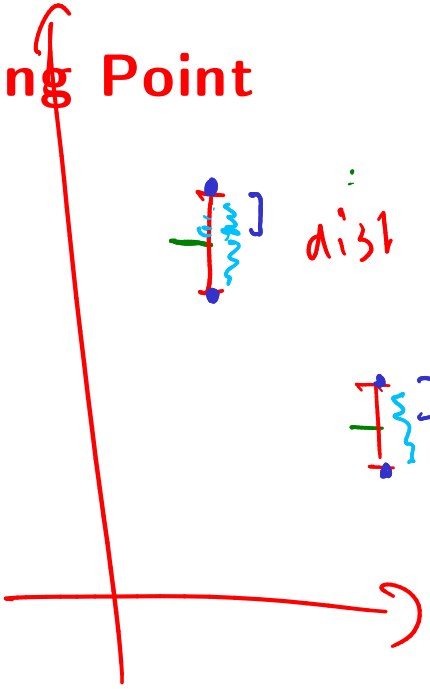
+ Working with rel. errors

5 Floating Point



dist 1 + dist 2

dist 1

dist 1 + err 1

dist 2

dist 2 + err 2

Wanted: Real Numbers... in a computer

- Computers can represent *integers*, using bits:

$$23 = 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = (10111)_2$$

How would we represent fractions, e.g. 23.625?

$17 = (10001)_2$

$17 = 16 + 0 + 0 + 0 + 1 =$

$2^4 \quad 2^3 \quad 2^2 \quad 2^1 \quad 2^0$

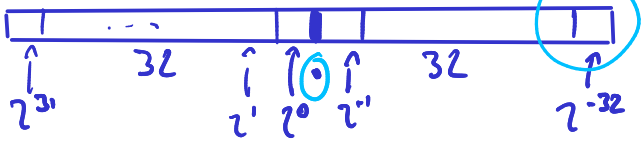
$23.625 = 2^4 + 0 + 0 + 0 + 2^0$ | $1 \quad 0 \quad 1$

$2^{-1} \quad 2^{-2} \quad 2^{-3}$

$0.5 \quad 0.25 \quad 0.125$

Fixed-Point Numbers

- Suppose we use units of 64 bits, with 32 bits for exponents $\geqslant 0$ and 32 bits for exponents <0. What numbers can we represent?



- How many 'digits' of relative accuracy (think relative rounding error) are available for the smallest vs. the largest number?

smallest. 2^{-32}
↳ next smaller: 0 rel.err. 100%
↳ next bigger: 2^{-31} rel.error 100%

biggest: $2^{31}+\dots+2^{-32} \approx 2^{31}$
↳ next smallest: $2^{30}+\dots+2^{-32} \approx 2^{31}$
↳ next bigger: $2^{31}+2^{-32}$
↳ $\frac{2^{-32}}{2^{31}}$ rel.error
$= 2^{-63}$ rel.error

10^{6}

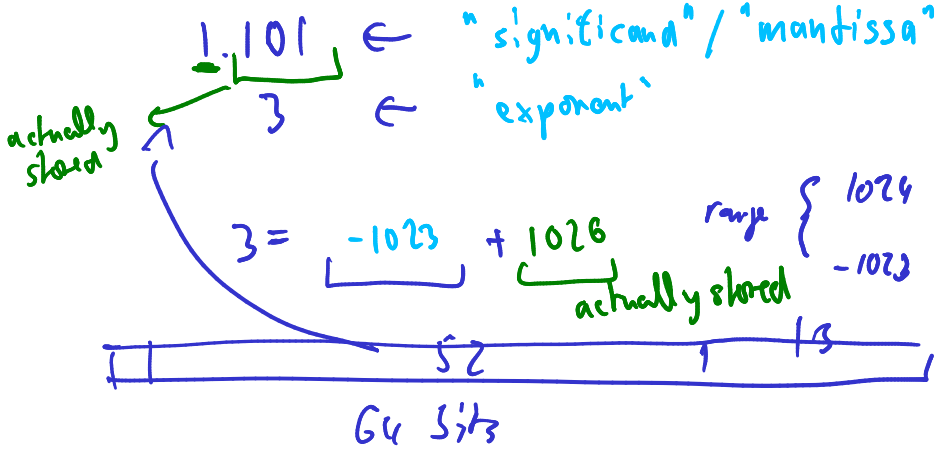
10^{-6} m

Floating Point numbers

- Convert $13 = (1101)_2$ into floating point representation.

$(1101)_2 = (110.1)_2 \cdot 2 = (1.101)_2 \cdot 2^3$

- What pieces do you need to store an FP number?



$$10.425 \approx (1010)_2 + (0.0110)_2 = (1010.0110)_2$$

$2^{-1}\ 2^{-2}\ 2^{-3}\ 2^{-4}$

$$= (1.01001101) \cdot 2^{3}$$

biggest possible number:

$$(1.11111111111)_2 \cdot 2^{1024}$$

smallest possible number:

$$\rightarrow (1.00000)_2 \cdot 2^{-1023}$$

$$(1.000000\ldots01)_2 \cdot 2^{-1023}$$

In-class activity: Floating Point

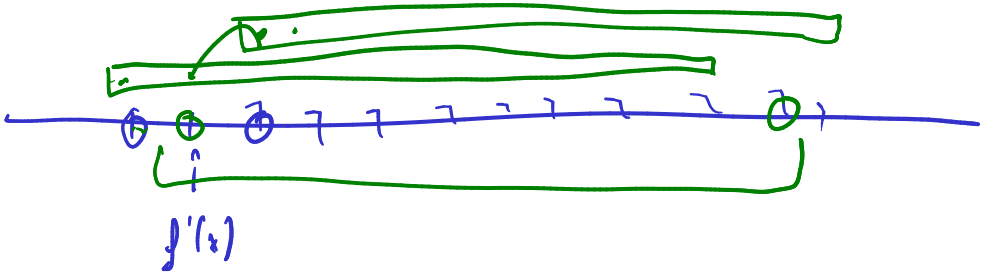
$1.000 \cdot 2^0$

$1.\square\square \cdot 2^{\square}$

$1 = (1.00)_2 \cdot 2^0$
$2 = (1.00) \cdot 2^1$
$3 = (1.10) \cdot 2^1$
$4 = (1.00) \cdot 2^2$
$5 = (1.01) \cdot 2^2$

$6 = (1.10)_2 \cdot 2^2$
$7 = (1.11) \cdot 2^2$
$8 = 1 \cdot 2^3$
$9 = (1.001) \cdot 2^3$

$f(x+h) - f(x-h)$



Unrepresentable numbers?

- Can you think of a somewhat central number that we cannot represent as

$$x = (1.\_\_\_\_\_)_2 \cdot 2^{-p}?$$

Demo: Picking apart a floating point number