How Numbers are represented in a computer!

Integers are exact but finite, dependent on number of bytes/bits

Char → 1 Byte = 2 Nibbles = 8 Bits

Short → 2 Bytes = 4 Nibbles = 16 Bits

Int → 4 Bytes = 8 Nibbles = 32 Bits

Long → 8 Bytes = 16 Nibbles = 64 Bits

Typical Integer Representations / Ranges

	Unsigned	Signed
1 Byte	$[0, 2^8 - 1] = 255$	$[-2^7, 2^7 - 1] \approx \pm 127$
2 Byte	$[0, 2^{16} - 1] \approx 65k$	$[-2^{15}, 2^{15} - 1] \approx \pm 32k$
4 Byte	$[0, 2^{32} - 1] \approx 4 \times 10^9$	$[-2^{31}, 2^{31} - 1] \approx \pm 2 \times 10^9$
8 Byte	$[0, 2^{64} - 1] = 18 \times 10^{18}$	$[-2^{63}, 2^{63} - 1] \approx \pm 9 \times 10^{18}$

Note: Range related to # of bits

$$2^{10} \approx 10^3$$

Real numbers are represented by floating data types. Never use == to compare floats! Only do the following

$$X == Y \leftarrow \text{Never}$$

$$|X - Y| < \varepsilon \leftarrow \text{Always}$$

$X - Y$ = difference

$|X - Y|$ = Absolute value of the difference

ε = Tolerance where

$$\varepsilon = |X| / 10^{SD}$$

SD = Significant Digits of the data type

Floating Data Types are <u>always inaccurate</u> due to finite number of bits.

Example $X \in Q$ Q is set of rational numbers

$$12/99 = .12121212 \ldots\ldots\ldots$$

forever

$$= .\underline{12}$$

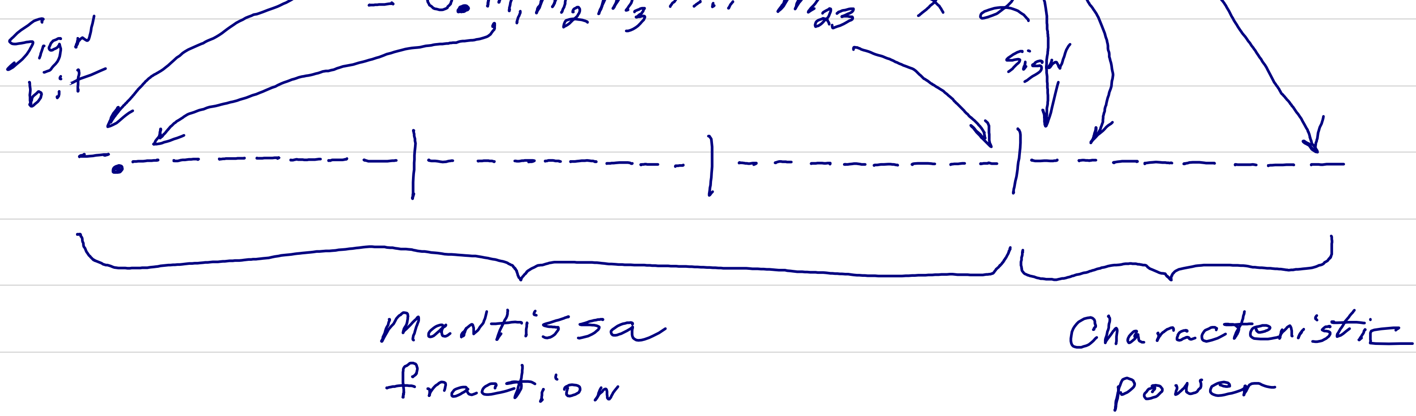
undersore means repeat for forever

So let's calculate the range of 2 Real Data types and how accurate they could be!

4 byte Real called float in C++

All numbers can be represented in scientific notation

$$\pm\, 0.m_1 m_2 m_3 \ldots\ m_{23} \times 2^{\pm c_1 c_2 \ldots c_7}$$



1 Sign bit for mantissa
23 bits for mantissa range
1 Sign bit for characteristic
7 bits for chacateristic range

So the range $\approx 2^{23} = 2^{3}\, 2^{10}\, 2^{10} \approx 8 \times 10^{3} \times 10^{3} \approx 10^{7}$

or $\approx$ 7 significant digits in Base 10

For the power

$$10^{x} = 2^{\pm(2^{7}-1)} = 2^{\pm 127}$$

$$\log_{10} 10^{x} = \log_{10} 2^{\pm 127}$$

$$x = \pm 127 \log_{10} 2$$

$$x \approx \pm 38$$

The limit of accuracy for a 4byte real called float is

$\underline{\underline{7}}$ significant digits Base 10

With the range

$$\underline{\underline{10^{\pm 38}}}$$

This datatype should be used for real values +99% of the time since it is rare that we know a number to this kind of accuracy.

Refer to any laboratory science and measurement accuracy.

Now define a real datatype with 2x the width!

8 bytes - 64 bits

53 bit mantissa

↑

1 bit sign

11 bit characteristic

↑

1 bit for sign

$$\pm 0.m_1 m_2 \cdots\cdots m_{52} \times 2^{\pm c_1 c_2 \cdots c_{10}}$$

Same analysis as before

Accuracy $2^{53} \approx 2^3 \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \approx 10 \times 10^3 \times 10^3 \times 10^3 \times 10^3 \times 10^3$

$\approx \underline{\underline{10^{16}}}$

So, $\underline{\underline{\text{16 significant}}}$ digits in Base 10

Now the range

The range in base 10

$$10^x = 2^{\pm(2^{10}-1)} = 2^{\pm 1023}$$

$$\log_{10} 10^x = \log_{10} 2^{\pm 1023}$$

$$x = \pm 1023 \log_{10} 2$$

$$= \pm 308$$

Then the range of an 8 Byte float, i.e. a Double is

$$10^{\pm 308}$$

with accuracy of 16 significant digits Base 10

When are double data types needed?

In CIS/CSC 5 or 17A ??

Never

Absolutely no need since no problem requires this range or accuracy.