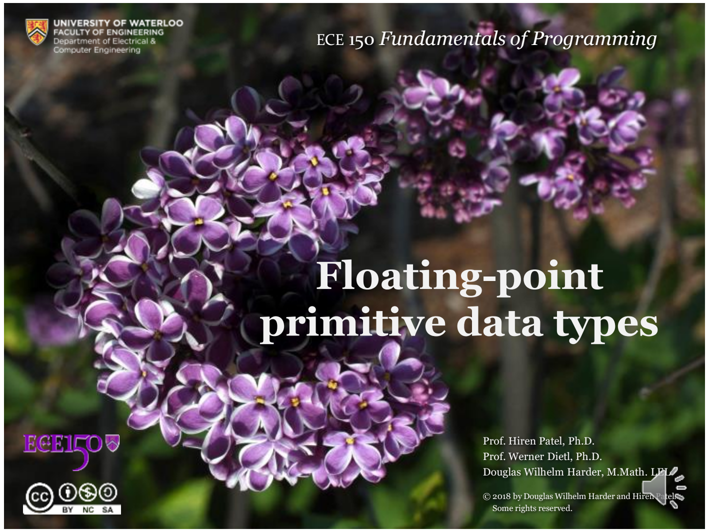
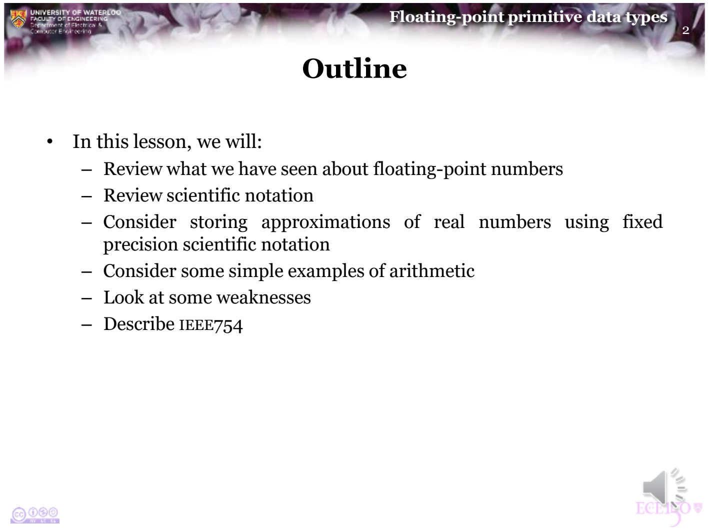


UNIVERSITY OF WATERLOO
FACULTY OF ENGINEERING
Department of Electrical & Computer Engineering

ECE 150 *Fundamentals of Programming*

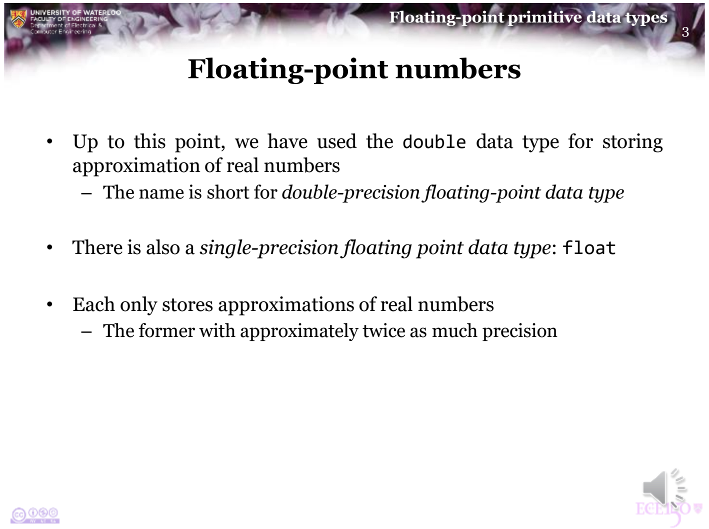
Floating-point primitive data types

Prof. Hiren Patel, Ph.D.
Prof. Werner Dietl, Ph.D.
Douglas Wilhelm Harder, M.Math. LEL



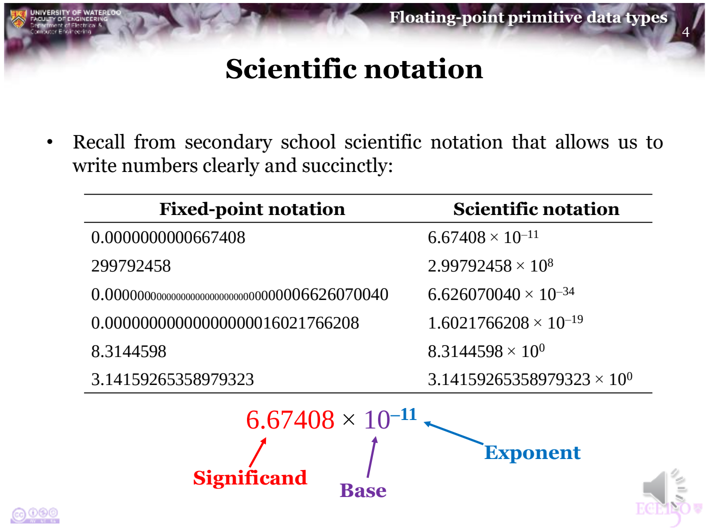
Outline

- In this lesson, we will:
  - Review what we have seen about floating-point numbers
  - Review scientific notation
  - Consider storing approximations of real numbers using fixed precision scientific notation
  - Consider some simple examples of arithmetic
  - Look at some weaknesses
  - Describe IEEE754



Floating-point numbers

- Up to this point, we have used the `double` data type for storing approximation of real numbers
  - The name is short for *double-precision floating-point data type*
- There is also a *single-precision floating point data type*: `float`
- Each only stores approximations of real numbers
  - The former with approximately twice as much precision



Scientific notation

- Recall from secondary school scientific notation that allows us to write numbers clearly and succinctly:

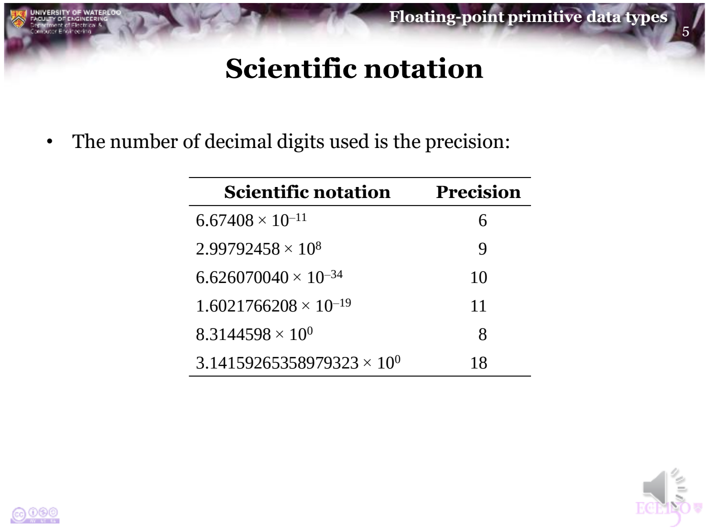
Fixed-point notation	Scientific notation
0.0000000000667408	6.67408×10^{-11}
299792458	2.99792458×10^{8}
0.0000000000000000000000000000000006626070040	$6.626070040 \times 10^{-34}$
0.00000000000000000016021766208	$1.6021766208 \times 10^{-19}$
8.3144598	8.3144598×10^{0}
3.14159265358979323	$3.14159265358979323 \times 10^{0}$

6.67408×10^{-11}

Significand

Base

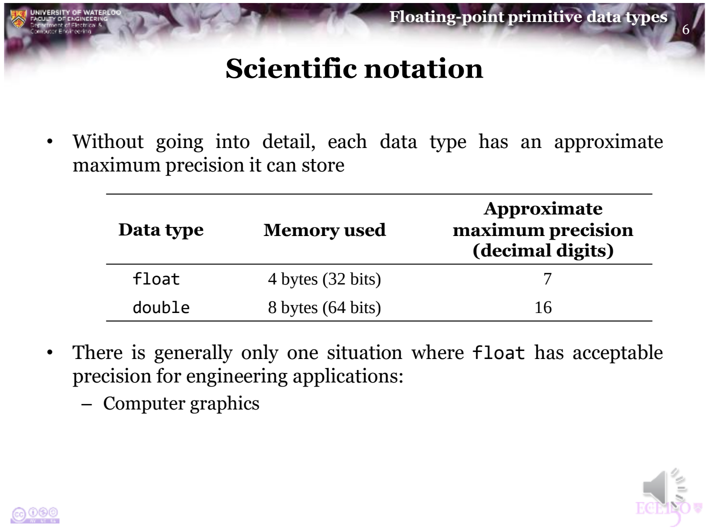
Exponent



Scientific notation

- The number of decimal digits used is the precision:

Scientific notation	Precision
6.67408×10^{-11}	6
2.99792458×10^{8}	9
$6.626070040 \times 10^{-34}$	10
$1.6021766208 \times 10^{-19}$	11
8.3144598×10^{0}	8
$3.14159265358979323 \times 10^{0}$	18

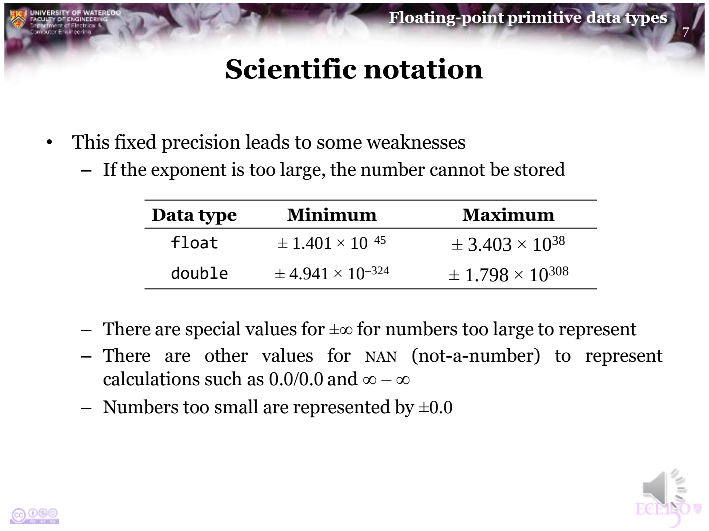


Scientific notation

- Without going into detail, each data type has an approximate maximum precision it can store

Data type	Memory used	Approximate maximum precision (decimal digits)
`float`	4 bytes (32 bits)	7
`double`	8 bytes (64 bits)	16

- There is generally only one situation where `float` has acceptable precision for engineering applications:
  - Computer graphics

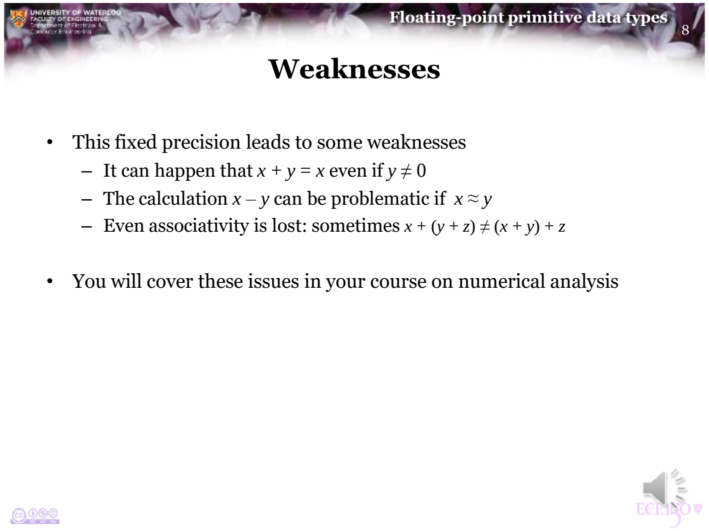


Scientific notation

- This fixed precision leads to some weaknesses
  - If the exponent is too large, the number cannot be stored

Data type	Minimum	Maximum
`float`	$\pm 1.401 \times 10^{-45}$	$\pm 3.403 \times 10^{38}$
`double`	$\pm 4.941 \times 10^{-324}$	$\pm 1.798 \times 10^{308}$

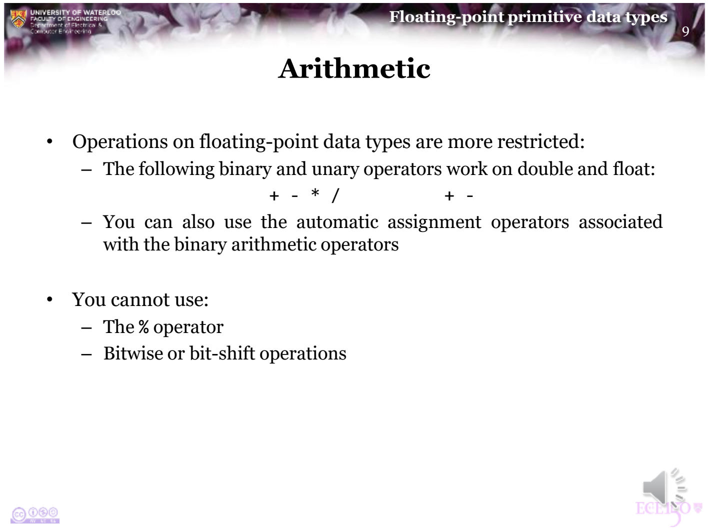
  - There are special values for ±∞ for numbers too large to represent
  - There are other values for NAN (not-a-number) to represent calculations such as 0.0/0.0 and $\infty - \infty$
  - Numbers too small are represented by ±0.0



Weaknesses

- This fixed precision leads to some weaknesses
  - It can happen that $x + y = x$ even if $y \neq 0$
  - The calculation $x - y$ can be problematic if $x \approx y$
  - Even associativity is lost: sometimes $x + (y + z) \neq (x + y) + z$

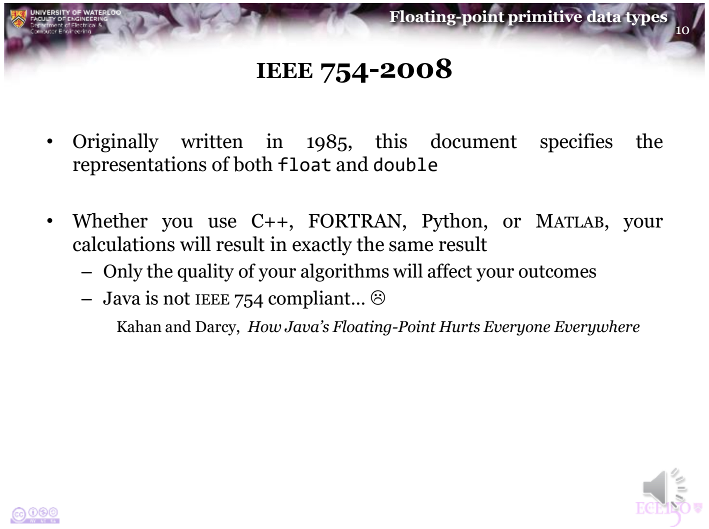
- You will cover these issues in your course on numerical analysis



Arithmetic

- Operations on floating-point data types are more restricted:
  - The following binary and unary operators work on double and float:

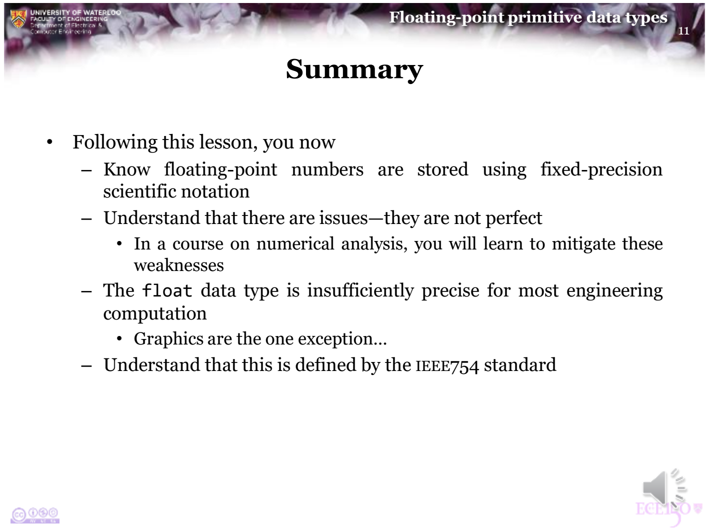
    `+ - * /`　　　`+ -`
  - You can also use the automatic assignment operators associated with the binary arithmetic operators
- You cannot use:
  - The `%` operator
  - Bitwise or bit-shift operations



IEEE 754-2008

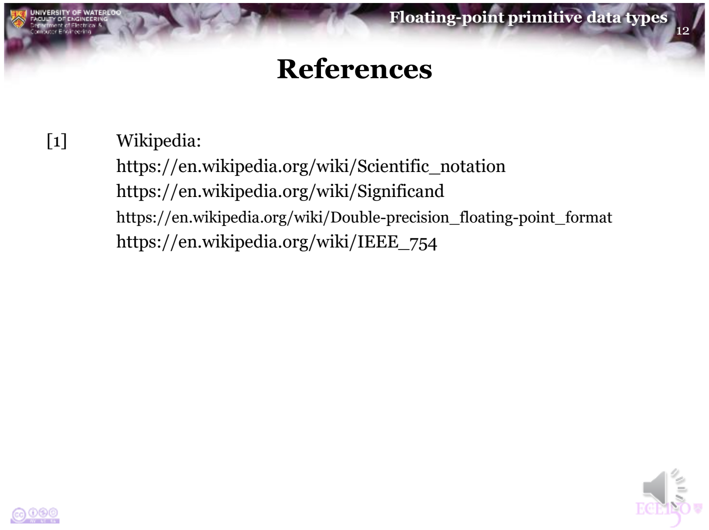
- Originally written in 1985, this document specifies the representations of both `float` and `double`
- Whether you use C++, FORTRAN, Python, or MATLAB, your calculations will result in exactly the same result
  - Only the quality of your algorithms will affect your outcomes
  - Java is not IEEE 754 compliant… ☹

    Kahan and Darcy, *How Java's Floating-Point Hurts Everyone Everywhere*



Summary

- Following this lesson, you now
  - Know floating-point numbers are stored using fixed-precision scientific notation
  - Understand that there are issues—they are not perfect
    - In a course on numerical analysis, you will learn to mitigate these weaknesses
  - The `float` data type is insufficiently precise for most engineering computation
    - Graphics are the one exception…
  - Understand that this is defined by the IEEE754 standard

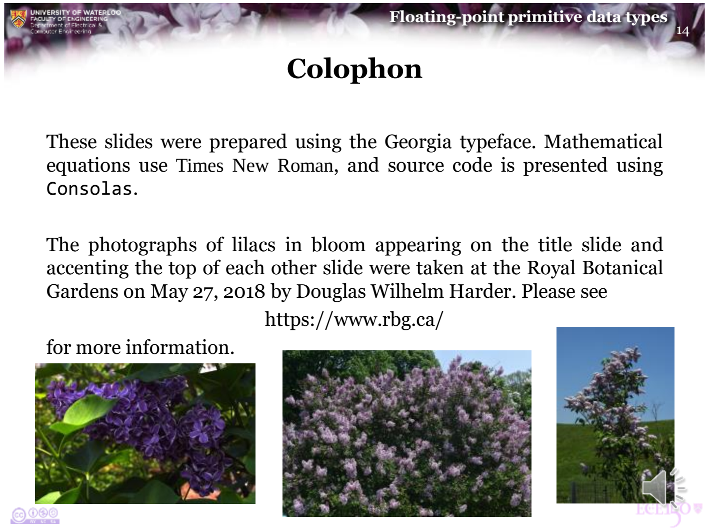


References

[1] Wikipedia:
https://en.wikipedia.org/wiki/Scientific_notation
https://en.wikipedia.org/wiki/Significand
https://en.wikipedia.org/wiki/Double-precision_floating-point_format
https://en.wikipedia.org/wiki/IEEE_754

Acknowledgements

[1] Zhuo En Dai



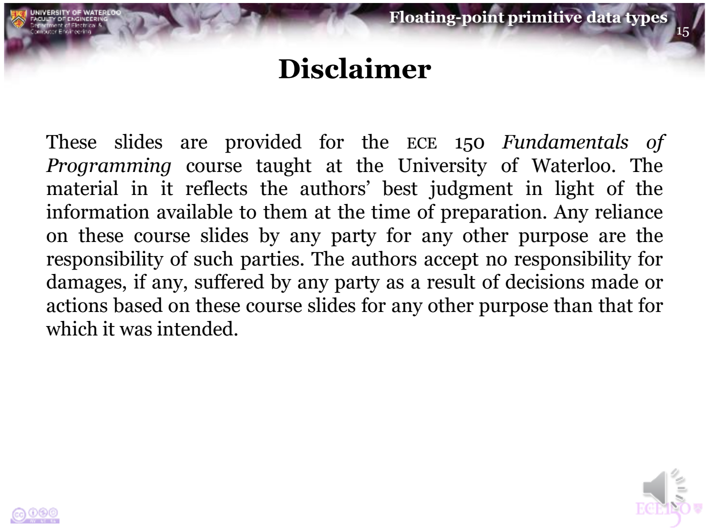
Colophon

These slides were prepared using the Georgia typeface. Mathematical equations use Times New Roman, and source code is presented using `Consolas`.

The photographs of lilacs in bloom appearing on the title slide and accenting the top of each other slide were taken at the Royal Botanical Gardens on May 27, 2018 by Douglas Wilhelm Harder. Please see

https://www.rbg.ca/

for more information.



Disclaimer

These slides are provided for the ECE 150 *Fundamentals of Programming* course taught at the University of Waterloo. The material in it reflects the authors' best judgment in light of the information available to them at the time of preparation. Any reliance on these course slides by any party for any other purpose are the responsibility of such parties. The authors accept no responsibility for damages, if any, suffered by any party as a result of decisions made or actions based on these course slides for any other purpose than that for which it was intended.