



Intro to Programming in Python

Computers out!
Internet connected!
Logged into GitHub!

Agenda for Monday

1. Intro to programming concepts
2. Download Spyder
3. Download Python
4. Practice using Spyder
5. Lecture 1 of text mining
6. Practice running on JupyterHub
7. Practice running on Spyder

Schedule

| | | | | Week 2 | | |
|--------|------|---------------------------|------------|---------------------------|----------|--|
| Day | Time | 9:30-11am | 11-11:20am | 11:20-1pm | 1pm-2:pm | 2-3pm |
| 26-Jun | M | intro to coding (AV & KN) | Break | Text Mining Overview (AV) | Lunch | Text Mining Module and Discussion (AV) |

1. Brief intro to programming
2. Download Spyder and Python
3. Python Basics
4. Practice running code in Spyder

1. Lecture of Text Preprocessing
2. Intro to JupyterHub

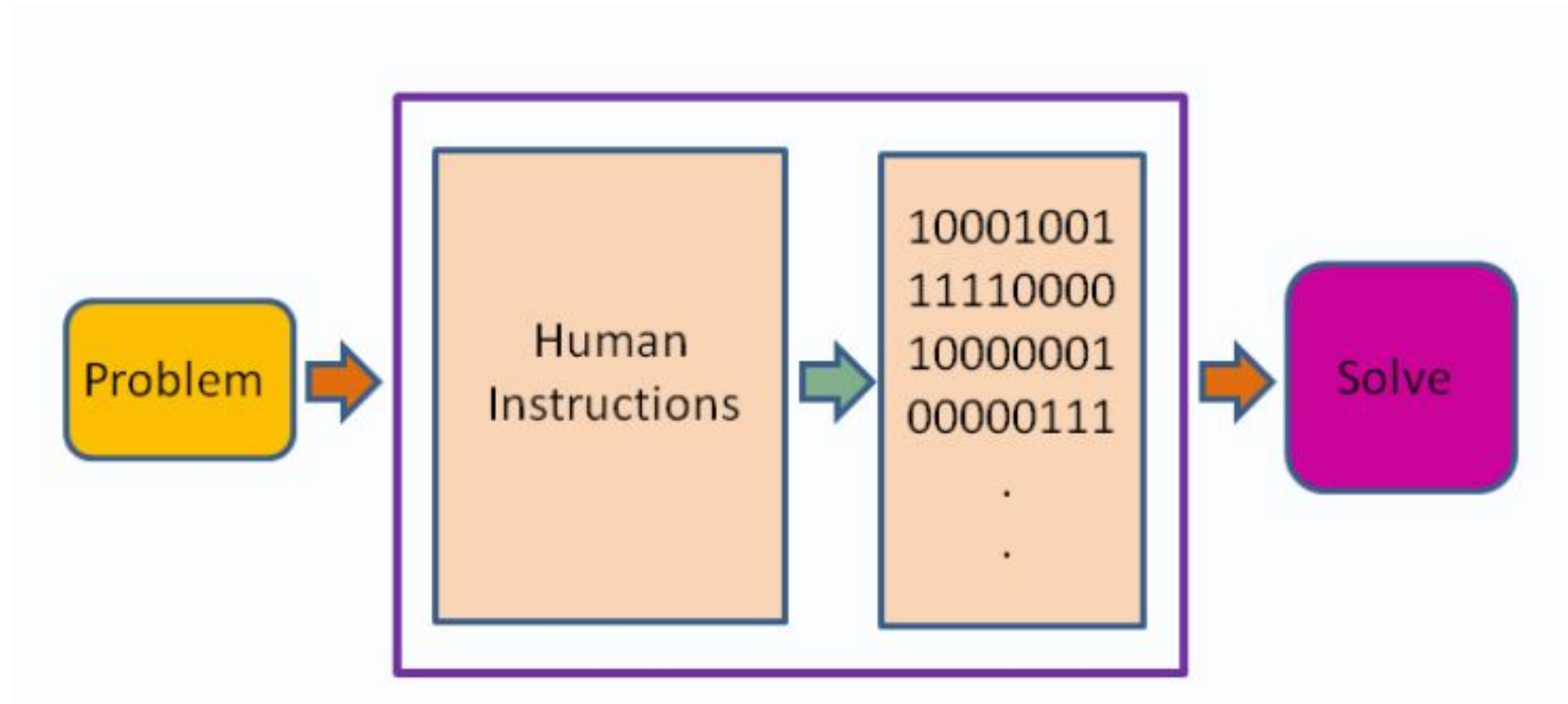
1. Working through Lecture 1 Text preprocessing code

What is programming?

“Programming is the act of instructing computers to carry out tasks.” It is often referred to as coding.

So then, what is a computer program? A computer program is a sequence of instructions that the computer executes.

What is programming?



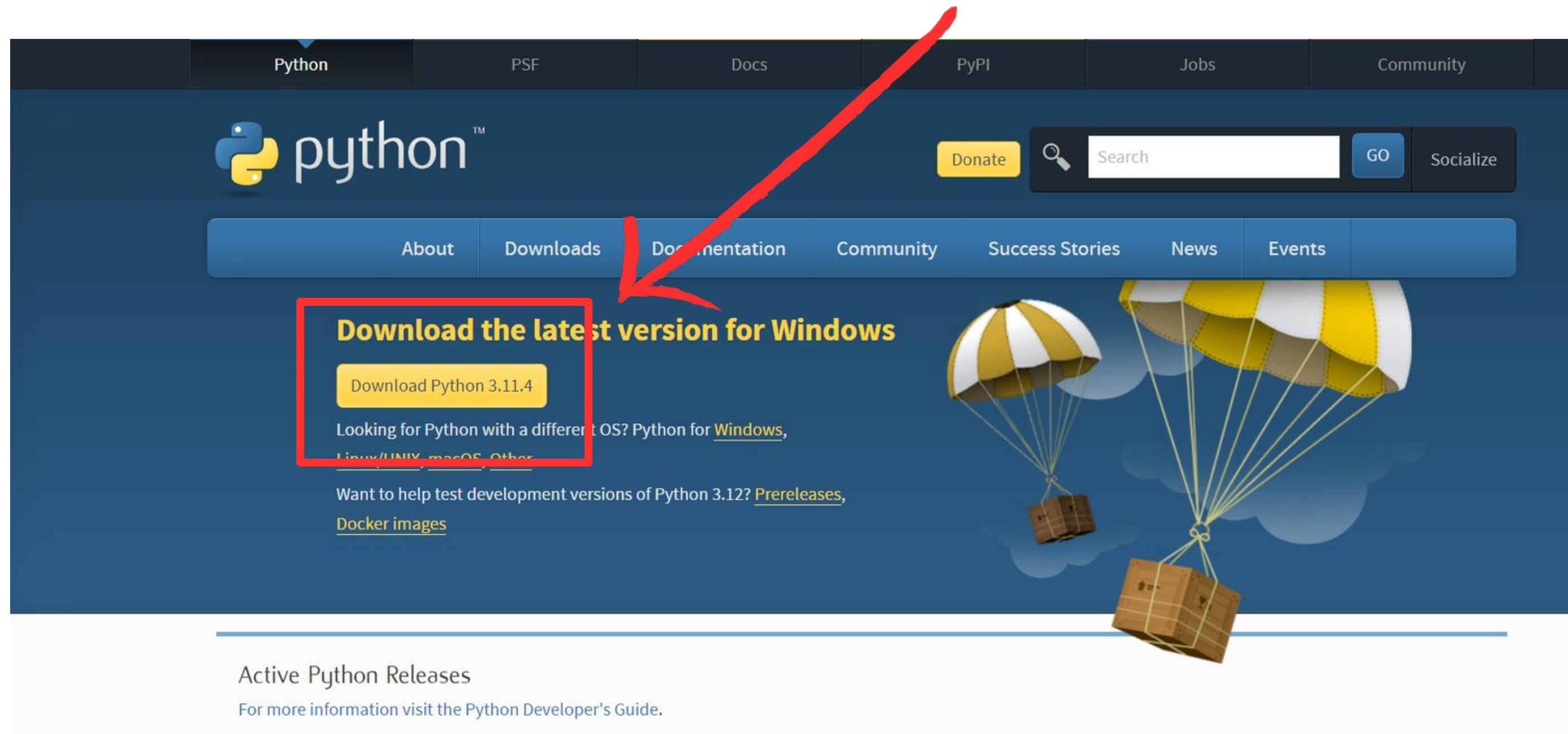
Instructions are given
through a programming
language (python)

Binary = language that
computers understand.

Downloading Python

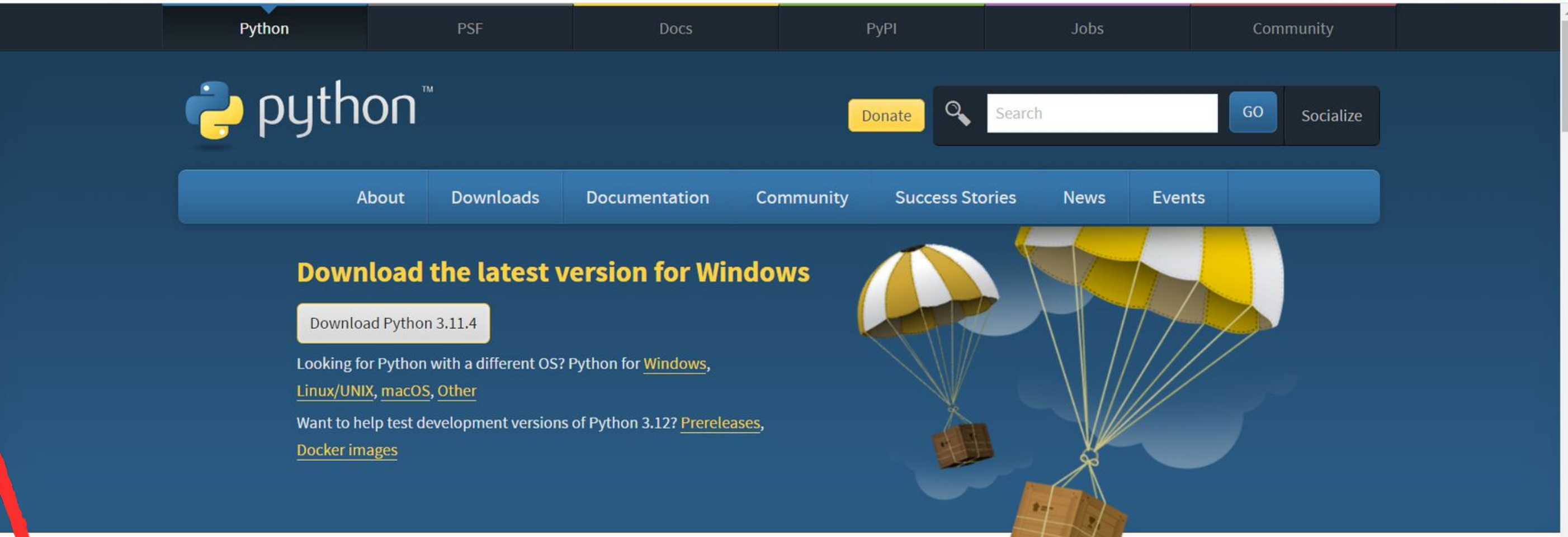
Go to <https://www.python.org/downloads/>

Click Download Python 3.11.4



Downloading Python

After the .exe file downloaded, double click to open



The screenshot shows the Python.org website with the following elements:

- Navigation bar: Python, PSF, Docs, PyPI, Jobs, Community
- Search bar: Search, GO, Socialize
- Buttons: About, Downloads, Documentation, Community, Success Stories, News, Events
- Main heading: **Download the latest version for Windows**
- Download button: Download Python 3.11.4
- Links: Looking for Python with a different OS? Python for [Windows](#), [Linux/UNIX](#), [macOS](#), [Other](#)
- Links: Want to help test development versions of Python 3.12? [Prereleases](#), [Docker images](#)
- Image: Two parachutes carrying crates.
- Section: Active Python Releases
- Text: For more information visit the Python Developer's Guide.
- Table:

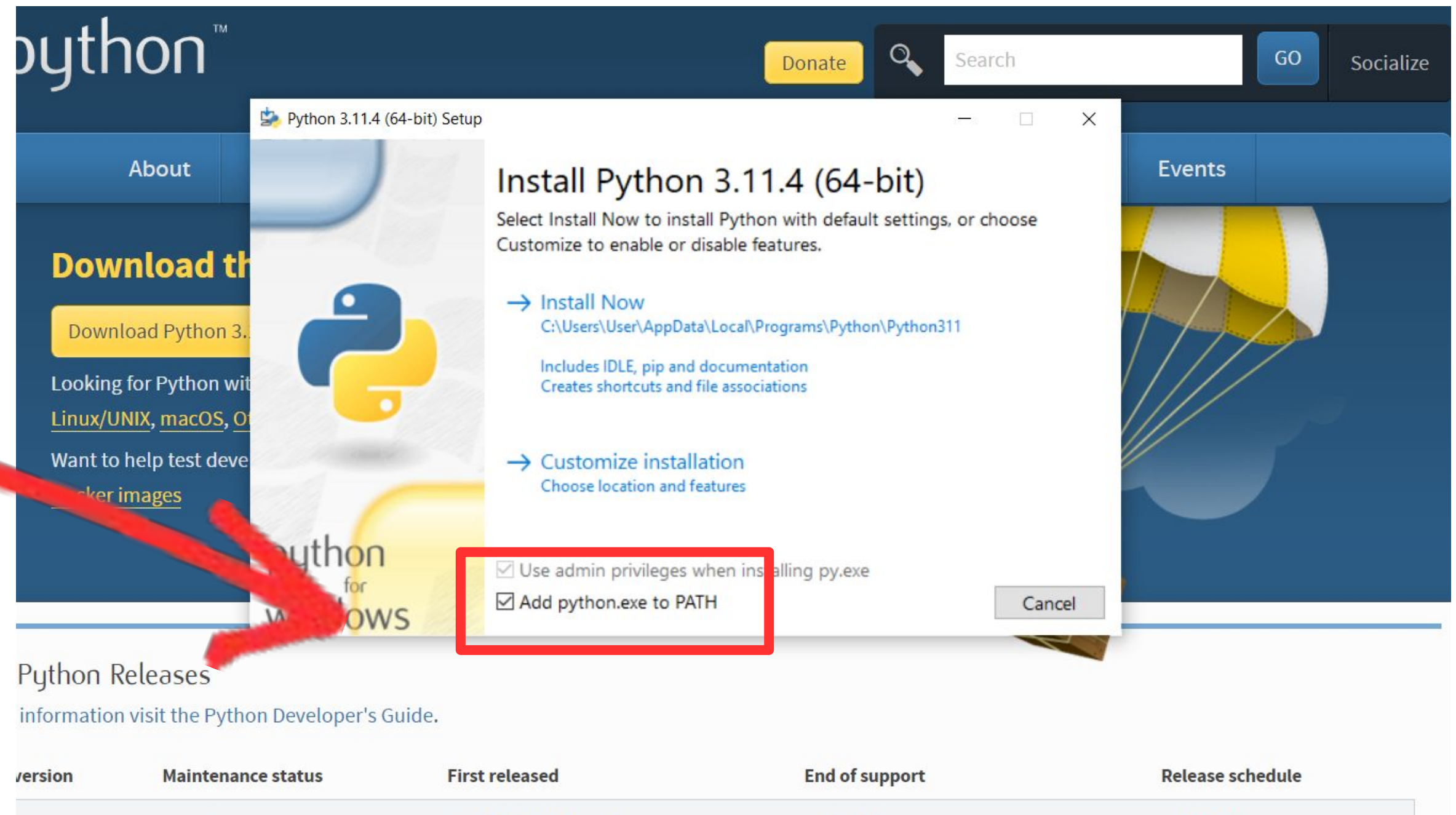
| Python version | Maintenance status | First released | End of support | Release schedule |
|----------------|--------------------|----------------------|----------------|------------------|
| 3.12 | prerelease | 2023-10-02 (planned) | 2028-10 | PEP 693 |

At the bottom, the Windows taskbar shows a download progress bar for 'python-3.11.4-am...exe' (4.8/24.2 MB, 12 secs left). A red arrow points from the text 'double click to open' to this file.

System tray: 11:44 AM, 6/13/2023, 80°F Partly sunny, 34% battery.

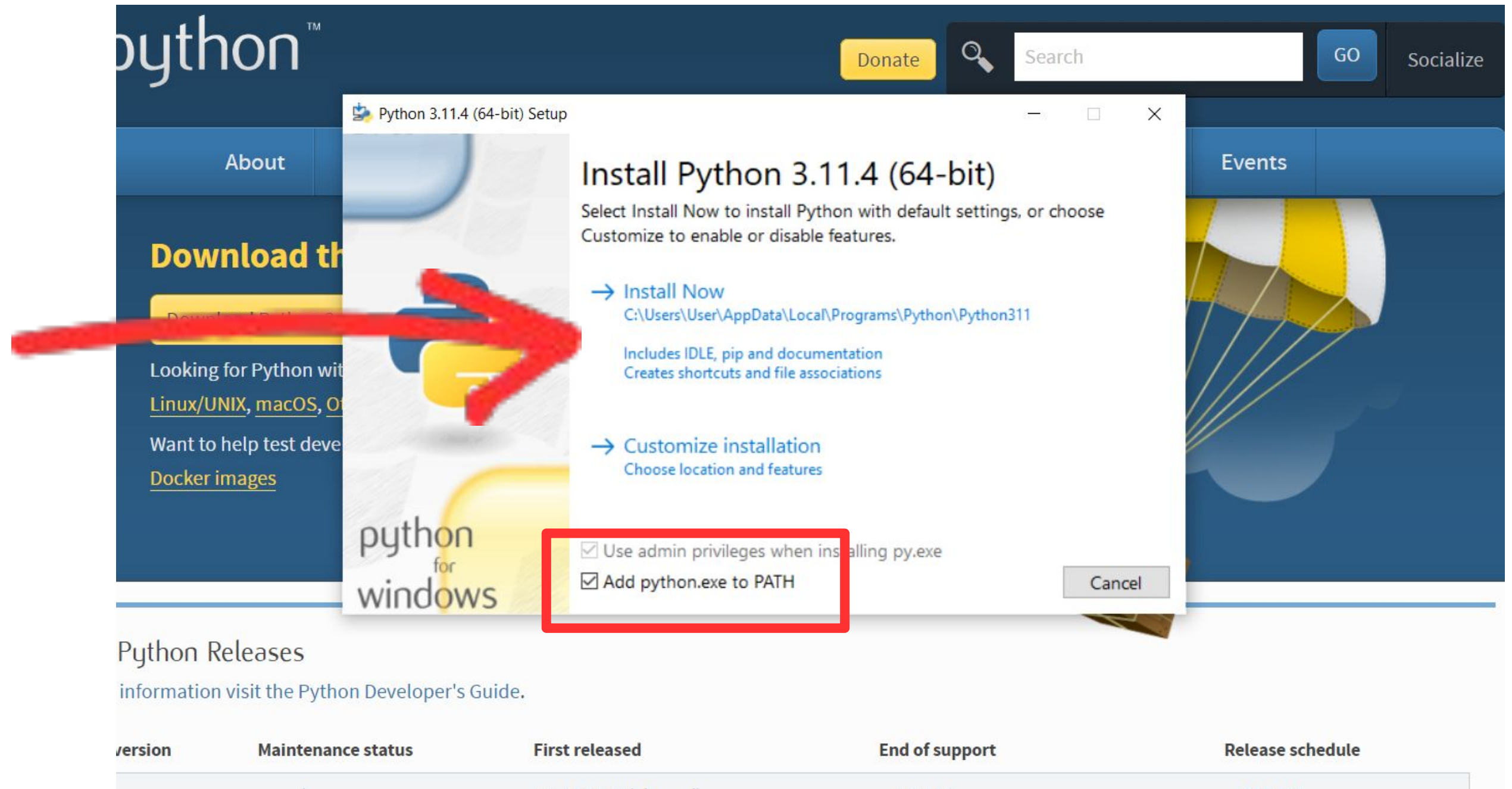
Downloading Python

Click "Add python.exe to
PATH"

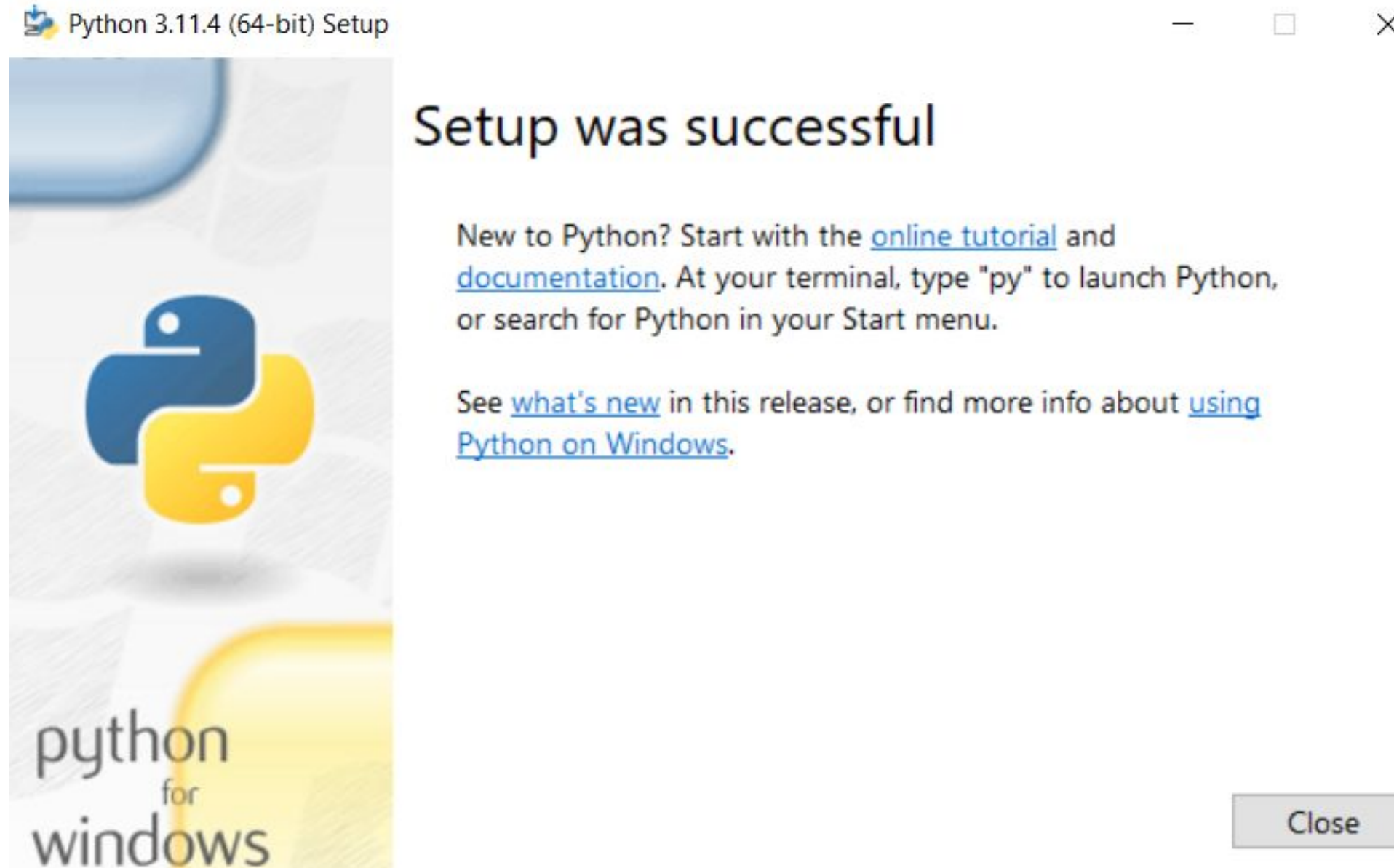


Downloading Python

Click "Install Now"



Downloading Python

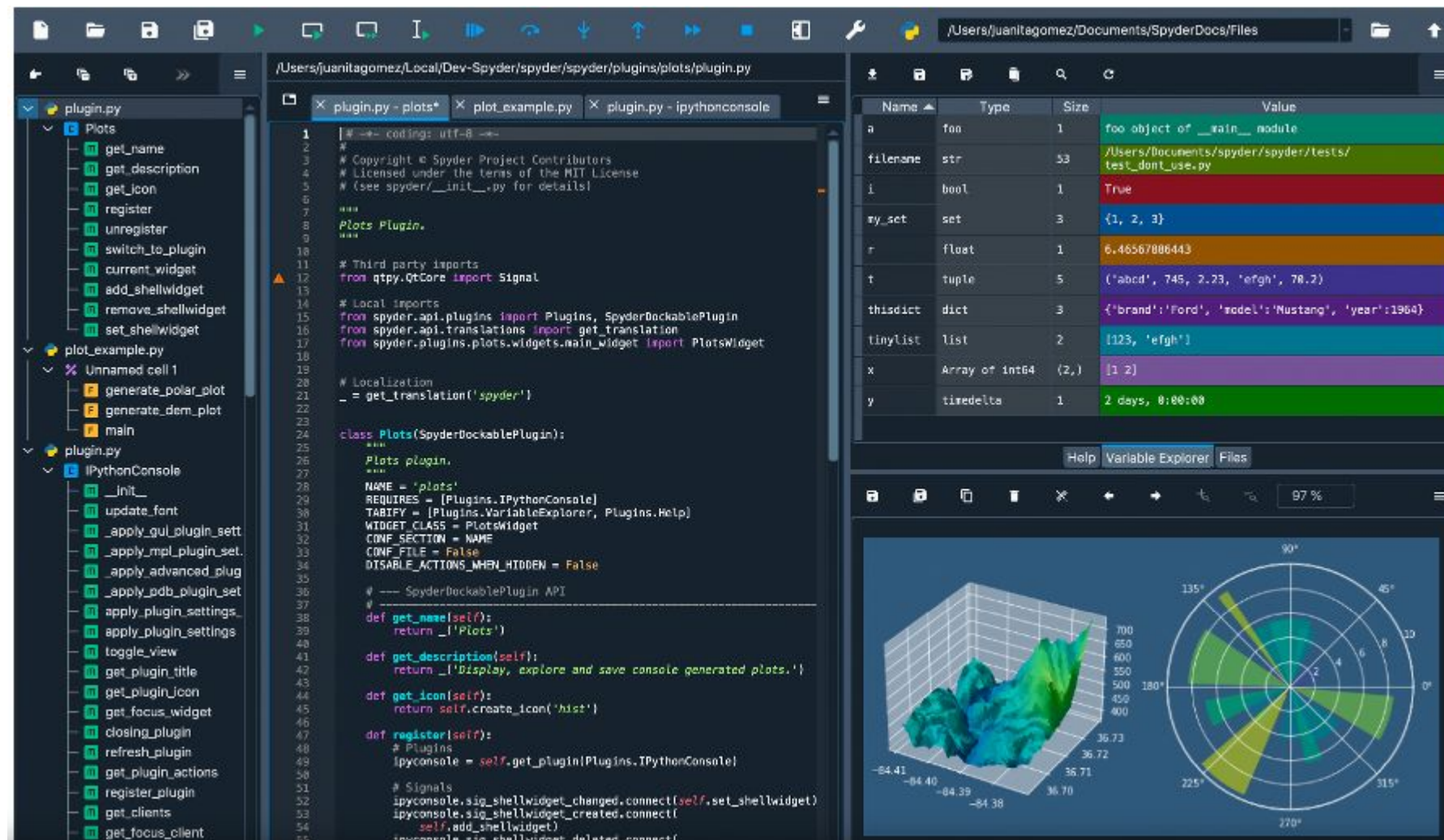


Downloading Spyder

Link: <https://www.spyder-ide.org/>

Spyder is a open source scientific environment written in Python.

Plots:
copy and
save
images
that you
create



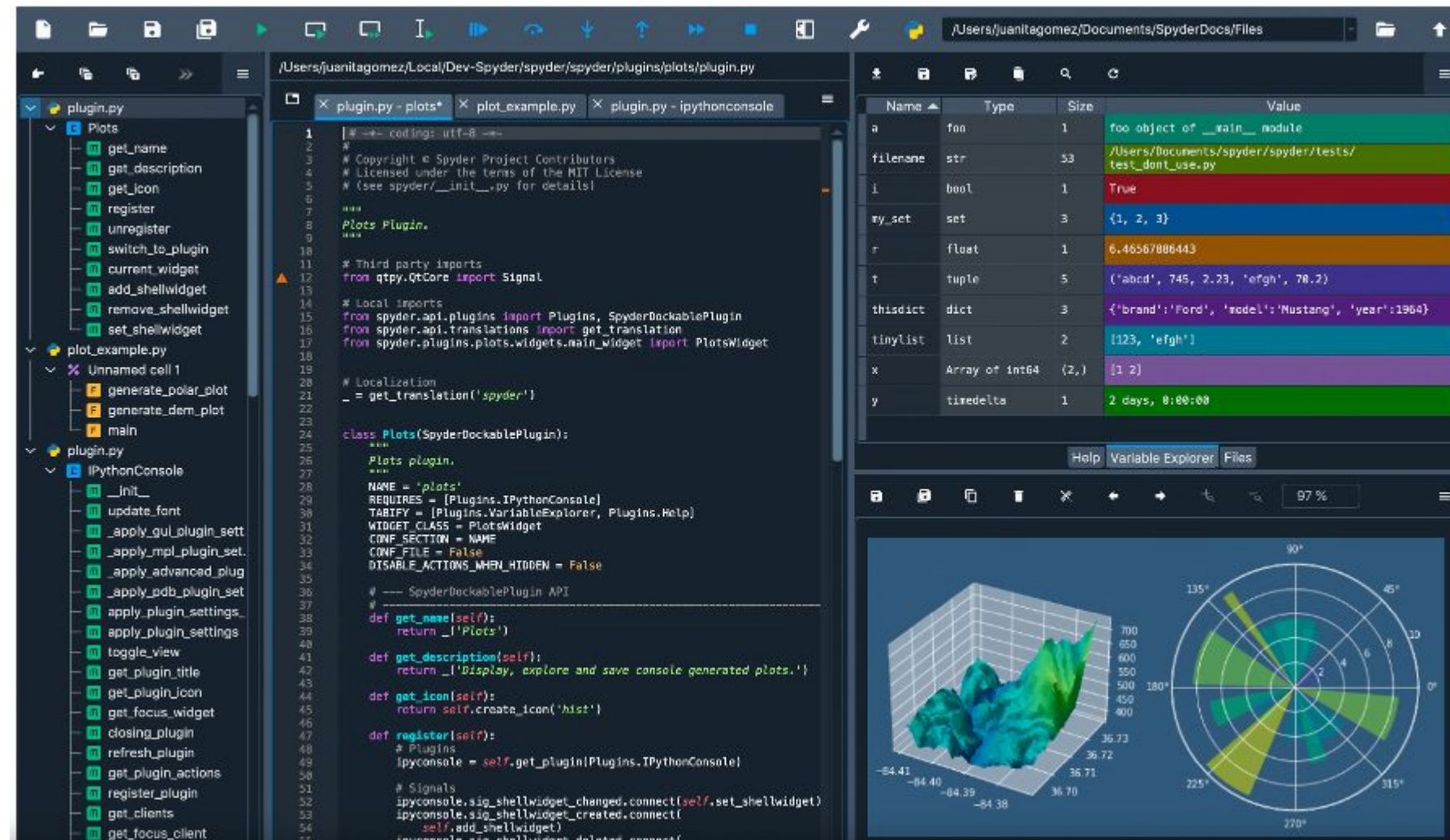
Debugger:
trace each
step of
your
code's
execution

Downloading Spyder

Link: <https://www.spyder-ide.org/>

Spyder is a open source scientific environment written in Python.

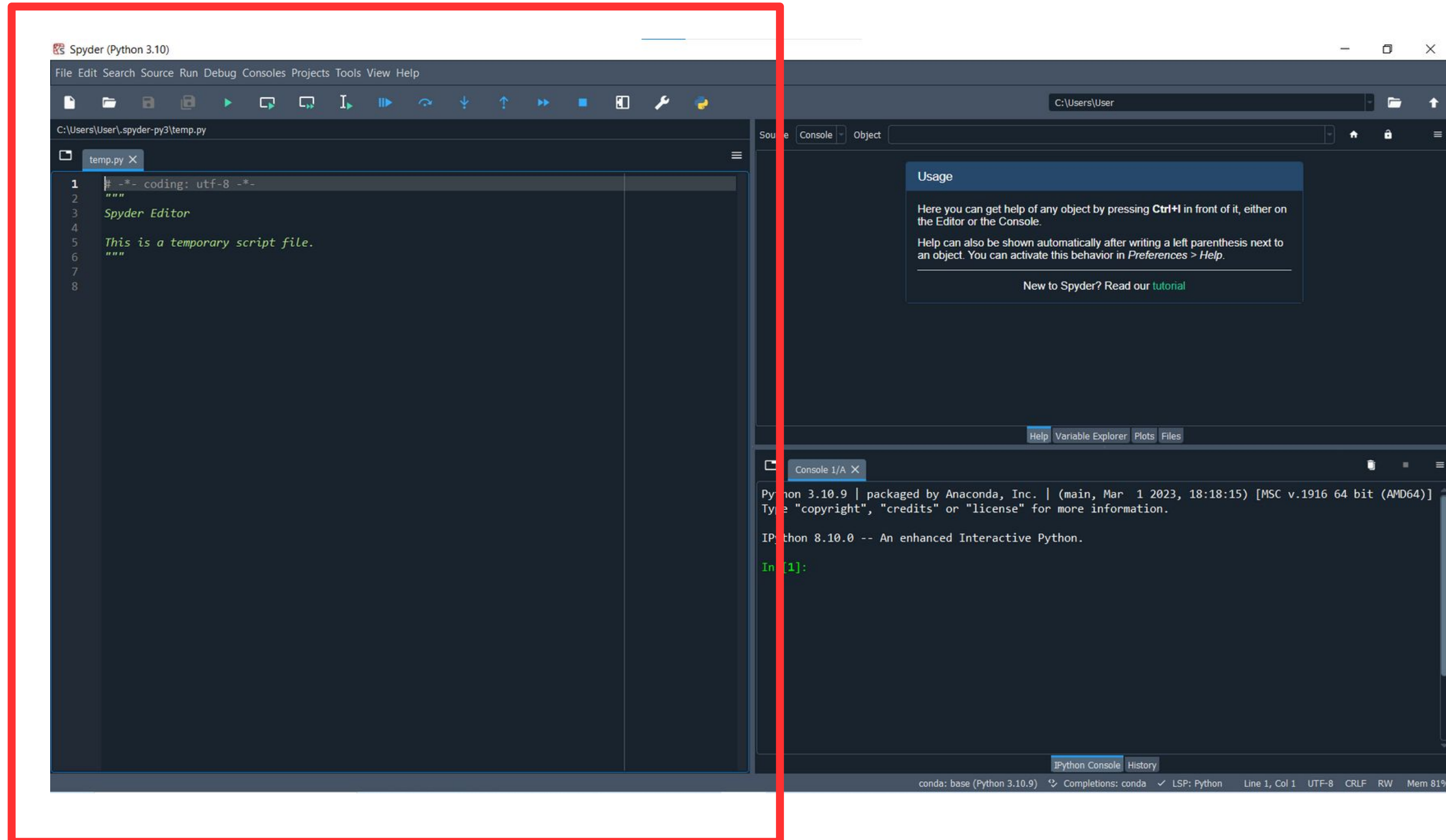
Plots:
copy and
save
images
that you
create



Debugger:
trace each
step of
your
code's
execution

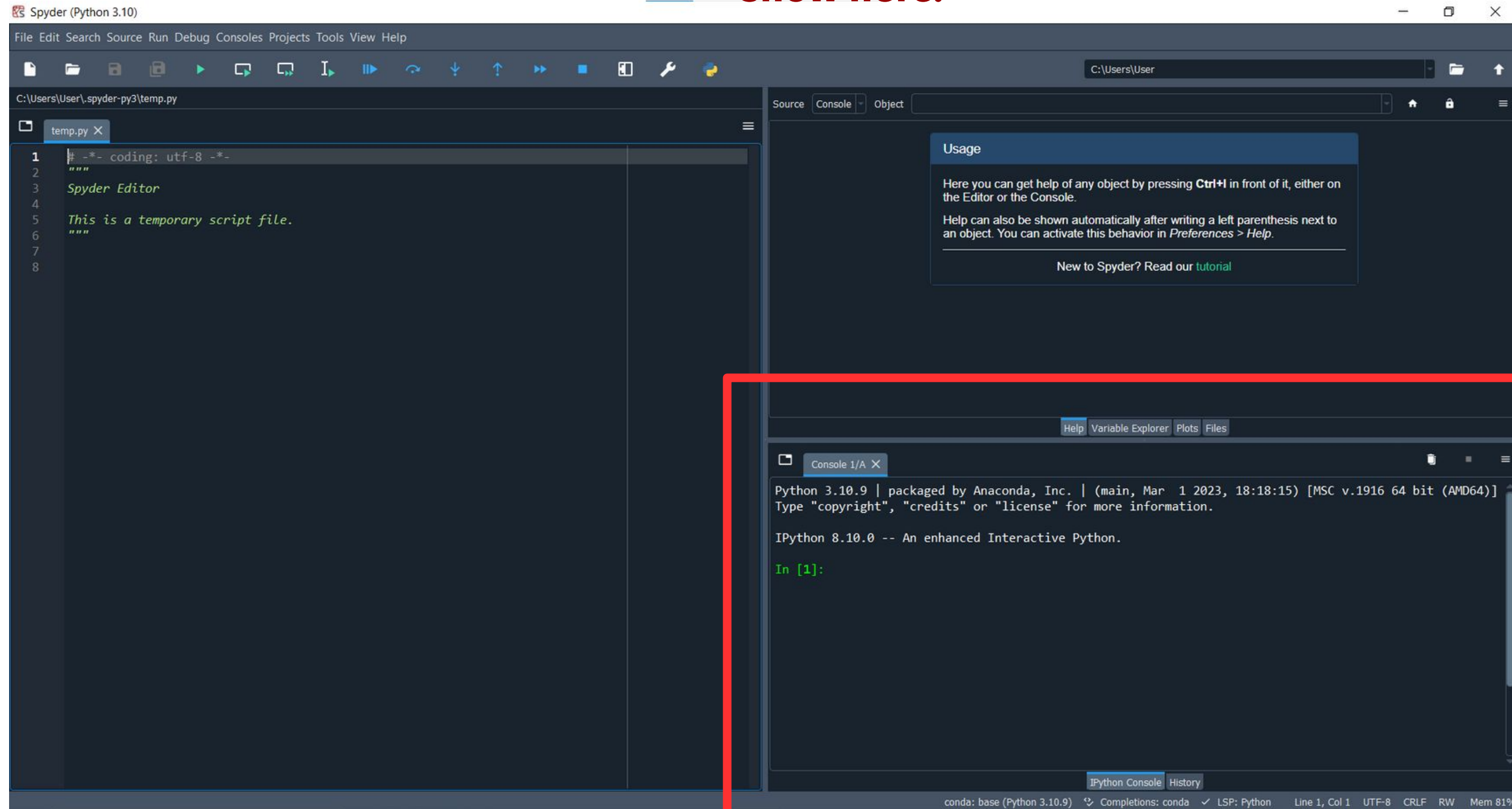
Spyder Python Environment

Text Editor: Write your python code here.



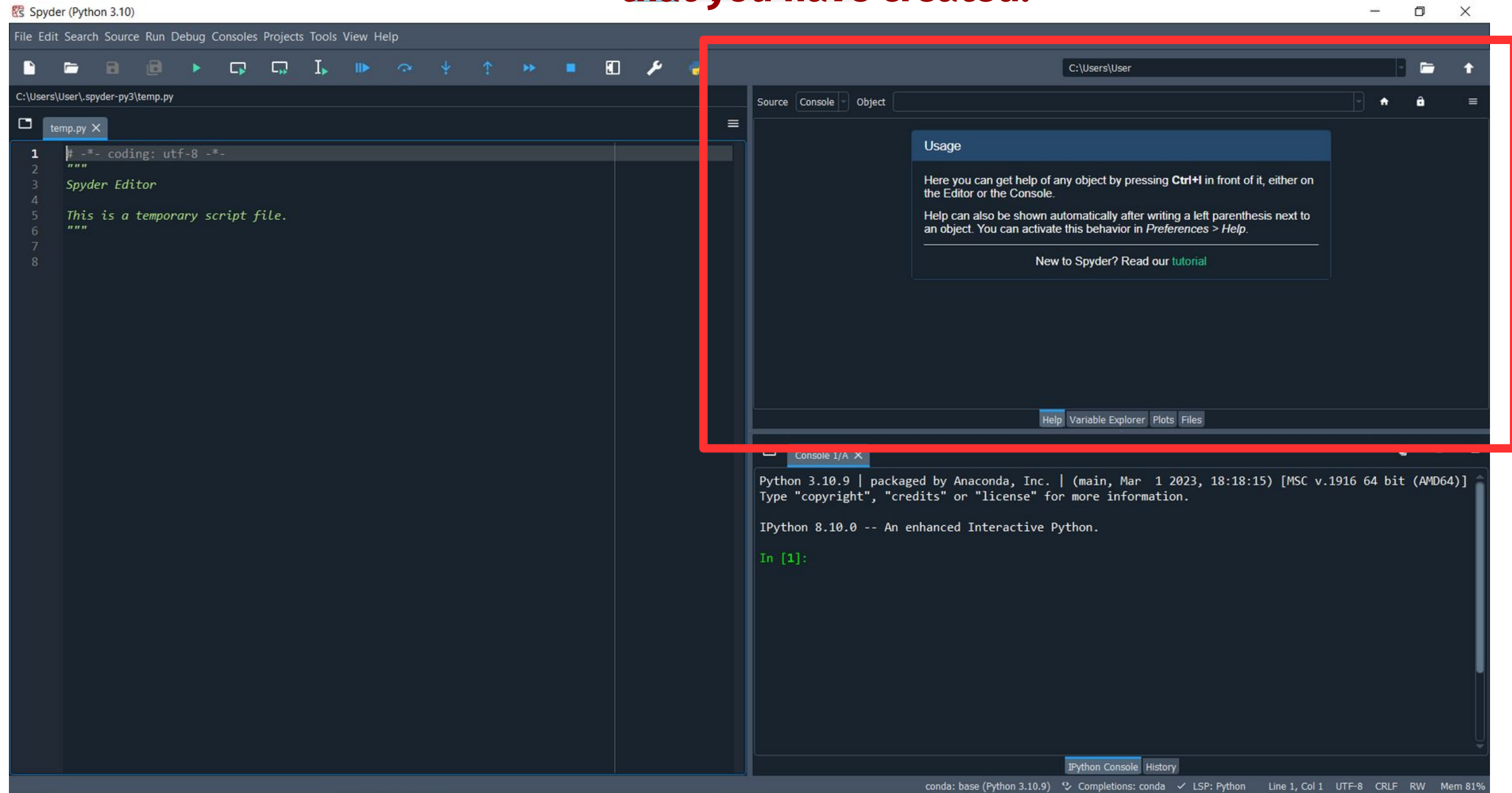
Spyder Python Environment

Console: Output after you run your code appears here. Error messages show here.



Spyder Python Environment

Variable Explorer: Allows you to see all the variables and functions that you have created.



Python Programming Basics

Variables

Variables are used to store information to be referenced and manipulated in a computer program.

- A variable name must start with a **letter or the underscore character**
- A variable name **cannot start with a number**
- A variable name can only contain **alpha-numeric characters and underscores** (A-z, 0-9, and _)
- Variable names are **case-sensitive** (age, Age and AGE are three different variables)

Exercise: Label the variable names as “legal” or “illegal”

name = “Dave”

2name = “Fred”

my-name = “Simon”

myname2 = “Fred”

my name = “Dave”

Variables

Variables are used to store information to be referenced and manipulated in a computer program.

Exercises:

#Exercise: Python program to print "Hello Python"

#Python program to find the area of a triangle.

#Python program to find the area of a circle

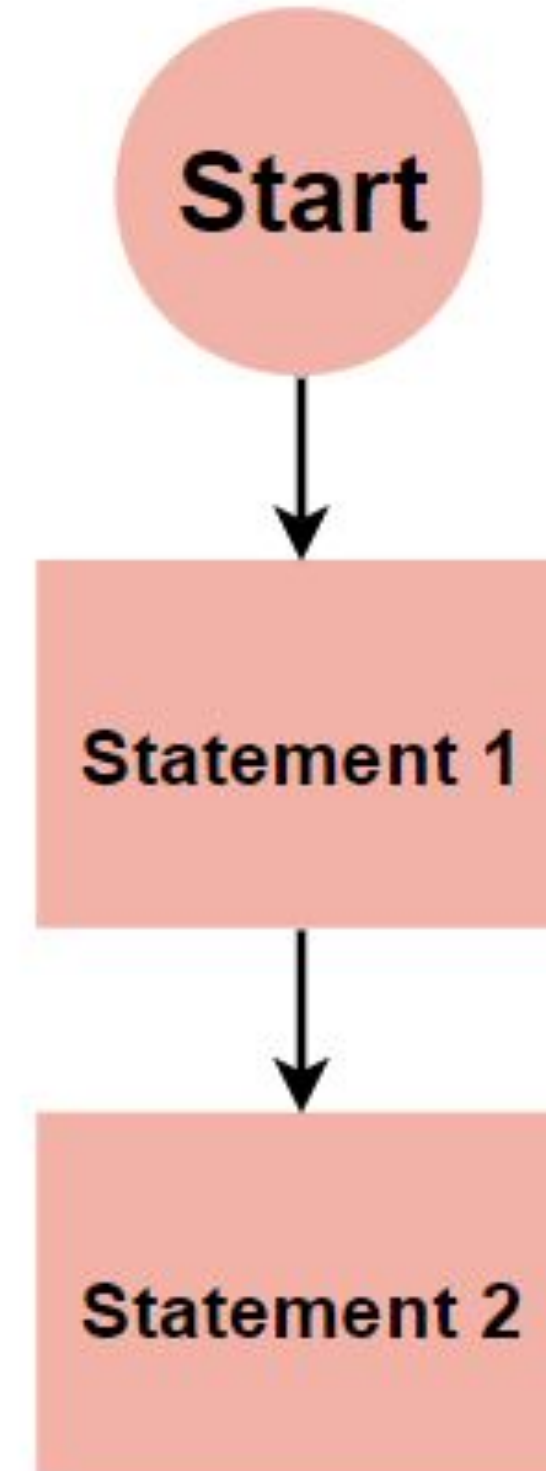
Flow Control

What order does the commands execute?

Sequential = one command gets run after the other.

```
print("Hello World")
```

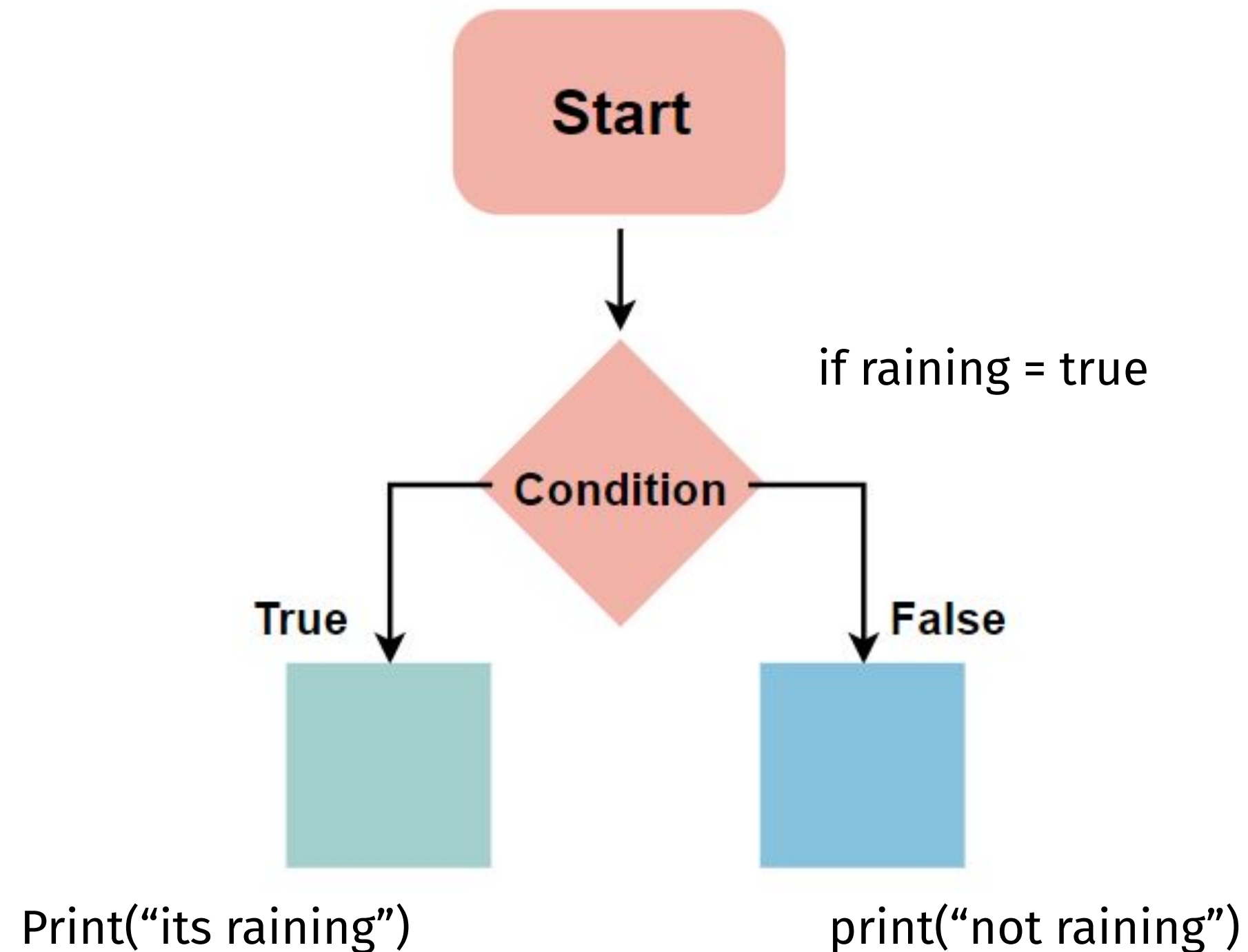
```
print("Hello Python")
```



Flow Control

What order does the commands execute?

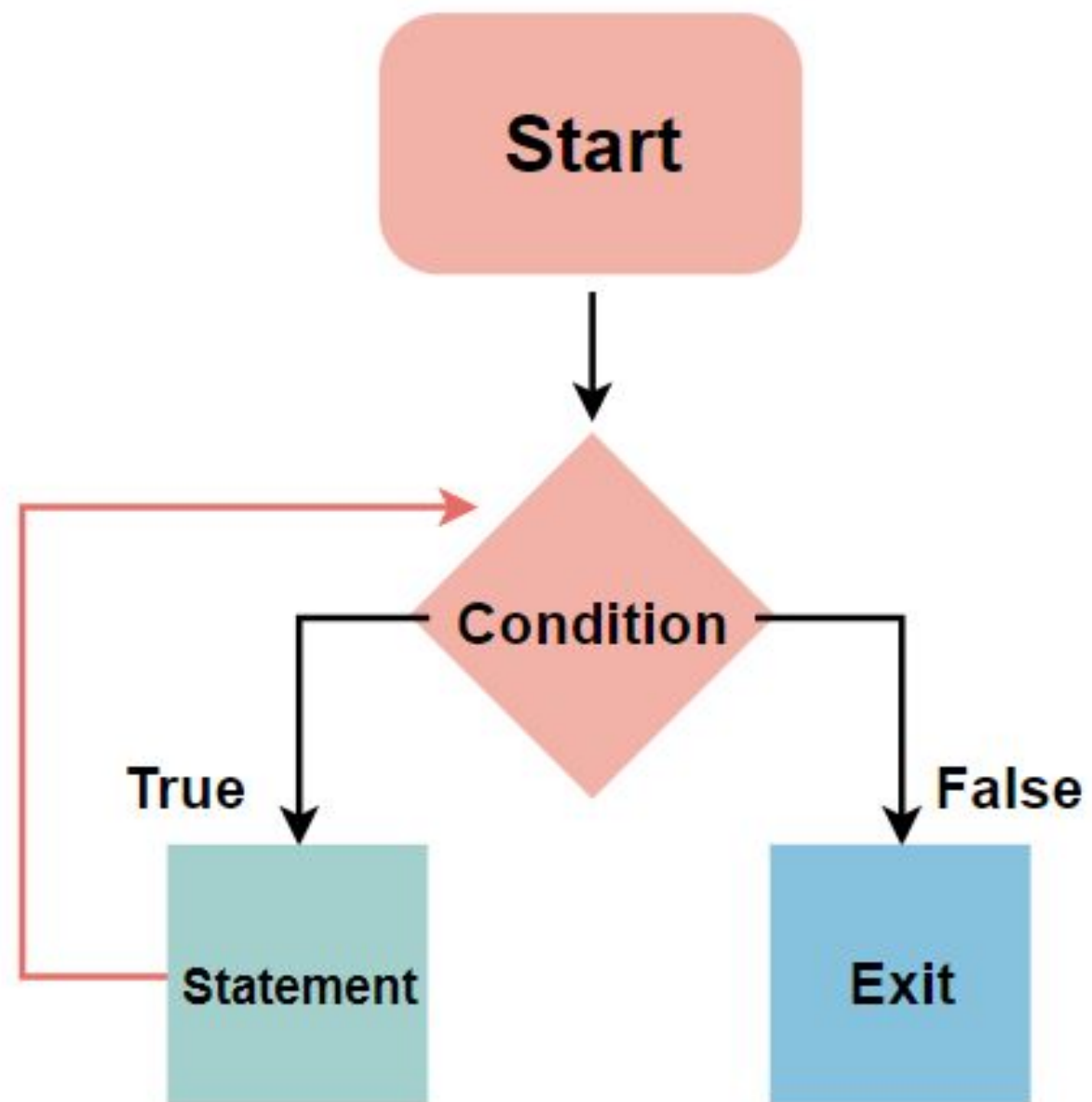
Selection = the computer decides what action to perform based on the result of a test or condition equalling true or false



Flow Control

What order does the commands execute?

Iteration (loop) = A loop is a programming structure that allows a statement or block of code to be run repeatedly until a specified condition is no longer true



Text Mining

The process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

What is the text mining process (pipeline)?

Today

Step #1: Text Preprocessing

- Special Character & Stopword Removal
- Stemming & Lemmatization

Step #2: Exploratory Analysis

- Preparing the Data
- Distribution of Target Class
- Word Frequencies

Step #3: Information Extraction

- Part of Speech Tagging (POS Tagging)
- Named Entity Recognition (NER)
- Relation Extraction

Step #4: Feature Representation

- Bag of Words Representation
- Vector and Vector Space
- Term Weighting

Step #5: Predictive Analysis

- Building a Bag-of-Word Matrix
- Predictive Analysis
- Logistic Regression (LR) Model
- KNN
- Concluding Visualization

Text Mining

The process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

Text Preprocessing

The process of cleaning and transforming unstructured text data to prepare it for analysis.

**Unstructured
data**



**Structured
data**



Insights

Text Preprocessing Pipeline

**Unstructured
Text**



- Remove Punctuation
- Remove Special Characters
- Remove Stopwords
- Lowercase
- Tokenization
- Stemming
- Lemmatization

**Structured
Text**



Overview of Text Data Preprocessing

Watch video 1

Follow up questions:

- a. What are some examples of the uses of text mining?
- b. What are examples of the unit of analysis in text mining?
- c. What are examples of the unit of analysis called in data mining?

Overview of Text Data Preprocessing

Answers for video 1

Follow up questions:

- a. What are some examples of the uses of text mining?
 - i. Sentiment analysis for hotel review
 - ii. Predicting the topic of news article
 - iii. Spelling error correction for search engines
 - iv. Find high risk patients based on patient notes
- b. What are examples of the unit of analysis in text mining?
 - i. “A”, “dog”, “tree”, “rock”...
- c. What are examples of the unit of analysis called in data mining?
 - i. Age, blood glucose level, race/ethnicity...

Text Data Preprocessing

Watch video 2

Follow up questions:

- a. What is Text Preprocessing? What are the three main components?
- b. What is part of the normalization process?
- c. According to the video, would an apostrophe be categorized as an independent token?
- d. Why do we remove special characters such as ^, *, # and stop words such as “a”, “on”, ...

Text Data Preprocessing

Answer to video 2

Follow up questions:

- a. What is Text Preprocessing? What are the three main components?
 - i. Transforms text into a more digestible form so machine learning algorithms can perform better.
 - ii. Tokenization & Segmentation, Normalization, Noise Removal
- b. What is part of the normalization process?
 - i. Removing Special Characters, Removing Stopwords, Stemming, and Lemmatization
- c. According to the video, would an apostrophe be categorized as an independent token?
 - i. Yes
- d. Why do we remove special characters such as ^, *, # and stop words such as “a”, “on”, ...
 - i. They add to the number of features and slow down computational speed.

Step 1. Removing Punctuation

Removing punctuation will help us treat text equally.

Data Data! Data. Data?

How many times does the word "Data" appear in the phrase above?

We would say there are **4 unique terms**. "Data" is different from "Data?", which is different from "Data!"

In reality, we see that it's just 1 word repeated 4 times.

Step 2. Removing Special Characters

Removing special characters will help us treat text equally.

Data@ Data& Data Data#

Data Data Data Data

Example

**Remove punctuation
& special characters**

text_data = "The original decaffeination process of coffee beans involved the use of a toxic chemical called benzene! Benzene is a carcinogen and very harmful to your health. New green chemistry techniques have been adopted to avoid the use of chemicals to remove caffeine from coffee beans."

Step 3. Removing Stopwords

What are stopwords?

text_data = "The original decaffeination process of coffee beans involved the use of a toxic chemical called benzene Benzene is a carcinogen and very harmful to your health New green chemistry techniques have been adopted to avoid the use of chemicals to remove caffeine from coffee beans"

Step 3. Removing Stopwords

What are **keywords**?
What are **stopwords**?

`text_data = "The original decaffeination process of coffee beans involved the use of a toxic chemical called benzene Benzene is a carcinogen and very harmful to your health New green chemistry techniques have been adopted to avoid the use of chemicals to remove caffeine from coffee beans"`

Step 3. Removing Stopwords

New Text without stopwords

```
new_text_data = "original decaffeination process coffee beans  
involved use toxic chemical called benzene Benzene carcinogen  
harmful health New green chemistry techniques adopted avoid  
use chemicals remove caffeine coffee beans"
```

Say that you are trying to look for your car keys in your messy bedroom. It's easier to find them after you have a clean room.

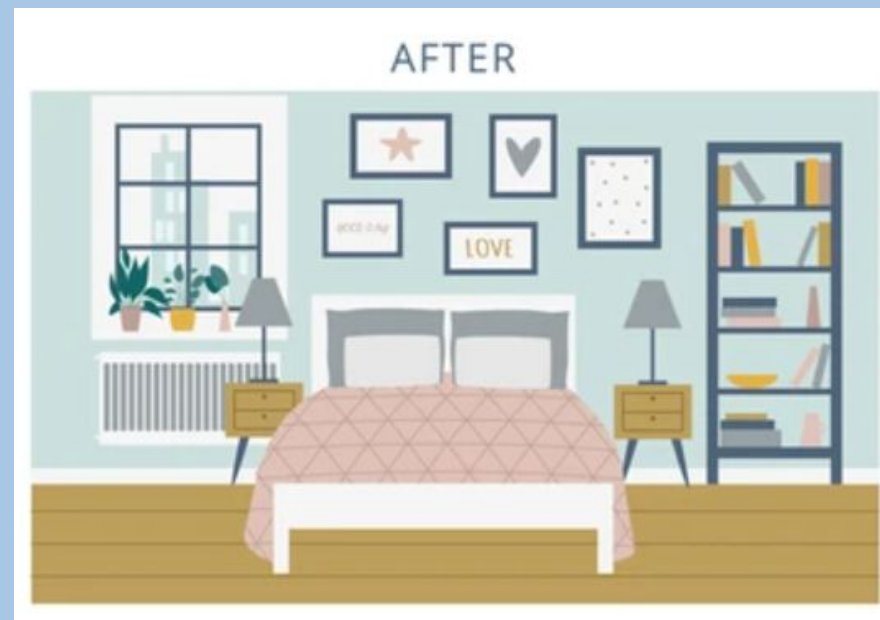
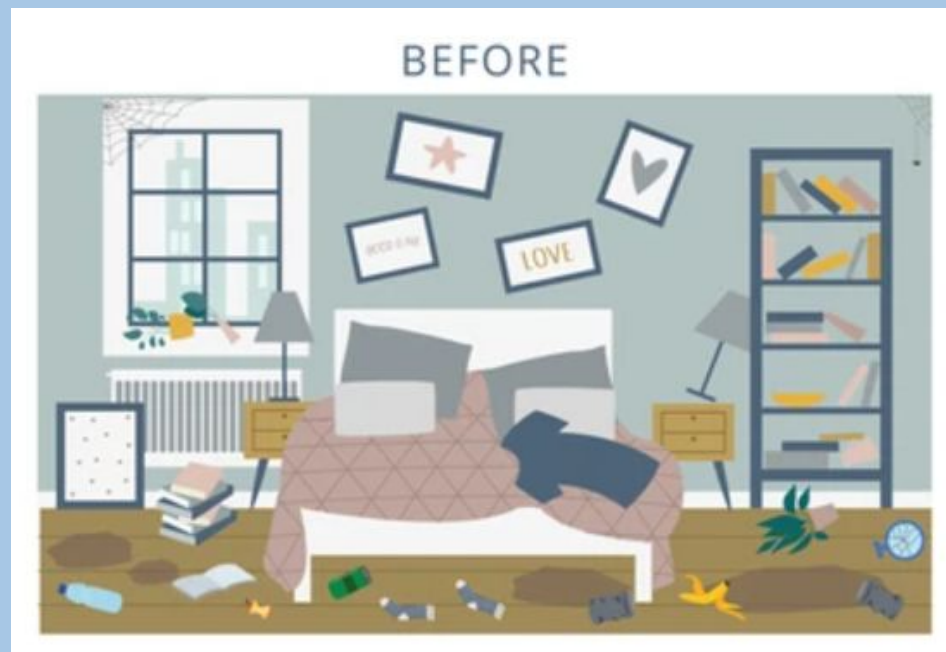
**Unstructured
data**



**Structured
data**



Insights



Room is CLEAN and ORGANIZED.

Step 4. Lowercasing

Lowercasing will help us treat text equally.

Data data DAta DaTA

How many times does the word "Data" appear in the phrase above?

**We would say there are 4 unique terms. "Data" is different from "data",
which is different from "DAta"**

In reality, we see that it's just 1 word repeated 4 times.

Step 4. Lowercasing

New Text without stopwords, lowercased

```
lc_text_data = "original decaffeination process coffee beans  
involved use toxic chemical called benzene benzene carcinogen  
harmful health new green chemistry techniques adopted avoid  
use chemicals remove caffeine coffee beans"
```

Step 5. Tokenization

Tokenization is the process of breaking up a given text into units called **tokens**.

Why do we need to **tokenize** text?

Step 5. Tokenization

New Text without stopwords, lowercased, tokenized

```
tokenized_text_data = [original, decaffeination, process, coffee,  
beans, involved, use, toxic, chemical, called, benzene, benzene,  
carcinogen, harmful, health, new, green, chemistry, techniques,  
adopted, avoid, use, chemicals, remove, caffeine, coffee, beans]
```

Step 6. Stemming

Stemming is the process that stems or removes last few characters from a word (removing suffixes).

What are the advantages of stemming? Why do we need to do this?

Stemming Examples

What is the stem of these words?

Chang**ing**



Chang

Chang**ed**



Chang

Do you see any limitations with stemming?

Porter Stemmer

It is one of the most popular stemming methods proposed in 1980.

It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes.

This stemmer is known for its speed and simplicity.

Advantage: It produces the best output as compared to other stemmers and it has less error rate.

Limitation: Morphological variants produced are not always real words.

Why do we need to do stemming?

Grouping similar words

Words with a similar meaning can be grouped together, even if they have distinct forms. This can be a useful technique in tasks such as document classification, where it's important to identify key topics or themes within a document

Improved model performance

Stemming reduces the number of unique words that need to be processed by an algorithm, which can improve its performance. Additionally, it can also make the algorithm run faster and more efficiently.

Easier to analyze and understand

Since stemming typically reduces the size of the vocabulary, it's much easier to analyze, compare, and understand texts. This is helpful in tasks such as sentiment analysis, where the goal is to determine the sentiment of a document.

Step 6. Stemming

What is the stem of these words?

```
text_data = [original, decaffeination, process, coffee, beans,  
             involved, use, toxic, chemical, called, benzene, benzene,  
             carcinogen, harmful, health, new, green, chemistry, techniques,  
             adopted, avoid, use, chemicals, remove, caffeine, coffee, beans]
```

Step 6. Stemming

What is the stem of these words?

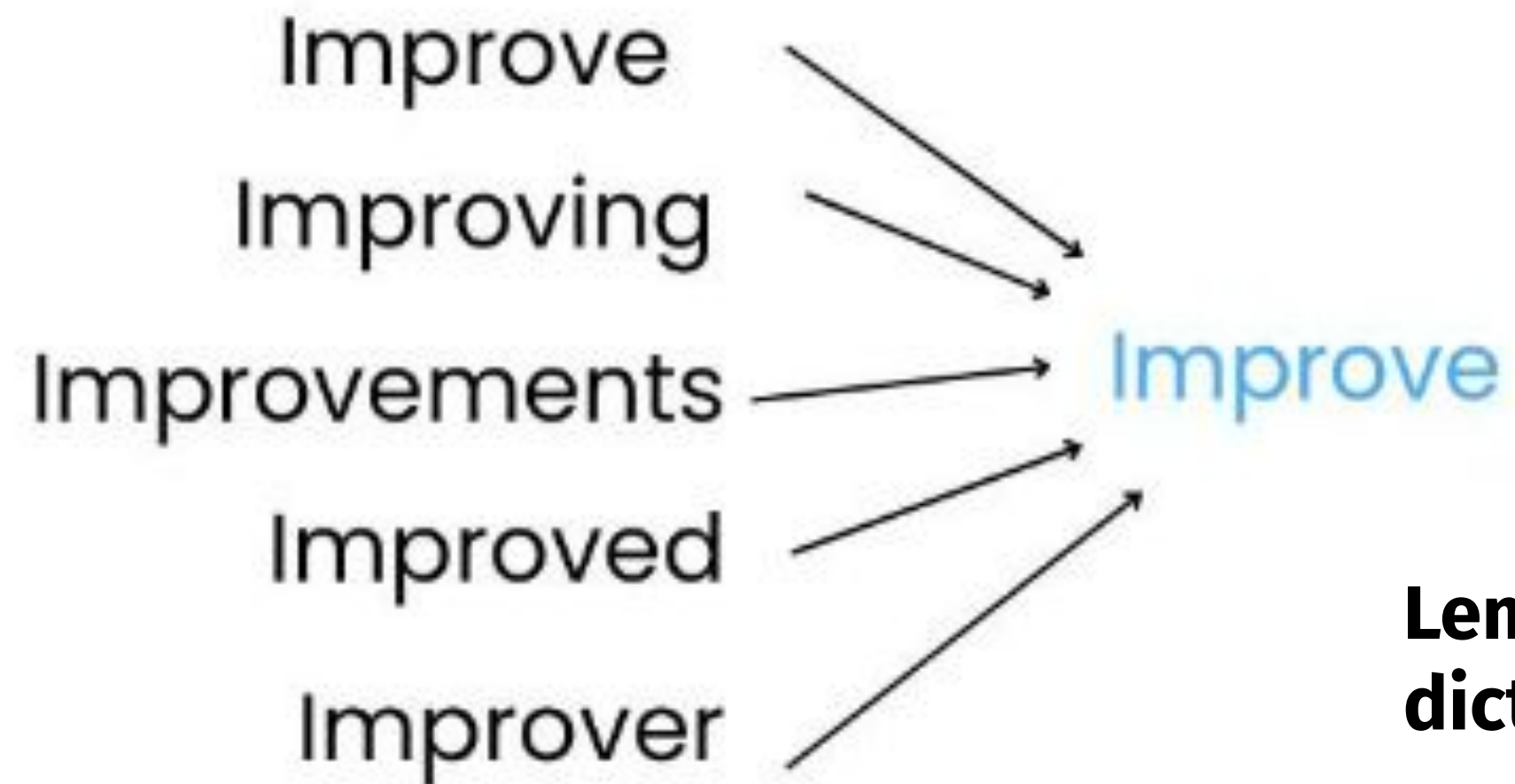
```
stemmed_text_data = [origin, decaffein, process, coffe, bean,  
involv, use, toxic, chemic, call, benzene, benzene, carcinogen,  
harm, health, new, green, chemistri, techniqu, adopt, avoid, use,  
chemic, remov, caffein, coffe, bean]
```

Step 7. Lemmatization

The process of reducing a word to its root form, called a lemma.

Lemmatization considers the context and converts the word to its **meaningful base form**, which is called lemma.

Lemmatization Examples



**Lemma =
dictionary word**

Stemming & Lemmatization

Watch video 3

Follow up questions:

- a. What is stemming?
- b. What is an Affix?
- c. What is the inflection of the word compute? What is the base form of the word, or stem?
- d. True or False, root stem can be found in the dictionary?
- e. What is the “lemma” of am?

Stemming & Lemmatization

Answers to video 3

Follow up questions:

- a. What is stemming?
 - i. Reducing inflectional forms in a word to the base form of a word
- b. What is an Affix?
 - i. An additional element at the beginning or end of a root, stem, or word, or in the body of a word, to modify its meaning. Ex) anti-, hyper-, sub-...
- c. What is the inflection of the word compute? What is the base form of the word, or stem?
 - i. Inflection: computes, computed, computing Stem: comput
- d. True or False, root stem can be found in the dictionary?
 - i. False, root words from lemmatization can be found in a dictionary but the root stem may not be
- e. What is the “lemma” of “am”?
 - i. Be

What is the difference between Stemming and Lemmatization?

Stemming

Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling.

For instance, stemming the word '**Caring**' would return '**Car**'.

Stemming is used in case of large dataset where performance is an issue.

Lemmatization

Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.

For instance, lemmatizing the word '**Caring**' would return '**Care**'.

Lemmatization is computationally expensive since it involves look-up tables and what not.

WordNet Lemmatizer

Python has a module called WordNet. WordNet performs lemmatization on the text we input.

Fixes the issue we had from stemming....After lemmatization, all the words have meaning (found in the English dictionary)

Step 7. Lemmatization

What is the lemma of these words?

lemmatization_text_data = [original, decaffeination, process, coffee, beans, involved, use, toxic, chemical, called, benzene, benzene, carcinogen, harmful, health, new, green, chemistry, techniques, adopted, avoid, use, chemicals, remove, caffeine, coffee, beans]

Python Tools

1. **NLTK (Natural Language Toolkit)** is a popular open-source library for natural language processing (NLP) tasks in Python. It provides a wide range of tools, resources, and algorithms for processing and analyzing human language data.

NLP stands for Natural Language Processing. It is a subfield of artificial intelligence (AI) and linguistics that focuses on the interaction between computers and human language. NLP involves the development of algorithms and models to enable computers to understand, interpret, and generate human language in a way that is meaningful and useful.