Floating Point Numbers

- Numerical analysis is the study of floating-point arithmetic.

- Floating-point arithmetic is unpredictable and hard to understand.

We intend to convince you that both of these assertions are false.

$$x = \pm(1 + f) \cdot 2^e$$

$$0 \le f < 1$$

$$f = (\text{integer} < 2^{52})/2^{52}$$

$$-1022 \le e \le 1023$$

$$e = \text{integer}$$

Finite f implies finite *precision*.

Finite e implies finite *range*

Floating point numbers have discrete spacing,
a maximum and a minimum.

`eps` *is the distance from* 1 *to the next larger floating-point number.*

```
eps = 2^(-52)
```

	Binary	Decimal
eps	2^(-52)	2.2204e-16
realmin	2^(-1022)	2.2251e-308
realmax	(2-eps)*2^1023	1.7977e+308

```
>> format hex
>> t = 1/10
t =
   3fb999999999999a
```

$$\frac{1}{10} = \frac{1}{2^4} + \frac{1}{2^5} + \frac{0}{2^6} + \frac{0}{2^7} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{0}{2^{10}} + \frac{0}{2^{11}} + \frac{1}{2^{12}} + ...$$

$$t = (1 + \frac{9}{16} + \frac{9}{16^2} + \frac{9}{16^3} + ... + \frac{9}{16^{12}} + \frac{10}{16^{13}}) \cdot 2^{-4}$$

Problem 1.34.

```
x = 1; while 1+x > 1, x = x/2, pause(.02), end

x = 1; while x+x > x, x = 2*x, pause(.02), end

x = 1; while x+x > x, x = x/2, pause(.02), end
```