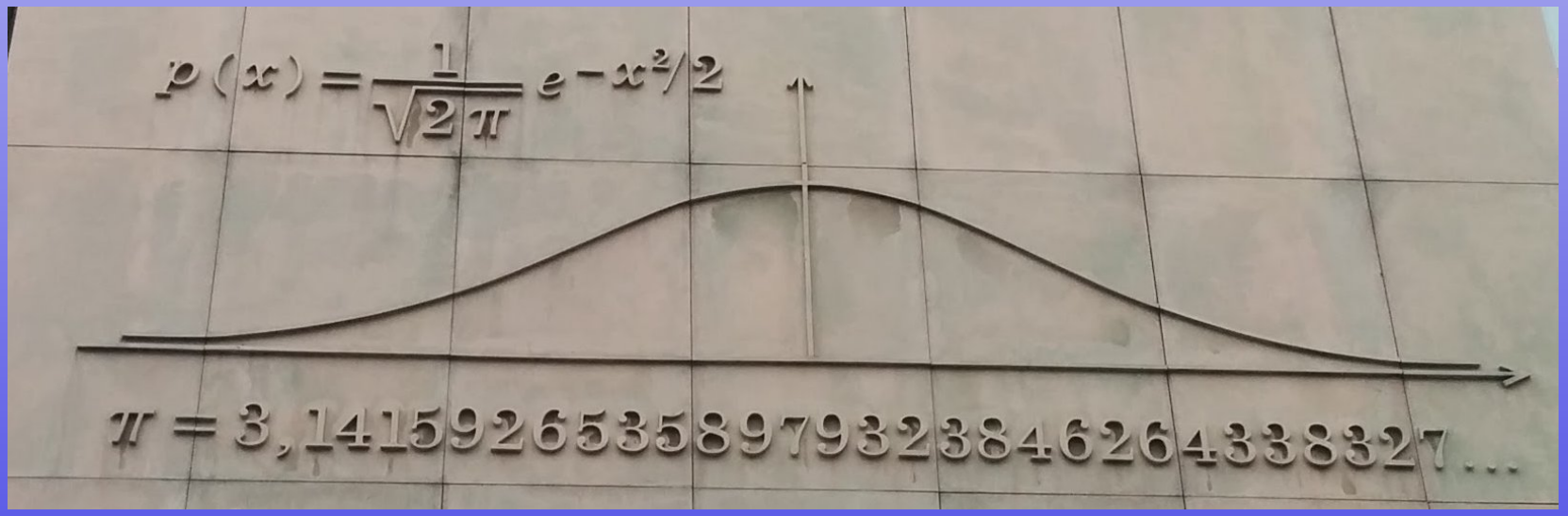
Regresja symboliczna

Dzień liczby π

Andrzej Odrzywołek

Zakład Teorii Względności i Astrofizyki, Instytut Fizyki UJ

3.14



Czy potrafisz rozpoznać liczby?

2.718281828459045

3.141592555900037

1.414213562373095

1.442695040888963

1.444667861009766

1.618033988749895

Zadanie wydaje się łatwe...

$$e \simeq 2.718281828459045$$

$$1 + e^{\frac{1}{1+\frac{1}{-1-\frac{1}{1+\frac{1}{e}}}}}{}^{2+e^e} \simeq 3.141592555900037$$

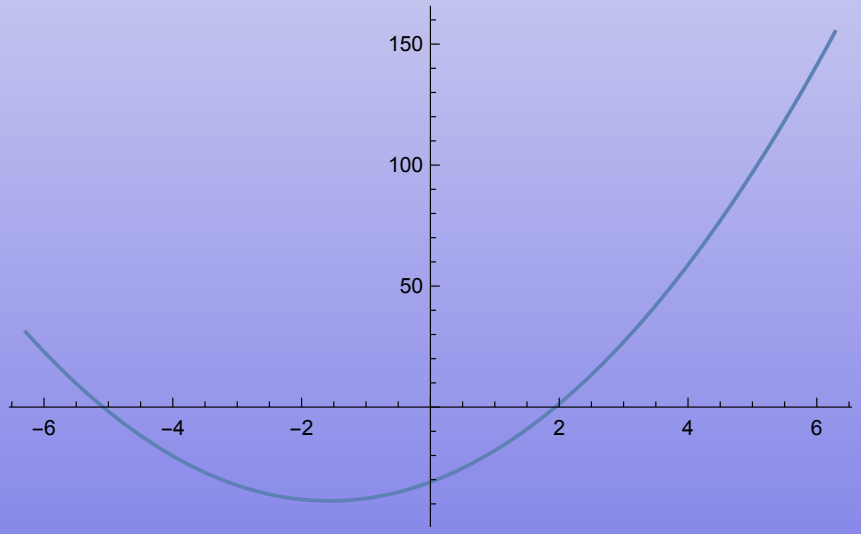
$$\sqrt{2} \simeq 1.4142135623730951$$

$$\frac{1}{\ln 2} \simeq 1.4426950408889634$$

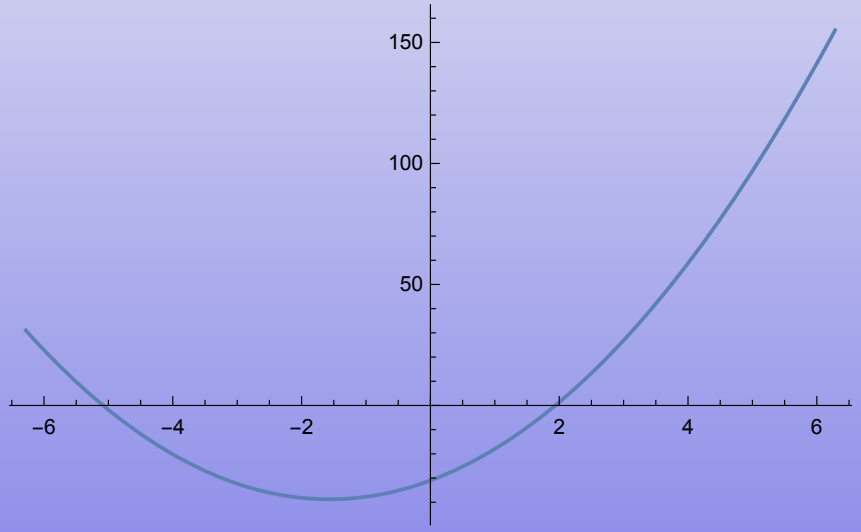
$$e^{1/e} \simeq 1.444667861009766$$

$$\varphi = \frac{1+\sqrt{5}}{2} \simeq 1.618033988749895$$

Czy można rozpoznać funkcję jednej zmiennej?

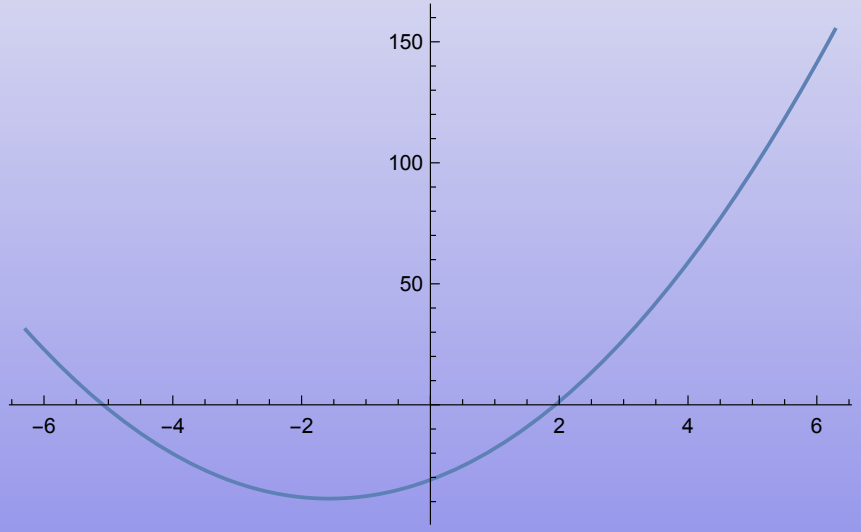


Czy można rozpoznać funkcję jednej zmiennej?



$$y = ax^2 + bx + c$$

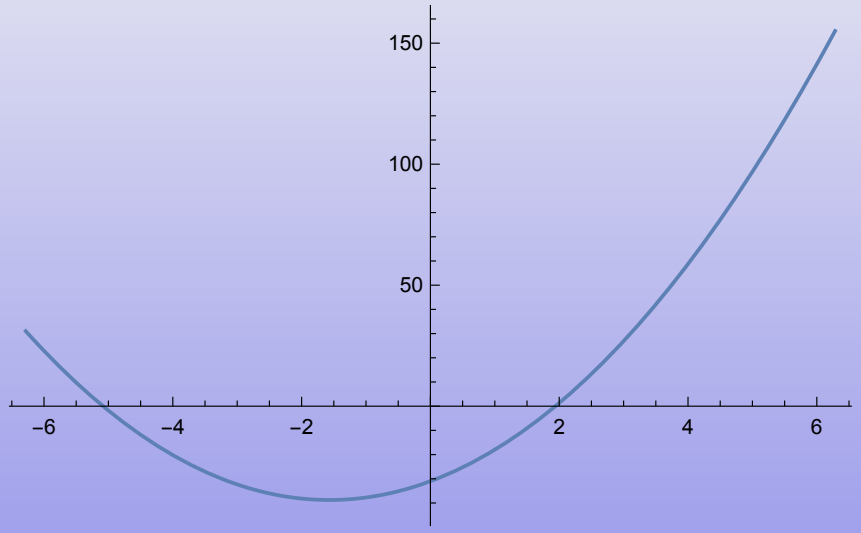
Czy można rozpoznać funkcję jednej zmiennej?



$$y = ax^2 + bx + c$$

$$y = 3.1416x^2 + 9.8696x - 31.0063$$

Czy można rozpoznać funkcję jednej zmiennej?

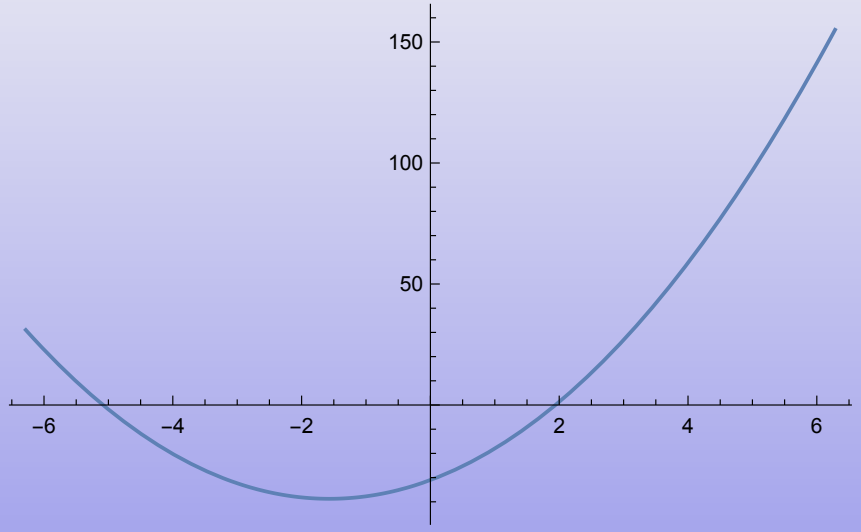


$$y = ax^2 + bx + c$$

$$y = 3.1416x^2 + 9.8696x - 31.0063$$

$$y = \pi x^2 + \pi^2 x - \pi^3$$

Czy można rozpoznać funkcję jednej zmiennej?



$$y = ax^2 + bx + c$$

$$y = 3.1416x^2 + 9.8696x - 31.0063$$

$$y = \pi x^2 + \pi^2 x - \pi^3$$

Powyższy proces, którego wynikiem jest **ścisły** matematycznie wynik, określamy mianem **regresji symbolicznej**.

Plan wykładu

1. liczby naturalne (trywialny przykład)
2. liczby wymierne (rozwiązany nietrywialny przykład)
3. wielomiany o współczynnikach całkowitych (nieoczywisty przykład)
4. przybliżanie i rozpoznawanie rozpoznawanie stałych przestępnych (działający trudny przykład)

Zadanie (trywialne)

Dla jakiego naturalnego n otrzymamy najlepsze przybliżenie liczby:

$$e^{\pi} - \pi$$

?

Oczywista metoda brute-force to wstawianie po kolei: 0,1,2,3,4, ...

UWAGA!

Gdyby chodziło o rozwiązania całkowite, wstawiamy po kolei:

$$0, 1, -1, 2, -2, 3, -3 \ldots$$

Idiotyczny błąd, formalnie możliwy do popełnienie, to wstawianie najpierw liczb dodatnich, potem ujemnych, lub wstawianie w kolejności, która pomija pewne liczby, np:

$$(-1)^k k \quad \text{zamiast} \quad \frac{1-(-1)^k(2k+1)}{4}.$$

O ile dla liczb całkowitych tego typu pomyłka byłaby absurdalna, w przypadku np: funkcji elementarnych podobny błąd łatwo popełnić i przeoczyć.

Liczby naturalne/całkowite : podejście funkcyjne

Zdefiniujmy trzy funkcje (zero, sukcesor, negacja):

$$Z(n) = 0, \qquad S(n) = n + 1, \qquad N(n) = -n.$$

Dowolne złożenie funkcji Z, S, N zawsze ma sens, i reprezentuje funkcję stałą, przyjmującą wartości całkowite.
Dowolny ciąg znaków ZSN reprezentuje liczbę całkowitą, np. cztery:

$$NSSSSZ.$$

Taka reprezentacja nie jest jednoznaczna, ale jest dobrą podstawą do metod genetycznych, gdyż po operacjach duplikacji, delecji, insercji czy odwrócenia nadal reprezentuje liczbę całkowitą. Pozwala także na wprowadzenie śmieciowego DNA, typu NNNNNNNNNN czy markerów np: ZZZ czy NSZZ pełniących rolę kodonu STOP/START.

Spostrzeżenia

- Zmiana znaku jest funkcją odwrotną do samej siebie. Funkcję $x + 1$ można zastąpić funkcją $-x + 1$, o tej samej własności.
- ciągi liter Z,S,N to liczby w systemie trójkowym
- można jednoznacznie powiązać liczbę całkowitą z numerem pierwszego pojawienia się (przy ustalonej kolejności „cyfr” Z, S, N)

Liczby wymierne

Zadanie

Dla jakiego wymiernego r otrzymamy najlepsze przybliżenie liczby π:

$$r = \frac{n}{m} \simeq \pi$$

?

Cantor podał metodę enumeracji liczb wymiernych $\frac{n}{m}$ o rosnących n, m. Liczby powtarzają się (np: $\frac{2}{4} = \frac{1}{2}$), ale redundancja jest niewielka, asymptotycznie dąży do złotego podziału.

Jak generować liczby wymierne bez powtórzeń?

Okazuje się, że od połowy XIX w. znane są sposoby enumeracji liczb wymiernych:

1. drzewo Sterna-Brocota (przeszukiwanie binarne „przedziału" $[0, \infty]$)
2. ciąg Fareya
3. ułamki łańcuchowe
4. składanie funkcji

Najciekawsze ponownie okazuje się podejście funkcyjne.

Funkcyjne generowanie liczb wymiernych

Zdefiniujmy cztery funkcje (one, minus sukcesor, negacja, odwrotność):

$$ONE(r) = 1, \qquad NSU(r) = -r + 1, \qquad NEG(r) = -r, \qquad INV(r) = 1/r.$$

Trzy ostanie są „samoodwrotne", i pozwalają poprzez składanie na generowanie wszystkich liczb wymiernych na zasadzie ułamka łańcuchowego.

Wszystko co powiedziano o liczbach całkowitych przenosi się na wymierne.

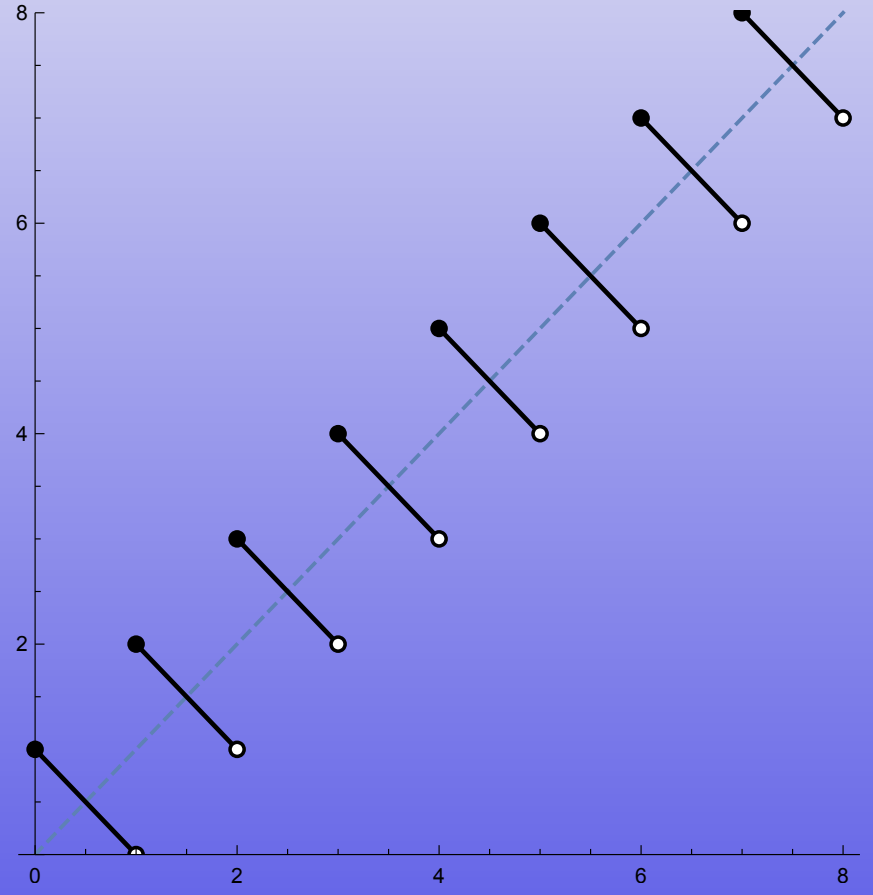
Niespodzianka

Istnieje funkcja, której składanie generuje po kolei wszystkie liczby wymierne bez powtórzeń. Jest ona złożeniem dwóch funkcji „self-inverse":

$$INV(r) = \frac{1}{r}, \qquad LAD(r) = Floor[x] + 1 - FractionalPart[x]$$

Wykres funkcji LAD(r)

$$LAD(x) = Floor[x] + 1 - FractionalPart[x]$$



Generowanie ułamków poprzez składanie funkcji

Okazuje się, że składając wielokrotnie funkcję:

$$f(x) = \frac{1}{Floor[x] + 1 - FractionalPart[x]}$$

otrzymujemy wszystkie liczby wymierne (w tym naturalne), w określonej kolejności, bez powtórzeń!

$$0, 1, \frac{1}{2}, 2, \frac{1}{3}, \frac{3}{2}, \frac{2}{3}, 3, \frac{1}{4}, \frac{4}{3}, \frac{3}{5}, \frac{5}{2}, \frac{2}{5}, \frac{5}{3}, \frac{3}{4}, 4, \frac{1}{5}, \frac{5}{4}, \frac{4}{7}, \frac{7}{3}, \frac{3}{8}, \frac{8}{5}, \frac{5}{7}, \frac{7}{2}, \frac{2}{7}, \frac{7}{5}, \frac{5}{8}, \frac{8}{3}, \frac{3}{7}, \frac{7}{4}, \frac{4}{5}, 5, \ldots$$

Można w ten sposób wygenerować wszystkie możliwe przybliżenia wymierne π:

Przybliżenie	Wartość numeryczna	Złożoność Kołmogorowa	Błąd bezwzględny
3/1	3.0	7	-0.1415926535
13/4	3.25	71	0.1084073464
16/5	3.2	135	0.0584073464
19/6	3.1666(6)	263	0.025074
22/7	3.142857142857143	519	0.00126449
179/57	3.14035087719	261 127	-0.0012417763
201/64	3.140625	523 271	-0.0009676535
223/71	3.1408450704225355	1 047 559	-0.0007475831
245/78	3.141025641025641	2 096 135	-0.0005670125
267/85	3.1411764705882352	4 193 287	-0.0004161830
289/92	3.141304347826087	8 387 591	-0.0002883057
311/99	3.1414(14)	16 776 199	-0.0001785121
333/106	3.141509433962264	33 553 415	-0.0000832196
355/113	3.141592920353982	67 107 847	0.0000002667

Własności kolejnych przybliżeń wymiernych π

Przybliżenie „zerowe”:

$$\pi \simeq \frac{3}{1}$$

Przybliżenie „pierwsze”: $\frac{13}{4}$ „plus” 3x najlepsze przybliżenie zerowe:

$$\pi \simeq \frac{22}{7} = \frac{13+3+3+3}{4\ \ +1+1+1}$$

Przybliżenie drugie: $\frac{179}{57}$ „plus” 8x najlepsze przybliżenie poprzednie:

$$\pi \simeq \frac{335}{113} = \frac{179+22+22+22+22+22+22+22+22}{57\ +\ \ 7+\ \ 7+\ \ 7+\ \ 7+\ \ 7+\ \ 7+\ \ 7+\ \ 7}.$$

Jak wyjść poza przybliżenia ułamkowe?

Jak szukać kolejnych przybliżeń zawierających np: $\sqrt{}$?

$$\pi \simeq \sqrt{2}+\sqrt{3}, \qquad \pi \simeq 2+\sqrt{\sqrt{1+\frac{1}{\sqrt{2}}}}, \qquad \pi \simeq 1+\sqrt{4+\frac{1}{1+\frac{1}{\sqrt{2}}}}$$

Lub logarytm naturalny $\ln$?

$$\pi \simeq 1+\frac{1}{\ln 2}+\ln 2, \qquad \pi \simeq \frac{1}{\log\left(\frac{11}{8}\right)}$$

PROJECT RAND

A DECISION METHOD FOR ELEMENTARY ALGEBRA AND GEOMETRY

ALFRED TARSKI

Prepared for Publication by J. C. C. McKinsey

This report, although published by the RAND Corporation, was written while the Project was a part of Douglas Aircraft Co., Inc.

August 1, 1948

(Revised May, 1951)

R-109

The RAND Corporation

1700 MAIN ST. • SANTA MONICA • CALIFORN

SECTION 1.

THE SYSTEM OF ELEMENTARY ALGEBRA

In this section we want to describe a formal system of elementary algebra — and in particular to define in a precise way the class of sentences of this system[7].

By a *variable* we shall mean any one of the following symbols:

$$x,\ x_1,\ x_2, \ldots;\ y,\ y_1,\ y_2, \ldots;\ z,\ z_1,\ z_2, \ldots\ .$$

We suppose that there are infinitely many variables and that they are arranged in a sequence, so that we can speak of the variable occupying the 1st, 2nd,..., nth place in the sequence. These variables are to be thought of intuitively as ranging over the set of real numbers.

By an *algebraic constant* we shall mean one of the following three symbols:

$$1,\ 0,\ -1\ .$$

By an *algebraic operation-sign* we shall mean one of the following two symbols:

$$+,\ \cdot\ .$$

The first is called the *addition sign*, and the second the *multiplication sign*.

By an *algebraic term* we understand any meaningful expression built up from variables and algebraic constants by means of the elementary operation-signs. Thus, for example,

$$x,\qquad x_1 + y,\ -1\cdot x,\qquad \left[\left(x_1 \cdot -1\right)\cdot x_1\right] + x_2$$

are algebraic terms. But

$$x + ,\qquad \sqrt{2} + x$$

are not algebraic terms: the first, because it is meaningless; the second, because it involves the sign " $\sqrt{2}$ ", which is neither a variable nor an algebraic constant (in the restricted meaning we have given to the latter term).

If one wants a precise definition of algebraic terms, they can be defined recursively as follows: An algebraic term of first order is simply a variable or one of the three algebraic constants. If α and β are algebraic terms of order at most k, and if the maximum of the orders of α and β is k, then $(\alpha\cdot\beta)$ and $(\alpha + \beta)$ are algebraic terms of order $k + 1$. An expression is called an algebraic term if, for some k, it is an algebraic term of order k.

According to the above definition, one should inclose in parentheses the results of performing operations on terms. Thus one should write, for example, always

Elementarna algebra wg. Tarskiego

Oryginalnie, minimalny (startowy) zbiór symboli Tarskiego to:

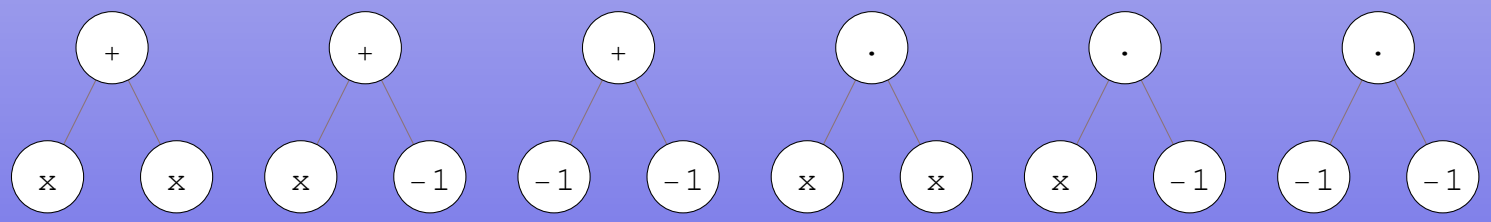
$$-1, 0, 1, x + y, x \cdot y, x, y, z, \ldots$$

Zero i 1 są redundantne:

$$1 = (-1) \cdot (-1), \qquad 0 = -1 + 1$$

Wygenerowanie wszystkich możliwych wyrażeń elementarnej algebry, polega na rekurencyjnym stosowaniu symboli i operacji bazowych. Tradycyjne symbole matematyczne traktujemy jako „skróty myślowe" np:

$$5 = 1 + 1 + 1 + 1 + 1, \qquad x^2 + 1 = x \cdot x + (-1) \cdot (-1), \ldots.$$

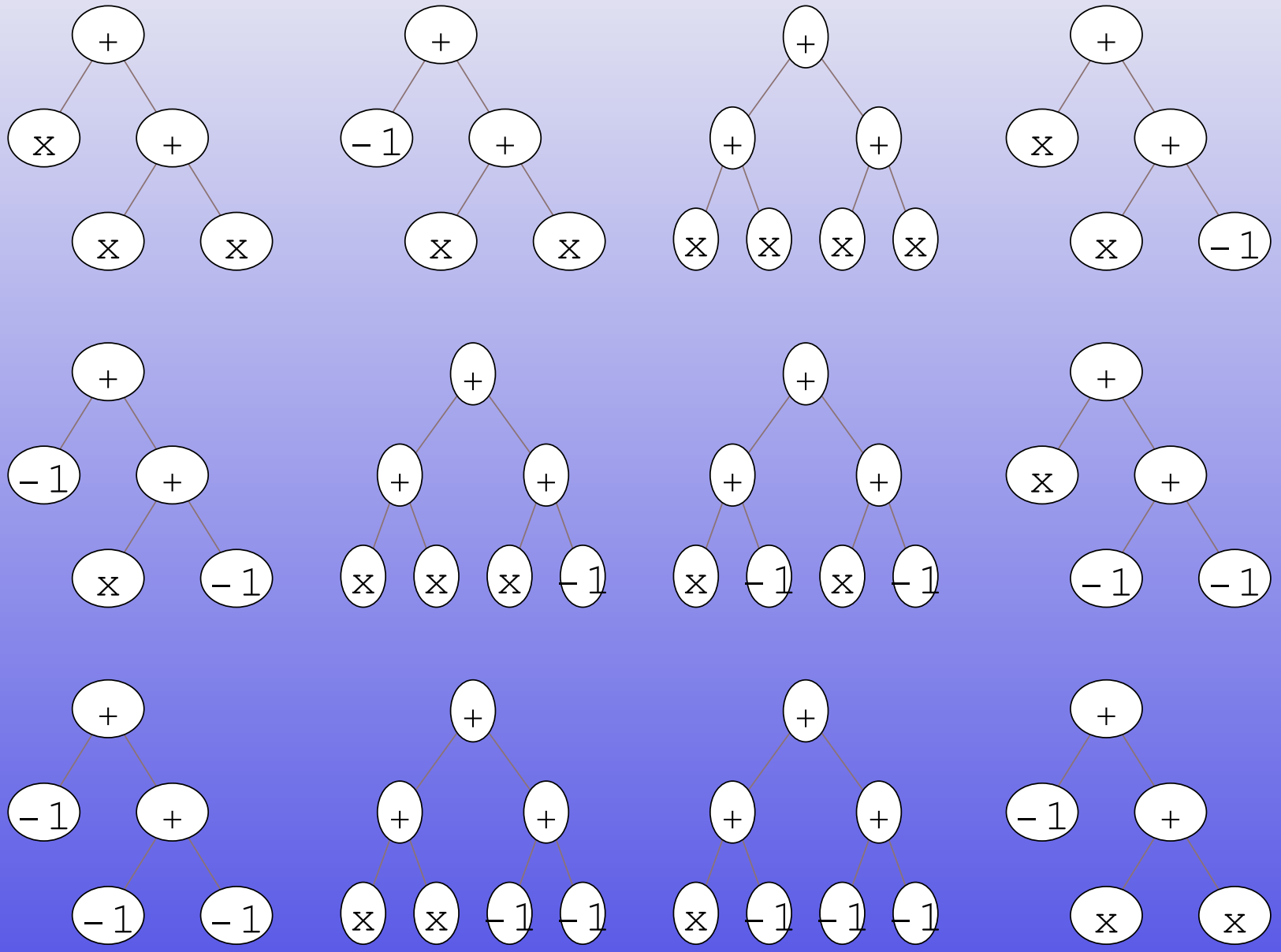


$$x+x = 2x, x+(-1) = x-1, (-1)+(-1) = -2, x \cdot x = x^2, (-1) \cdot x = -x, (-1) \cdot (-1) = 1$$

Lista pierwszych wyrażeń:

$$x, -1, 2x, x-1, -2, x^2, -x, 1, 3x, 2x-1, 4x, 2x-1, x-2, 3x-1, 2(x-1), x-2, -3, 2(x-1), x-3,$$

Elementarna algebra wg. Tarskiego



Zagadnienie zepsutego kalkulatora

„Broken calculator problem"

Czy jeżeli w kalkulatorze zostały nam tylko klawisze $+$, EXP, LN możemy nadal mnożyć i potęgować?

Praktyka pokazuje, że klawiszy nie można wyrywać w zupełnie dowolny sposób. Jest to główne źródło niepowodzeń i kiepskiej/losowej skuteczności istniejących metod regresji symbolicznej.
Przykłady zadań:

- czy dysponując klawiszami $+, \exp, \log, -, 1$ możemy obliczyć π ?
- czy dysponując symbolami $+, \times, \wedge, \ln, e$ możemy uzyskać -1 ?

Konieczne jest wydzielenie spośród zbioru „wszystkich" symboli tych, których nie wolno usunąć, pod groźbą ostatecznego zepsucia naszego kalkulatora.

Przykłady minimalnego zbioru funkcji/operatorów

Zbiór minimalny „botton-up”

1. dodawanie $+$, $x + y$
2. f. wykładnicza $\exp$, e^x
3. logarytm naturalny $\ln$, lnx
4. odwrotność (brak symbolu) $\frac{1}{x}$

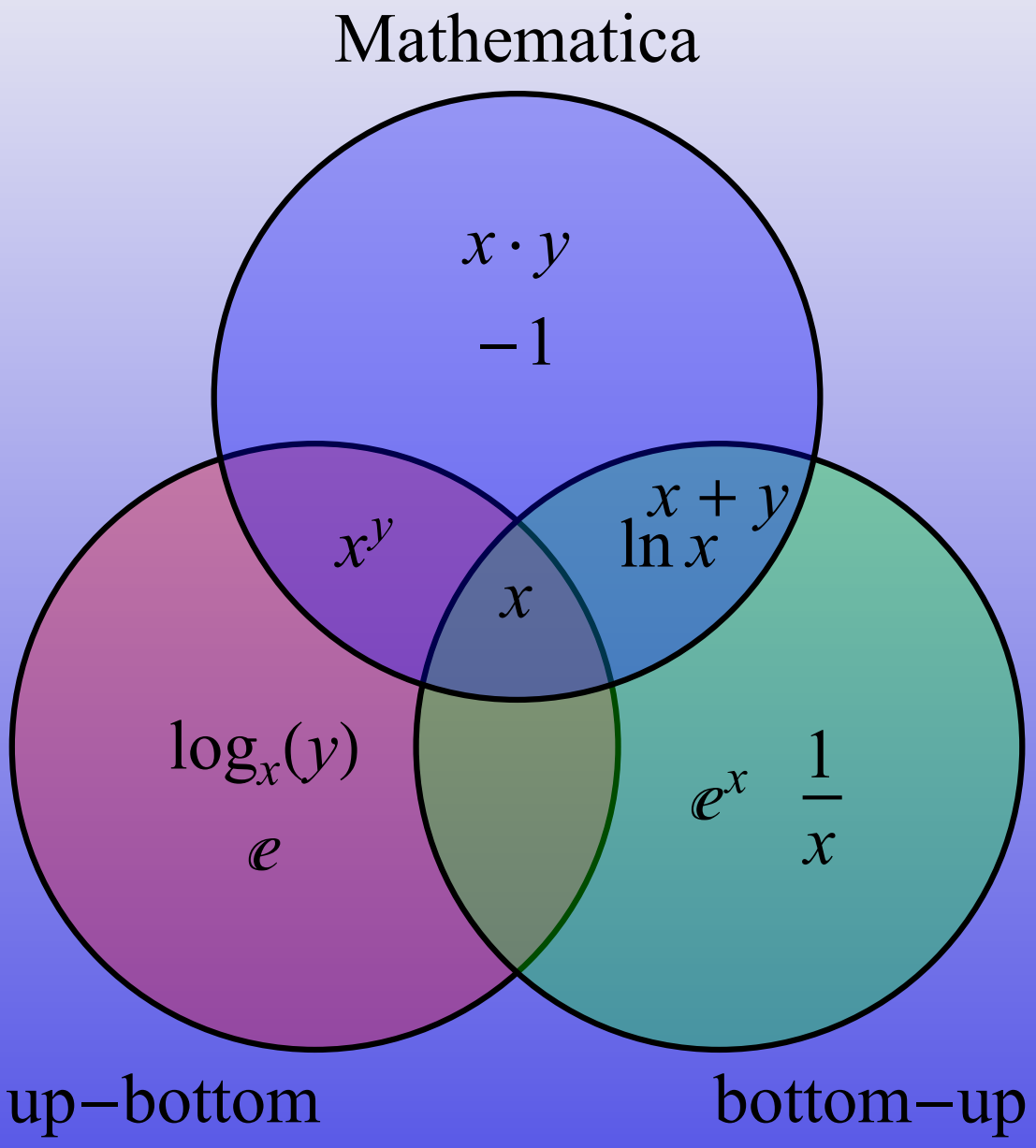
Zbiór minimalny „up-bottom”

1. potęgowanie $**$, x^y
2. logarytm dwuargumentowy (o dowolnej podstawie) $\log_x y$
3. liczba E (podstawa logarytmu naturalnego) lub π

Dla porównania: *Mathematica*

1. potęgowanie **Power**, mnożenie **Times**, dodawanie **Plus** (wieloargumentowe!)
2. logarytm naturalny **Log**
3. liczba **-1** oraz spory zestaw redundantnych stałych przestępnych (**E**, **Pi**, ...)
4. liczby całkowite **Integer**, wymierne **Rational**, zespolone **Complex**, algebraiczne **Root**
5. „skróty” typu **Sin**, **ArcTan**, **LegendereP** ... rozwijane przez **FunctionExpand**, **TrigToExp**, ...

Powyższe pozwala generować liczby. Aby generować funkcje, trzeba dodać zmienne i parametry $x, y, z, \ldots a, b, c \ldots$. Ewentualnie dodać stałe, np: $\pi, 2, \sqrt{2}, \ln 2, \ldots$.


Mathematica
$x \cdot y$
-1
x^y
$x + y$
$\ln x$
x
$\log_x(y)$
e
e^x
$\frac{1}{x}$
up−bottom
bottom−up

Pierwsze przybliżenia π w bazie $+, \ln, \exp, 1/x, 1$

Pierwsze przybliżenia π

$$\begin{pmatrix} 1 & 0 & 1 & \{\text{ONE}\} & 1. \\ e & 5 & 2 & \{\text{EXP}, \text{ONE}\} & 2.71828 \\ 3 & 2600 & 5 & \{\text{PLUS}, \text{ONE}, \text{PLUS}, \text{ONE}, \text{ONE}\} & 3. \\ \frac{1}{e} + e & 13180 & 6 & \{\text{PLUS}, \text{EXP}, \text{ONE}, \text{INV}, \text{EXP}, \text{ONE}\} & 3.08616 \\ e^{e^{\frac{1}{e^2}}} & 20225 & 7 & \{\text{EXP}, \text{EXP}, \text{INV}, \text{EXP}, \text{PLUS}, \text{ONE}, \text{ONE}\} & 3.14219 \end{pmatrix}$$

Reprezentacja operacji arytmetycznych

$$x \cdot y = e^{\ln x + \ln y}, \qquad x^y = \exp\left(e^{\ln y + \ln \ln x}\right)$$

Pierwsze przybliżenia w bazie $x^y, \log_x y, e$

Pierwsze przybliżenia π

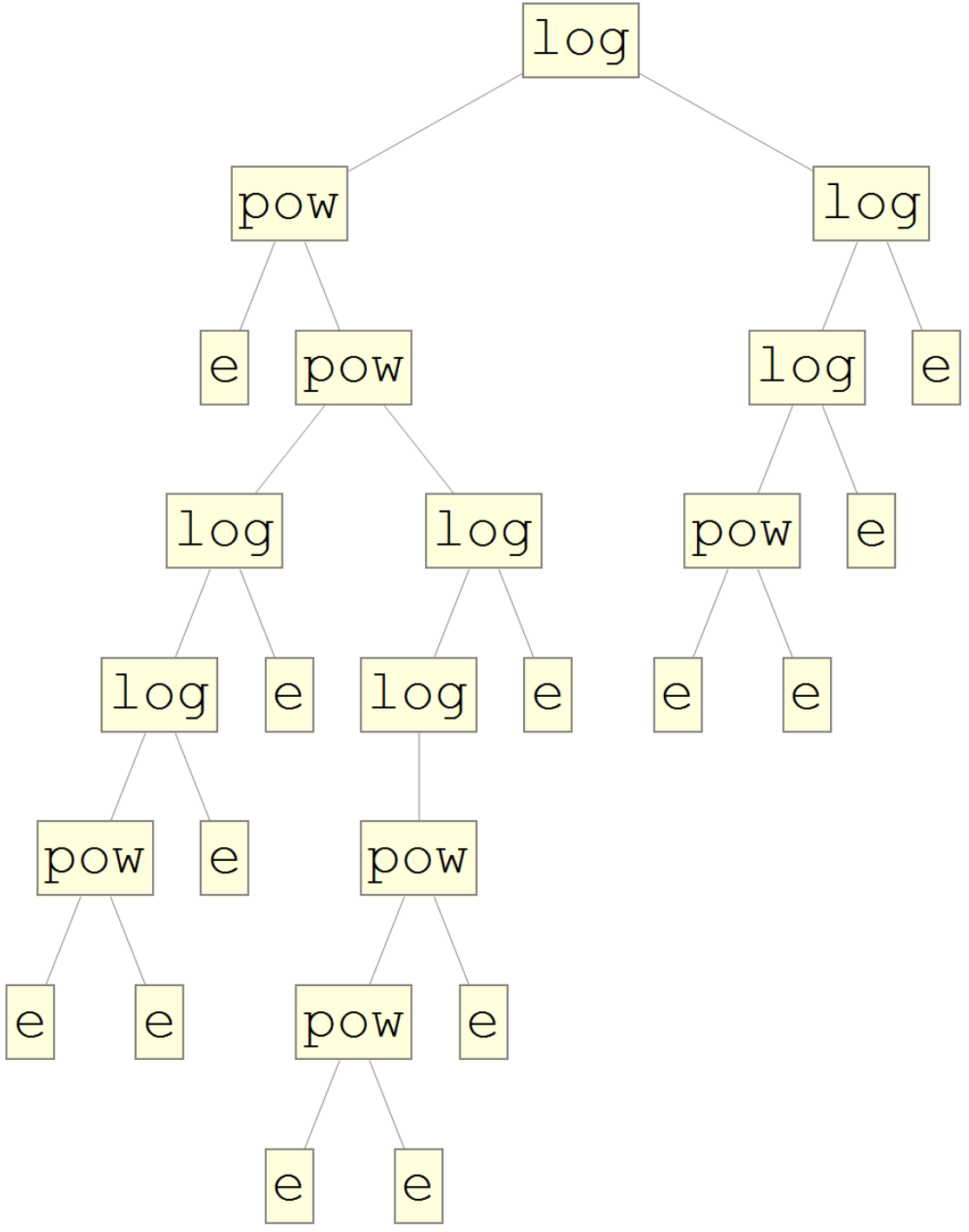
e	0	1	{E}	2.71828
3	67716	11	{LOG, E, LOG, E, POW, POW, POW, E, E, E, E}	3.
$e^{e^{\frac{1}{e^2}}}$	132597	11	{POW, E, POW, E, LOG, POW, POW, E, E, E, E}	3.14219

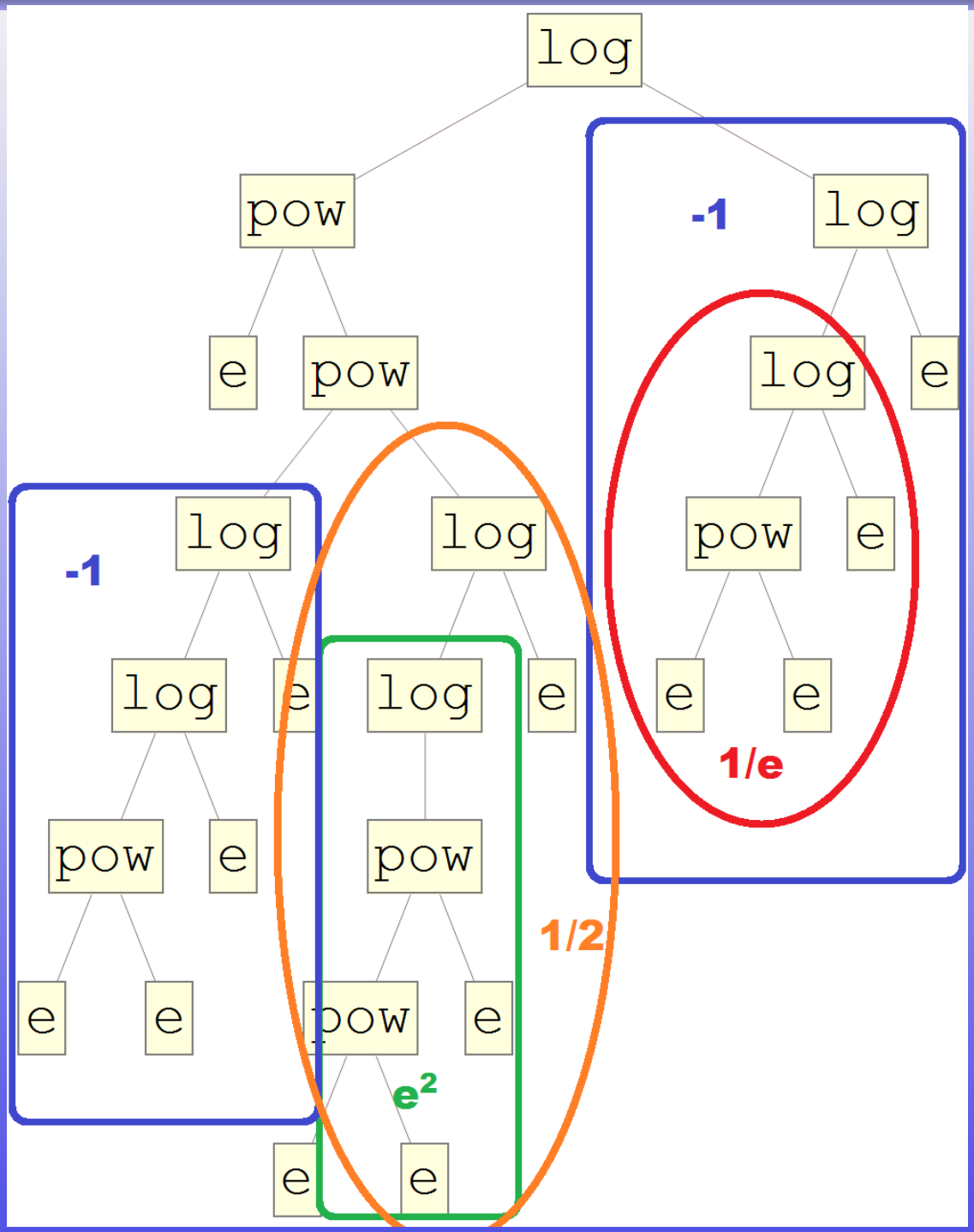
Reprezentacja operacji arytmetycznych

$$x \cdot y = \log_x \left((x^x)^y\right), \qquad x + y =$$

Liczba π wyliczona dokładnie z liczby e

$$\pi = \log_{\left[e^{\left(\log_{(\log_e e\, e)}\, e\right)^{\log_{[\ln\,(e^e)\,e]}\,[e]}}\right]} \left[\log_{(\log_e e\, e)}\, e\right]$$


log
pow
log
e
pow
log
e
log
log
pow
e
log
e
log
e
e
e
pow
e
pow
e
e
pow
e
e
e


log
pow
-1
log
e
pow
log
e
log
log
-1
pow
e
log
e
log
e
e
e
1/e
pow
e
pow
1/2
e
e
pow
e
e²
e
e

$$\pi = \log_{e^i}(-1), \qquad i = \sqrt{-1}$$

$$\pi = \log_{e^i}(-1) = \frac{\ln -1}{\ln e^i} = \frac{i\pi}{i \ln e} = \pi$$

Hierarchia operatorów: top-bottom czy bottom-up ?

Pokazano, że zbiór symboli startowych i operacji jest w zasadzie dowolny, jeżeli zawiera zbiór minimalny.

1. bramka logiczna NAND
2. zeracja (sukcesor) $x + 1 + 1 + 1 + 1 + 1 + \ldots + 1 + 1 = x + n$
3. dodawanie/odejmowanie $x + x + x + x + \ldots + x + x + x = nx$
4. mnożenie/dzielenie $x \cdot x \cdot x \cdot \ldots \cdot x = x^n$
5. potęgowanie/logarytm $x^{x^{\cdot^{\cdot^{x}}}} =^n x$
6. tetracja/superlogarytm/superpierwiastek

Dla teoretyka informatyki cyfrowej operacje logiczne i dodawanie są pierwotne, dla fizyka czy cybernetyka: potęgowanie i logarytmy.

- Operatory wyższe niż potęgowanie są nadal słabo zbadane
- Nie ma zgody co do kontynuacji analitycznej na płaszczyźnie liczb zespolonych
- Nie wiadomo np: czy ${}^4\pi$ jest liczbą naturalną !

Potencjał generowania $\pi, e, \ldots$ lub/i funkcji specjalnych, liczb algebraicznych itp?

Rozpoznawanie stałych: praktyka

Najpopularniejszym aktualnie zastosowaniem regresji symbolicznej jest dopasowanie wzoru analitycznego do zadanej liczby zmiennoprzecinkowej.

1. Maple (identify)
2. Mathematica (FindFormula, FindSequenceFunction, RootApproximant), WolframAlpha
3. Inverse Symbolic Calculator (https://isc.carma.newcastle.edu.au/), PSLQ
4. RIES (http://mrob.com/pub/ries/)
5. nsimplify (SymPy)
6. A.O. https://github.com/VA00/SymbolicRegressionPackage

Całki oznaczone: testy

$$\int_0^{\pi/2} \ln \sin x \, dx, \qquad \int_0^1 \frac{arctg\sqrt{x^2+2}}{(x^2+1)\sqrt{x^2+2}} dx, \qquad \int_0^1 \left(\frac{1}{\ln x} + \frac{1}{1-x}\right)^2 dx,$$

$$\int_{-1}^{1} \frac{1}{x}\sqrt{\frac{1+x}{1-x}} \ln\left(\frac{2\,x^2+2\,x+1}{2\,x^2-2\,x+1}\right) \, \mathrm{d}x$$

	WolframAlpha	identify	ISC	RIES	Google
$x \simeq -2.17758609$	none	$-\pi \ln(2)$	-ln(2)/ln(exp(1/Pi))	$-x/\pi = \ln 2$ $\ln(2^{-\pi})$	TAK
$x \simeq 0.51404189589$	$\frac{5\zeta(2)}{16}$	$\frac{5}{96}\pi^2$	$\frac{5\pi^2}{96}$	$x = \left(\frac{\pi}{4}\right)^2 - \frac{x}{5}$	TAK
$x \simeq 0.3378770664$	$\frac{857261070\,\pi}{7970843089}$	—	x+2 = 1/2+ln(2)+ln(Pi)	$x = \ln(2\pi) - \frac{3}{2}$	TAK
$x \simeq 8.3722116266$	matem. bełkot	—	Wow, really found nothing	$\cos\left(\frac{x}{\pi}\right) + 1 = \frac{2}{\phi^6}$	NIE

Podsumowanie: zastosowania regresji symbolicznej

- matematyka rekreacyjna
- generowanie losowych zadań matematycznych
- testowanie oprogramowania matematycznego
- szukanie rozwiązań specjalnych równań różniczkowych, funkcyjnych, całek, itp.
- przemysłowe generowanie modeli matematycznych i teorii