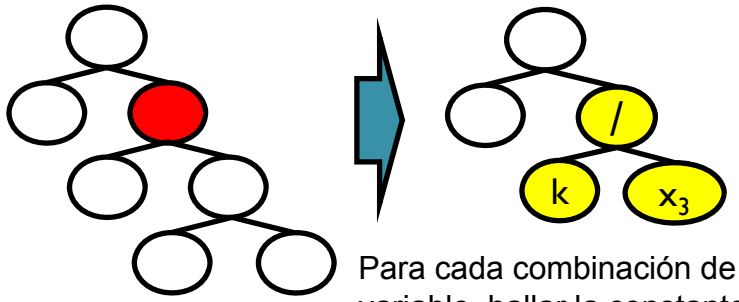
      - Valor de k que minimice el MSE:
        - Se deriva esta expresión y se iguala a 0:

$$\frac{2}{N}\sum_{i=0}^{N}(b_i \cdot c_i - a_i \cdot d_i)\frac{a_i \cdot k - b_i}{(c_i \cdot k - d_i)^3} = 0$$

        - Se despeja el valor de k
        - Existen expresiones analíticas para los casos más comunes

# REGRESIÓN SIMBÓLICA

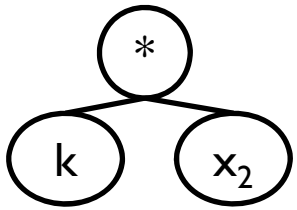
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-variable search*
    - Se busca sustituir el nodo por un subárbol que consta de 3 nodos:
      - <constante (nodo terminal)> <operación> <variable (nodo terminal)>




Para cada combinación de operación y variable, hallar la constante óptima k
Se elimina todo el subárbol

# Regresión Simbólica

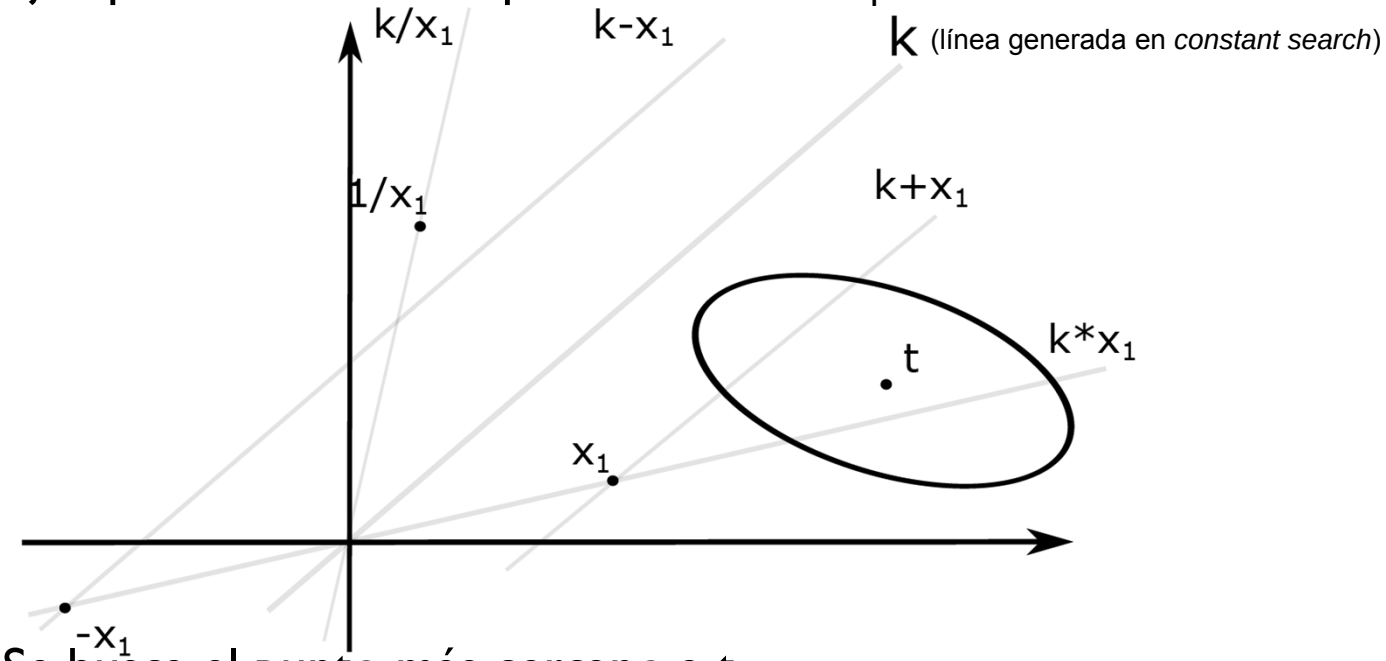
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-variable search*
    - Se busca sustituir el nodo por un subárbol que consta de 3 nodos:
      - <constante (nodo terminal)> <operación> <variable (nodo terminal)>
      - Por ejemplo:




    - Se crean 4 líneas (una por operador) **para vada variable**, y para cada línea se escoge el punto que minimice el MSE
      - Si algún punto mejora el MSE, el nodo puede ser sustituido por el subárbol formado por estos 3 nodos
      - *Constant search* creaba una línea y escogía un punto de esa línea

# Regresión Simbólica

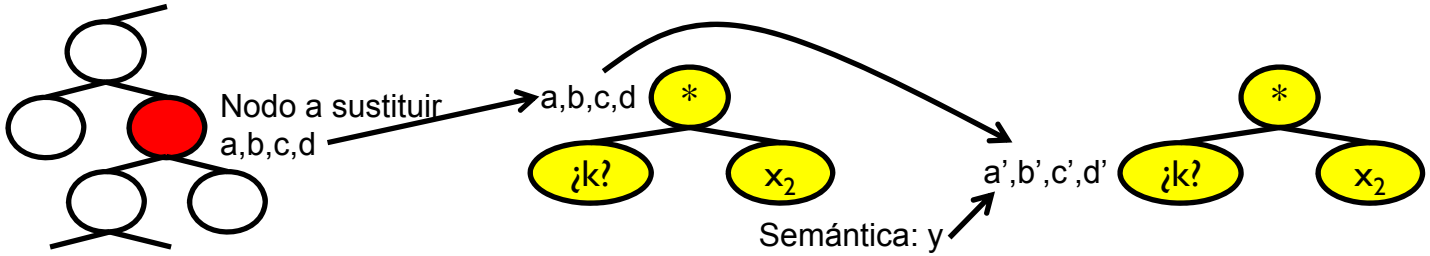
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-variable search*
    - Ejemplo de las 4 líneas para la variable $x_1$:




    - Se busca el punto más cercano a t
      - El que minimice el MSE
    - Se realiza este proceso para el resto de variables

# REGRESIÓN SIMBÓLICA

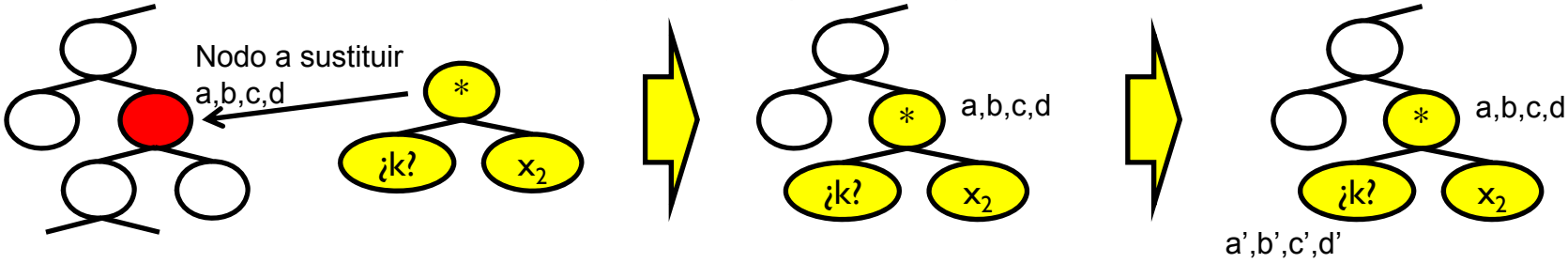
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-variable search*
    - Para calcular el nuevo MSE con una operación y una variable:
      - Se toman los vectores a,b,c,d de ese nodo (nodo padre)
      - Se toma la semántica y de esa variable (segundo hijo)
      - Se calculan nuevos vectores a',b',c',d' correspondientes al primer hijo como se ha descrito anteriormente mediante a,b,c,d,y (y: semántica del segundo hijo)




      - Se calcula el valor óptimo de k de la misma forma que se hizo en *constant search*
      - Se aplica la fórmula con a',b',c',d',k: devuelve el nuevo MSE
      - Si este MSE es inferior al anterior, se puede realizar la sustitución
    - Se realiza este proceso para cada operación y cada variable

# REGRESIÓN SIMBÓLICA

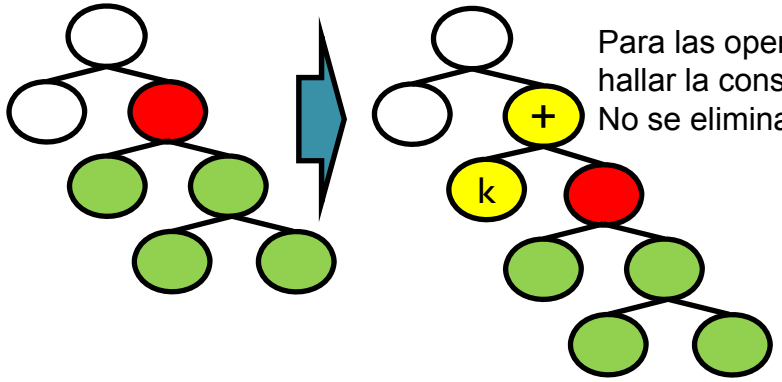
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-variable search*
    - Otra forma de ver este proceso:
      - Se sustituye temporalmente el nodo por el nuevo subárbol
        - Los vectores a,b,c,d se mantienen (son independientes del subárbol)
      - Se calculan nuevos vectores a',b',c',d' correspondientes al primer hijo como se ha descrito anteriormente mediante:
        - Los vectores a,b,c,d del nodo padre
        - La semántica del segundo hijo (la variable)




      - Se calcula el valor de k que minimiza el MSE y el nuevo valor de MSE
        - Si es inferior al anterior, este cambio se puede dejar permanente
        - Si no, se deshace el cambio

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-expression search*
    - Se busca sustituir el nodo por un subárbol que consta de:
      - Nodo padre: Operación
        - Sólo suma y multiplicación
      - Primer hijo: constante a hallar
      - Segundo hijo: el subárbol designado por el nodo seleccionado a ser sustituido
        - Tiene una semántica



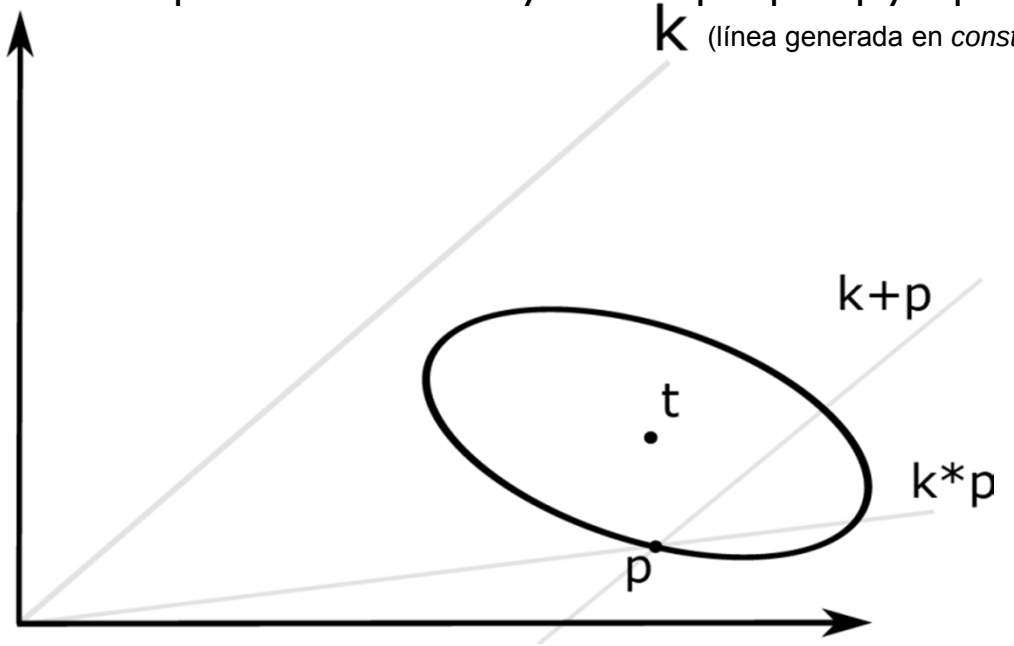

Para las operaciones + y *,
hallar la constante óptima k
No se elimina el subárbol

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-expression search*
    - Se busca sustituir el nodo por un subárbol que consta de:
      - Nodo padre: Operación
        - Sólo suma y multiplicación
      - Primer hijo: constante a hallar
      - Segundo hijo: el subárbol designado por el nodo seleccionado a ser sustituido
        - Tiene una semántica
    - Se aprovecha la semántica del nodo para acercar el árbol al objetivo
      - **Esa semántica está en el límite de la forma**
        - Parece fácil desplazarla hacia el interior con una línea como las anteriores
      - No se elimina el subárbol, sino que se desplaza «hacia abajo» dentro del árbol

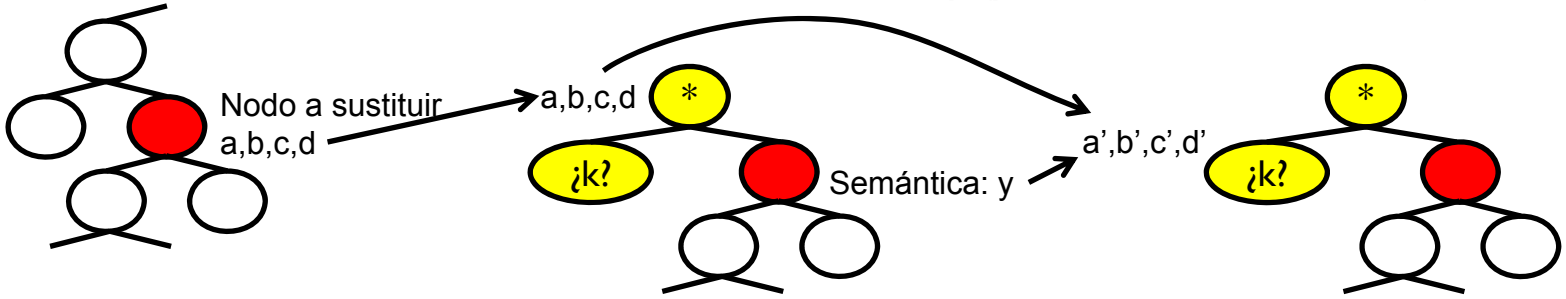
# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-expression search*
    - Ejemplo:
      - Nodo a sustituir es el punto p (en el límite de la forma)
      - Se puede desplazar al interior mediante las operaciones de suma y multiplicación
        - Usando las constantes adecuadas
      - No se usan las operaciones de resta y división porque –p y 1/p no están en el límite



# REGRESIÓN SIMBÓLICA

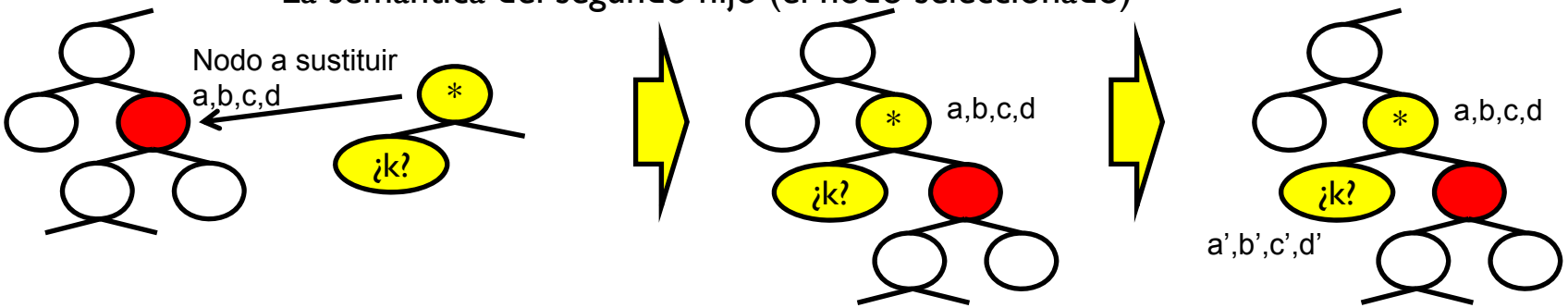
- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-expression search*
    - Para realizar este proceso, similar a *constant-variable search*:
      - Para las operaciones de suma y multiplicación:
        - Se toman los vectores a,b,c,d de ese nodo (nodo padre)
        - Se toma la semántica y de ese nodo (segundo hijo)
        - Se calculan nuevos vectores a',b',c',d' correspondientes al primer hijo como se ha descrito anteriormente mediante a,b,c,d,y (y: semántica de ese nodo)




        - Se calcula el valor óptimo de k de la misma forma que se hizo en *constant search*
        - Se aplica la fórmula con a',b',c',d',k: devuelve el nuevo MSE
        - Si este MSE es inferior al anterior, se puede realizar la sustitución

# REGRESIÓN SIMBÓLICA

- Development of Mathematical Expressions (DoME)
  - Búsqueda local en un nodo: *constant-expression search*
    - Otra forma de ver este proceso:
      - Se desplaza el subárbol incluyendo dos nodos con la operación y una constante
        - Los vectores a,b,c,d se mantienen (son independientes del subárbol)
      - Se calculan nuevos vectores a',b',c',d' correspondientes al primer hijo como se ha descrito anteriormente mediante:
        - Los vectores a,b,c,d del nodo padre
        - La semántica del segundo hijo (el nodo seleccionado)




      - Se calcula el valor de k que minimiza el MSE y el nuevo valor de MSE
        - Si es inferior al anterior, este cambio se puede dejar permanente
        - Si no, se deshace el cambio

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Restricciones:
    - Número máximo de nodos
      - Limita la complejidad del modelo desarrollado
      - Valores muy altos:
        - MSE en entrenamiento muy bajo, pero posible *overfitting*
      - Valores muy bajos:
        - Buena generalización, pero posible *underfitting*
      - Limita la posibilidad de usar *constant-variable search* y *constant-expression search*
        - Pueden aumentar el número de nodos en 2 si se aplican en un nodo terminal
        - *Constant search* y *variable search* no aumentan el número de nodos

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Algoritmo:
    - Árbol inicial:
      - Formado por un nodo terminal con una constante
      - *Constant search* usando como vectores los propios de la raíz del árbol:
        - $a_i=1$, $b_i=t_i$, $c_i=0$, $d_i=-1$
    - A partir de ese árbol, se inicia un proceso iterativo
      - En cada iteración, se sustituye un nodo por otro nodo o subárbol si se mejora el MSE
        - Resultado de una de las 4 búsquedas locales en algún nodo
        - ¿Cómo saber en qué nodos buscar y qué búsquedas hacer en cada iteración?
          - Recorrer el árbol aplicando las búsquedas siguiente una **estrategia**
            - Ejemplos:
              - Exhaustiva
              - Selectiva

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Estrategias: Exhaustiva:
    - Recorrer todos los nodos, aplicando las 4 búsquedas en cada nodo
      - Reemplazar el nodo con el resultado de la estrategia que disminuya más el MSE
    - Computacionalmente costoso
      - Muchos cálculos que no van a llevar a mejoras
      - Se pueden ahorrar muchos cálculos realizando las búsquedas locales en los nodos donde suelen tener éxito, y realizar el resto sólo si estas no tienen éxito
        - Estrategia selectiva

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Estrategias: Selectiva:
    - Selectiva. En cada iteración, realizar los siguientes pasos. Si alguno tiene éxito, aplicar el cambio correspondiente y comenzar la siguiente iteración:
      1. Recorrer los nodos terminales con constantes realizando en ellos *constant search*
      2. Recorrer los nodos terminales con constantes realizando en ellos *variable search*
      3. Recorrer los nodos no terminales realizando en ellos *constant-expression search*
      4. Recorrer los nodos terminales realizando en ellos *constant-variable search*
      5. Realizar de forma conjunta:
         - *Constant search* en nodos terminales con variables y nodos no terminales
         - *Variable search* en nodos terminales con variables y nodos no terminales
         - *Constant-variable search* en nodos no terminales

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Algoritmo:
    - Criterios de parada:
      - Se ha alcanzado el valor de MSE objetivo
      - En la iteración actual la estrategia no ha tenido éxito
        - No se han encontrado cambios en el árbol que mejoren el MSE en una cantidad determinada
    - Parámetros:
      - Número máximo de nodos
      - Estrategia a utilizar
      - Valor de MSE a alcanzar. Por defecto: 0
      - Mejora mínima en el MSE para que una búsqueda tenga éxito
        - Multiplicada por el valor del MSE actual
        - Por ejemplo, un valor de $10^{-2}$ implica que es necesaria una mejorar el MSE como mínimo en un 1%
        - Valores típicos: $10^{-5}$, $10^{-6}$, $10^{-7}$

# Regresión Simbólica

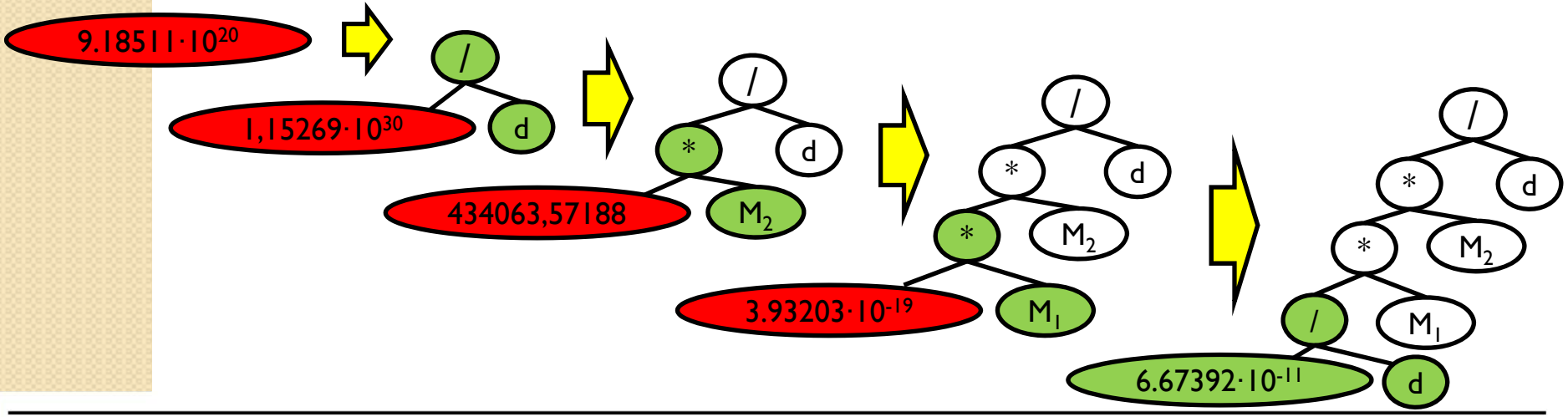
- Development of Mathematical Expressions (DoME)
  - Ejemplo: Ley de la Gravitación Universal de Newton:

$$F = G \cdot \frac{M_1 \cdot M_2}{d^2}$$

    - G = $6.67392 \cdot 10^{-11}$
    - Base de datos artificial
      - 1000 instancias, cada una con:
        - $M_1$: masa del primer planeta
          - Valor aleatorio entre $10^{23}$ y $10^{25}$
        - $M_2$: masa del segundo planeta
          - Valor aleatorio entre $10^{23}$ y $10^{25}$
        - d: distancia entre ambos planetas
          - Valor aleatorio
        - Target: resultado de la ecuación anterior

# Regresión Simbólica

- Development of Mathematical Expressions (DoME)
  - Ejemplo: Ley de la Gravitación Universal de Newton:
    - Resultados:




| Iteration | Node selected for substitution | Resulting expression | $MSE$ |
|---|---|---|---|
| 0 | - | $9.185106275464827 \cdot 10^{20}$ | $2.7645 \cdot 10^{43}$ |
| 1 | $9.185106275464827 \cdot 10^{20}$ | $(1.1526861538137104 \cdot 10^{30}/d)$ | $2.1717 \cdot 10^{43}$ |
| 2 | $1.1526861538137104 \cdot 10^{30}$ | $((434063.57187533064 \cdot M_2)/d)$ | $1.9183 \cdot 10^{43}$ |
| 3 | 434063.57187533064 | $(((3.93202929320367 \cdot 10^{-19} \cdot M_1) \cdot M_2)/d)$ | $5.1733 \cdot 10^{42}$ |
| 4 | $3.93202929320367 \cdot 10^{19}$ | $((((6.673920000000007 \cdot 10^{-11}/d) \cdot M_1) \cdot M_2)/d)$ | $3.1828 \cdot 10^{13}$ |

# Regresión Simbólica

- Ventajas de la Regresión Simbólica
  - Al contrario que las técnicas de regresión clásicas como RR.NN.AA. o SVR (SVM para Regresión), el resultado es una ecuación matemática explícita. Esto conlleva una serie de ventajas importantes:
    - Modelo explicable
    - Ofrece información sobre los datos
      - Permite ser analizada para obtener nuevo conocimiento
      - Permite descubrir relaciones ocultas entre los datos
    - Modelo muy portátil: no es necesario implementarlo en ningún lenguaje ni cargar librerías para utilizarlo
      - Caso muy habitual: se puede usar fácilmente en una hoja de cálculo

# Regresión Simbólica

- Ventajas de la Regresión Simbólica
  - Muchas técnicas intentan reducir el tamaño de las expresiones que devuelve
    - Modelos más sencillos → menos sobreajuste
  - Los resultados no son peores a los obtenidos por el resto de las técnicas de Regresión como RR.NN.AA. o SVR (SVM para Regresión)
    - Ninguna técnica de Regresión va a ser mejor que el resto en promedio
      - Teoremas No Free Lunch (NFL)
    - Los resultados parecen indicar que las técnicas de Regresión Simbólica ofrecen mejores resultados que las técnicas de regresión clásicas
      - Pero muchas veces llevan a sobreajustar → Importante limitar la complejidad

# Regresión Simbólica

- Desventajas de la Regresión Simbólica
  - Espacio de búsqueda mucho mayor que el resto de técnicas de Regresión
    - Formado por todas las posibles ecuaciones matemáticas que aproximen una relación entrada/salida
    - En otras técnicas como RR.NN.AA. o SVR (SVM para Regresión) se parte de una arquitectura fijada de antemano, con lo que el número de parámetros es fijo
    - Este espacio de búsqueda mucho mayor hace que el proceso de búsqueda sea generalmente más lento
      - Sobre todo en las técnicas basadas en poblaciones (como Programación Genética y técnicas derivadas de ella)
        - Se desperdicia mucho tiempo evaluando individuos que no van a contribuir en la solución final

# Regresión Simbólica

- Ventajas de DoME
  - Al contrario que las técnicas de Regresión basadas en poblaciones, esta tiene una base matemática que explica su funcionamiento y convergencia
  - Más rápida que las técnicas basadas en poblaciones
  - Se puede limitar explícitamente la complejidad de las expresiones que devuelve
    - Controlar el sobreajuste
  - Número muy bajo de hiperparámetros que controlan el proceso
    - Fácil experimentar

# Regresión Simbólica

- Desventajas de DoME
  - En cada iteración se realizan búsquedas en muchos nodos pero sólo se modifica uno
    - Se realizan muchos cálculos que no llevan a ninguna modificación
    - Esto se acentúa conforme el árbol se va haciendo más grande
    - Aún así, es más eficiente que las técnicas basadas en poblaciones
      - En general, que el resto de técnicas de Regresión Simbólica
    - Estos cálculos innecesarios se pueden controlar escogiendo una estrategia adecuada para recorrer los nodos del árbol
  - Por ahora la formulación sólo permite usar operadores aritméticos
    - Apropiados para problemas del mundo real, pero pueden tener limitaciones en otros tipos de problemas donde los datos tengan una naturaleza donde claramente se necesite alguna función determinada (por ejemplo, con datos periódicos)