Co-funded by the
Erasmus+ Programme
of the European Union

DeCAIR

Data Science

Course Introduction

Prof. Gheith Abandah
أ.د. غيث علي عبندة

Developing Curricula for Artificial Intelligence and Robotics (DeCAIR)
618535-EPP-1-2020-1-JO-EPPKA2-CBHE-JP

Outline

- Instructor Information
- Video: What is Data Science?
- Data Science Hierarchy of Needs
- Data Science Skills
- Textbook and References
- Course Objectives and Outcomes
- Course Outline
- Policies and Grading
- Important Dates

Instructor Information

- **Instructor**: Prof. Gheith Abandah
- **Email**: abandah@ju.edu.jo
- **Office**: CPE 406
- **Home page**: http://www.abandah.com/gheith
- **MS Team**: Link
- **Office hours**: Sun – Thu: 13:00 – 14:00

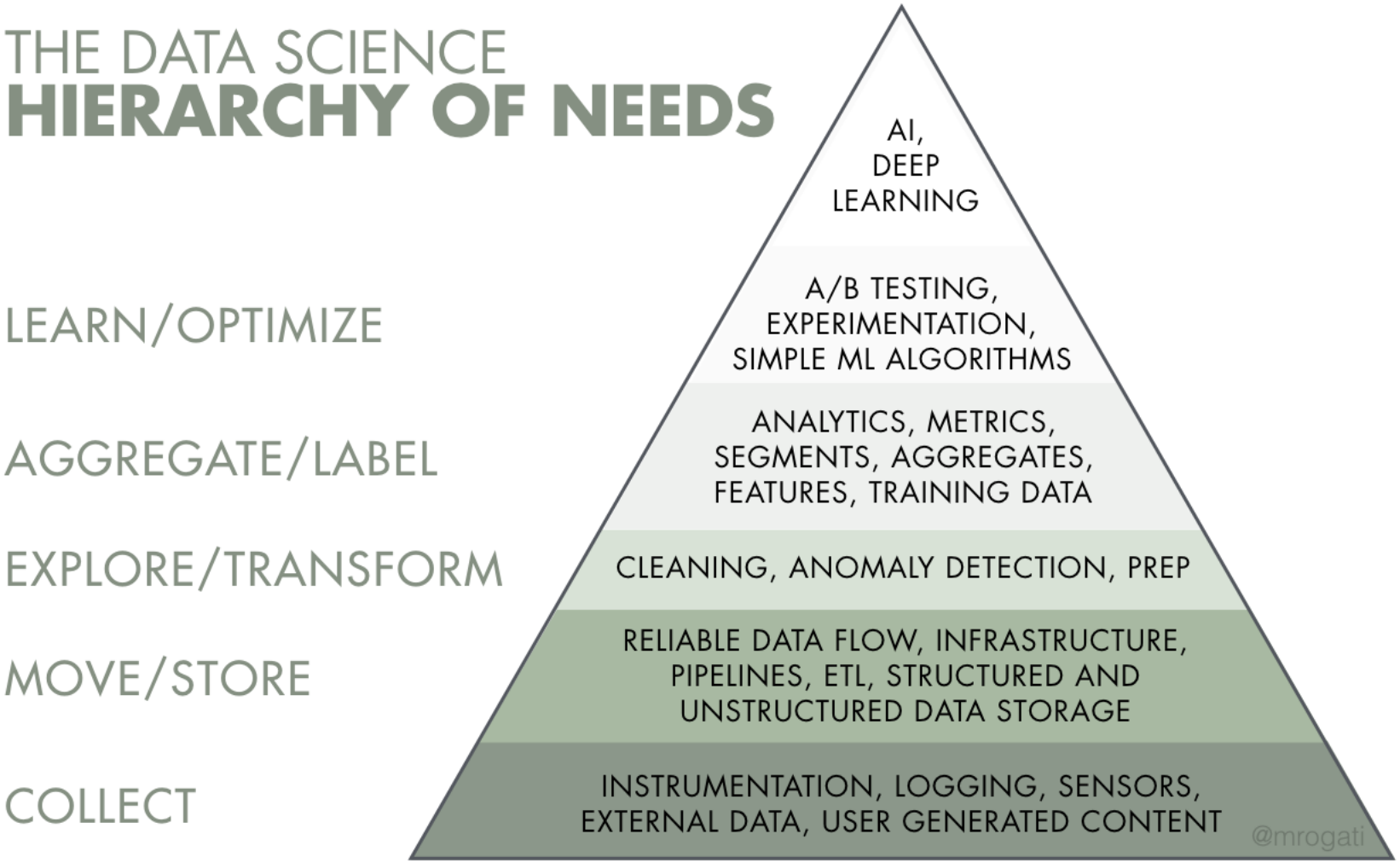
What is Data Science?

- YouTube Video from **Joma Tech**

What REALLY is Data Science? Told by a Data Scientist

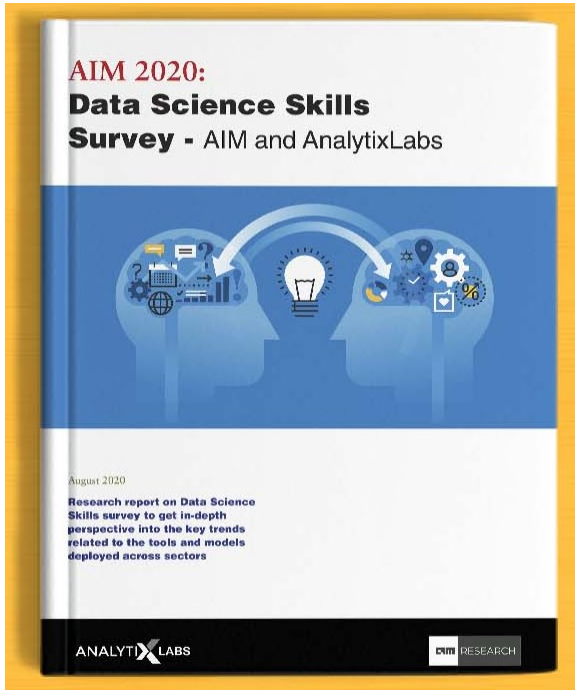
https://youtu.be/xC-c7E5PK0Y

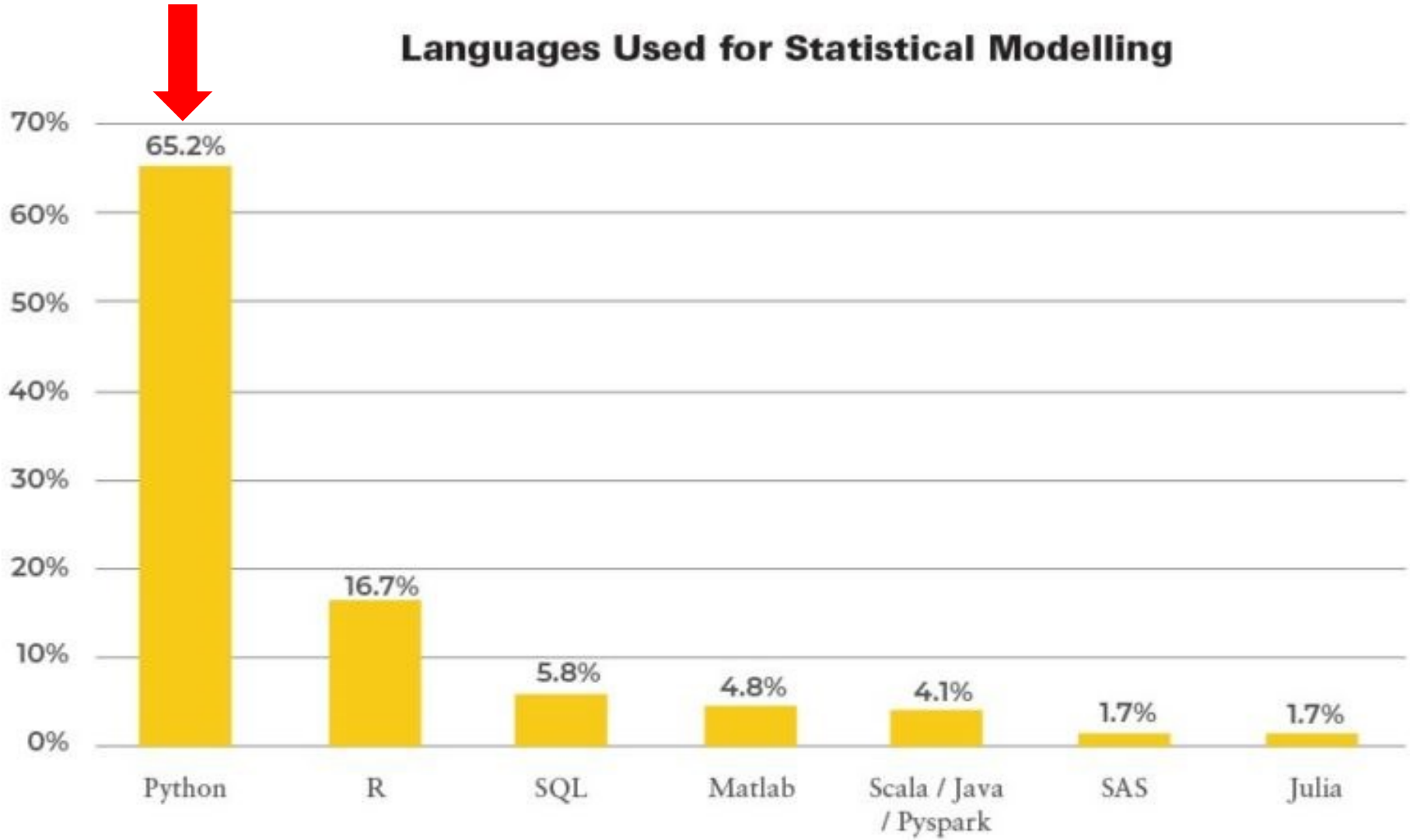
Data Science Hierarchy of Needs

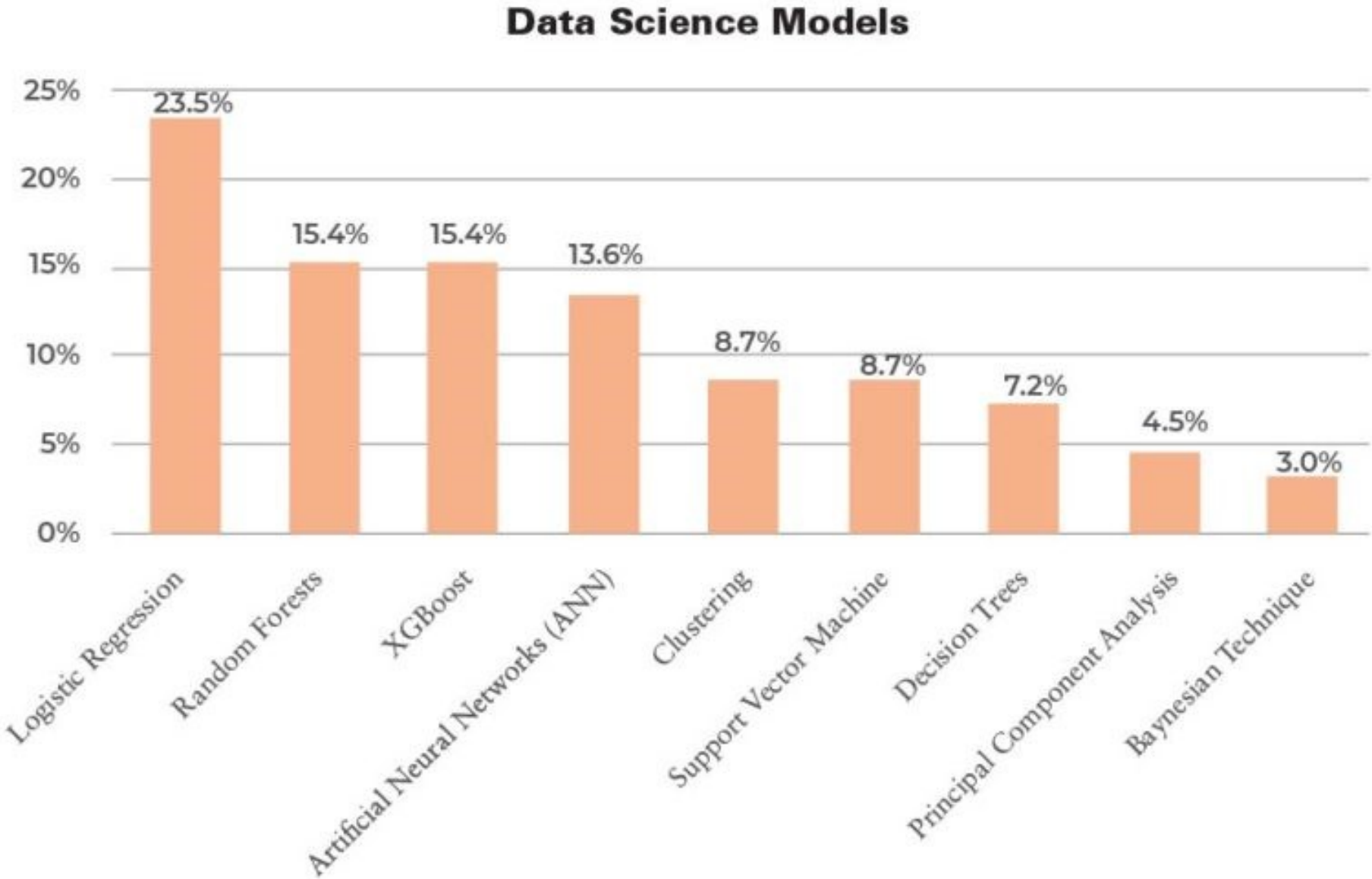


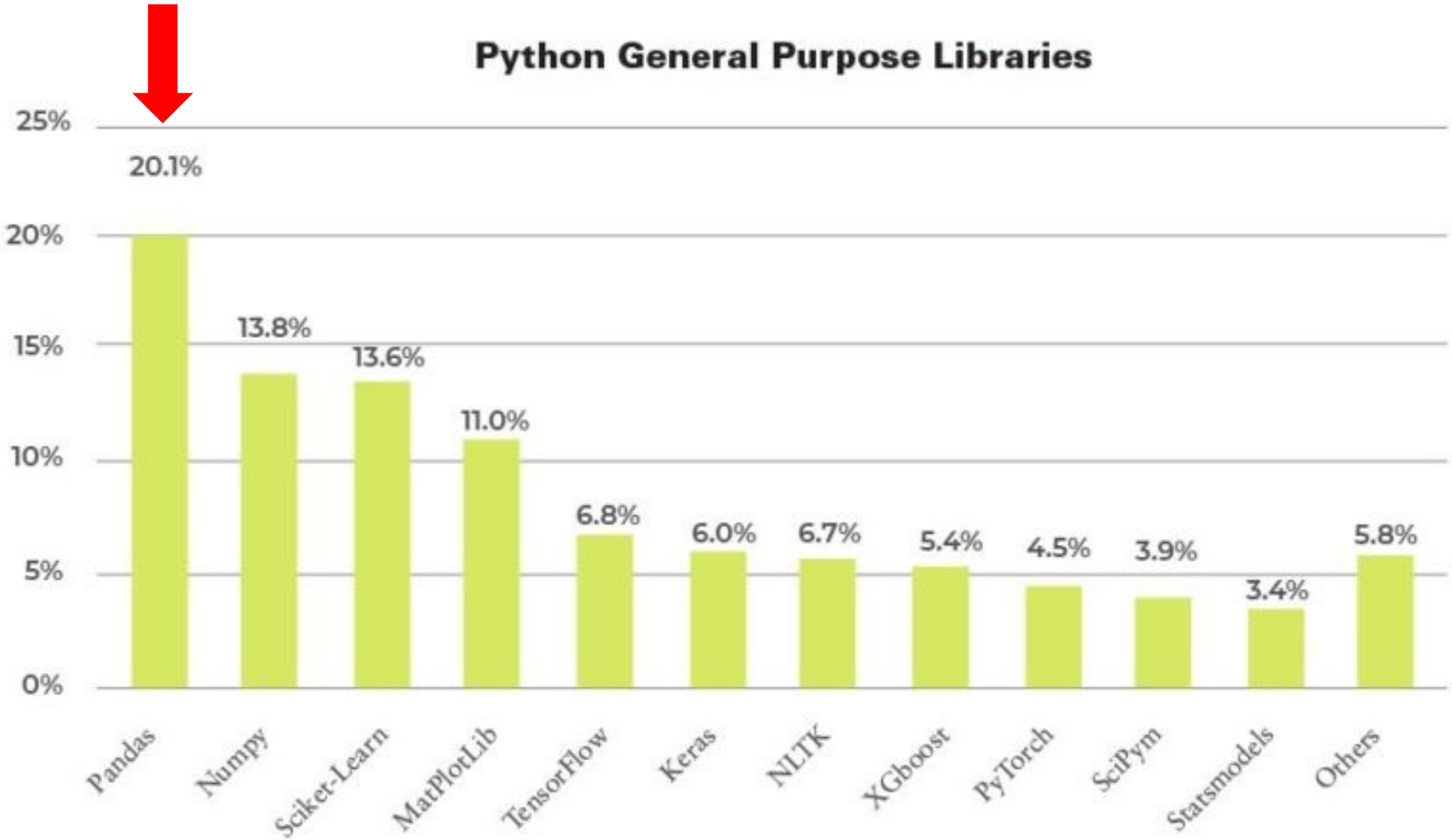
Data Science Skills

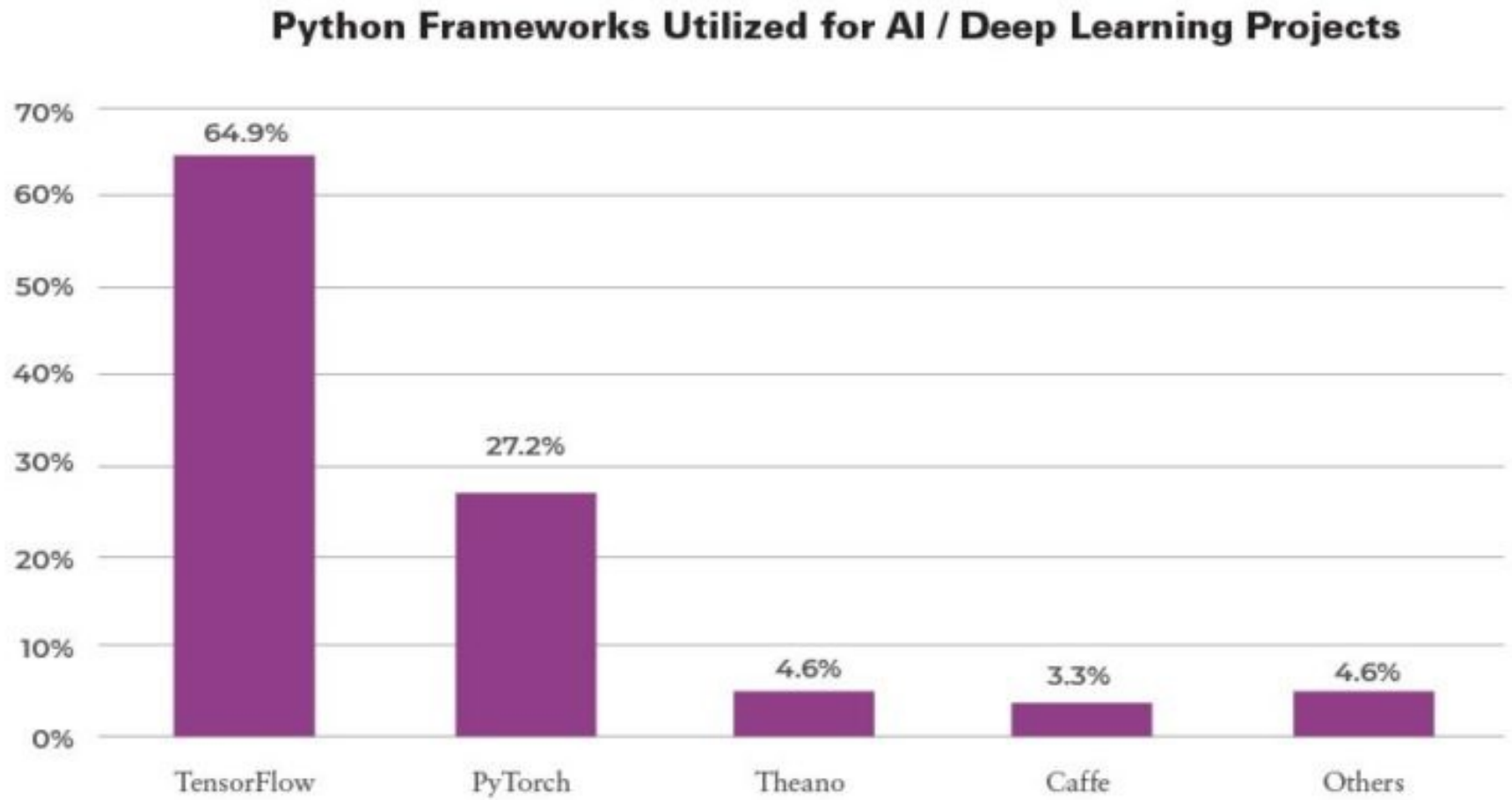
- **Data Science Skills Study 2020**
- https://analyticsindiamag.com/aim-2020-data-science-skills-survey-aim-and-analytix-labs/
- By AIM and AnalytixLabs
- Released on 17/8/2020

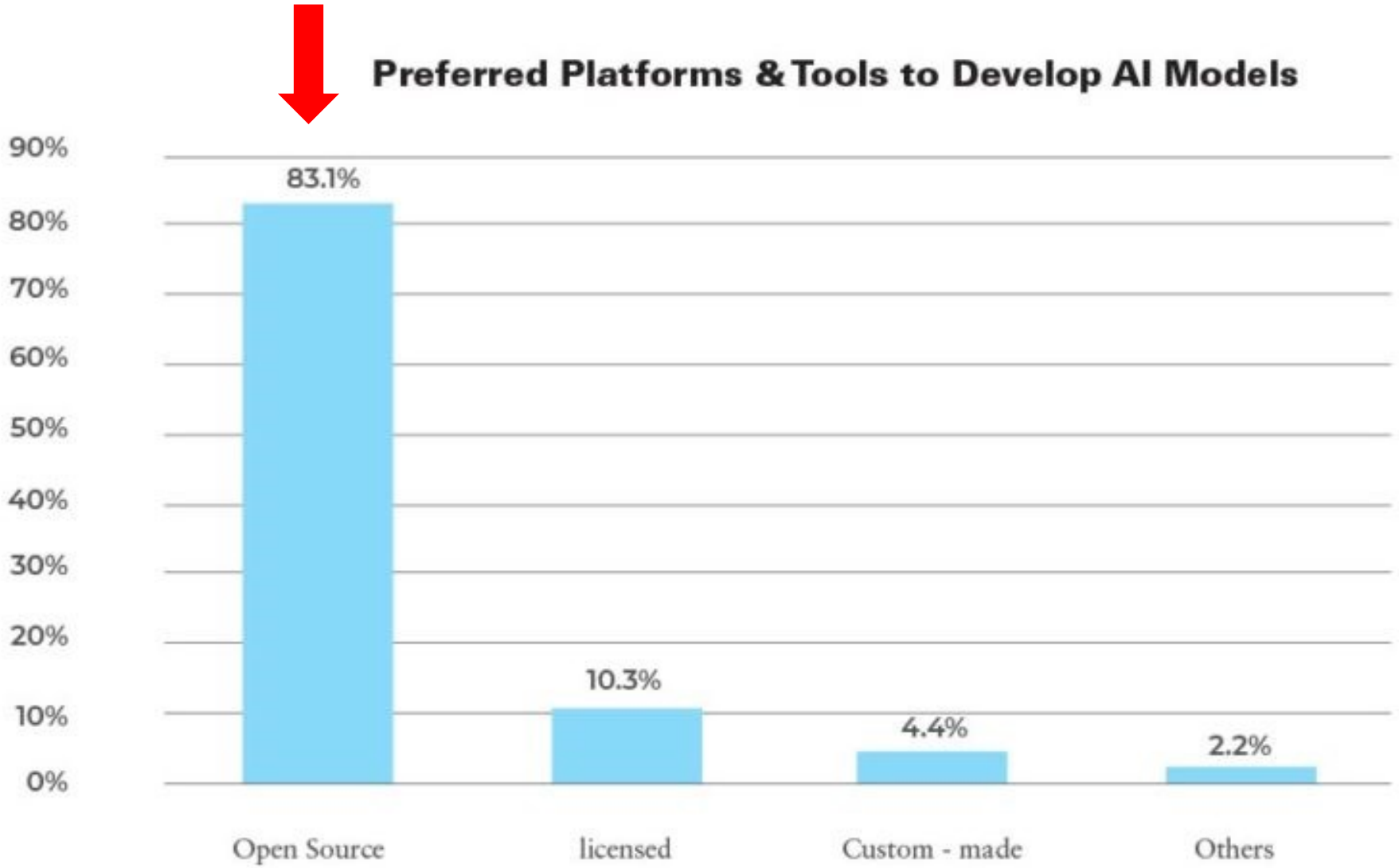


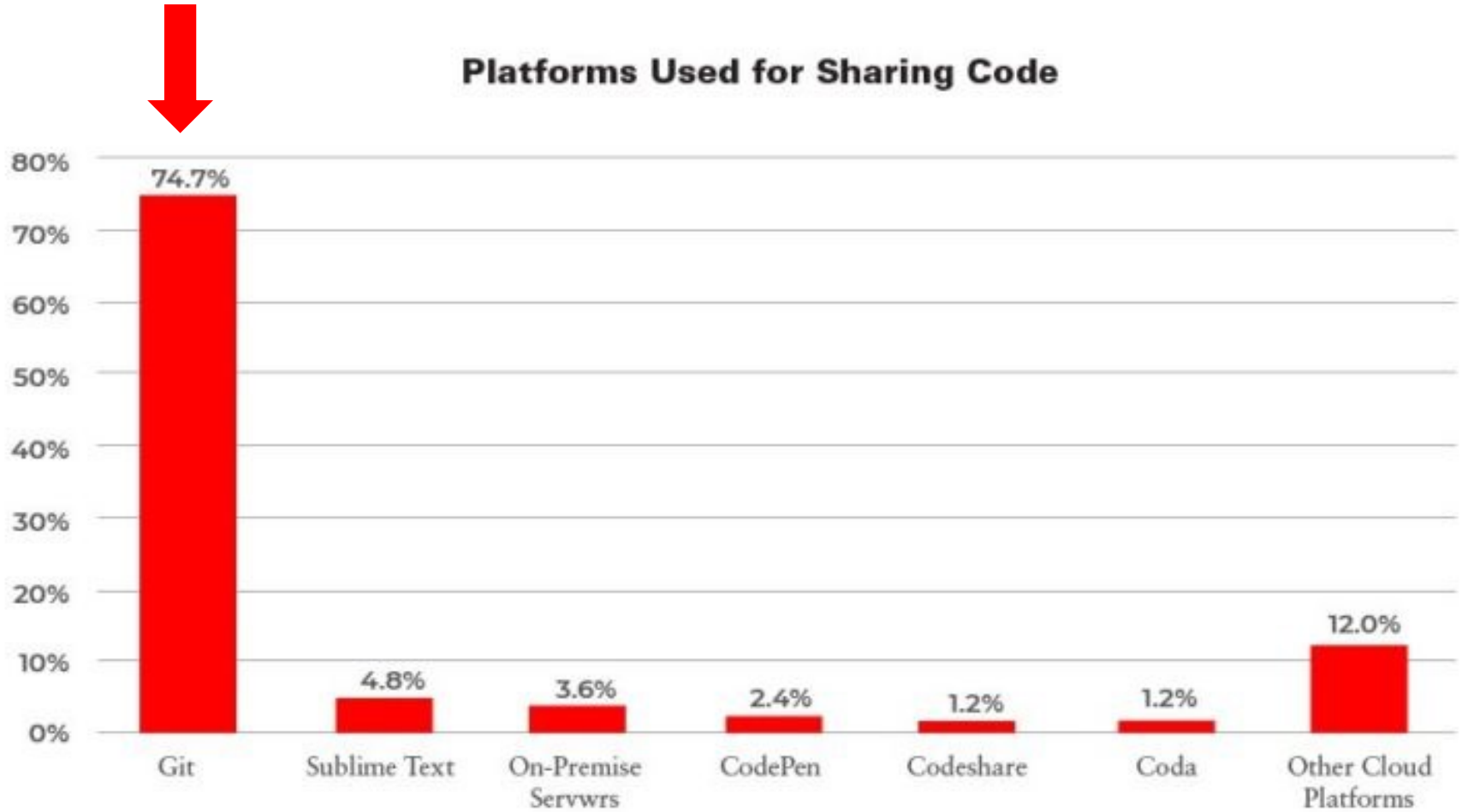

Languages Used for Statistical Modelling
70%
60%
50%
40%
30%
20%
10%
0%
65.2%
16.7%
5.8%
4.8%
4.1%
1.7%
1.7%
Python
R
SQL
Matlab
Scala / Java / Pyspark
SAS
Julia

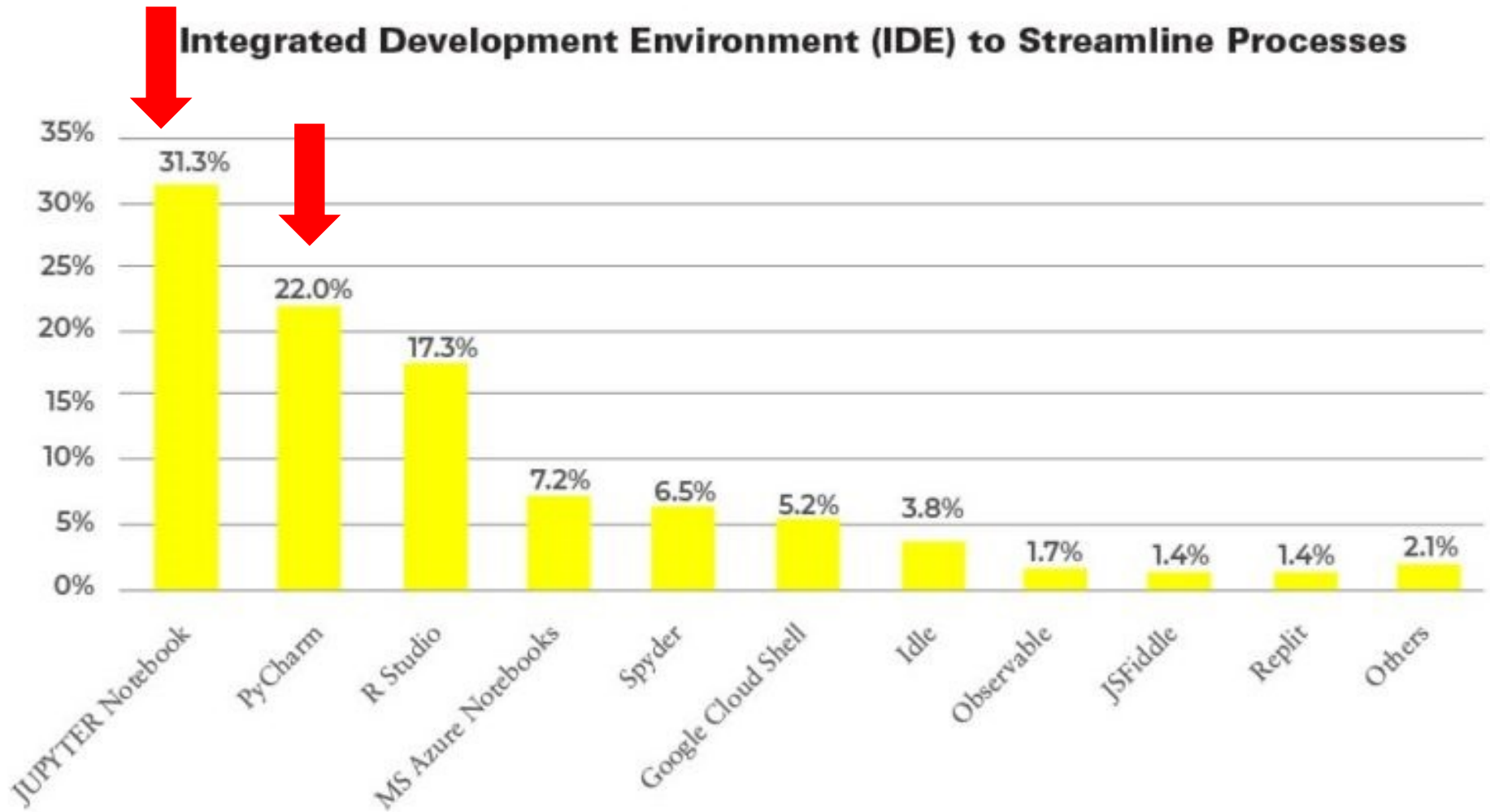

Data Science Models
25%
20%
15%
10%
5%
0%
23.5%
15.4%
15.4%
13.6%
8.7%
8.7%
7.2%
4.5%
3.0%
Logistic Regression
Random Forests
XGBoost
Artificial Neural Networks (ANN)
Clustering
Support Vector Machine
Decision Trees
Principal Component Analysis
Baynesian Technique

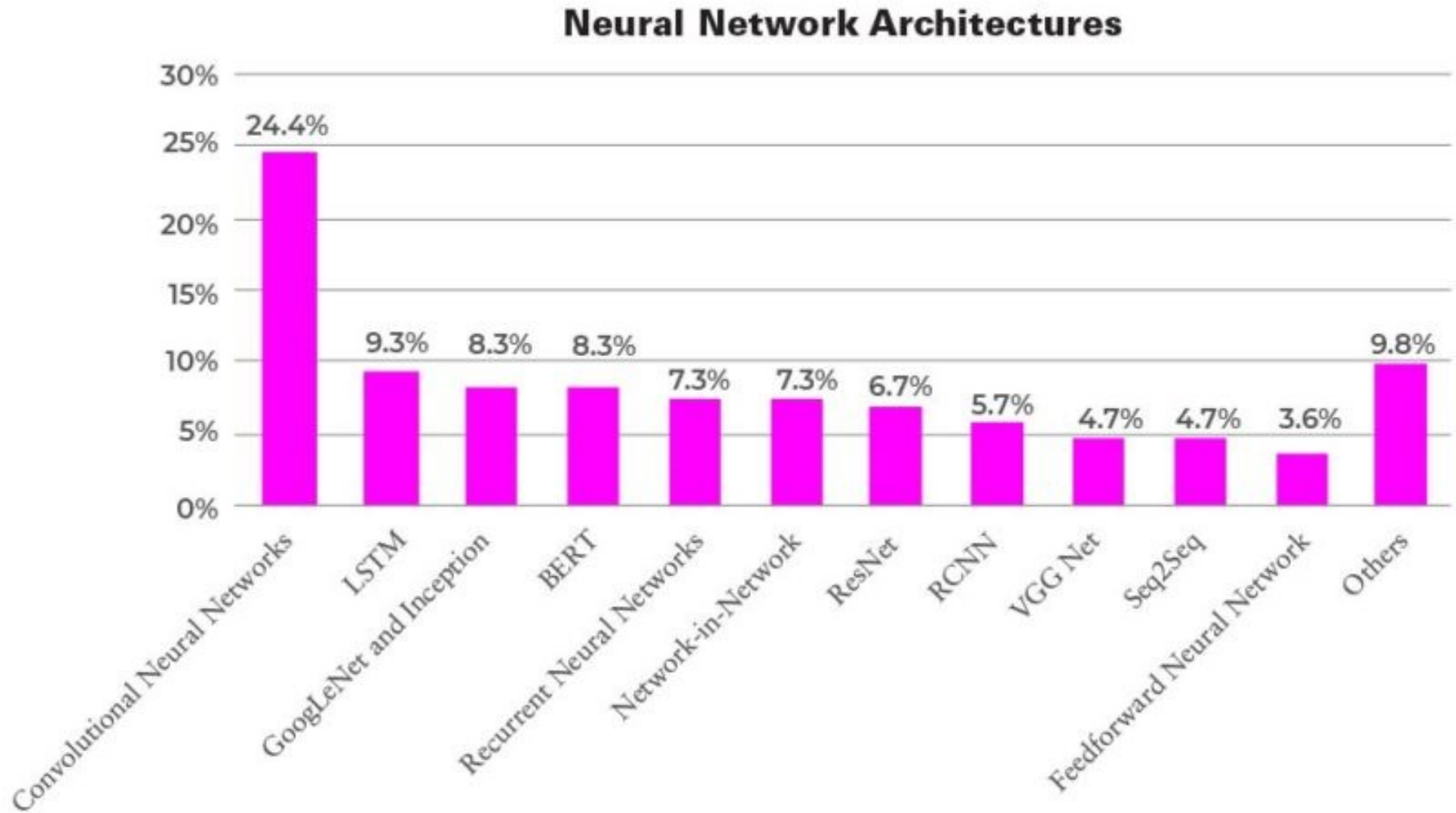

Python General Purpose Libraries
25%
20%
15%
10%
5%
0%
20.1%
13.8%
13.6%
11.0%
6.8%
6.0%
6.7%
5.4%
4.5%
3.9%
3.4%
5.8%
Pandas
Numpy
Sciker-Learn
MatPlotLib
TensorFlow
Keras
NLTK
XGboost
PyTorch
SciPym
Statsmodels
Others

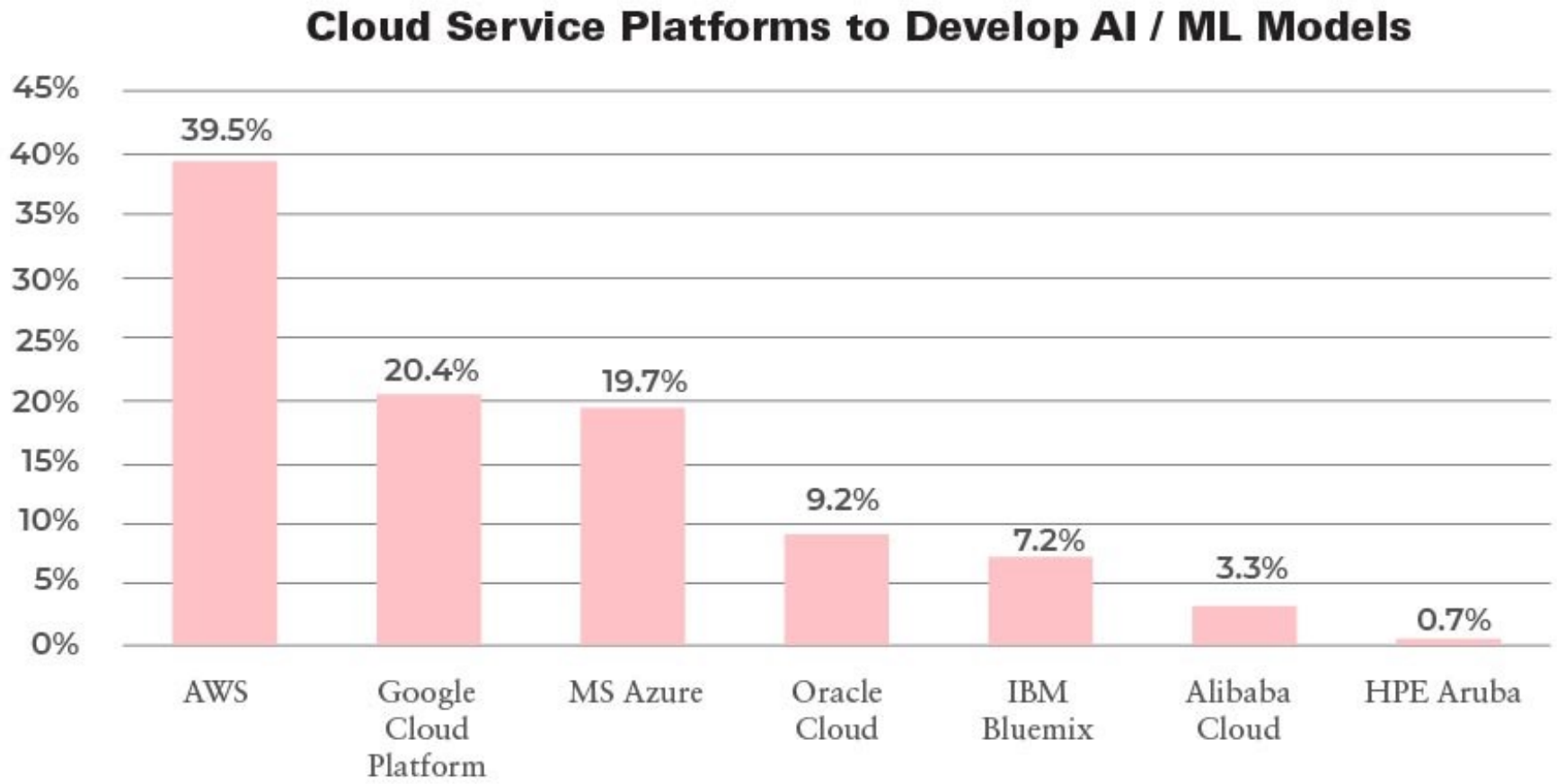

Python Frameworks Utilized for AI / Deep Learning Projects
70%
60%
50%
40%
30%
20%
10%
0%
64.9%
27.2%
4.6%
3.3%
4.6%
TensorFlow
PyTorch
Theano
Caffe
Others

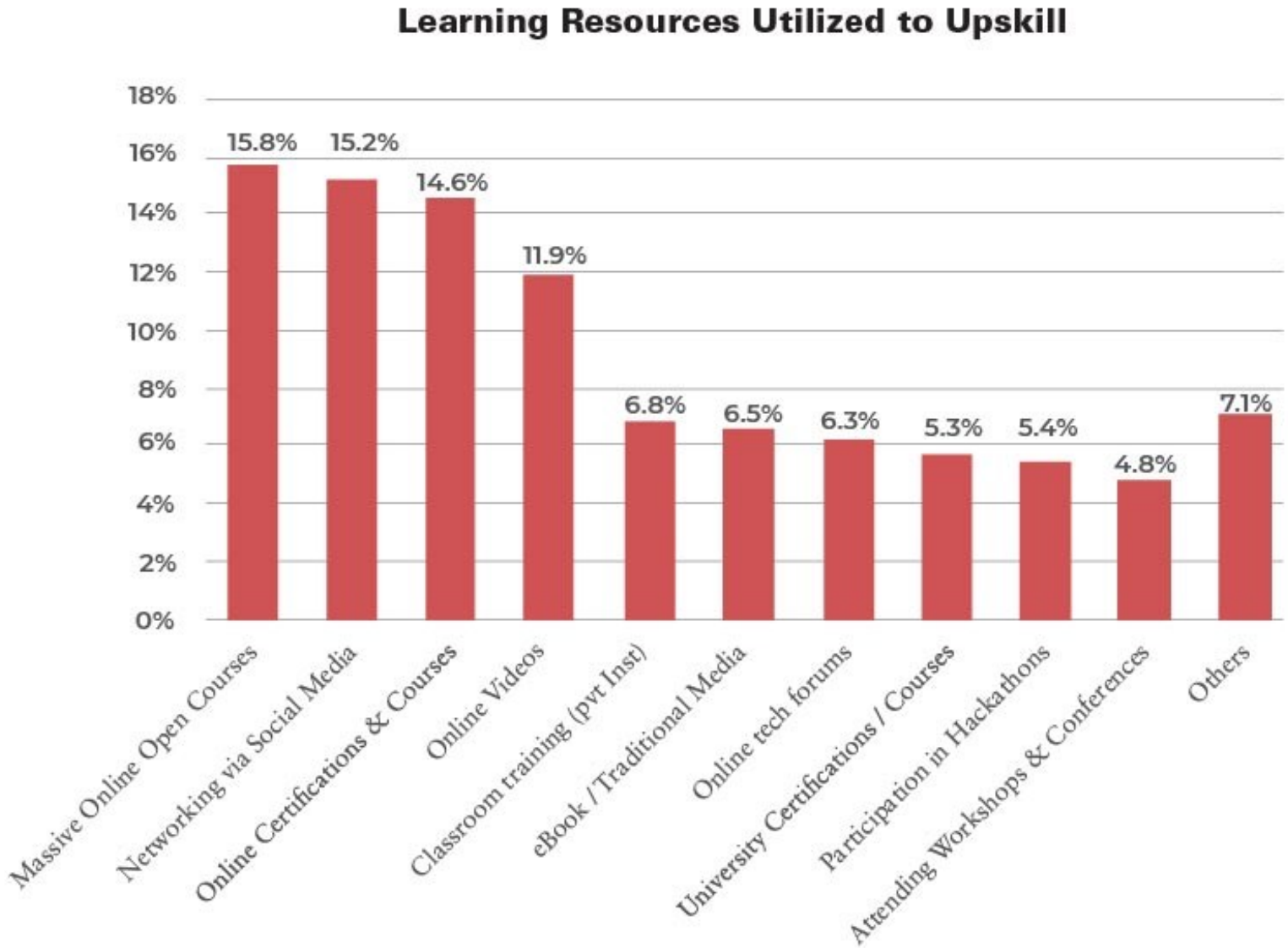

Preferred Platforms & Tools to Develop AI Models
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
83.1%
10.3%
4.4%
2.2%
Open Source
licensed
Custom - made
Others


Platforms Used for Sharing Code
80%
70%
60%
50%
40%
30%
20%
10%
0%
74.7%
4.8%
3.6%
2.4%
1.2%
1.2%
12.0%
Git
Sublime Text
On-Premise Servwrs
CodePen
Codeshare
Coda
Other Cloud Platforms


Integrated Development Environment (IDE) to Streamline Processes
35%
30%
25%
20%
15%
10%
5%
0%
31.3%
22.0%
17.3%
7.2%
6.5%
5.2%
3.8%
1.7%
1.4%
1.4%
2.1%
JUPYTER Notebook
PyCharm
R Studio
MS Azure Notebooks
Spyder
Google Cloud Shell
Idle
Observable
JSFiddle
Replit
Others

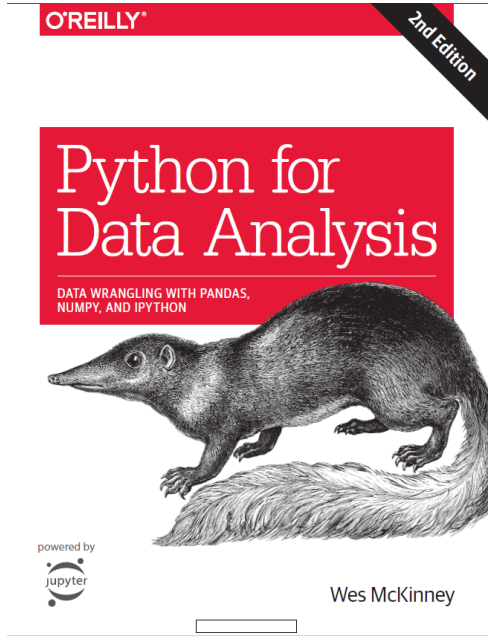

Neural Network Architectures
30%
25%
20%
15%
10%
5%
0%
24.4%
9.3%
8.3%
8.3%
7.3%
7.3%
6.7%
5.7%
4.7%
4.7%
3.6%
9.8%
Convolutional Neural Networks
LSTM
GoogLeNet and Inception
BERT
Recurrent Neural Networks
Network-in-Network
ResNet
RCNN
VGG Net
Seq2Seq
Feedforward Neural Network
Others

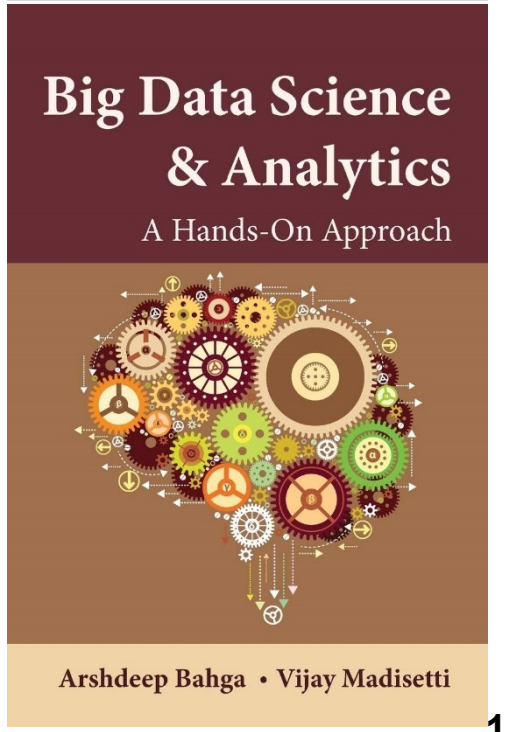

Cloud Service Platforms to Develop AI / ML Models
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
39.5%
20.4%
19.7%
9.2%
7.2%
3.3%
0.7%
AWS
Google Cloud Platform
MS Azure
Oracle Cloud
IBM Bluemix
Alibaba Cloud
HPE Aruba


Learning Resources Utilized to Upskill
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
15.8%
15.2%
14.6%
11.9%
6.8%
6.5%
6.3%
5.3%
5.4%
4.8%
7.1%
Massive Online Open Courses
Networking via Social Media
Online Certifications & Courses
Online Videos
Classroom training (pvt Inst)
eBook / Traditional Media
Online tech forums
University Certifications / Courses
Participation in Hackathons
Attending Workshops & Conferences
Others

Textbooks

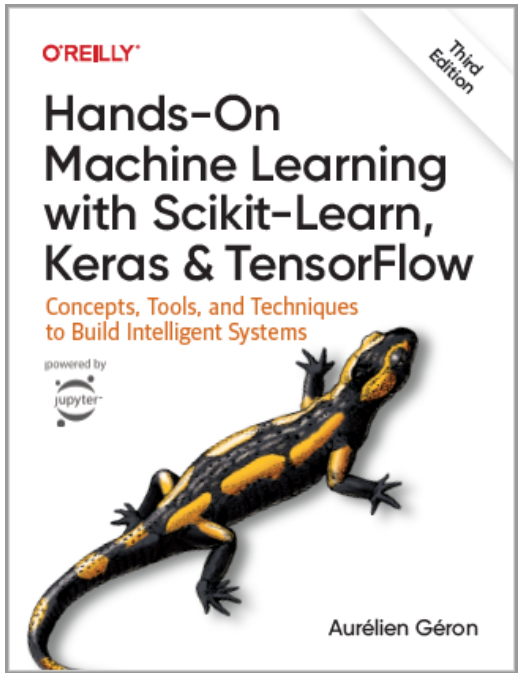
1. Wes McKinney, **Python for Data Analysis**: Data Wrangling with Pandas, NumPy, and Ipython, O'Reilly Media, 2nd Edition, 2018.
2. Arshdeep Bahga and Vijay Madisetti, **Big Data Analytics**: A Hands-On Approach, 2019.
3. Course web page: http://www.abandah.com/gheith/?page_id=3022





References

1. Jake VanderPlas, **A Whirlwind Tour of Python**, O'Reilly Media, 2016.
2. Joel Gurs, **Data Science from Scratch**, O'Reilly Media, 2015.
3. Aurélien Géron, **Hands-On Machine Learning** with Scikit-Learn, Keras and TensorFlow: Concepts: Tools, and Techniques to Build Intelligent Systems, 3rd Edition, O'Reilly Media, Oct 2022.



Course Objectives

- Introduce students to the practical techniques used in data analytics including loading, cleaning, preparation, wrangling, visualization, and analysis.
- Introduce students to the basic concepts and techniques in big data.

Course Outcomes

- Use Python and its specialized libraries to gain insight from data and solve problems.
- Know the main concepts and techniques used in handling big data and performing data analytics.

Course Outline

Week	Topic	ILO	Resources
1	Course Introduction	1	3
2+3	Pandas Data Structures, Essential Functionality & Descriptive Statistics	1	1
4+6	Data Loading, Storage and File Formats	1	1
6+7	Data Cleaning and Preparation	1	1
8	Data Wrangling: Join, Combine and Reshape	1	1
9+10	Plotting and Visualization with Matplotlib and Seaborn	1	1
11	Data Aggregation and Group Operations	1	1
12	Time Series	1	1
13	Introduction to Big Data	2	2
14	Big Data Architectures and Patterns	2	2
15	MapReduce Patterns	2	2

Policies

- Attendance is required
- Makeup exams need acceptable absence cause
- Late penalty is 25%
- All submitted work must be yours
- Cheating will not be tolerated
- Open-book exams
- Join the Microsoft Team at: Link
- Check department announcements at:
  - http://www.facebook.com/pages/Computer-Engineering-Department/369639656466107

Grading

Assessment tool	Mark	Topic(s)	Time
Quizzes and Homework	20%	Programming aspects	W2-W14
Midterm exam	30%	First 8 weeks	W8
Final exam	50%	All material	W16
Total	100%		

Important Dates

Mon 27 Feb 2023	First Lecture
16-30 Apr 2023	Midterm Exam Period
Sun 4 Jun 2023	Last Date to Withdraw
Mon 5 Jun 2023	Last Lecture
8-20 Jun 2023	Final Exam Period