FTDS // INTRODUCTION TO DATA SCIENCE & TOOLBOX

Hacktiv8 DS Curriculum Team

Phase 0
Day 1 AM
2021

Phase 0

2021

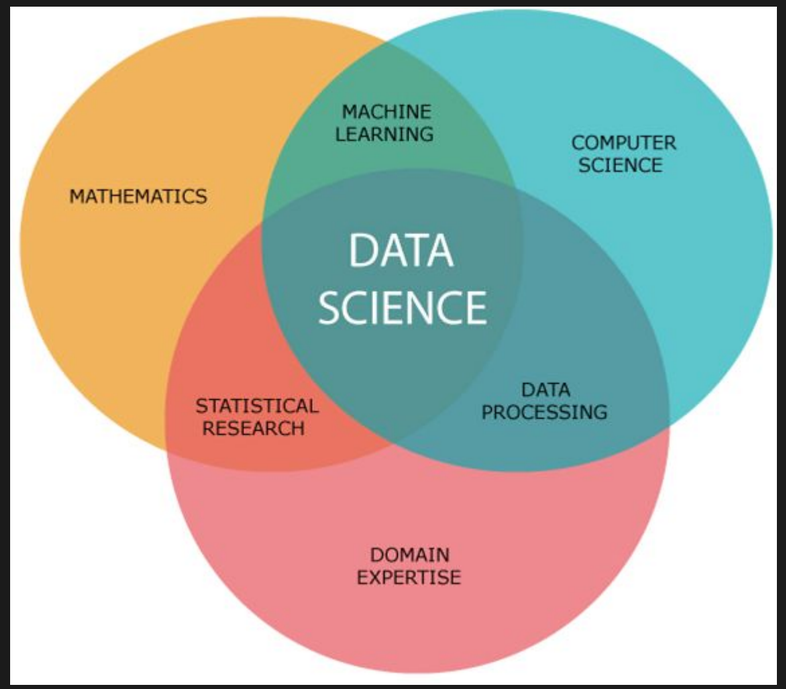
Contents

WEEK 1

Introduction: Data Science Toolbox

What is data science?

Data Science is about extraction, preparation, analysis, visualization, and maintenance of information. It is **a cross-disciplinary field** which uses scientific methods and processes to draw insights from data.



WEEK 1

Introduction: Data Science Toolbox

What is data science?

In short, we can say that data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

What can be done with data?

- Data can help detect anomalous events, such as fraudulent purchases.
- Data can predict future events, such as estimating population size.
- Make better decisions.
- Innovate products, services, and processes
- Improve quality, eliminate costs, and build trust.

Types of Data Science Job

The main job roles are given below:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect
6. Data Administrator
7. Business Intelligence

WEEK 1

Introduction: Data Science Toolbox

Skill-Sets

Data Analyst	Data Engineer	Data Scientist
Data Warehousing	Data Warehousing & ETL	Statistical & Analytical skills
Adobe & Google Analytics	Advanced programming knowledge	Data Mining
Programming knowledge	Hadoop-based Analytics	Machine Learning & Deep learning principles
Scripting & Statistical skills	In-depth knowledge of SQL/ database	In-depth programming knowledge (SAS/R/ Python coding)
Reporting & data visualization	Data architecture & pipelining	Hadoop-based analytics
SQL/ database knowledge	Machine learning concept knowledge	Data optimization
Spread-Sheet knowledge	Scripting, reporting & data visualization	Decision making and soft skills

WEEK 1

Introduction: Data Science Toolbox

Road Map

Mathematics

- Vector
- Matrix
- Linear algebra
- Eigen Vector & Eigen Value
- analytics geometry
- Calculus
- Probability
- Statistics
- Regression
- Classification
- Dimensionality Reduction

Programming

- Python:
  - Python Data types
  - Conditional
  - Looping
  - Function
  - Pandas
  - Matplotlib/seaborn
  - etc
- Database:
  - SQL
  - No SQL
- Other
  - Linux
  - GIt
  - Web Scraping

Machine Learning

- Introduction:
  - How model works
  - Basic EDA
  - First model ML
  - Train test split
  - Model Validation
  - Scikit-learn
- Intermediate:
  - Feature Engineering
  - Feature Selection
  - Ensemble Learning
  - Cross Validation
  - Pipelines
  - End to End ML

WEEK 1

Introduction: Data Science Toolbox

Road Map

Deep Learning

- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- Tensorflow
- Pytorch
- Stochastics Gradient Descent
- Dropout
- Batch Normalization
- Optimizer

Deployment

- Flask
- Streamlit
- Django
- Microsoft Azure
- Amazon Web Services
- Google Cloud Platform

Other Points

- Domain Knowledge
- Communication Skill
- Reinforcement Learning
- Tableau
- Power BI
- Case Studies:
  - TIme series (stock price prediction)
  - Recomendation System
  - NLP (sentiment analysis)
  - Fraud Detection
  - Computer Vision

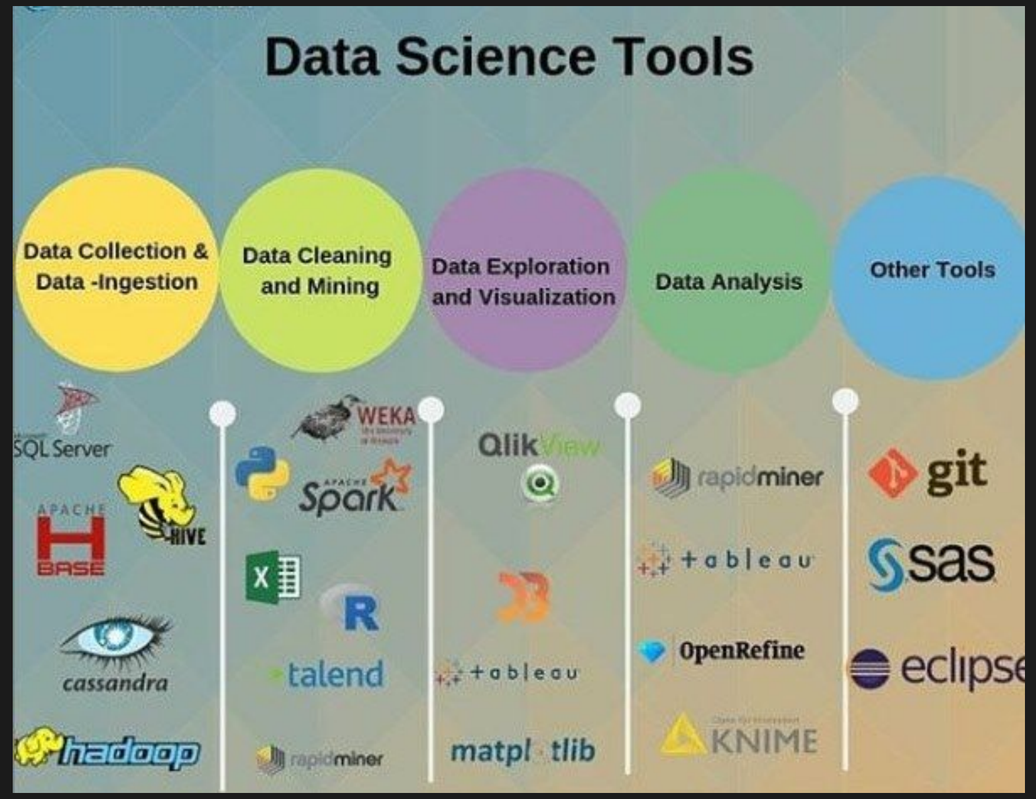
WEEK 1

Introduction: Data Science Toolbox

Tools for Data Science

Knowing how to use Data Science tools can help you build a bright and promising career in Data Science. Here are some of the top Data Science tools available in the market:

- Python
- TensorFlow
- Hadoop
- Microsoft SQL Server
- PowerBI
- MongoDB
- Tableau
- Knime
- Excel, etc.

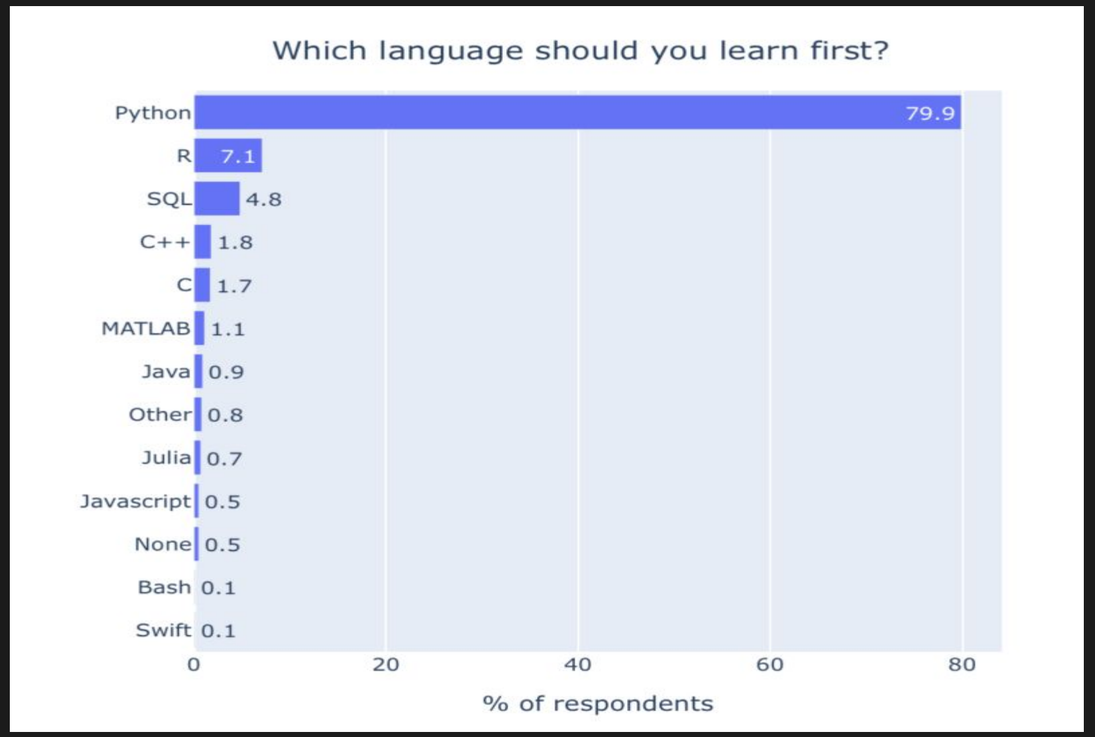


WEEK 1

Introduction: Data Science Toolbox

Tools for Data Science

The language you choose to learn will depend on the things you need to accomplish and the problems you need to solve. It will also depend on what company you work for, what role you have, and the age of your existing application.



Source: 2020 Kaggle Data Science and Machine Learning Survey

Python Vs R

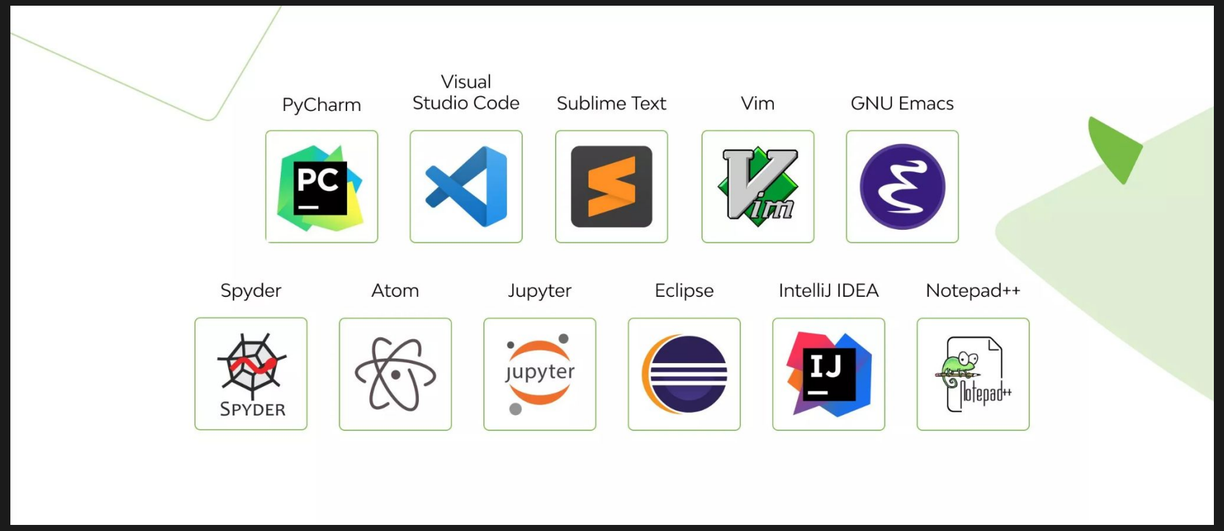
Parameter	R	Python
Objective	Data Analysis and Statistical Modeling	Data Science, Web Development, Embedded Systems
Workability	Consists of many easy to use packages	Can easily perform matrix computation as well as optimization
Integration	Locally Run Programs	Programs integrated with web-app for easy deployment
Database Handling Capacity	Poses problem for handling large dataset	Can handle large data easily without any fault
IDE	Rstudio, R GUI	Spyder, IPython, Jupyter Notebook
Essential Packages and library	ggplot2, tidyverse, caret	Numpy, pandas, scipy, scikit-learn, TensorFlow

WEEK 1

Introduction: Data Science Toolbox

IDE

- An integrated development environment (IDE) is a software application that provides comprehensive facilities to computer programmers for software development.
- IDE and text editors are special development environments that programmers use to write code, test it, debug it, and upload it to GitHub — or any other Git hosting website.
- Here are the best IDEs and text editors:



Anaconda Basic

Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

Managing Conda:

- **Verify that conda is installed and running on your system by typing:**
  conda –version
- **Update conda to the current version. Type the following:**
  conda update conda
- **If a newer version of conda is available, type y to update:**
  Proceed ([y]/n)? y

Managing Environment

Conda allows you to create separate environments containing files, packages, and their dependencies that will not interact with other environments. You can create separate environments to keep your programs isolated from each other.

- **We will name the environment Hacktiv8 and install the package pandas. At the Anaconda Prompt or in your terminal window, type the following:**
  conda create --name Hacktiv8 pandas

- **Conda checks to see what additional packages ("dependencies") pandas will need, and asks if you want to proceed:**
  Proceed (y]/n)? y
  Type "y" and press Enter to proceed.

- **To use, or "activate" the new environment, type the following:**
  Windows: conda activate Hacktiv8
  macOS and Linux: conda activate Hacktiv8

Managing Environtment

- **To see a list of all your environments, type:**
  conda info –envs
- **A list of environments appears, similar to the following:**
  conda environments:

  base /home/username/Anaconda3
  snowflakes * /home/username/Anaconda3/envs/snowflakes
- **The active environment is the one with an asterisk (*).**
- **Change your current environment back to the default (base): conda activate**

WEEK 1

Introduction: Data Science Toolbox

Managing Python & Packages

If you want to use a different version of Python, for example Python 3.6, simply create a new environment and specify the version of Python that you want.

- Create a new environment named "fox" that contains Python 3.6:
  `conda create --name fox python=3.6`
  When conda asks if you want to proceed, type "y" and press Enter.

- if you want to check a package (eg ‘beautifulsoup4’) whether it is installed or not, you can use the command:
  `conda search beautifulsoup4`

- Install this package into the current environment:
  `conda install beautifulsoup4`

- Check to see if the newly installed program is in this environment:
  `conda list`

The Notebooks

- **JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.**

- **Google Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with:**

  a) **Zero configuration required**
  b) **Free access to GPUs**
  c) **Easy sharing Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier.**

WEEK 1

Python: Basic Syntax, Variables, Data Types

What is Python?

- **Python is a high-level, interpreted scripting language developed in the late 1980s by Guido van Rossum at the National Research Institute for Mathematics and Computer Science in the Netherlands. The initial version was published at the alt.sources newsgroup in 1991.**

- **Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.**

Why Choose Python?

If you're going to write programs, there are literally dozens of commonly used languages to choose from. Why choose Python? Here are some of the features that make Python an appealing choice:

1. **Python is Popular**
2. **Python is Interpreted**
3. **Python is Free**
4. **Python is Simple**

Python Quickstart

Example of running python in terminal:

1. **Create a new file containing python syntax, for example `print('hello world')`.**
2. **Then save it with `file_name.py`.**
3. **Run it in terminal with the command:**
   `python file_name.py`

We can also use the following methods:

```
C:\Users\Dell>python
Python 3.9.4 (tags/v3.9.4:1f2e308, Apr  6 2021, 13:40:21) [MSC v.1928 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print('irfansyah')
irfansyah
>>> exit()

C:\Users\Dell>
```

External References

Colab Link — Visit Here