Python与机器学习

——Python简介

华算科技 黄老师

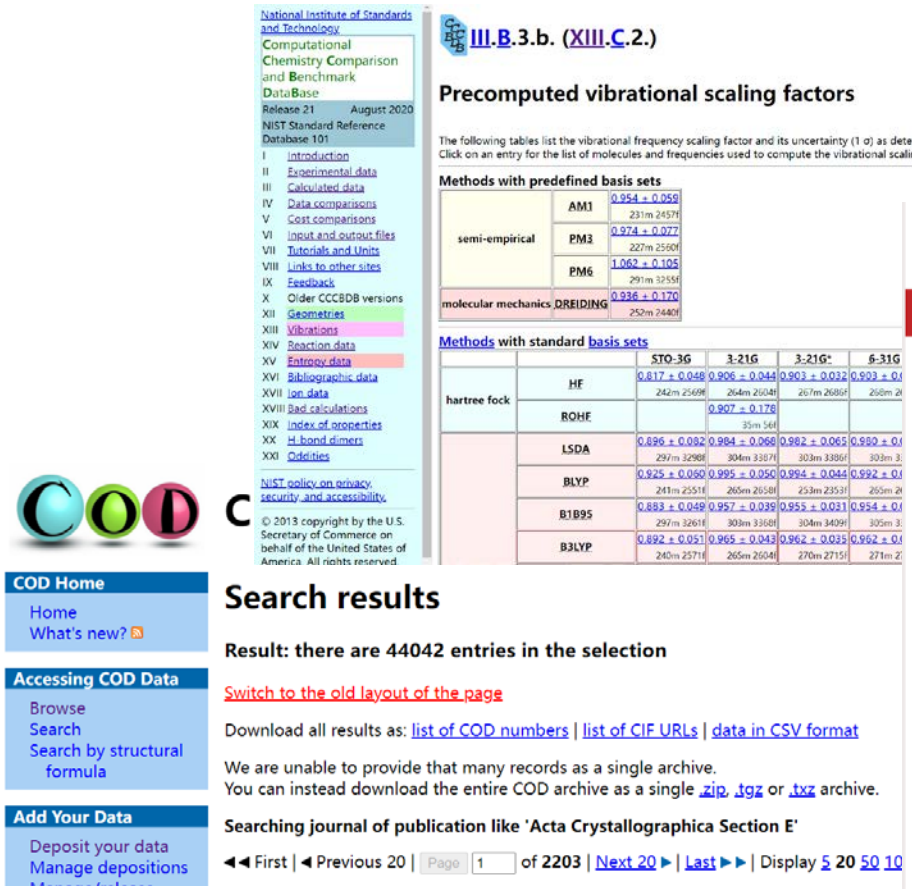
2022年1月17日

华算科技
HUASUAN TECHNOLOGY

课程简介

这是一个数据爆炸的时代

近30年来，人类生产的信息已超过过去5000年信息生产的总和。



调查

您实验室的研究中使用数据科学最大的障碍是什么？（单选）

A. 了解可用的工具
B. 缺乏数据科学的专业知识
C. 该领域变化过于迅速
D. 优先级不够
E. 其它

调查

您实验室的研究中使用数据科学最大的障碍是什么？（单选）

A. 了解可用的工具	34%
B. 缺乏数据科学的专业知识	24%
C. 该领域变化过于迅速	14%
D. 优先级不够	17%
E. 其它	11%

数据来源：DASSAULT SYSTÈMES

课程安排

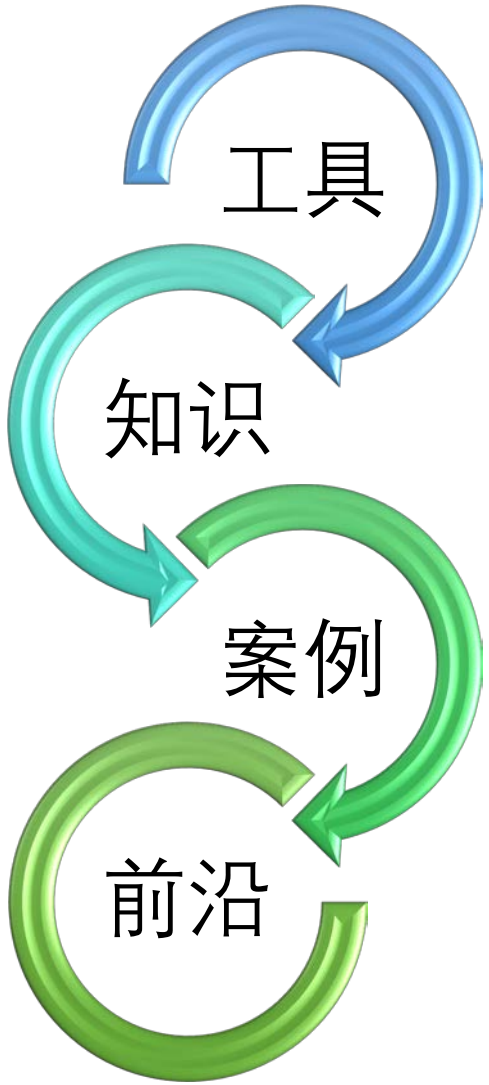
1.17
- 上午：Python简介，编写Python程序，变量，数据结构，条件语句
- 下午：循环语句，函数，Python处理数据，文件读写

1.18
- 上午：numpy库，pandas库，谱数据平滑，matplotlibs库
- 下午：机器学习简介，sklearn库，最小二乘原理，线性回归

1.19
- 上午：交叉验证，决策树分类，支持向量机算法
- 下午：案例1——预测d带中心

1.20
- 下午：高通量筛选，matminer库

1.21
- 上午：案例2——预测体模量，机器学学习前沿

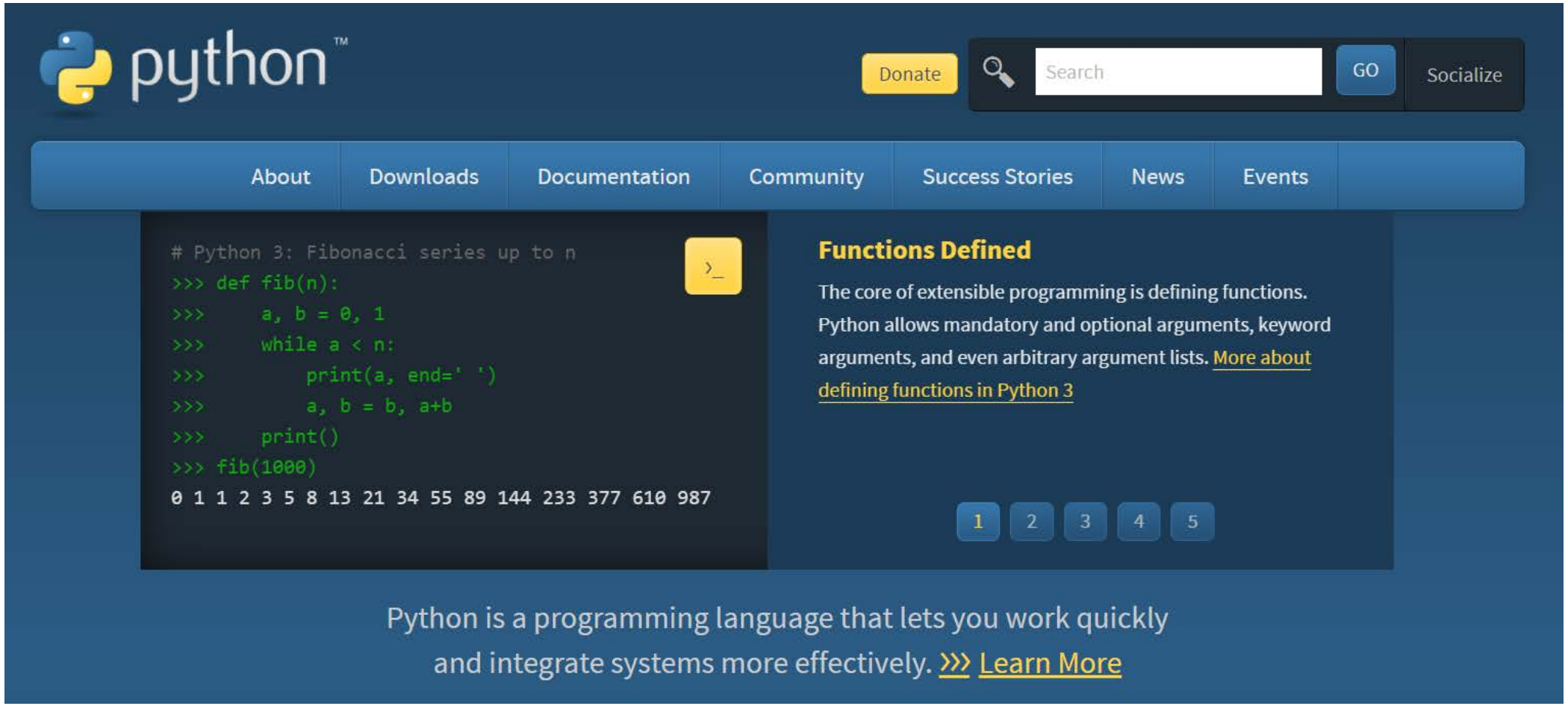


目录

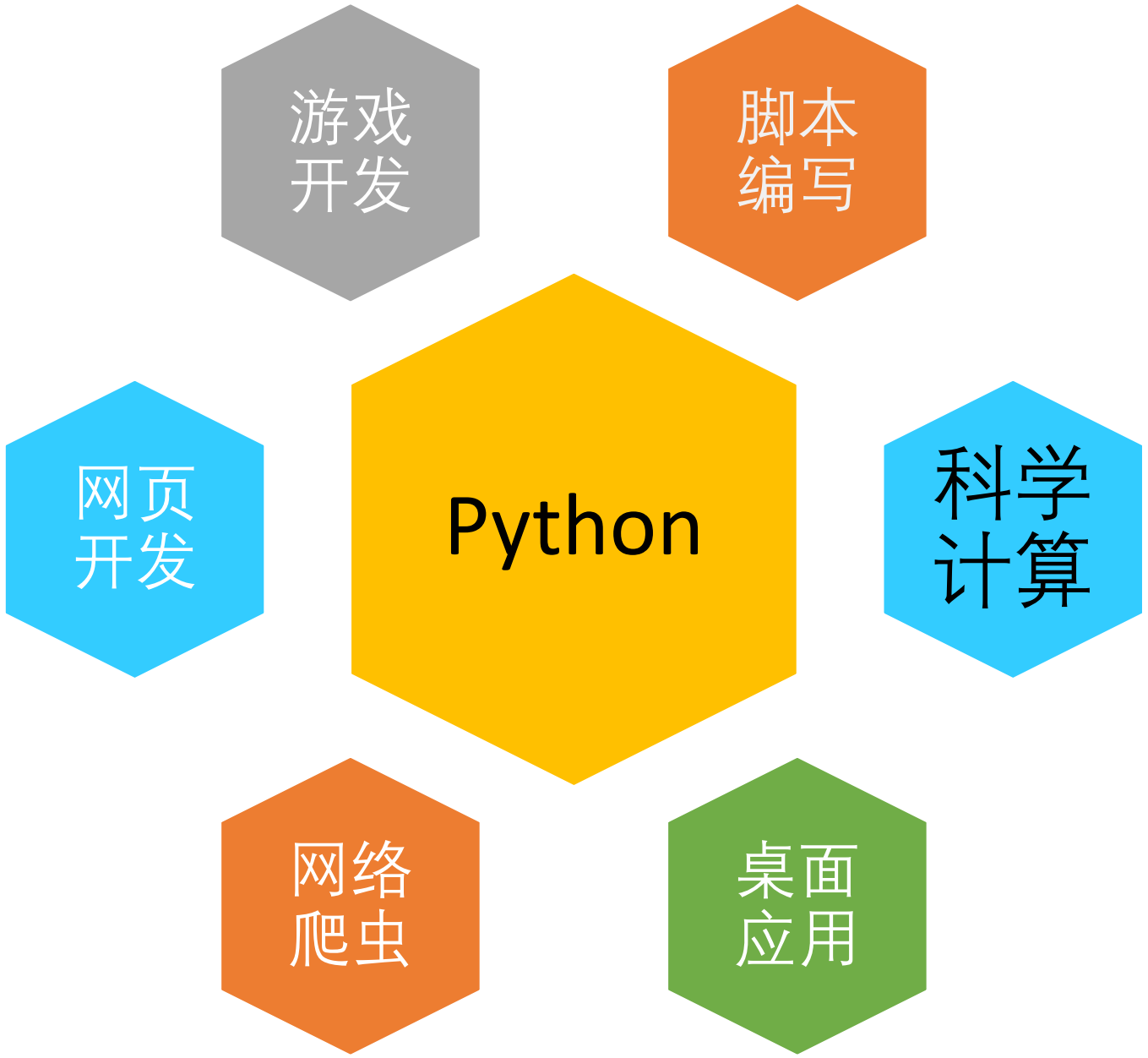
目录

Python

Python 是一种编程语言，可以加快我们的工作并提高系统的效率。



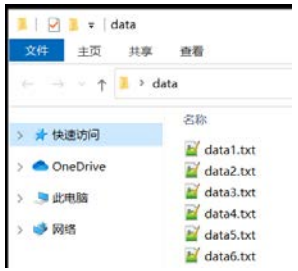
Python功能



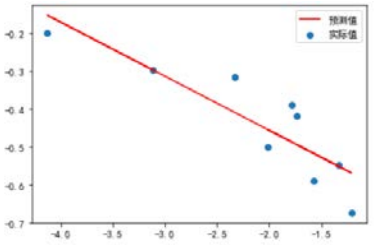
- 数据处理

```
# For loop on a list
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
The product is: 384
```

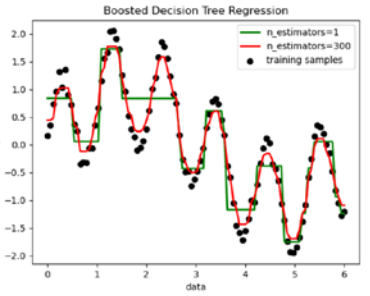
- 文件批处理



- 可视化



- 机器学习



为什么叫Python?

python 的翻译

名词

蟒蛇 python, boa

蟒 python, boa

精 essence, extract, demon, daemon, fiend, python

Python

Apple

天猫

Anaconda：免费开源的Python和R语言的发行版本

Anaconda（中文名：大蟒蛇）

为什么叫Python?

吉多·范罗苏姆(1956-)

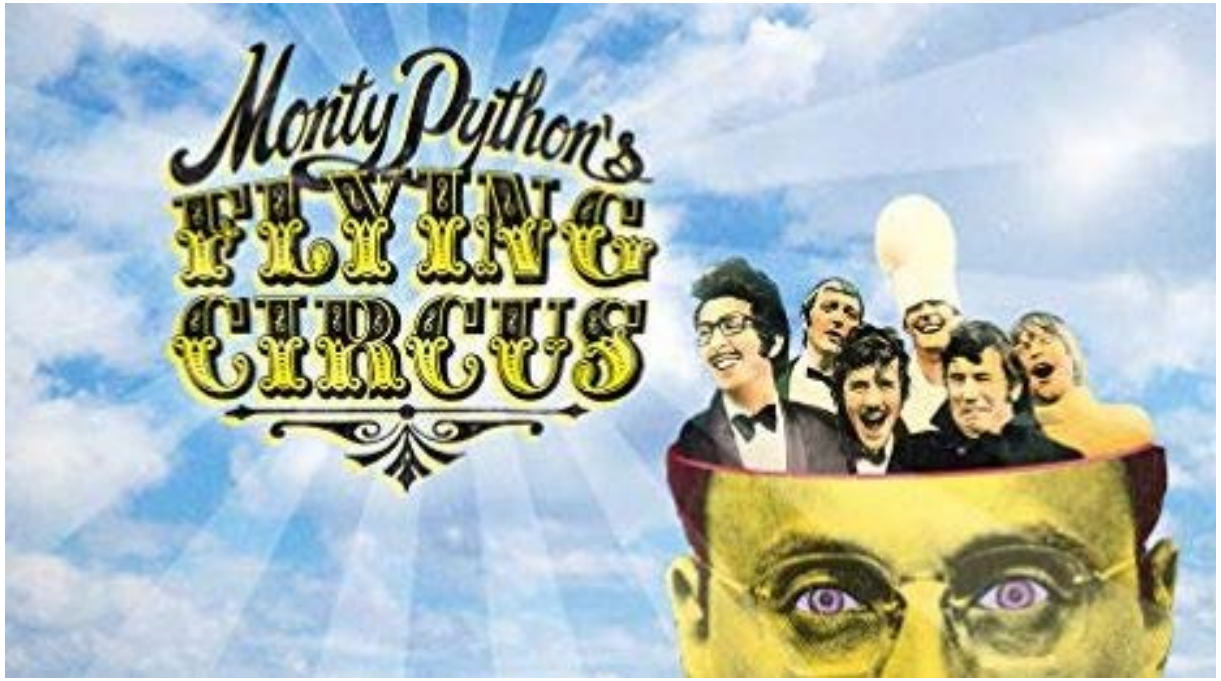
1982年 阿姆斯特丹大学
数学和计算机科学硕士

1991年 Python

2005年 Google

2013年 Dropbox　　2019年 退休

2020年 Microsoft



Monty Python's Flying Circus(1969-1974)

Python版本

1991年，第一个Python解释器诞生，由C语言实现

1994年1月，Python1.0版本发布

2000年10月，Python2.0发布

2008年12月，Python3.0发布
注意，Python3并不向后兼容Python2。

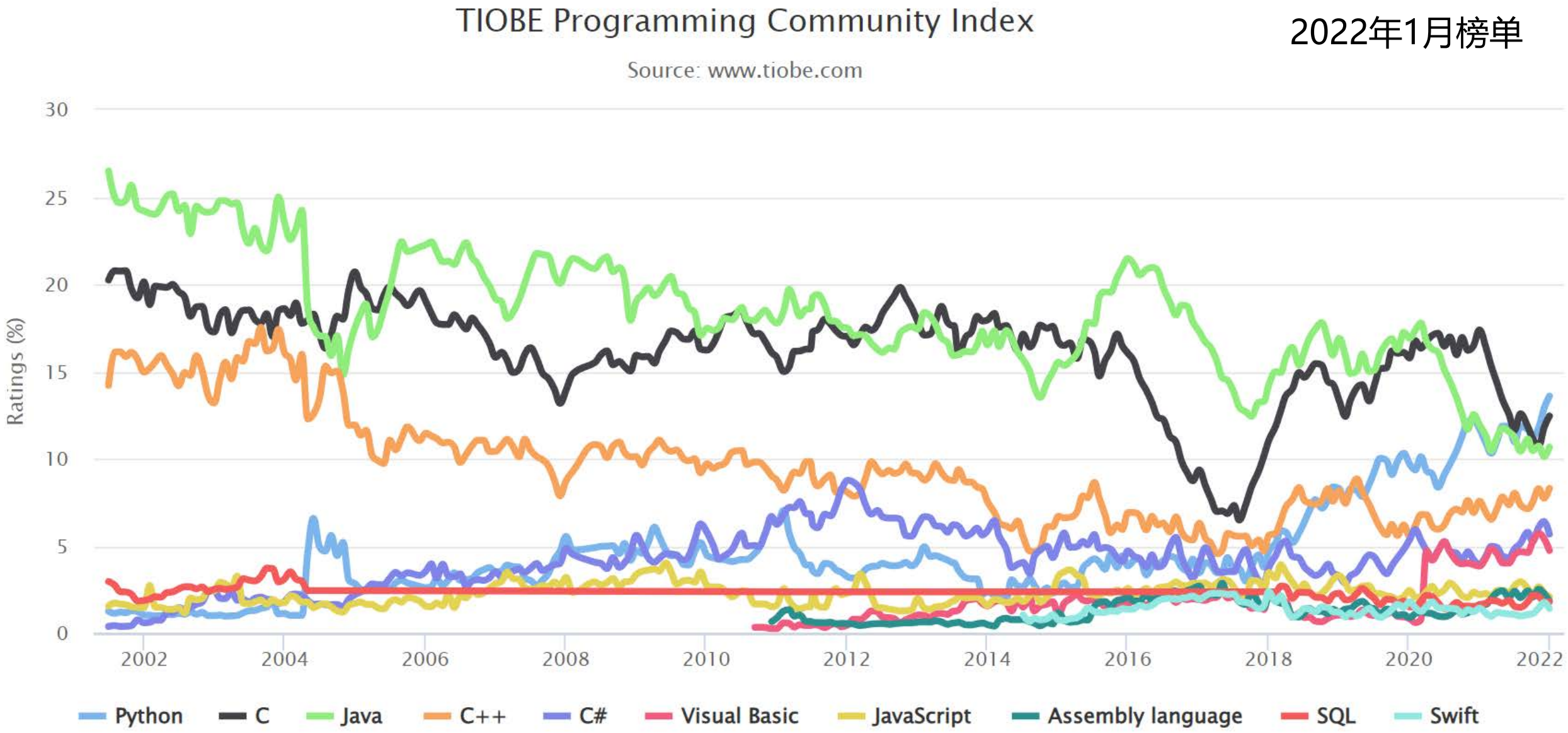
2020年1月1日，Python2停止更新和维护。

简洁
优美
容易使用

目录

TIOBE

2022年1月榜单



TIOBE——Python

TIOBE Index for January 2022

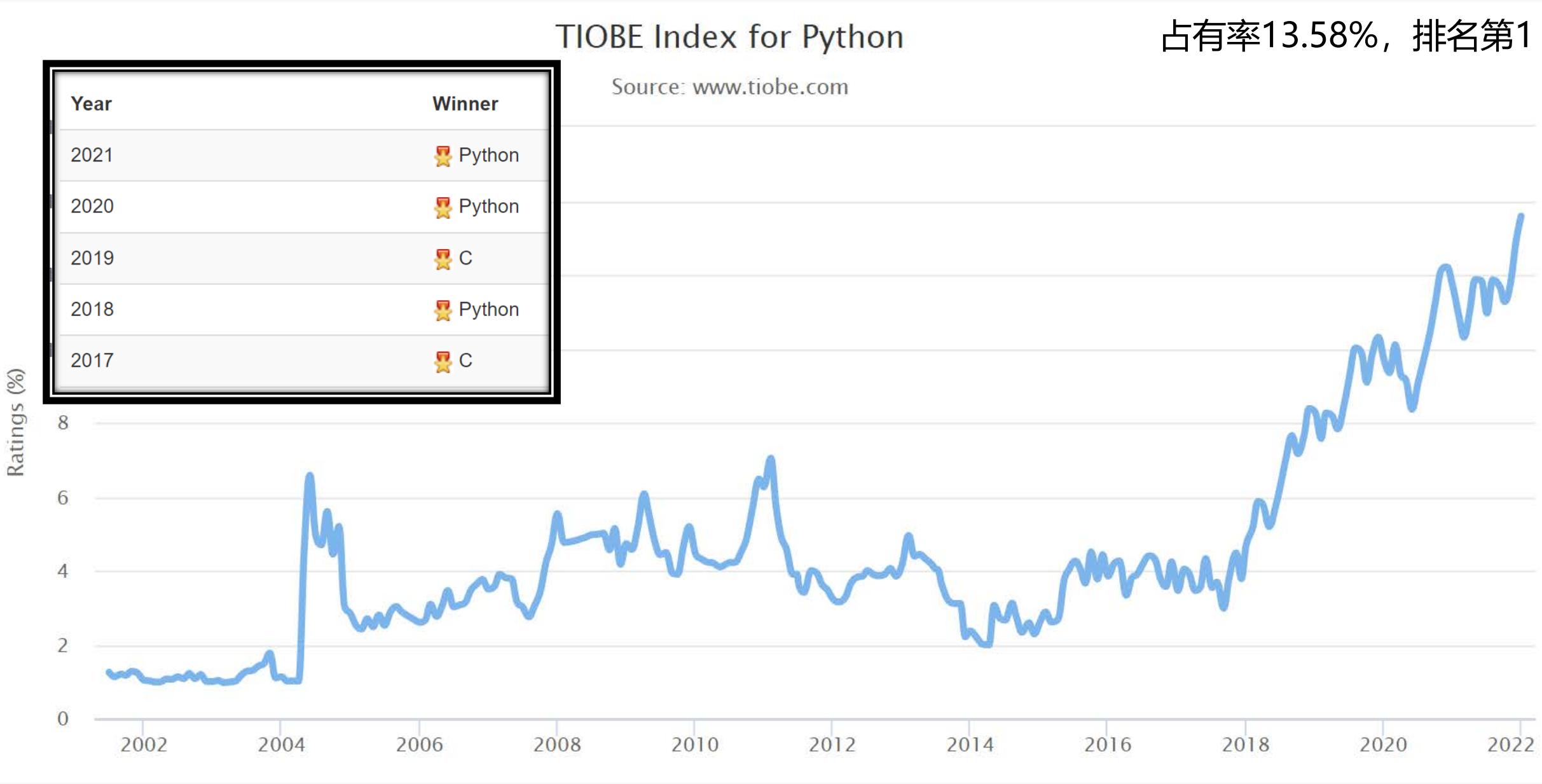
January Headline: Python Programming Language of the Year 2021

Jan 2022	Jan 2021	Change	Programming Language	Ratings	Change
1	3	▲	Python	13.58%	+1.86%
2	1	▼	C	12.44%	-4.94%
3	2	▼	Java	10.66%	-1.30%
4	4		C++	8.29%	+0.73%
5	5		C#	5.68%	+1.73%
6	6		Visual Basic	4.74%	+0.90%
7	7		JavaScript	2.09%	-0.11%

2021年年度编程语言

占有率13.58%，排名第1

Year	Winner
2021	Python
2020	Python
2019	C
2018	Python
2017	C

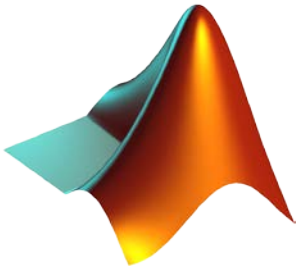


2021年10月榜单题词

For the first time in more than 20 years we have a new leader of the pack: the Python programming language. The long-standing hegemony of Java and C is over. Python, which started as a simple scripting language, as an alternative to Perl, has become mature. Its ease of learning, its huge amount of libraries, and its widespread use in all kinds of domains, has made it the most popular programming language of today. Congratulations Guido van Rossum! Proficiat!

-- Paul Jansen CEO TIOBE Software

Python与其它编程语言



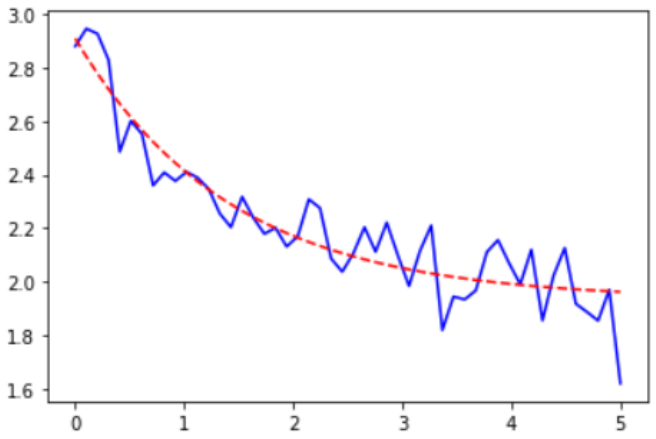
	Python	C / C++	Fortran	MATLAB
占有率	13.58%	12.44% / 8.29%	0.77%	0.96%
排行	1	2 / 4	19	16
是否免费	免费	免费	免费	收费
难易程度	★	★★★★★	★★★	★
平台大小	★	★	★	★★★★★
编写效率	★★★★★	★	★★★	★★★★★

Python优势——编写效率

Python

In [5]:
```
from scipy.optimize import curve_fit
import matplotlib.pyplot as plt
import numpy as np

def func(x, a, b, c):
    return a * np.exp(-b * x) + c

x_val = np.linspace(0, 5, 50)
y_val = func(x_val, 1, 1, 2) + 0.1 * np.random.normal(size = len(x_val))
plt.plot(x_val, y_val, 'b-')
pfit, pcov = curve_fit(func, x_val, y_val)
y_fit = [func(i, pfit[0], pfit[1], pfit[2]) for i in x_val]
plt.plot(x_val, y_fit, 'r--')
```

Out[5]: [<matplotlib.lines.Line2D at 0x259ed1a68e0>]



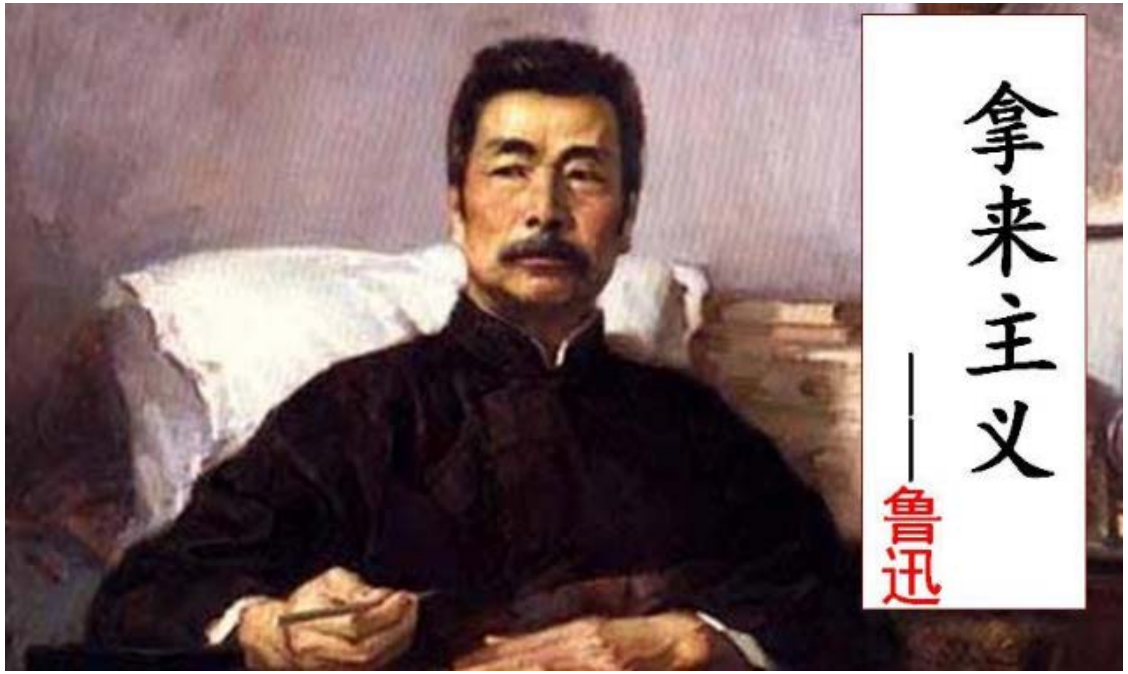
C++

```
void Iz::fit_GN(){
	cout<<setiosflags(ios::fixed)<<setprecision(4);

	para = para_initial(xvals, yvals);
	Vector2d para_d(0.0, 0.0);
	VectorXd r(xvals.size());
	double cod = cal_cod(xvals, yvals, para(0), para(1));
	cout<<"LOOP:"<<endl;
	cout<<"n\tA\tt\tcod"<<endl;
	cout<<"0\t"<<para(0)<<"\t"<<para(1)<<"\t"<<cod<<endl;
	for(int i = 0; i < 50; i++){
		for(int j = 0; j < xvals.size(); j++){
			r(j) = f_err(para(0), para(1), xvals[j], yvals[j]);
		}
		MatrixXd jac = cal_jac(para(0), para(1), xvals);
		para_d = - (jac.transpose() * jac).inverse() * (jac.transpose() * r);
		para = para + para_d;
		cod = cal_cod(xvals, yvals, para(0), para(1));
		cout<<i + 1<<"\t"<<para(0)<<"\t"<<para(1)<<"\t"<<cod<<endl;
		if(cod < 0){
			cout<<"Bad cod value..."<<endl;
			break;
		}
		if(abs(para_d(0)) < 0.0001 && abs(para_d(1)) < 0.0001){
			cout<<"Accuracy achieved, leave the loop..."<<endl;
			break;
		}
		if(i == 49){
			cout<<"Unable to converge..."<<endl;
			break;
		}
	}
	if(cod > cod_opt){
		para_opt(0) = para(0);
		para_opt(1) = para(1);
		step_opt = step;
		cod_opt = cod;
	}

	cout<<resetiosflags(ios::fixed);
}
```

Python优势——库

Python开发效率高的重要原因：有非常强大的第三方库

```
C:\Windows\system32>pip3 list
Package                           Version
--------------------------------- ------------------
alabaster                         0.7.12
anaconda-client                   1.7.2
anaconda-navigator                2.0.3
anaconda-project                  0.9.1
anyio                             2.2.0
appdirs                           1.4.4
argh                              0.26.2
argon2-cffi                       20.1.0
asnlcrypto                        1.4.0
astroid                           2.5
astropy                           4.2.1
async-generator                   1.10
atomicwrites                      1.4.0
attrs                             20.3.0
autopep8                          1.5.6
Babel                             2.9.0
backcall                          0.2.0
backports.functools-lru-cache     1.6.4
backports.shutil-get-terminal-size 1.0.0
backports.tempfile                1.0
backports.weakref                 1.0.post1
bcrypt                            3.2.0
beautifulsoup4                    4.9.3
bitarray                          1.9.2
```



目录

开始使用Python

集成了Python编译器

Python编写、学习平台

环境变量：在操作系统中用来指定操作系统运行环境的一些参数

Jupyter的使用

命令模式（蓝色）

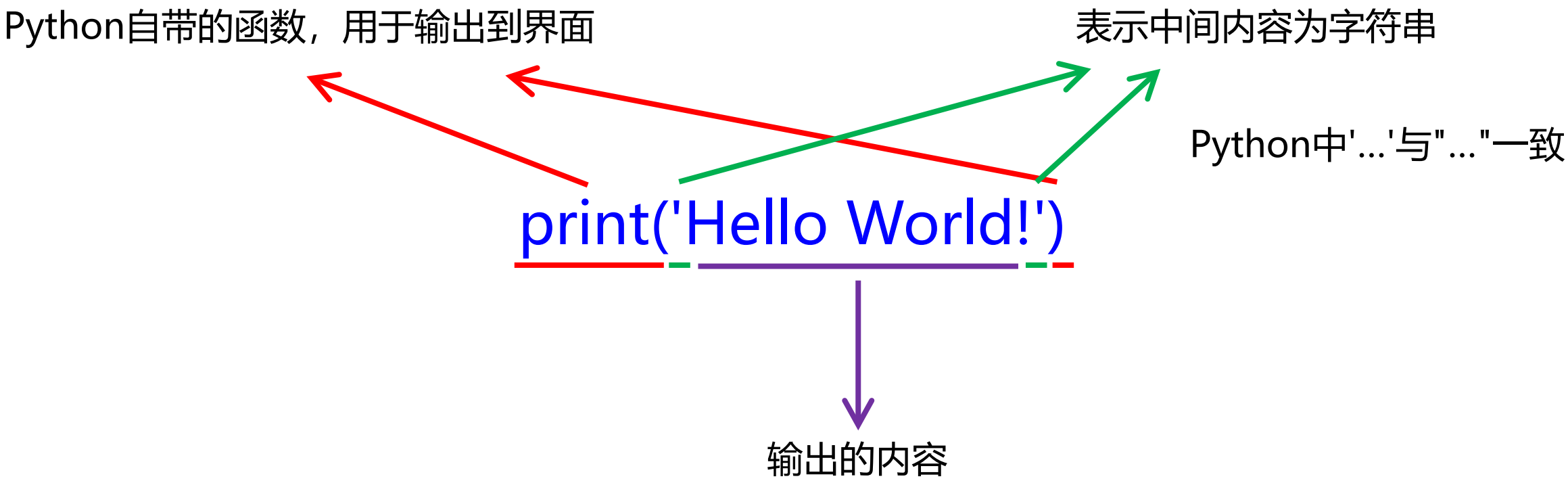
Enter:	切换到编辑模式
A:	在代码块前插入空白代码块
B:	在代码块后插入空白代码块
X:	剪切当前代码块
C:	复制当前代码块
V:	粘贴当前代码块
DD:	删除代码块
Z:	取消删除代码块

编辑模式（绿色）

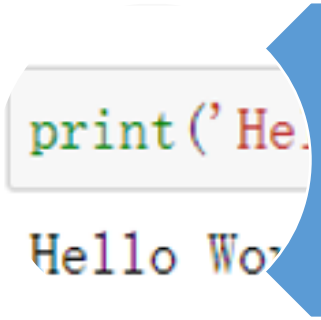
Ctrl + Enter:	运行当前代码块
Shift + Enter:	运行当前代码块并选定下一代码块
Alt + Enter:	运行当前代码块并在后面插入新代码块

第一个Python程序

```
In [1]: print('Hello World!')
        Hello World!
```



编写方式

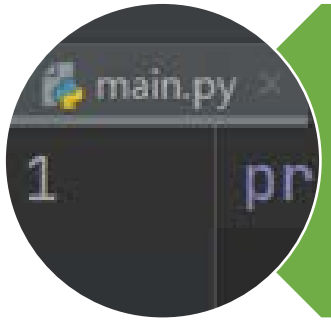


Jupyter

适合课程学习，交流

命令提示符(cmd)

不推荐



集成开发环境(IDE)

适合用于开发，做项目

错误提示

程序编写过程中，报错是非常常见的，可根据报错提示对代码进行修改

```
In [2]: print(Hello World)
  File "C:\Users\26093\AppData\Local\Temp/ipykernel_8484/4293340409.py", line 1
    print(Hello World)
                ^
SyntaxError: invalid syntax
```

错误位置

错误说明

Hello World ⟶ 'Hello World'

最常见错误：拼写错误

避免方法：注意关键词高亮提示

多个字符串输出

```
In [3]: print('Hello!','Machine','Learning!')
Hello! Machine Learning!
```

print('Hello!', 'Machine', 'Learning! ')

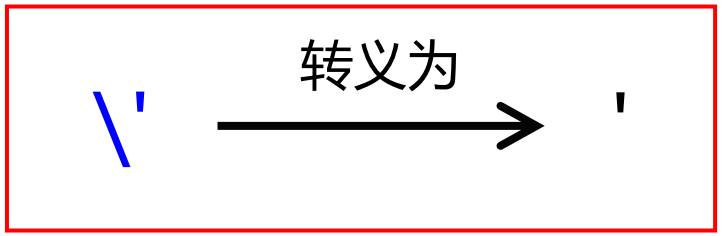
Hello! Machine Learning!

转义字符

输出：I'm fine.

```
In [4]: print('I'm fine.')
  File "<ipython-input-4-470215fd291d>", line 1
    print('I'm fine.')
          ^
SyntaxError: invalid syntax
```

```
In [5]: print('I\'m fine.')
I'm fine.
```



常用转义字符

\' → '

\'' → ''

\\ → \

\n → 换行

\b → 退格

\r → 退回本行开头

\t → 横向制表符