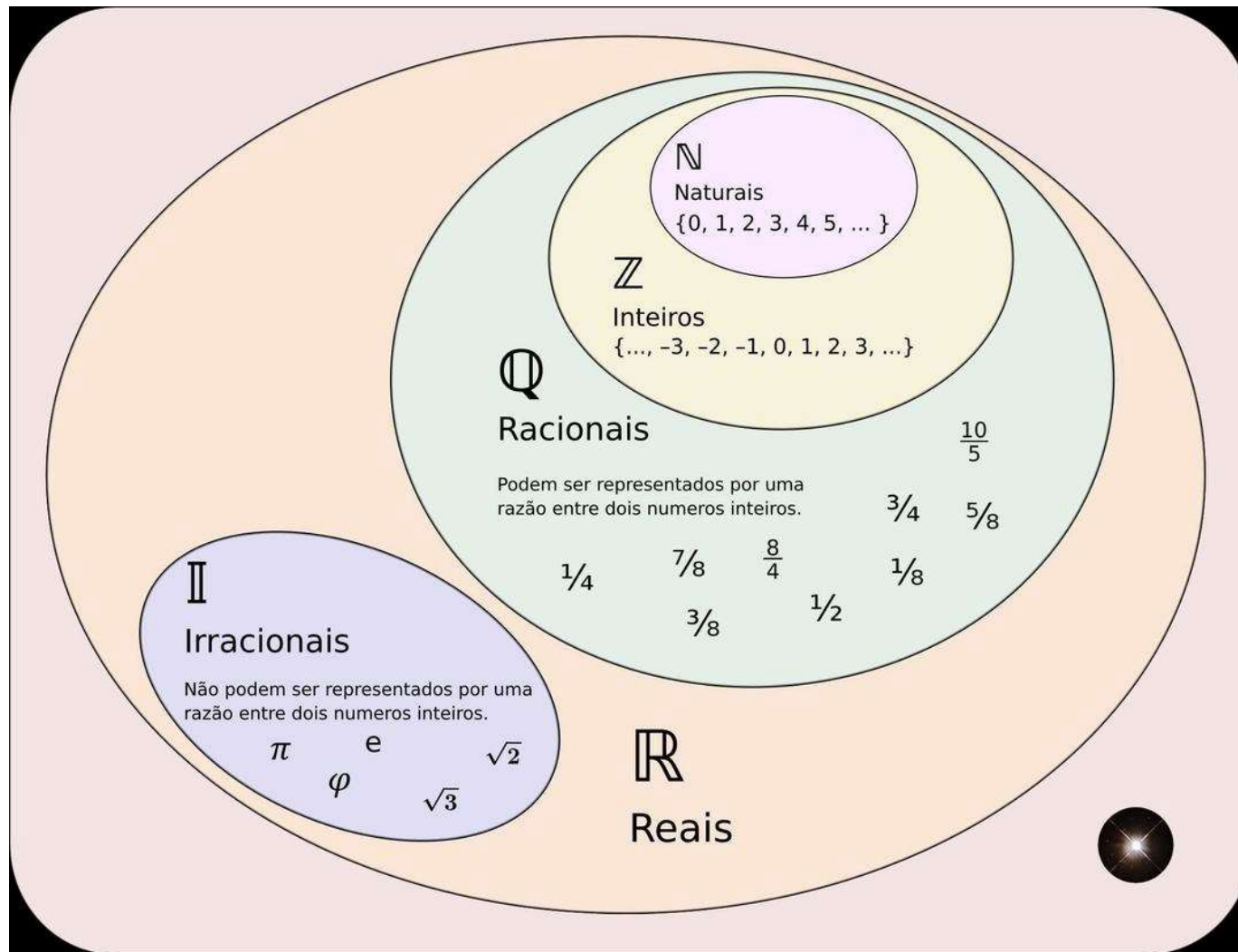


Conjuntos numéricos



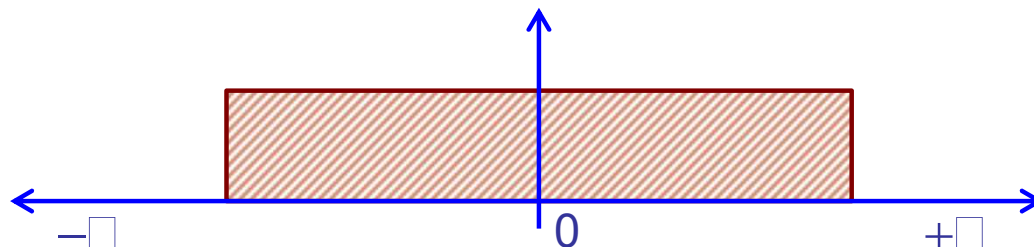
Representação de reais em vírgula flutuante (1)



- **Gama de valores**

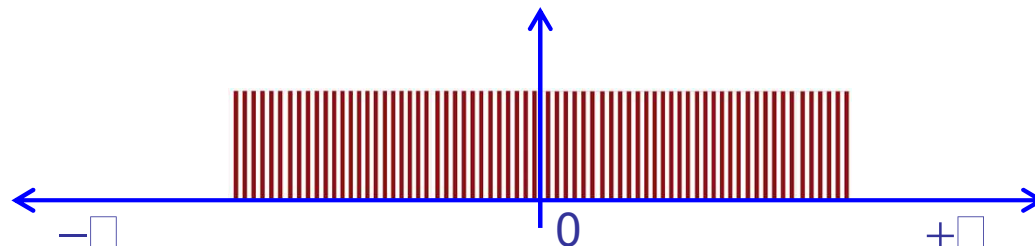
- esta gama é viável?
- não, porque...

entre quaisquer 2 n^os reais há um conjunto infinito de valores



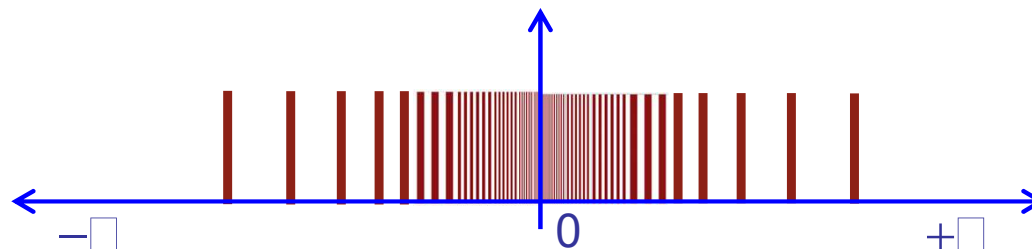
- então faz sentido a seguinte representação?

Notação vírgula fixa



- ou faria mais sentido a seguinte representação?

Notação científica



Representação de reais em vírgula flutuante (2)



- **Notação em vírgula fixa vs. notação científica**

- **Exemplos na base 10 usando apenas 5 dígitos**

- **Vírgula fixa: $\pm X X X . X X$**

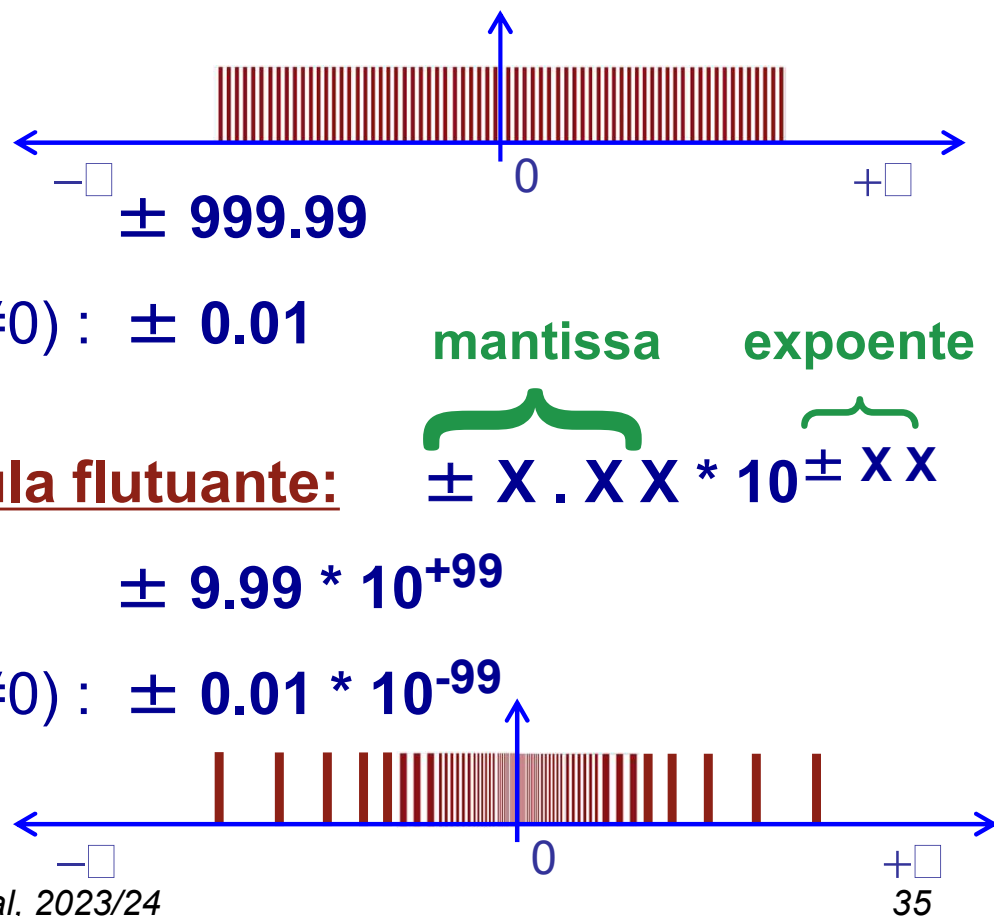
maior valor representável : ± 999.99

menor valor representável ($\neq 0$) : ± 0.01

- **Notação científica ou vírgula flutuante:**

maior valor representável : $\pm 9.99 * 10^{+99}$

menor valor representável ($\neq 0$) : $\pm 0.01 * 10^{-99}$



Representação de reais em vírgula flutuante (3)



- **Normalização na representação**

- Para garantir que cada n° a representar usa apenas uma única sequência de dígitos, há que fixar o local para o ponto decimal
- Nº normalizado: à esquerda do ponto decimal apenas 1 dígito ≠ 0
$$\pm Y . XXX \dots X * \text{Radix}^{\text{Exp}} \quad 0 < Y \leq \text{maior_dígito}$$
- Menor valor normalizado representável:
$$\pm 1.000 \dots 0 * 10^{-9\dots 9}$$
- Nota_1: a normalização não permite representar o zero
Nota_2: a normalização desperdiça muitos dígitos → subnormais
- Nº subnormal: expoente fixo (menor normalizado), 0 à esq do ponto
de $\pm 0.000 \dots 0 * 10^{-9\dots 9}$ a $\pm 0.999 \dots 9 * 10^{-9\dots 9}$

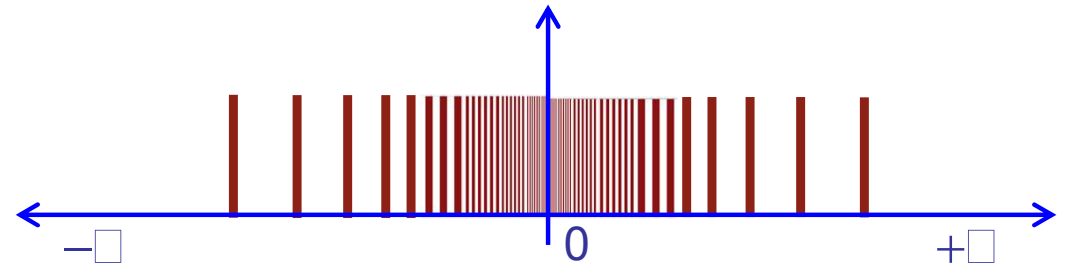
Representação binária de reais em vírgula flutuante (4)



- **Formato binário dum valor em fp**
 - formato inclui mantissa e expoente; radix é implícito (=2)
 - mantissa e expoente podem ser valores positivos ou negativos
 - formato tem n° de bits fixo, variável:
 - **reais com precisão simples: 32 bits**
 - **reais com precisão dupla: 64 bits**
 - **reais com meia precisão : 16 bits**

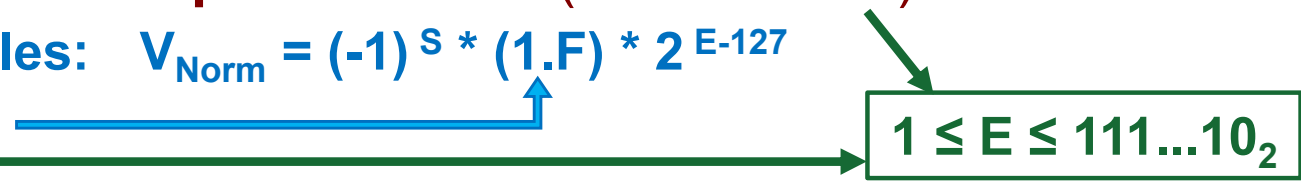
▪ mantissa vs. expoente ⇔

precisão vs. intervalo valores representáveis



A norma IEEE 754-2008 para valores em fp (1)



- **Representação do sinal na mantissa e no expoente**
 - Mantissa: $S + M$
 - Expoente: Excesso $2^{n-1} - 1$
- **Valor decimal de um fp em binário (normalizado)**
 - Precisão simples: $V_{\text{Norm}} = (-1)^S * (1.F) * 2^{E-127}$
 - Bit escondido 
- Exceções (zero, subnormais, ...): $E = 0$ ou $E = 1111...1_2$
- **Representação do zero:** $E = 0$ e $F = 0$
- **Representação e valor decimal de um fp (subnormal)**
 - Representação: $E = 0$ e $F \neq 0$
 - Precisão simples: $V_{\text{SubN}} = (-1)^S * (0.F) * 2^{-126}$
- **Representação de $\pm\infty$:** $E = 1111...1_2$ e $F = 0$
- **Representação de n.º não real:** $E = 1111...1_2$ e $F \neq 0 \rightarrow \text{NaN}$

A norma IEEE 754-2008 para valores em fp (2)



Format of Floating points IEEE754

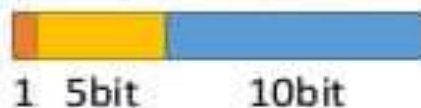
64bit = double, double precision



32bit = float, single precision



16bit = half, half precision

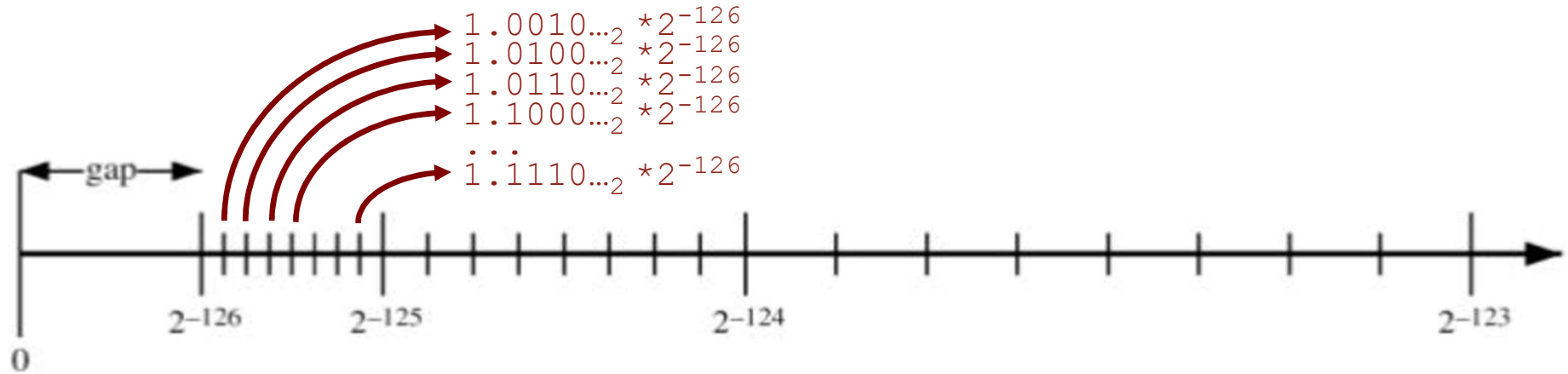


Signal

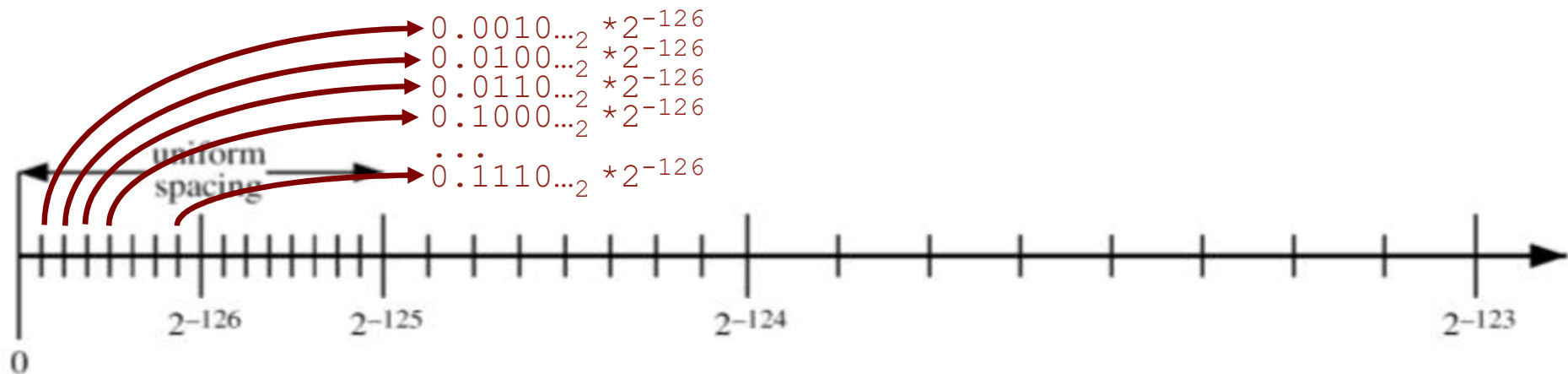
Exponent

Mantissa

O papel dos subnormais na norma IEEE 754



(a) 32-bit format without denormalized numbers



(b) 32-bit format with denormalized numbers

Tipo de dados suportados pela linguagem C

(table based on <https://en.cppreference.com/w/cpp/language/types>)



Type	Size in bits	Format	Value range	
			Approximate	Exact
character	8 (char)	signed		-128 to 127
		unsigned		0 to 255
	16	UTF-16		0 to 65535
	32	UTF-32		0 to 1114111 (0x10ffff)
Integer (LP64)	16 (short int)	signed		-32 768 to 32 767
		unsigned		0 to 65 535
	32 (int)	signed	$\pm 2.14 \cdot 10^9$	-2 147 483 648 to 2 147 483 647
		unsigned	0 to $4.29 \cdot 10^9$	0 to 4,294,967,295
	64 (long)	signed	$\pm 9.22 \cdot 10^{18}$	-9 223 372 036 854 775 808 to 9 223 807
		unsigned	0 to $1.84 \cdot 10^{19}$	0 to 18 446 744 073 709 551 615
floating-point	32 (float)	IEEE-754	•min subnormal: $\pm 1.401 \cdot 10^{-45}$ •min normal: $\pm 1.175 \cdot 10^{-38}$ •max: $\pm 3.402 \cdot 10^{38}$	•min subnormal: $\pm 0x1p-149$ •min normal: $\pm 0x1p-126$ •max: $\pm 0x1.fffffep+127$
	64 (double)	IEEE-754	•min subnormal: $\pm 4.940 \cdot 10^{-324}$ •min normal: $\pm 2.225 \cdot 10^{-308}$ •max: $\pm 1.797 \cdot 10^{308}$	•min subnormal: $\pm 0x1p-1074$ •min normal: $\pm 0x1p-1022$ •max: $\pm 0x1.ffffffffffffp+1023$
	80 (long double?)			
	128	IEEE-754		