Data Science Toolkit on Ubuntu 18.04

Anaconda Installation Instruction

This document provides a step-by-step instruction to install useful data science tools on your Ubuntu 18.04 Linux distro. If you would like to try scripts to automatically install those tools, you can download bash shell files at DSIA Team SharePoint site, and then execute it with sudo user privileges.

1. Prepare Ubuntu for Installing Data Science Toolkit

In your Ubuntu Linux laptop, Ubuntu virtual machine or Windows Subsystem for Linux (WSL), open terminal and start typing and executing the following commands to install some basic software.

```
$ sudo apt update

$ sudo apt install wget curl

$ sudo apt install git-all

$ sudo apt install git
```

2. Install Anaconda

Designed for data science and machine learning workflows, Anaconda is an open-source package manager, environment manager, and distribution of the Python and R programming languages.

This section will guide you through installing Anaconda on an Ubuntu 18.04 server. From a web browser, go to the Anaconda Distribution page via the following link: `https://www.anaconda.com/distribution/` to find the latest Linux version and copy the installer

Please Note: Uninstall Anaconda (`rm -rf~/anaconda3)` if you found your machine has an old version Anaconda. If you want to keep your existing Anaconda, you can ignore the steps of "Download and Install Anaconda" and "Run the Anaconda Script" to avoid any conflictions.

Download and Install Anaconda

Logged into your Ubuntu 18.04 as a sudo non-root user, move into the `/tmp` directory and use `curl` to download the link you copied from the Anaconda website:

```
$ cd /tmp
```

```
$ curl -O -k https://repo.anaconda.com/archive/Anaconda3-2019.10-Linux-x86_64.sh
```

Run the Anaconda Script

```
$ bash Anaconda3-2019.10-Linux-x86_64.sh
```

You'll receive the following output to review the license agreement by pressing `ENTER` until you reach the end. When you get to the end of the license, type `yes` if you agree to the license to complete installation.

```
Welcome to Anaconda3 5.2.0 (by Continuum Analytics, Inc.)

In order to continue the installation process, please review the license agreement.
Please, press ENTER to continue
Do you approve the license terms? [yes|no]
[no] >>> yes
Anaconda will now be installed into this location:
/home/username/anaconda3

- Press ENTER to confirm the location
- Press CTRL-C to abort the installation
- Or specify an different location below
[/home/username/anaconda3] >>>
Python 3.6.5 :: Continuum Analytics, Inc.
creating default environment...
installation finished.
Do you wish the installer to prepend the Anaconda install location
to PATH in your /home/username/.bashrc ? [yes|no]
[no] >>> yes

Prepending PATH=/home/username/anaconda/bin to PATH in /home/username/.bashrc
A backup will be made to: /home/username/.bashrc-anaconda.bak


Do you wish to proceed with the installation of Microsoft VSCode? [yes|no]
>>> no

For this change to become active, you have to open a new terminal.

Thank you for installing Anaconda!
```

Activate Installation

You can now activate the installation with the following command:

```
$ sudo chmod -R 777 ~/anaconda3/

$ PATH=~/anaconda3/bin:$PATH

$ echo export PATH=~/anaconda3/bin:\$PATH >> ~/.bashrc

$ . ~/.bashrc
```

Test Installation

Use the `conda` command to test the installation and activation:

```
$ conda list
```

You'll receive output of all the packages you have available through the Anaconda installation.

Check python version via python command.

```
username@ubuntu:~$ python
Python 3.7.3 (default, Mar 27 2019, 22:11:17)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> quit()
```

Update Installation (Optional)

You can easily update Anaconda to the latest version.

```
username@ubuntu:~$ conda update --all --yes
```

Additional Installation of Python Libraries (Optional)

You can easily install some python popular data science libraries, such as tensorflow, keras, etc. by using conda install or pip install. For example, the following commands will install those useful libraries:

```
$ pip install --user tensorflow pymc3 keras

$ pip install --user fbprophet

$ pip install --user clarify

$ pip install --user pandas-profiling

$ pip install --user koalas

$ pip install --user ipython-sql

$ pip install --user jupyter_contrib_nbextensions

$ jupyter contrib nbextension install --sys-prefix

$ pip install --user autopep8

$ pip install --user findspark

$ pip install --user spark-df-profiling
```

You can use conda and pip to manage many python libraries/packages. The official documentation can be found here.

Install R Kernel and R Packages (Optional)

To use R in an anaconda environment, all you need to do is to install the r-essentials bundle, which includes over 80 of the most popular scientific R packages.

```
$ conda install -c conda-forge r-essentials

$ conda install -c conda-forge r-irkernel
```

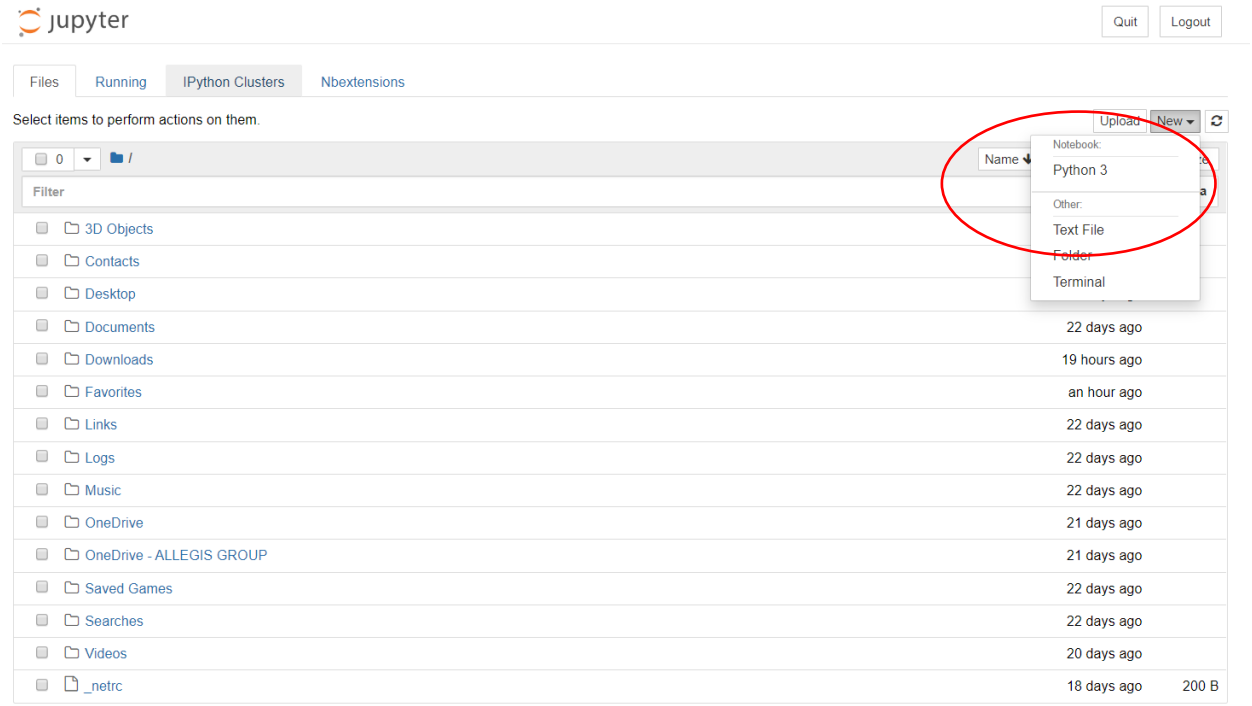
The R language packages are available to install with conda at http://repo.anaconda.com/pkgs/r/. You can install any of these R language packages into your current environment with the conda command conda install -c conda-forge package-name. For more information, you can check this link.

Start Jupyter Notebook

Type the command to start Jupyter Notebook.

```
$ jupyter notebook
```

If Jupyter notebook cannot be open in a web browser, you can copy the server url (e.g. http://localhost:8888/?token=0fa79120798f795452ebc143ce11d7f1f8ac73e5f860d551) and paste it on a open web browser, such as google chrome. After you can access the jupyter notebook web page, you should see the content like the following screen shot.

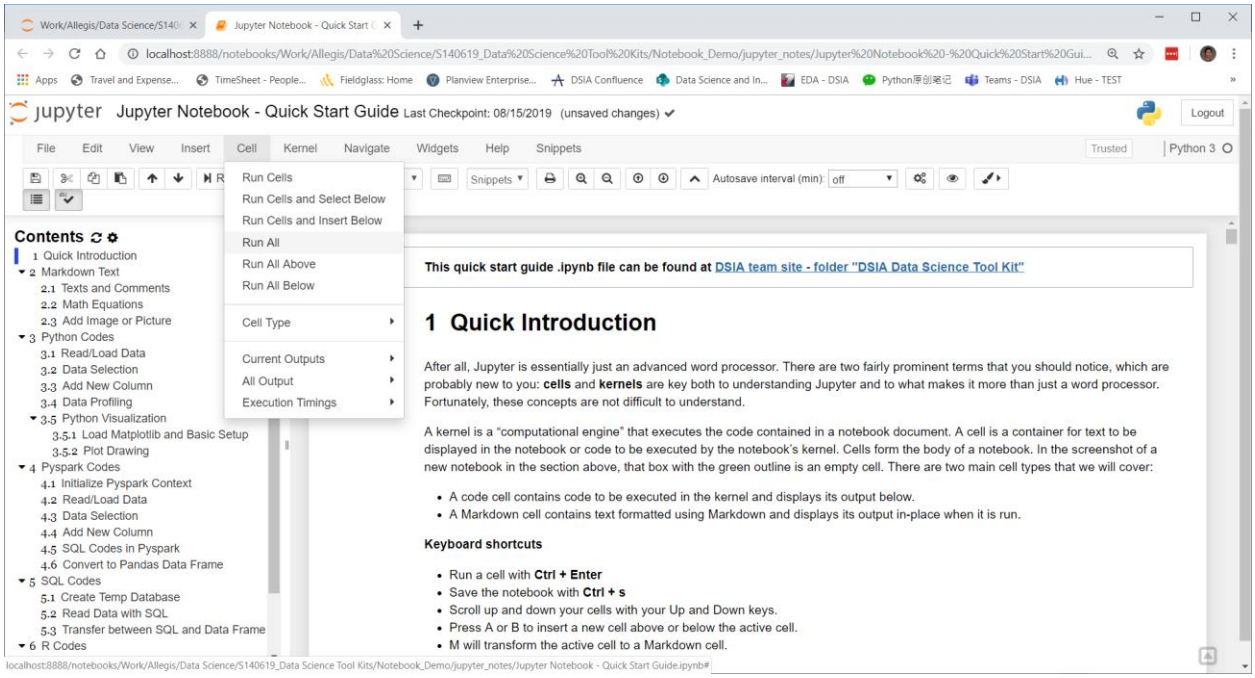


You can create a new Python 3 notebook to start coding your Python codes by clicking one of options on the right-hand side "New" dropdown button.

3. Run Quick Start Guide (Optional)

After all the steps above have been properly performed, you can run a jupyter notebook, called Jupyter Notebook – Quick Start Guide, to test and explore what you can do with this data science environment.

The jupyter notebook .ipynb file of the quick start guide can be downloaded from the Team site here. After it is downloaded, open jupyter notebook, navigate to the folder you have downloaded the .ipynb file, and then click it to open. When this jupyter notebook is open in a separate tab, you can run all the contents/cells by clicking **Cell → Runn All** menu item (see the following screen shot).



Please notify Data Science team or manager if you see any errors when running this jupyter notebook for trouble shooting. Otherwise, you can play this jupyter notebook a little bit to learn how you are able to perform data science and data engineering work with jupyter notebook and the environment. Enjoy! 😊

4.Appendix – Install Windows Subsystem for Linux (WSL)

The Windows Subsystem for Linux lets developers run a GNU/Linux environment -- including most command-line tools, utilities, and applications -- directly on Windows, unmodified, without the overhead of a virtual machine.

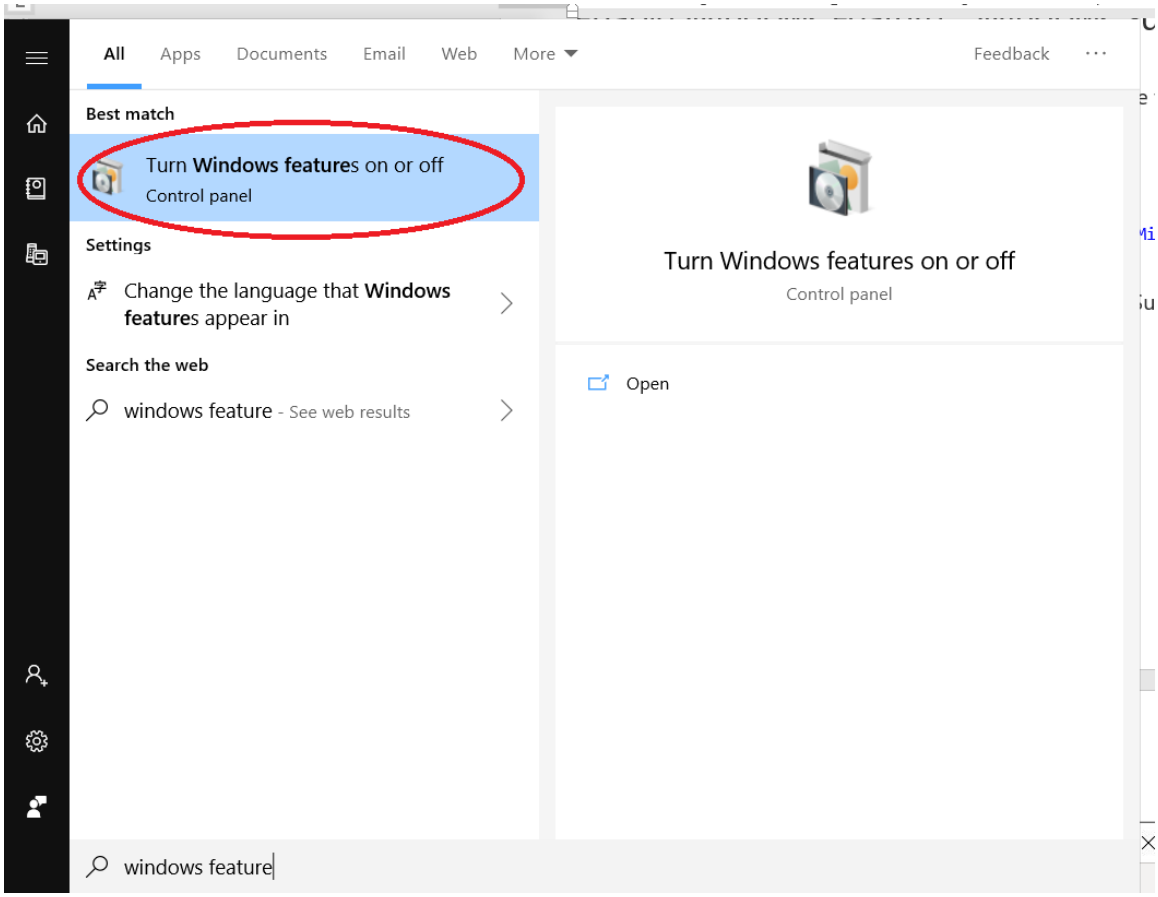
Enable Windows Feature - Windows Subsystem for Linux

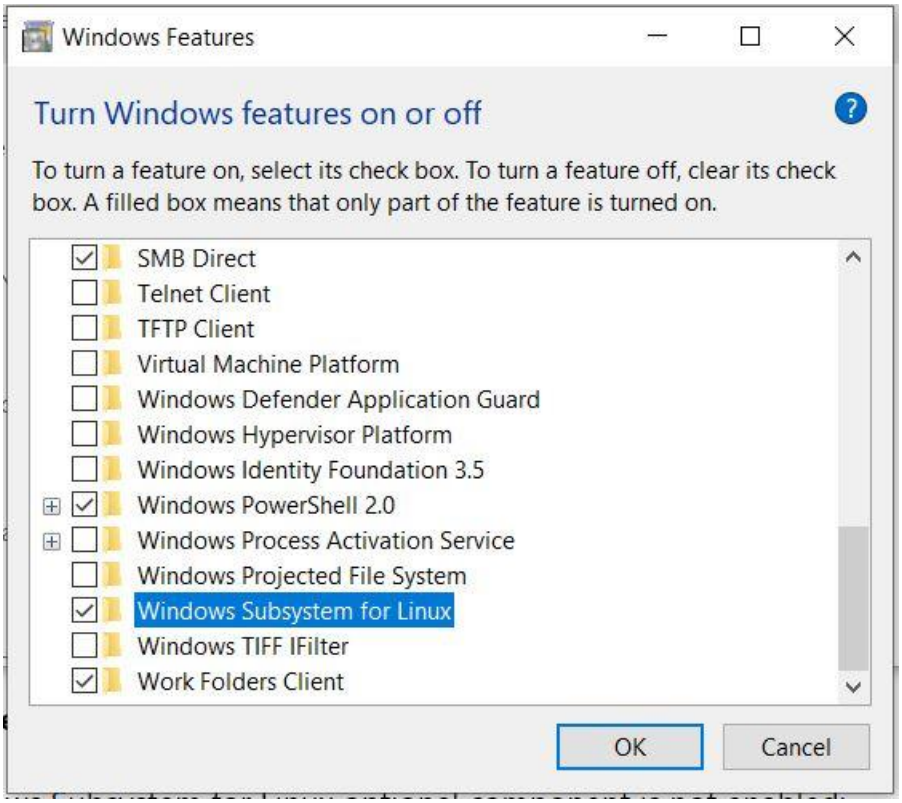
Before installing any Linux distros for WSL, you must ensure that the "Windows Subsystem for Linux" optional feature is enabled:

1. Open PowerShell as Administrator and run:

```
Enable-WindowsOptionalFeature -Online -FeatureName Microsoft-Windows-Subsystem-Linux
```

Or you can turn Windows features on to enable Windows Subsystem for Linux by launching the “Turn Windows features on or off” like following screen shots:





2. Restart your computer when prompted.

Download and Install Windows Subsystem for Linux

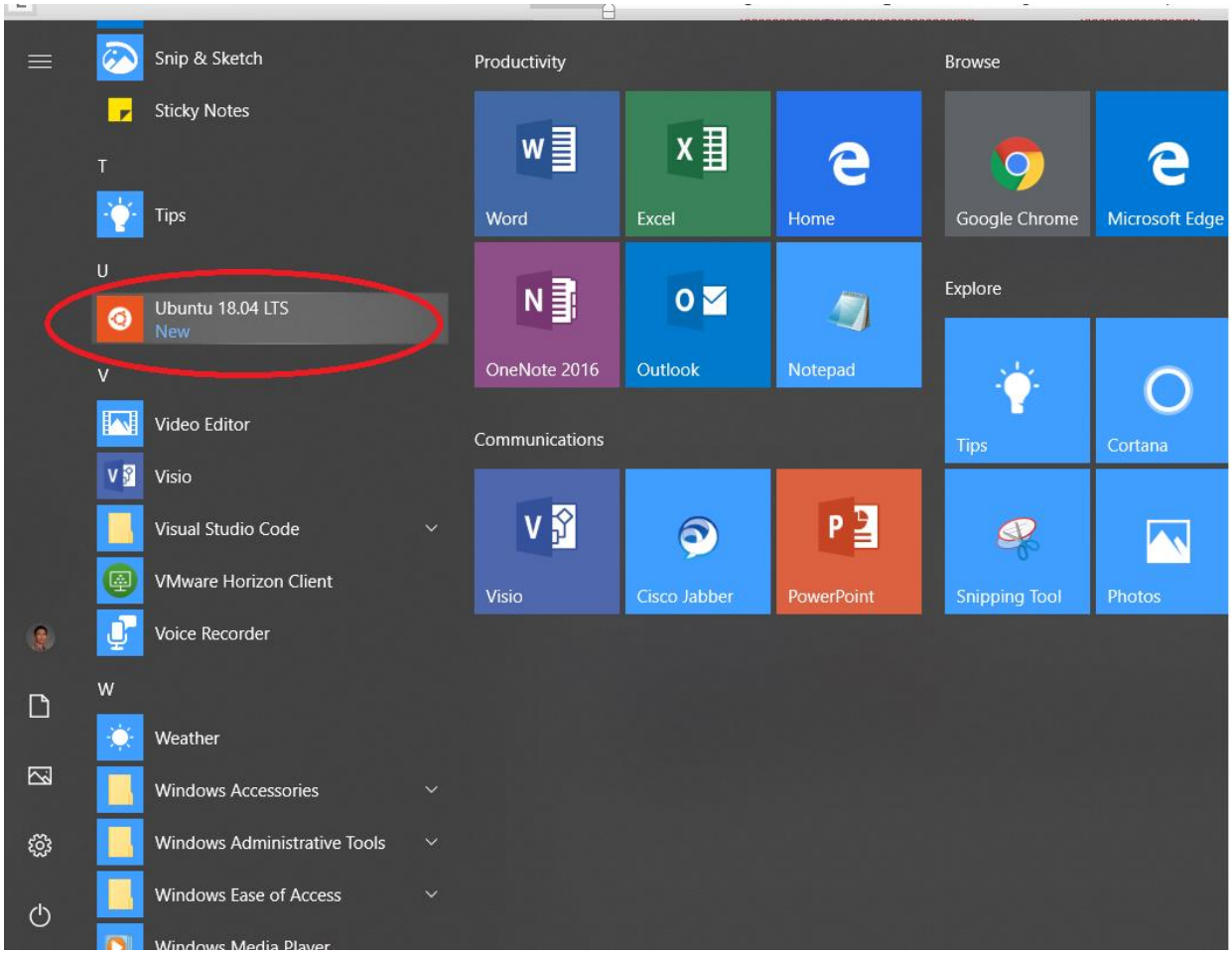
After your computer restarted, you need to manually download and install Ubuntu 18.04 distro. You can use the link here to download the Ubuntu 18.04 WSL or use PowerShell command:

```
Invoke-WebRequest -Uri https://aka.ms/wsl-ubuntu-1804 -OutFile Ubuntu.appx -UseBasicParsing
```

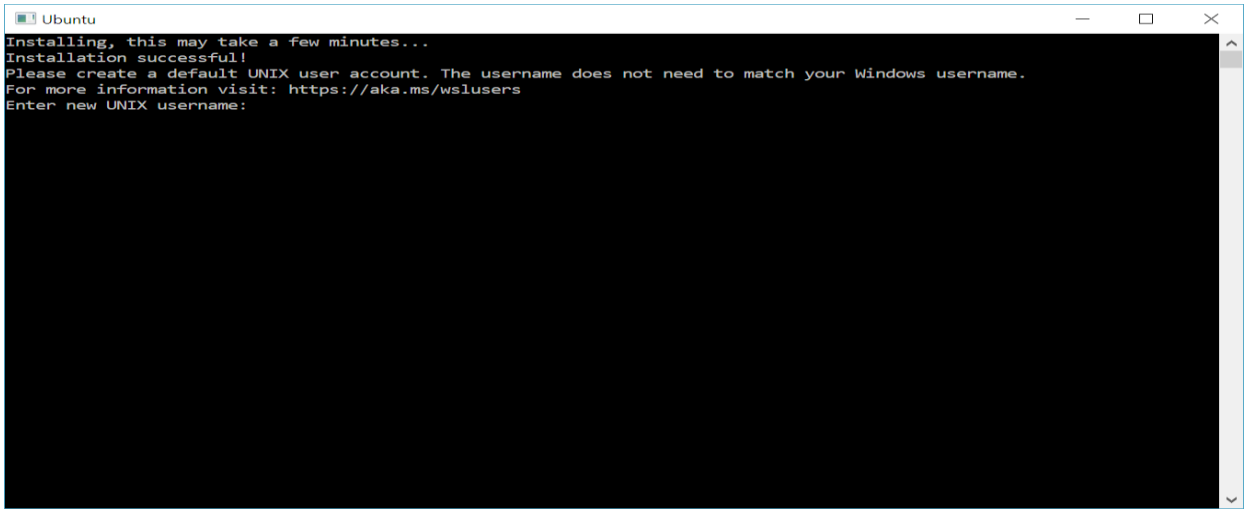
After the WSL distro was downloaded, you can use PowerShell command to install it if you're using Windows 10:

```
Add-AppxPackage .\Ubuntu.appx
```

Once your distro is installed you need to initialize your new distro. To complete the initialization of your newly installed distro, launch a new instance. You can do this by launching the distro from the Start menu:



The first time a newly installed distro runs, a Console window will open, and you'll be asked to wait for a minute or two for the installation to complete. Once installation is complete, you will be prompted to create a new user account (and its password).



This user account is for the normal non-admin user that you'll be logged-in as by default when launching a distro. You can choose any username and password you wish - they have no bearing on your Windows username.

When you open a new distro instance, you won't be prompted for your password, but **if you elevate a process using** sudo**, you will need to enter your password**, so make sure you choose a password you can easily remember! See the User Support page for more info.

Update & upgrade your distro's packages

Most distros ship with an empty/minimal package catalog. We strongly recommend regularly updating your package catalog, and upgrading your installed packages using your distro's preferred package manager. On Ubuntu, you use apt:

```
$ sudo apt update && sudo apt upgrade
```

Windows does not automatically update or upgrade your Linux distro(s). You're done! Enjoy using your new Linux distro on WSL! To learn more about WSL, review the other WSL docs, or the WSL learning resources page.