



中国科学院深圳先进技术研究院
SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Course Title:

Data Science (数据科学)

(Semester: Fall 2021)

Dr. Oluwarotimi W. SAMUEL

**Research Center for Neural Engineering
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences**

Contact: (Email: samuel@siat.ac.cn & timitex92@gmail.com)

Phone: +86-15814491870

(2021.09.30)



□ Outline for today's lecture

- ✓ Importance of Programming skills in DS
- ✓ Popular DS Programming Languages
- ✓ Pros and Cons of DS Programming Languages



- ❑ **Objective:** This lecture will focus on discussing the relevance of programming and top Data Science programming languages, as well as their strength and weakness
- ❑ **Expectation:** At the end of this lecture, students are expected to understand the advantages and disadvantages of the most commonly adopted Data Science programming languages.



- ❑ DS is among the top popular technologies of the 21th Century.
- ❑ Hence, for people who want to pursue a carrier in DS, there is need to be proficient in the field. And one of the core skills required is programming.



Programming: A Core DS Skill

- ❑ **Aspiring** Data Scientist should be able to make the right decision about the type of **Programming Language** required for the job.



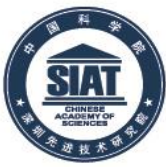


□ What is a Programming Language?

- ✓ A Programming Language is a formal language comprising of a *set of instructions* that produce various kinds of output.
- ✓ They are used to implement algorithms that run on computers and have multiple applications.



- ❑ There are a number of programming languages tailored towards Data Science tasks.
- ❑ Data Scientists should master at **least one** DS language, as it is essential for their job.
- ❑ We shall study some programming languages required to be a proficient Data Scientist.



❑ Top Data Science Programming Languages

- 1 Python
- 2 R
- 3 SAS
- 4 Julia
- 5 Scala
- 6 SQL

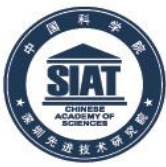




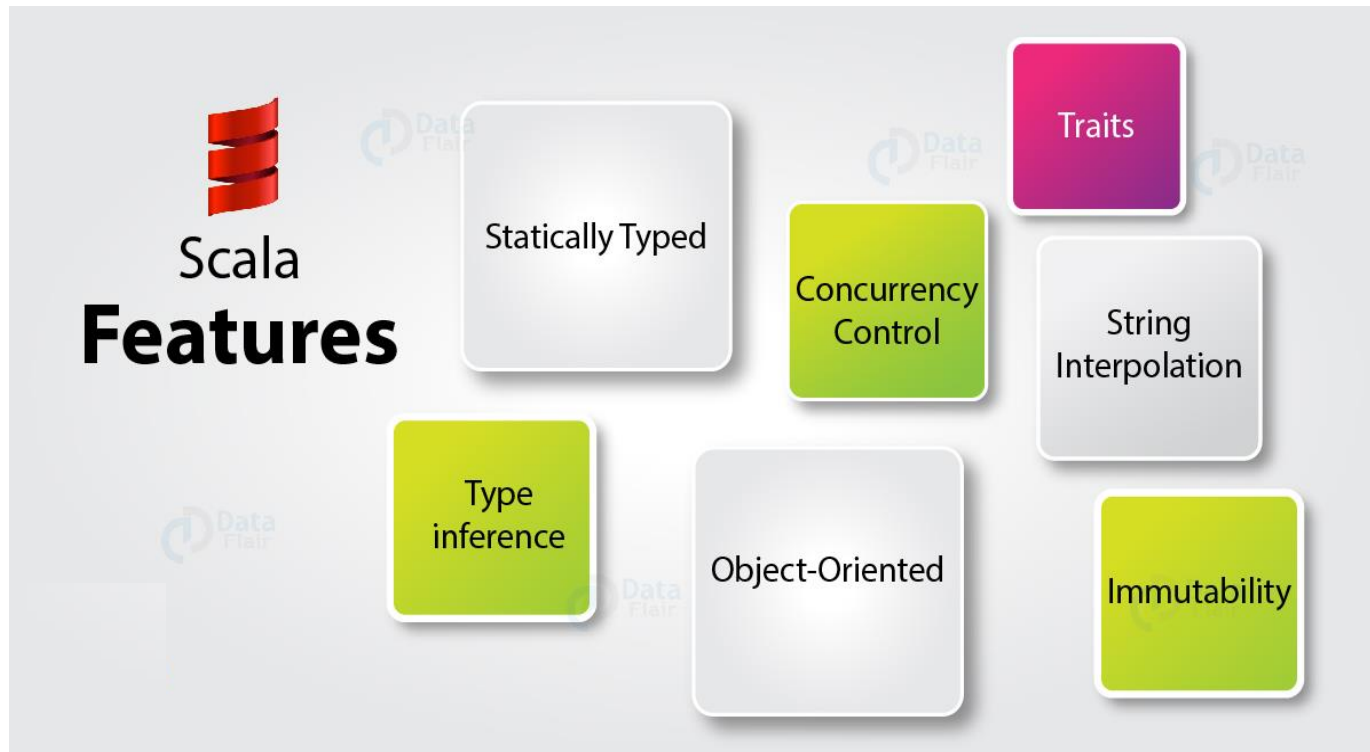
□ Scala Programming



- ✓ Scala is an extension of Java programming language operating on JVM
- ✓ It can be used in conjunction with Spark, a big data platform
- ✓ And this makes it ideal when dealing with large volumes of data
- ✓ Importantly, Scala has the ability to facilitate parallel processing on a large scale



□ Scala Programming



NOTE: If your preference as a Data Scientist is dealing with a large volume of data, then Scala + Spark is your best option.

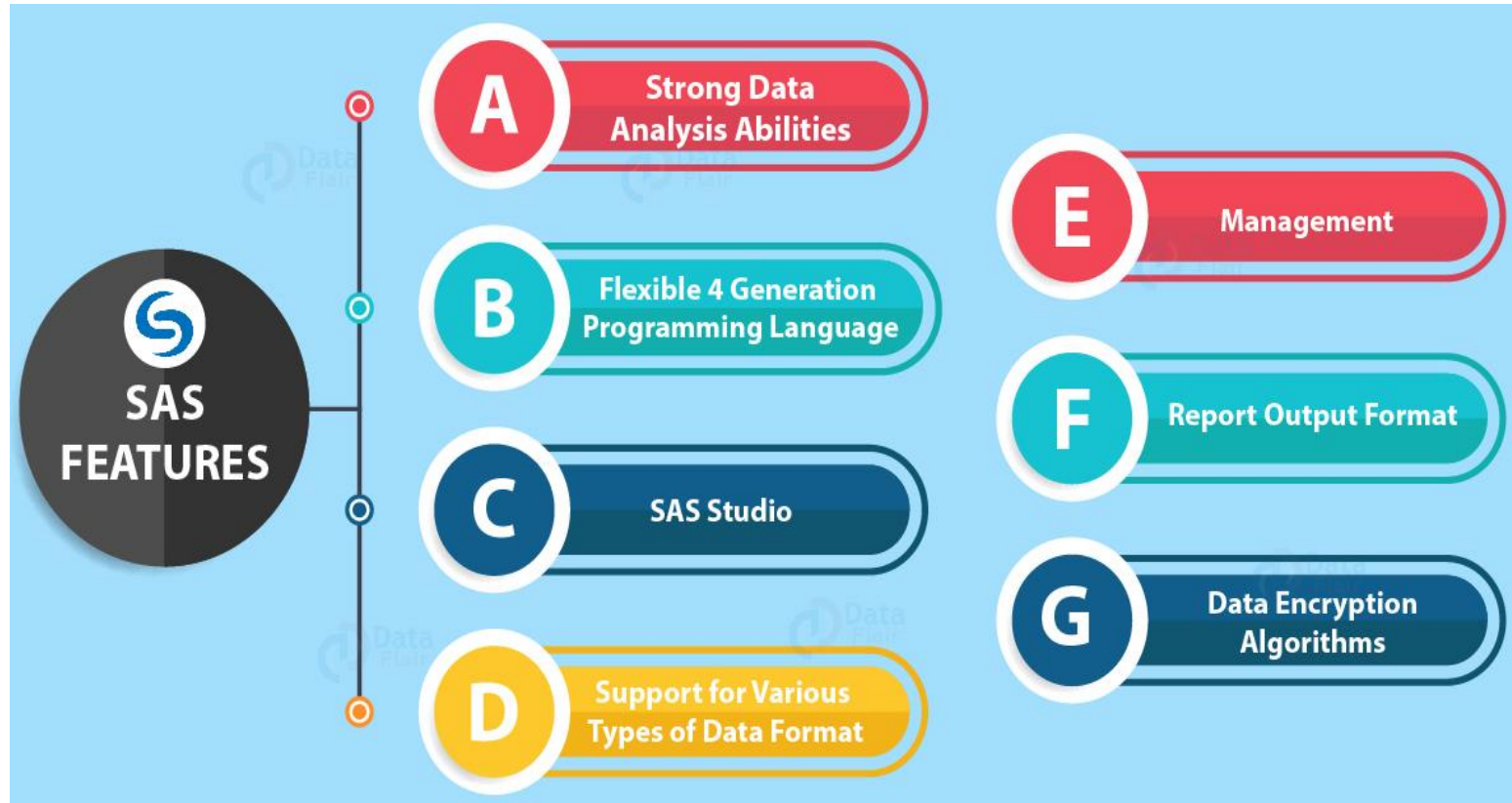


❑ SAS Programming



- ✓ SAS is built mainly for Statistical Analysis
- ✓ Unlike R, SAS is a non open-source programming language
- ✓ It supports advanced analytics, predictive modeling, and business intelligence
- ✓ It is highly reliable and has been **very well approved** by professionals and analysts

❑ SAS Programming



It offers a wide range of libraries and packages for statistical analysis and machine learning.



□ Julia Programming



- ✓ Julia is built based on C language, and specifically for *scientific computing*
- ✓ It is an ideal language for areas requiring complex mathematical operations.
- ✓ As a Data Scientist, you may work on problems requiring complex mathematics.
- ✓ Julia is capable of solving such problems at a very high speed



□ Julia Programming

What Makes Julia Great?

- 1 When coded well, it is very fast
- 2 Clear concise code that can easily be changed
- 3 Great ability to mix loop based & matrix/vector operations

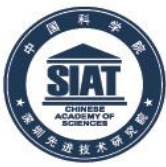
It has support for a wide variety of operators, allowing scientists to create domain specific equations that can be elegantly expressed as code.



□ R Programming



- ✓ It is an open source programming language developed in 1991 (by Ross Ihaka and Robert Gentleman)
- ✓ For statistically oriented DS tasks, R is the perfect language
- ✓ Hence, it is very popular among statisticians.



❑ R Programming



R-Programming

R-Programming

Programming features of



01

R is an interpreted language

02

Supports matrix arithmetic

03

Supports procedural programming with functions

04

Supports object oriented programming with generic functions



□ R Programming

- ✓ With over 10k packages in the open-source repository of CRAN, R caters for all statistical applications
- ✓ Can handle complex linear algebra
- ✓ Other than just statistical analysis, R can also be used for **building neural networks models**



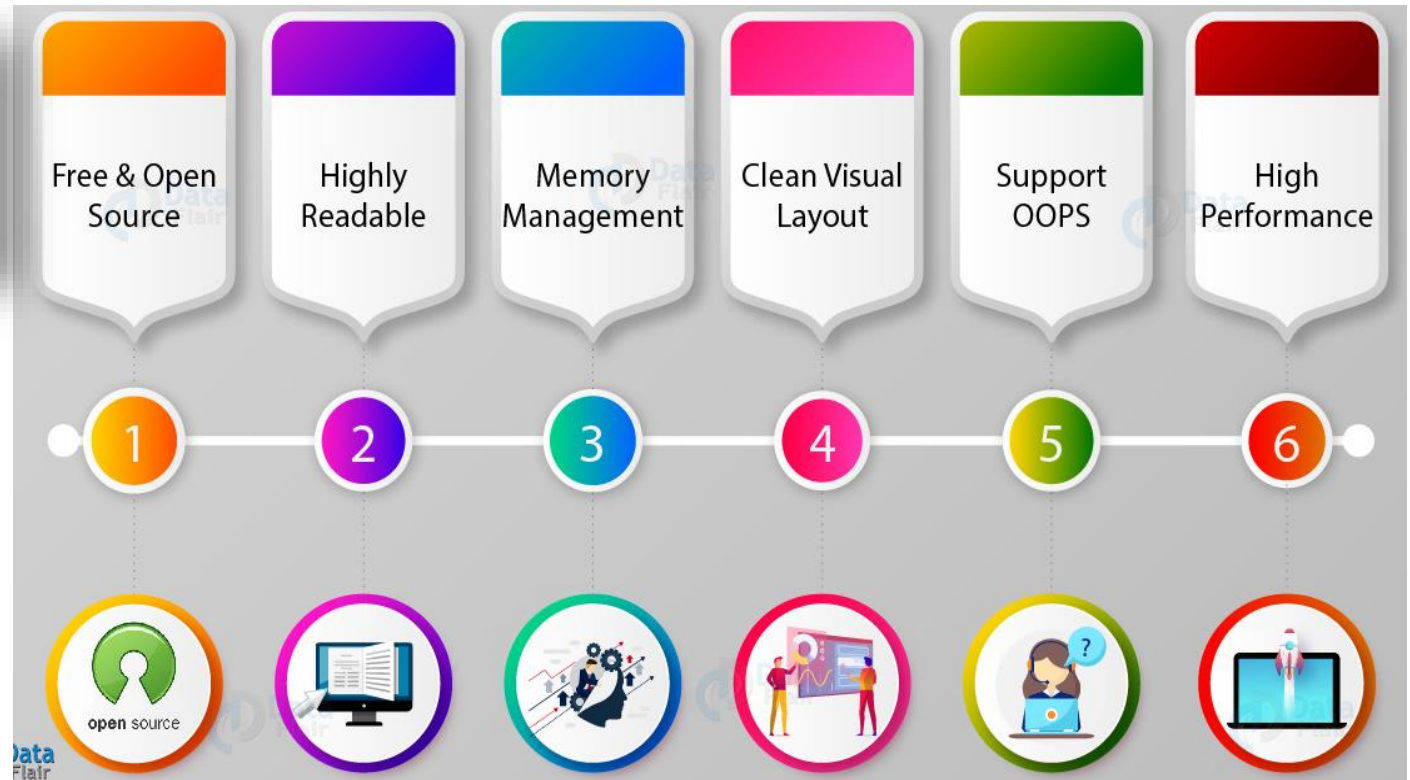
□ Python Programming



- ✓ It is an open source, easy-to-use language that has been around since 1991
- ✓ Supports multiple paradigms (Functional, structured, & procedural programming)
- ✓ **Most applied** language for Data Science tasks today



□ Python Programming





□ Python Programming

Support for several Libraries required for DS tasks



NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.



A free software machine learning library that features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, and k-means and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.



Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.



□ Python Programming

Support for several Libraries required for DS tasks



matplotlib

A plotting library for the Python programming language and its numerical mathematics extension NumPy



TensorFlow

TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.



Keras

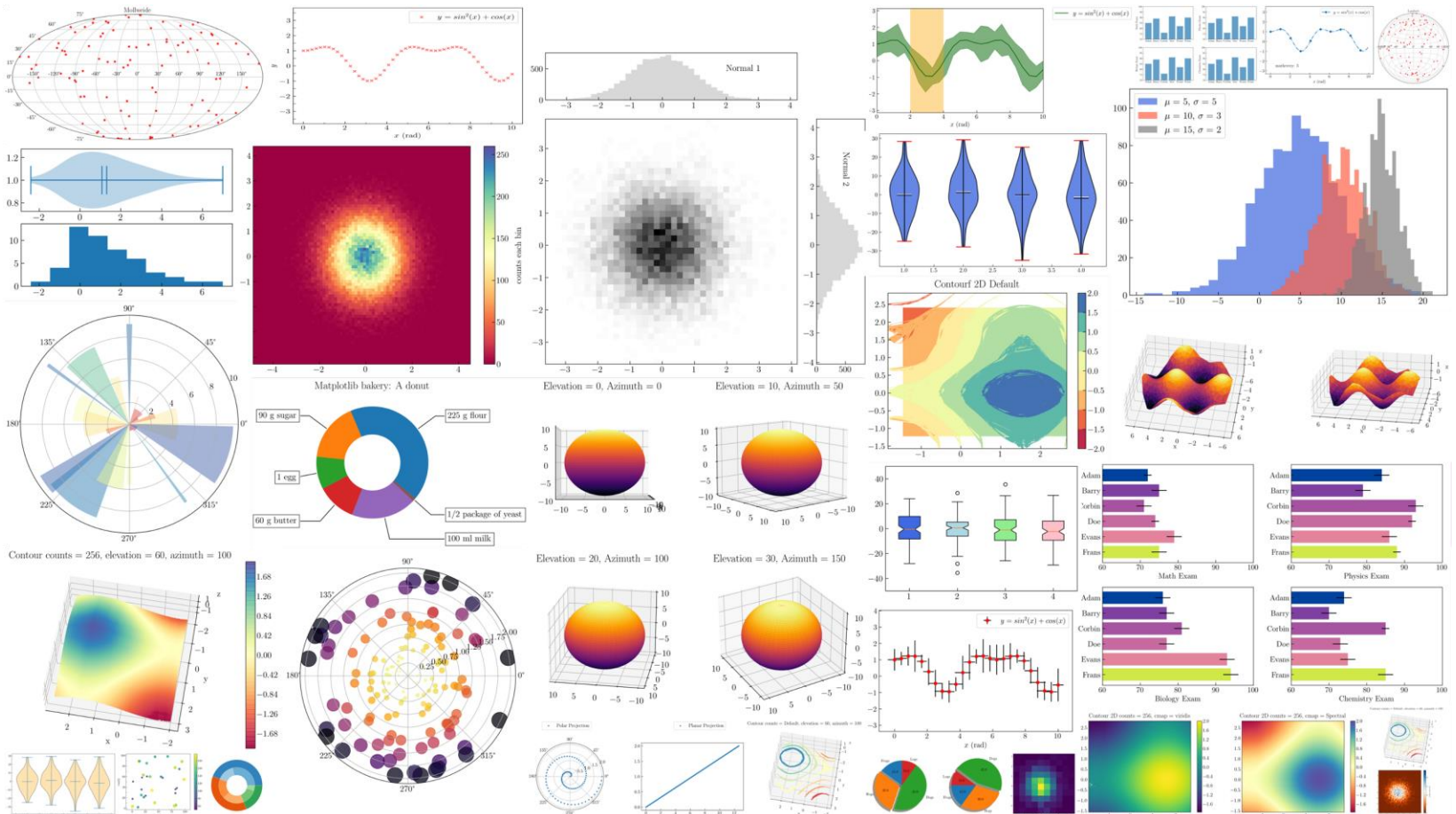
Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or MXNet. It was developed with a focus on enabling fast experimentation



PyTorch




□ Python Programming





□ Python Programming

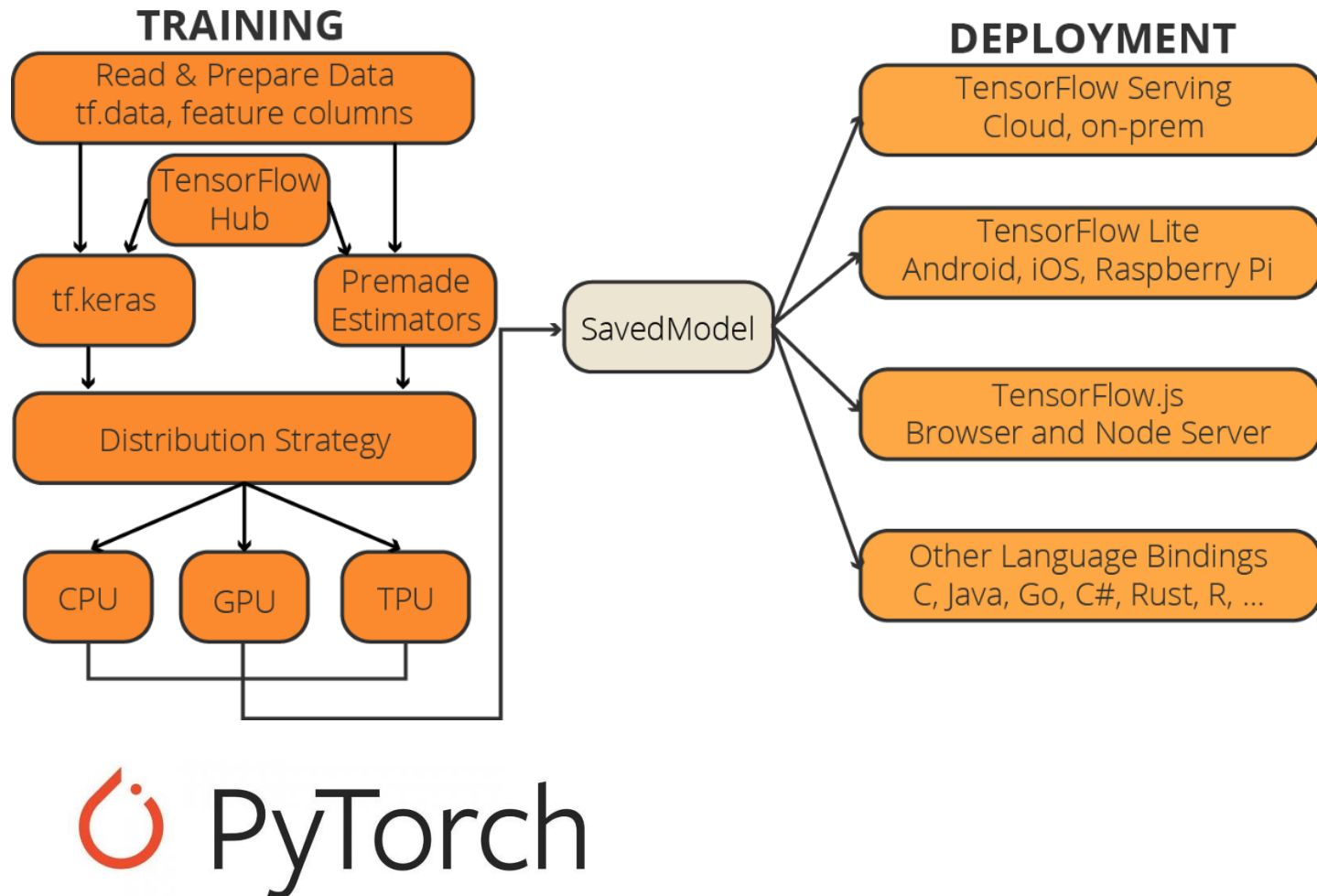


TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.



Keras

Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or MXNet. It was developed with a focus on enabling fast experimentation





□ Python Programming

Could be supported by several tools



Open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text
<http://jupyter.org/>



Similar to Jupyter Notebook, but with the added benefit of "google doc" type sharing and collaboration
<https://colab.research.google.com>

Crestle

Effortless infrastructure for deep learning

Crestle is your GPU-enabled Jupyter environment in the cloud.

<https://www.crestle.com/>



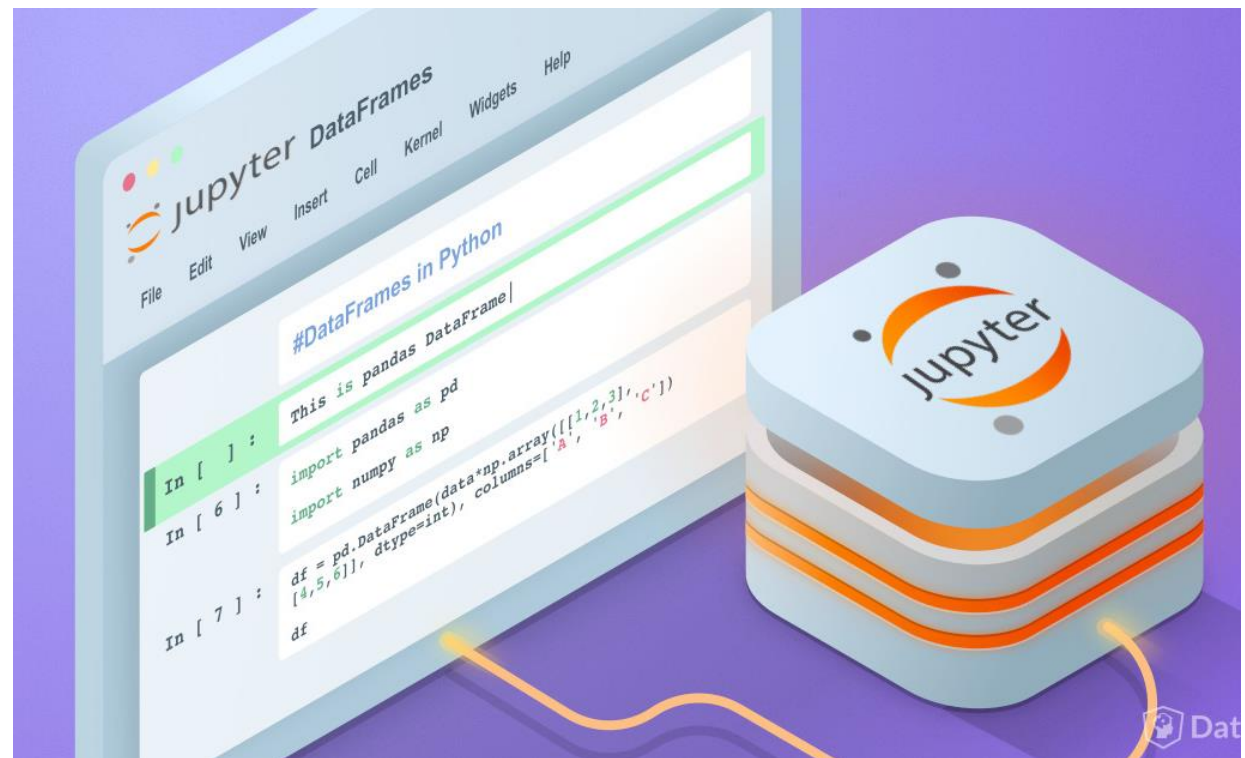
□ Python Programming

Supports Jupyter



Open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text

<http://jupyter.org/>





□ Python Programming

Supports Colab



Similar to Jupyter Notebook, but
with the added benefit of "google
doc" type sharing and
collaboration

<https://colab.research.google.com>

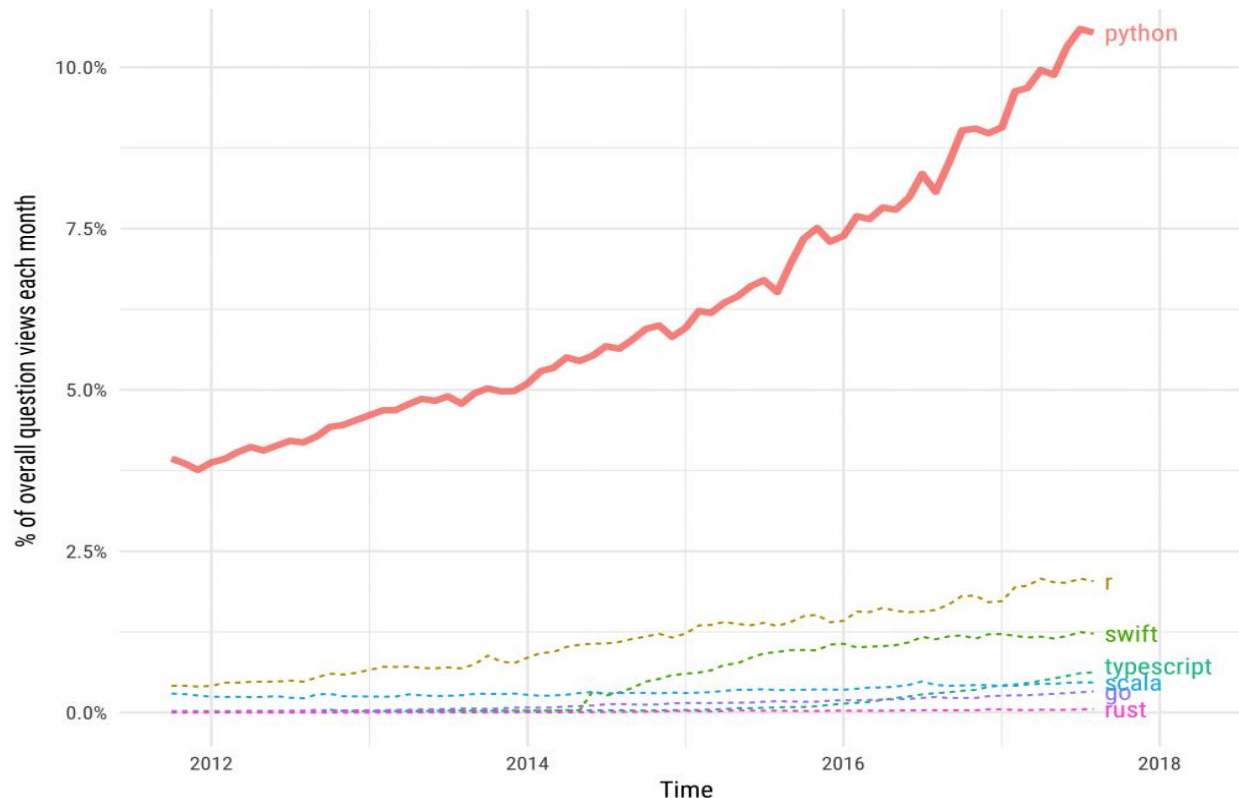
Google

colab



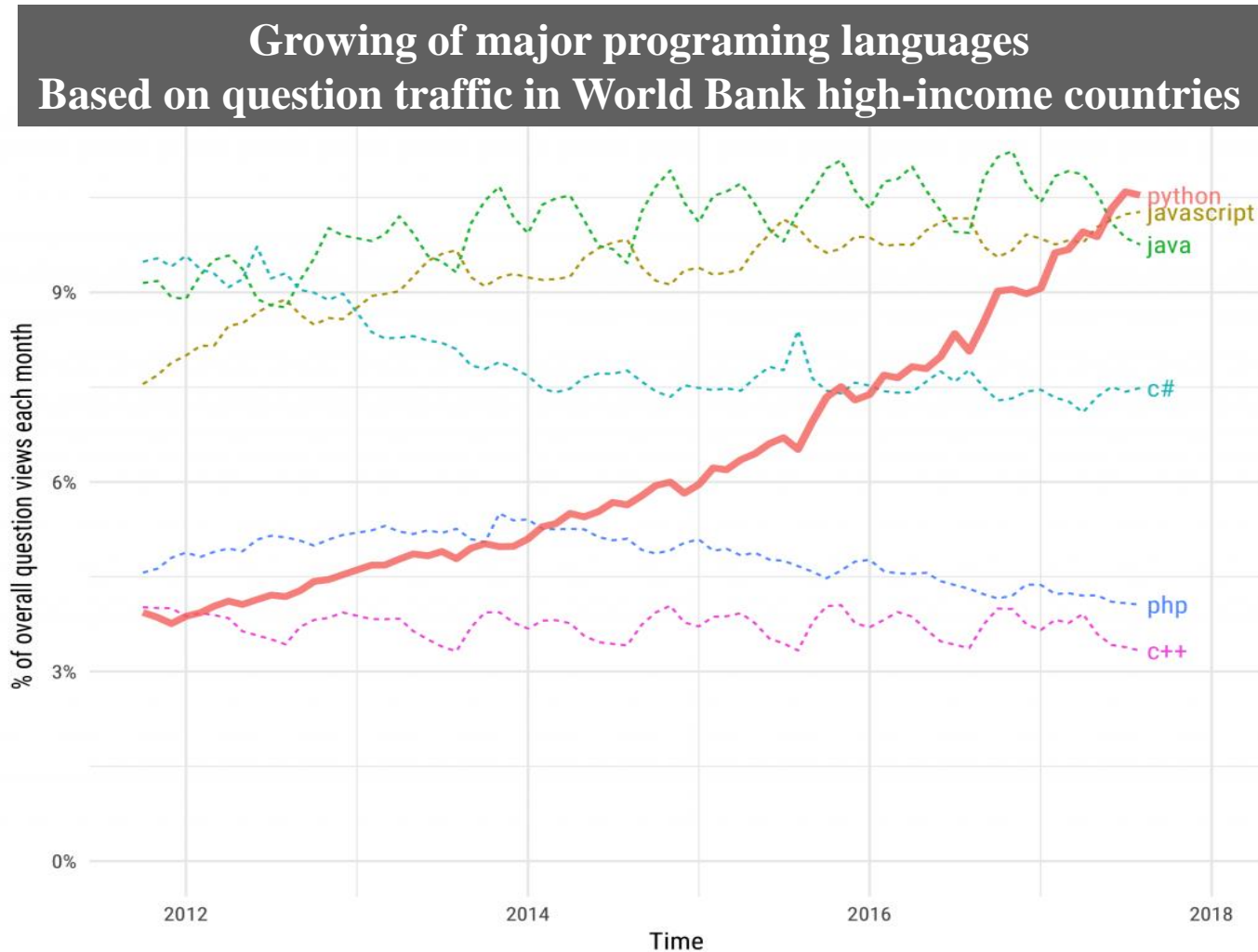
□ The Incredible Growth of Python:

Comparison of Python to smaller growing technologies
Based on question traffic in World Bank high-income countries





□ Procedure for developing in Data Science driven solution:





□ Top 2 Data Science Languages



VS

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main = "Density Plot",
       xlab = "x", ylab = "Density",
       if(orientati == "y")
         dx2 <- (dx - min(dx)) / (max(dx) - min(dx))
         x[1.]
       dy2 <- (dy - min(dy)) / (max(dy) - min(dy))
       y[1.]
  seqbelow <- rep(y[1.], length(dx))
  if(Fill == T)
    confshade(dx2, seqbelow, dy2)
```

Each has its Supporters (Community) & Opponents

Basic difference between Python and R for DS tasks



Differences (Python and R programming)

Python

- Data analysis integrated in web applications
- Statistics code in production of database
- Productivity
- Code readability

R

- Data analysis using standalone computing on individual servers
 - Better Graphical models



Differences (Python and R programming)

Python

- Used by programmers that wants to delve into data analysis or statistics and developers that delve into DS
- Easy-to-use syntax
- Easy to learn

R

- Used primarily in academics and research. However, it is gradually expanding into the enterprise market
- Statistical models can be written only with few lines of codes in R
- Steep learning curve



□ Conclusion

Since DS is a dynamic field with ever growing technologies and tools, the programming language that best suited for it should be considered.

Though Data Scientists are encouraged to explore new languages as well based on their needs/requirements



❖ Summary for today's lecture

- ✓ We have learned about the significance of acquiring programming skills as a Data Scientist.
- ✓ And the characteristics of the top Data Science Programming Languages required in accomplishing Data Science tasks.



中国科学院深圳先进技术研究院
SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Questions and Comments!

Thank You



中国科学院深圳先进技术研究院
SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES