- Form invites
- One question/concern per post
- \( \Phi \)
- Numerical analysis
  \( \rightarrow \) norms/waves/cont'd.
  \( \frac{2^{-21}}{0.01} \)
  \( \frac{2^{-52}}{0.001} \)

\[ 0.000 \overline{147777} \]
\[ 0.000 \overline{314777} \]
Rounding Modes

How is rounding performed? (Imagine trying to represent $\pi$.)

$\left(1.110101011\right)_2$

representable

What is done in case of a tie? $0.5 = (0.1)_2$ ("Nearest"?)

Demo: Density of Floating Point Numbers [cleared]
Demo: Floating Point vs Program Logic [cleared]
Smallest Numbers Above... 

What is smallest FP number > 1? Assume 4 bits in the significand.

What’s the smallest FP number > 1024 in that same system?

Can we give that number a name?
Unit roundoff or *machine precision* or *machine epsilon* or \( \varepsilon_{\text{mach}} \) is the smallest number such that
\[
\text{float}(1 + \varepsilon) > 1.
\]

- Assuming round-to-nearest, in the above system, \( \varepsilon_{\text{mach}} = (0.00001)_2 \).
- Note the extra zero.
- Another, related, quantity is *ULP*, or *unit in the last place*.
  \( \varepsilon_{\text{mach}} = 0.5 \text{ ULP} \)
FP: Relative Rounding Error

What does this say about the relative error incurred in floating point calculations?

\[
\left| \frac{\hat{x} - \tilde{x}}{x} \right| = \left| \frac{x - x(1 + 3\epsilon_{\text{mach}})}{x} \right| \\ \\
\leq |3| = 3_{\text{mach}}.
\]

\[
|\hat{x}| = |x(1 + 3\epsilon_{\text{mach}})|
\]
FP: Machine Epsilon

What’s that same number for double-precision floating point? (52 bits in the significand)

Demo: Floating Point and the Harmonic Series [cleared]
In-Class Activity: Floating Point

In-class activity: Floating Point
Implementing Arithmetic

How is floating point addition implemented? Consider adding $a = (1.101)_2 \cdot 2^1$ and $b = (1.001)_2 \cdot 2^{-1}$ in a system with three bits in the significand.
Problems with FP Addition

What happens if you subtract two numbers of very similar magnitude?
As an example, consider $a = (1.1011)_2 \cdot 2^0$ and $b = (1.1010)_2 \cdot 2^0$.

Demo: Catastrophic Cancellation [cleared]
Supplementary Material

- Josh Haberman, *Floating Point Demystified, Part 1*
- David Goldberg, *What every computer programmer should know about floating point*