Modules and File Handles

Eric Franzosa (franzosa@hsph.harvard.edu)

http://franzosa.net/bst273

Overview

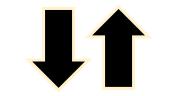
- Working with modules
- File I/O (input/output) and file handles
- Practice
- Revisit operator precedence (time permitting)

Modules

A little detour...



The "Learning Python" highway



modules

What is a module?

- A module (or package) is Python code designed to be re-used in other code
 - Sometimes referred to as "Libraries" in other languages
- Contributes to our "Second Rule of Programming" from Day 1:
 - Be appropriately lazy
 - Don't reinvent the wheel
- Also helps with organizing code related to a particular domain

What is a module?

- A module can be as small as a single Python script
 - In fact, every Python script can be used as a module (more on that shortly...)
- Some modules come bundled with Python
 - The Python "Standard Library"
- Other (more specialized) modules can be installed separately
 - For example, the Python "scientific stack": scipy, numpy, pandas, and matplotlib
 - Anaconda installed a bunch of these for us automatically
- You can write your own modules
 - https://bitbucket.org/franzosa/zopy

Using modules

- We use modules by **import**ing them into our code
- We got a sneak peak of this earlier -- does anyone remember when?

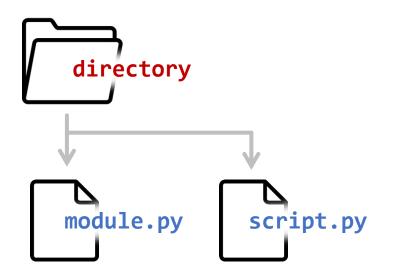
```
import time
for i in range( 5 ):
    i2 = i ** 2
    print( "i=", i, "i2=", i2 )
    time.sleep( 1 )
```

- time is a built-in module for dealing with matters related to, well, time
- <u>https://docs.python.org/3/library/time.html</u>
 - We'll look at module documentation shortly

Using modules

- Modules contain the same sort of elements as other Python code
- Modules can contain variables
 - math.pi contains the value of pi (to many decimal places)
 - **string.uppercase** contains the uppercase English alphabet
- Modules can contain functions
 - **time.sleep** pauses computation for N seconds
 - **math.sqrt** returns the square root of a number
- Modules can contain classes defining other data types
 - **collections.Counter** a special dictionary for counting
 - **Bio.Seq** special strings for representing biological sequences

- The following example assumes I have two Python files in the same folder
 - script.py is a new script I am working on
 - module.py is some existing code that I want to re-use



<pre>module.py (open in Atom)</pre>	<pre>script.py (open in Atom)</pre>	(a terminal)
<pre># an approximation of pi</pre>	<pre>print(pi)</pre>	<pre>\$ python script.py</pre>
<pre>pi = 3.14 # area of a circle def area(r): return pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11]</pre>		NameError: name 'pi' is not defined

<pre>module.py (open in Atom)</pre>	<pre>script.py (open in Atom)</pre>	(a terminal)
<pre># an approximation of pi</pre>	<pre>import module</pre>	<pre>\$ python script.py</pre>
<pre>pi = 3.14 # area of a circle def area(r): return pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11]</pre>	<pre>print(pi) Module "import" statements are among the first lines in a Python file</pre>	NameError: name 'pi' not defined

is

<pre>module.py (open in Atom)</pre>	<pre>script.py (open in Atom)</pre>	(a)
<pre># an approximation of pi pi = 3.14</pre>	<pre>import module</pre>	\$ py
<pre># area of a circle def area(r): return pi * r ** 2</pre>	<pre>print(module.pi)</pre>	3
<pre># first few primes primes = [2, 3, 5, 7, 11]</pre>	"module" is a namespace: a collection of previously defined objects (variables, functions, etc.) We request individual objects using "." syntax	

(a terminal)

\$ python script.py

3.14

module.py (open in Atom) # an approximation of pi pi = 3.14# area of a circle **def** area(r): **return** pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11] script.py (open in Atom)

import module
import math

print(module.pi)
print(math.pi)

"math" is a separate namespace: using namespaces helps to avoid "collisions" (redefining a second variable with the same name) (a terminal)

\$ python script.py
3.14
3.14159265359

module.py (open in Atom) # an approximation of pi pi = 3.14# area of a circle **def** area(r): **return** pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11]

```
import module
import math
pi = 3
print( module.pi )
print( math.pi )
print( pi )
    The running script has it's
     own namespace (called
    "___main___"). Here, pi is
     also defined in the that
          namespace.
```

script.py (open in Atom)

(a terminal)

\$ python script.py 3.14 3.14159265359

module.py (open in Atom) # an approximation of pi pi = 3.14# area of a circle **def** area(r): **return** pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11] script.py (open in Atom)

from module import area, pi

We can also import *specific*

variables/functions from a

module into the main

namespace as a comma-

separated list.

```
print( pi )
print( area( 2 ) )
```

(a terminal)

\$ python script.py 3.14 12.56

module.py (open in Atom)

an approximation of pi pi = 3.14# area of a circle def area(r): **return** pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11]

<pre>script.py (open in Atom)</pre>	
<pre>from module import area, pi</pre>	i
pi = 3 print(pi) print(area(2))	

Once we import a name, we can overwrite it (this is often something you want to avoid).

(*Tricky!*) Note that area still uses module.pi.

(a terminal)

	\$ python	<pre>script.py</pre>	
	3 12.56		

module.py (open in Atom)

an approximation of pi
pi = 3.14

area of a circle
def area(r):
 return pi * r ** 2

first few primes
primes = [2, 3, 5, 7, 11]

script.py (open in Atom)

from module import primes

```
primes.append( 13 )
print( primes )
```

from module import primes

print(primes)

Nothing we do in this box actually changes module.py (the original is always loaded as a fresh copy). (a terminal)

\$ python script.py
[2, 3, 5, 7, 11, 13]
[2, 3, 5, 7, 11]

<pre>module.py (open in Atom)</pre>	<pre>script.py (open in Atom)</pre>	(a terminal)
<pre># an approximation of pi pi = 3.14 # area of a circle def area(r): return pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11]</pre>	<pre>from module import * Imports everything from module into the main namespace. You will see code that does this, but it is sloppy!</pre>	<pre>\$ python script.py</pre>

module.py (open in Atom) # an approximation of pi pi = 3.14# area of a circle **def** area(r): **return** pi * r ** 2 # first few primes primes = [2, 3, 5, 7, 11] script.py (open in Atom) import module as m print(m.pi) If a module has a particularly long name, we can use "import...as" to give it an alias in our code. Some "famous" modules have common nicknames, for example: **import** numpy **as** np

(a terminal)

\$ python script.py
3.14

<pre>module.py (open in Atom)</pre>	<pre>script.py (open in Atom)</pre>	(a terminal)
<pre># an approximation of pi</pre>	<pre>import module</pre>	<pre>\$ python script.py</pre>
<pre>pi = 3.14 # area of a circle def area(r): return pi * r ** 2 # first few primes</pre>	print (module.pi)	"Hello, World!" 3.14
<pre>primes = [2, 3, 5, 7, 11] # say hello print("Hello, World!")</pre>	Module code is executed when it's imported. This will cause "Hello, World!" to print before pi.	

module.py (open in Atom) script.py (open in Atom) (a terminal) # an approximation of pi import module \$ python script.py pi = 3.143.14 print(module.pi) # area of a circle **def** area(r): **return** pi * r ** 2 We can use a special conditional to indicate # first few primes primes = [2, 3, 5, 7, 11]that some code should only be run when the # say hello in script mode module is run as a if __name__ == "__main_": script! print("Hello, World!") (We'll come back to this next week)

module.py (open in Atom)

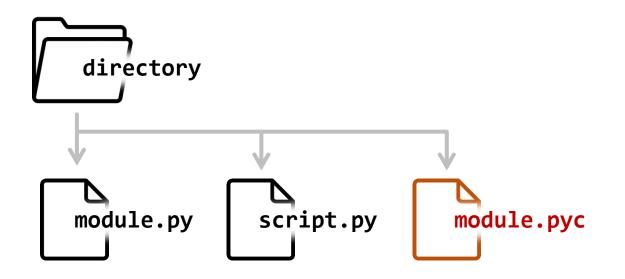
```
# an approximation of pi
pi = 3.14
# area of a circle
def area( r ):
    return pi * r ** 2
# first few primes
primes = [2, 3, 5, 7, 11]
# say hello in script mode
if __name__ == "__main_":
    print( "Hello, World!" )
```

This module is just a Python script and can also be executed.

(a terminal)

\$ python module.py
"Hello, World!"

- You'll find a new file in the working directory called module.pyc
- This is "compiled" Python code
- Compiled code can be acted on more directly by the computer
- Rather than compiling "on the fly" (as with script-code), module code might not change as often, so we save time by keeping a compiled version



- Not all modules live in your working directory
- Most of them live in the folder for your Python installation
 e.g. C:\Users\Franzosa\Anaconda3
- New packages will be installed there too
 - e.g.\$ conda install scipy
- You can also tell Python to look in other places for modules
 - Via the **\$PYTHONPATH** environment variable
- The Python Standard Library + the Anaconda install from day 1 have all the modules you're likely to need for this course

- Complex modules are organized in a nested structure
 - scipy.stats.wilcoxon is a function in the module stats in the module scipy
 - In reality, scipy would be a folder, stats would be a file within that folder, and wilcoxon would be a function in that file
- Modules may also be bundled with useful datasets
 - e.g. genetic code translation tables in the **biopython** module
- Python modules may also be bundled with compiled C code
 - For operations that need to be done very fast

- Modules contain documentation that we can access directly, similar to using man on the command line.
- Within the Python interpreter (or a Jupyter Notebook), import a module, then execute help(module_name)

```
$ python
Python 3.6.5 | Anaconda, Inc.
Type "help", "copyright", "credits" or "license" for more information.
>>> import math
>>> help( math )
```

- Modules contain documentation that we can access directly, similar to using man on the command line.
- Within the Python interpreter (or a Jupyter Notebook), import a module, then execute help(module_name)

```
Help on built-in module math:
NAME
  math
DESCRIPTION
  This module is always available. It provides access to the
  mathematical functions defined by the C standard.
FUNCTIONS
  acos(...)
```

- Unlike man pages, I rarely consult help() from the command line
- It is much more convenient to read about modules online:
 - <u>https://docs.python.org/3/library/math.html</u>
 - (or Google "python module math" like I did)

🥐 Python » English 🔻 3.7.0	 Documentation » The Python Standard Library » Numeric and Mathematical Modules » 	previous next modules index
		Quick search Go
Table Of Contents math — Mathematical	math — Mathematical functions	
functions Number-theoretic and representation functions 	This module is always available. It provides access to the mathematical functions defined by the 0	C standard.
 Power and logarithmic functions Trigonometric functions Angular conversion Hyperbolic functions Special functions Constants « 	These functions cannot be used with complex numbers; use the functions of the same name module if you require support for complex numbers. The distinction between functions which s numbers and those which don't is made since most users do not want to learn quite as much required to understand complex numbers. Receiving an exception instead of a complex resu detection of the unexpected complex number used as a parameter, so that the programmer can and why it was generated in the first place.	support complex mathematics as Ilt allows earlier

- Visit module help pages to 1) find a function that performs a data transformation that you need and 2) determine "how to use it."
- "How to use it" = "What arguments does it take?" + "What does it return?"
- For example:

math.log(x[, base])

With one argument, return the natural logarithm of x (to base e).

With two arguments, return the logarithm of x to the given base, calculated as log(x)/log(base).

• In function help, [,...] indicates an <u>optional</u> positional argument.

• We'll get into the nuances of function arguments next week.

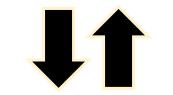
- Starting next week, we will devote large chunks of individual lectures to exploring important Python modules
 - **argparse** for command-line interfaces
 - **re** for pattern matching
 - numpy for fast numerical computation
 - subprocess for calling other programs from Python
- You may end up needing/wanting to use other modules for final projects
- Unlike Python grammar (for, if/else, etc.), modules and their most useful components are a vocabulary you build up slowly over time

File I/O

A little detour...



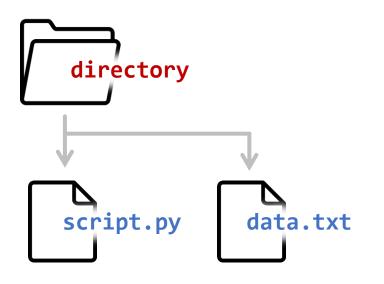
The "Learning Python" highway



modules

File I/O and the sys module

- Python uses a special **file** data type to read data into a program
 - For when you get bored of looking through lists of numbers...
 - I'll refer to **file**s as "file handles" to differentiate them from "files on your computer"
- We create **file**s using the **open** function
 - **Open** takes a path to a file as a mandatory argument
- Let's look at an example



File I/O and the sys module

data.txt (open in Atom) script.py (open in Atom) (a terminal) fh = open("data.txt") Come gather 'round people \$ python script.py Wherever you roam And admit that the waters *No output yet* Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon open returns a file object For the wheel's still in spin that we store in a variable. And there's no tellin' who That it's namin'. I use "fh" as a convention For the loser now (for "file handle"). Will be later to win For the times they are a-changin'.

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

fh = open("data.txt")
for line in fh:
 print(line)

Each iteration of the for loop gives us the next line of the file (stored here in the aptly named line variable).

(a terminal)

\$ python script.py

Come gather 'round people

Wherever you roam

And admit that the waters

Around you have grown

And accept it that soon

You'll be drenched to the bone.

If your time to you

Is worth savin'

• • •

Then you better start swimmin'

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

fh = open("data.txt")
for line in fh:
 print(line)

Each line of the file ends with a newline (\n) character.

print() adds its own
 newline by default,
resulting in double spacing.

(a terminal)

\$ python script.py

Come gather 'round people Wherever you roam

And admit that the waters

Around you have grown

And accept it that soon

You'll be drenched to the bone.

If your time to you

Is worth savin'

• • •

Then you better start swimmin'

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

fh = open("data.txt")
for line in fh:
 line = line.strip()
 print(line)

The string method .strip() will remove white space (including newlines) at the beginning and end of a string, solving this problem.

(a terminal)

\$ python script.py

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

fh = open("data.txt")
for line in fh:
 print(line, end=""")

Another solution

(a terminal)

\$ python script.py

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

```
fh = open( "data.txt" )
for line in fh:
    line = line.strip( )
    print( line )
fh.close( )
```

It is a good convention to

close a file when we are

done with it.

(a terminal)

\$ python script.py

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

import sys

```
print( sys.argv )
```

```
fh = open( "data.txt" )
for line in fh:
    line = line.strip( )
    print( line )
fh.close( )
```

The sys module (among other things) provides us with a list of arguments passed on the command line: sys.argv (a terminal)

\$ python script.py x y z
['script.py', 'x', 'y', 'z']
The first element of
sys.argv, in other words
sys.argv[0], is the name
of the script that is running

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

import sys

```
fh = open( sys.argv[1] )
for line in fh:
    line = line.strip( )
    print( line )
fh.close( )
```

This is a better way to pass a file into a program than "hard-coding" the path.

Can you guess why?

(a terminal)

\$ python script.py data.txt

data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

import sys

for line in sys.stdin: line = line.strip()
 print(line)

sys.stdin is a file handle
 that returns information
 piped into the script. We
read it just like a normal file.

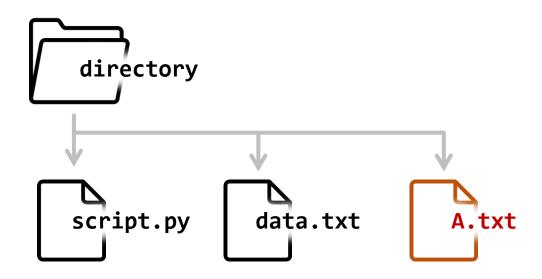
(a terminal)

\$ cat data.txt | python script.py

- We can also use open() to create file handles for writing new files
- The syntax for this is open(*file_name*, "w")

• *!!!* Warning: this will overwrite file_name if it already exists *!!!*

- We can write lines to such file handles using print()
 - By default, print() writes to sys.stdout, i.e. the terminal itself



data.txt (open in Atom)

Come gather 'round people Wherever you roam And admit that the waters Around you have grown And accept it that soon You'll be drenched to the bone. If your time to you Is worth savin' Then you better start swimmin' Or you'll sink like a stone For the times they are a-changin'. Come writers and critics Who prophesize with your pen And keep your eyes wide The chance won't come again And don't speak too soon For the wheel's still in spin And there's no tellin' who That it's namin'. For the loser now Will be later to win For the times they are a-changin'.

script.py (open in Atom)

import sys

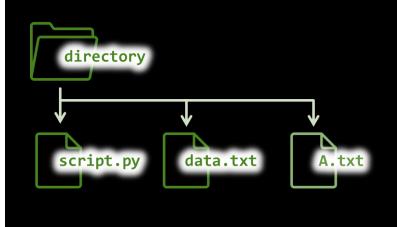
```
fh = open( "A.txt", "w" )
for line in sys.stdin:
    line = line.strip( )
    if line[0] == "A":
        print( line, file=fh )
fh.close( )
```

We use the optional "file=" argument of print to tell the function where to put the data.

(a terminal)

\$ cat data.txt | python script.py
\$ cat A.txt

And admit that the waters Around you have grown And accept it that soon And keep your eyes wide And don't speak too soon And there's no tellin' who



In perspective

• This is very powerful

- We now have the ability to read a data file into a program, operate on that data, and then write some results as a new file.
- We can use this to make programs that behave like command-line tools:
 - Solve problems on their own...
 - ... or as part of command-line chains
- Actually, that sounds like a good idea...

Activity: grep.py

grep.py

- Write a Python script called grep.py
- The script should take two positional, command-line arguments:
 - \$ python grep.py pattern file_name
 - The 1st argument (*pattern*) is a string to search for
 - The 2nd argument (*file_name*) is a path to a file you want to search
- The script should print (to the screen) lines from the file that contain the matching pattern
- An example execution might look like:
 - s python grep.py "e" data.txt

grep.py

- Compose your script in a plain-text editor (such as Atom)
- Be careful about copying and pasting code directly from the slides
 - The text sometimes has extra formatting embedded that can cause errors
 - Better to retype the commands you want (good for muscle memory too!)
- Other suggestions:
 - $_{\circ}$ Use <code>sys.argv[1]</code> and <code>sys.argv[2]</code> to access the pattern and file name
 - Use the in operator to see if a pattern occurs in a line
 - The code from the slide with the star (like the one above) is a good starting point
- Alternatively...
 - Download my grep() function from Canvas and import it into your script
 - You'll still use sys.argv[1] and sys.argv[2] to get the pattern and file name
 - Canvas also contains a data.txt file to experiment with

(a quick word about...) Operator Precedence

Python operator precedence

Higher precedence / Tighter binding

()
**
*,/,//,%
+, -
in, <, <=, >, >=, !=, ==
not
and
or
Lower precedence / Weaker binding

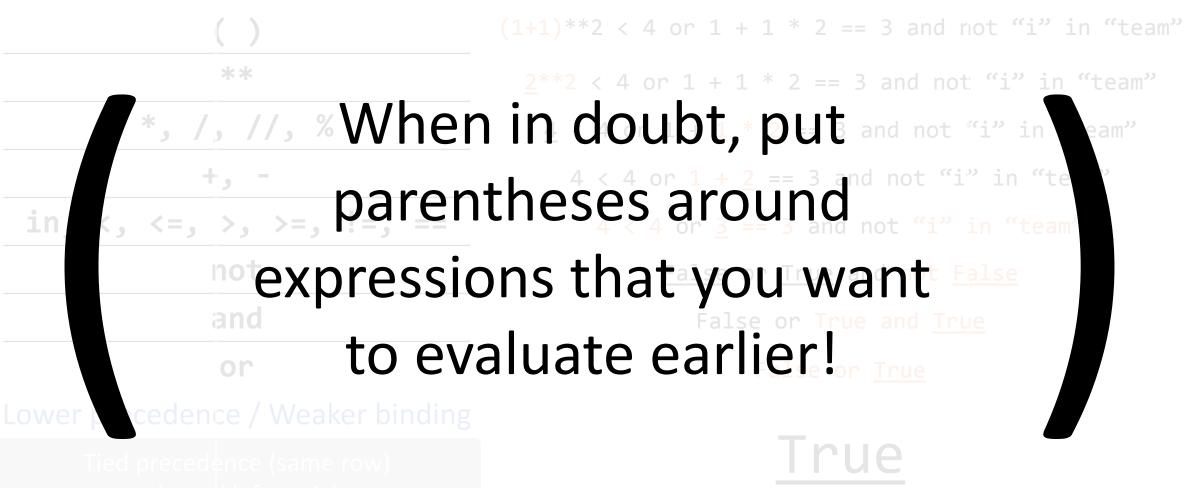
Tied precedence (same row) evaluated left to right

1+1)**2 < 4 or 1 + 1 * 2 == 3 and not "i" in "team"
<u>2</u> **2 < 4 or 1 + 1 * 2 == 3 and not "i" in "team"
<u>4</u> < 4 or 1 + 1 * 2 == 3 and not "i" in "team"
4 < 4 or 1 + 2 == 3 and not "i" in "team"
$4 < 4$ or $\underline{3} == 3$ and not "i" in "team"
False or True and not False
False or True and <u>True</u>
False or <u>True</u>
Тюно
Inie

https://docs.python.org/3/reference/expressions.html

Python operator precedence

Higher precedence / Tighter binding



evaluated left to right

https://docs.python.org/3/reference/expressions.html