Number Representation Methods - Fixed Point vs. Floating Point

Fixed Point

Integer with implicit decimal position

- ✓ Fast computation
- ✓ Limited range
- ✓ Fixed precision

Floating Point

Sign + Exponent + Mantissa

- ✓ Flexible representation
- ✓ Wide range
- ✓ Slower processing

Floating Point Formats

FP32 (32 bits)

Sign: 1 bit
Exponent: 8 bits
Mantissa: 23 bits

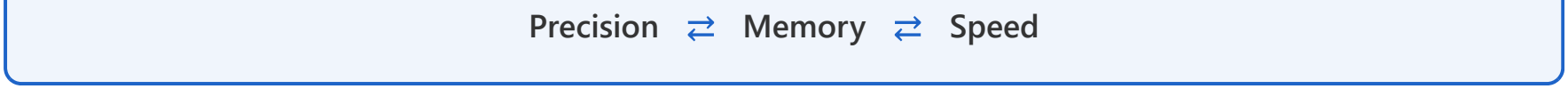
Range: **$\pm 3.4 \times 10^{38}$**

FP16 (16 bits)

Sign: 1 bit
Exponent: 5 bits
Mantissa: 10 bits

Range: **$\pm 6.5 \times 10^{4}$**

Key Trade-off


Precision ⇄ Memory ⇄ Speed