Introduce

JCBioinformatics-2019 Python tutorial
HZAU

JOURNAL CLUB

BBDM Lab

Cao Lab

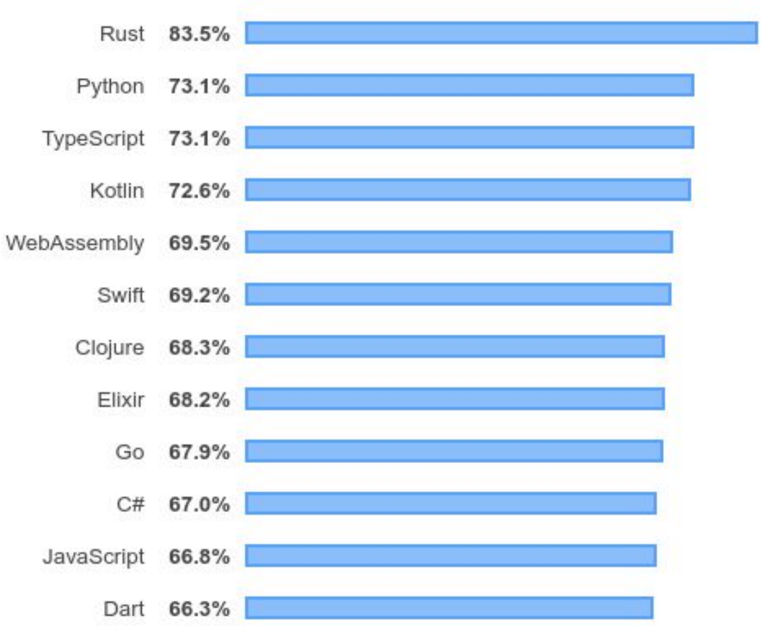
python™

- 一个由社区(Python Software Foundation)开发维护的开源、动态类型、解释执行的通用编程语言。
- was conceived in the late 1980s by by Guido van Rossum.
- Python 3.0, released 2008
- 目前已经发展为具有丰富生态的通用编程语言。

Developer Survey Results 2019 by stackoverflow

数据显示 Python 受到广大开发者的喜爱，流行度与满意度都比较高。

- Most popular: Rank 4
- Most loved: Rank 2
- Most wanted: Rank 1

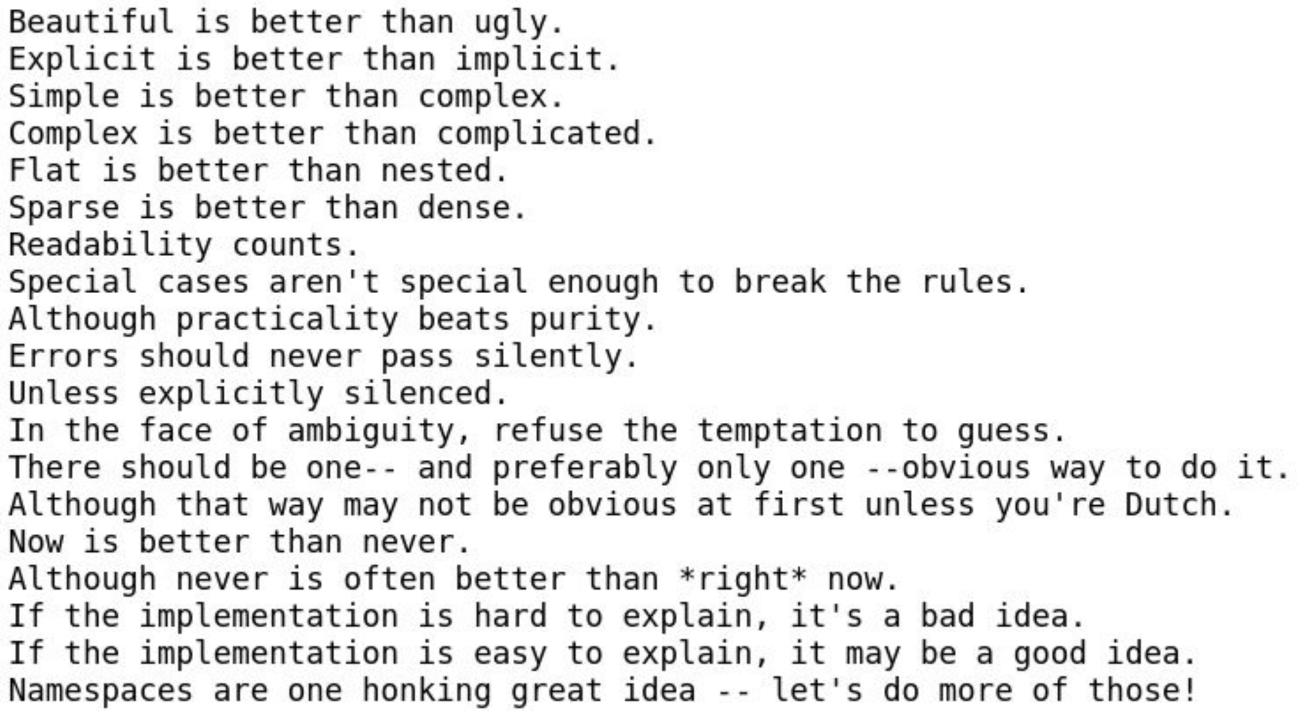


[1] https://insights.stackoverflow.com/survey/2019

The Zen of Python by Tim Peters

Python 的哲学

```
>>> import this
```



```
Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one-- and preferably only one --obvious way to do it.
Although that way may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than *right* now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking great idea -- let's do more of those!
```

Python解释器

- 官方主流实现:CPython
- 其他实现:
  - PyPy: with JIT
  - Jython: on JVM
  - IronPython: on .NET framework
  - CircuitPython
  - MicroPython
  - Stackless Python

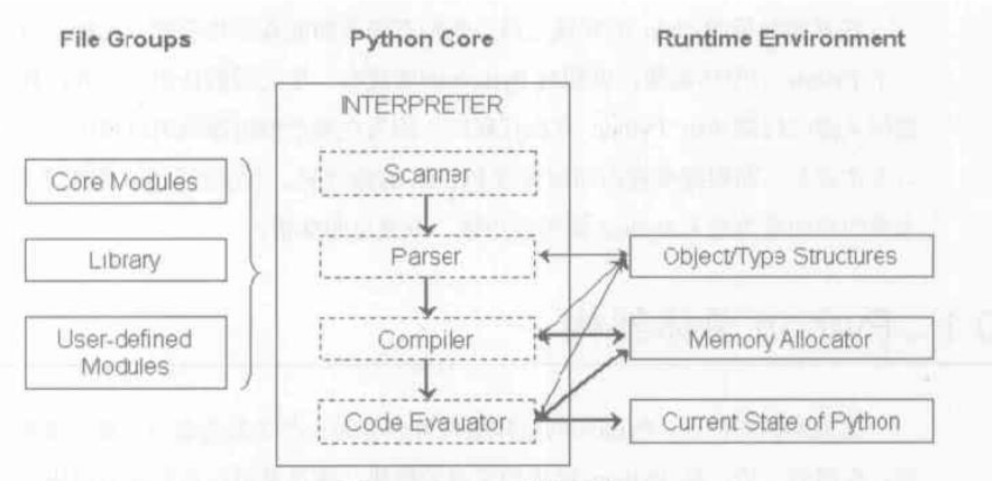


图 0-1 Python 总体架构

截自《Python与源码剖析》

Python 很慢吗？

是的，的确很慢。

但有很多方案让它跑的飞快：

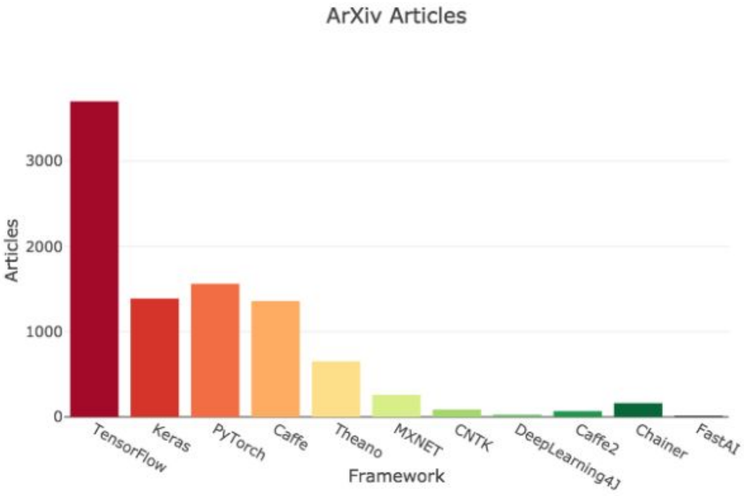
- 矩阵化: Numpy, Cupy
- C extension, Cython
- JIT:
  - Numba
  - PyPy

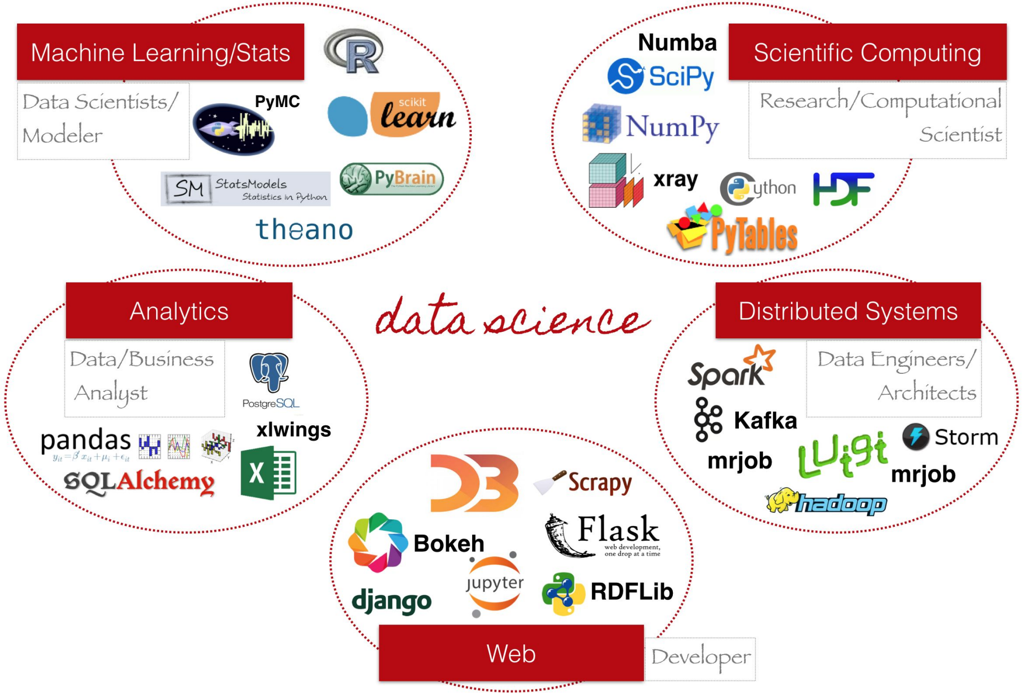
并行化：

- Ray
- modin
- dask

Ecosystem

非常丰富的数据科学、机器学 习、深度学 习生态。



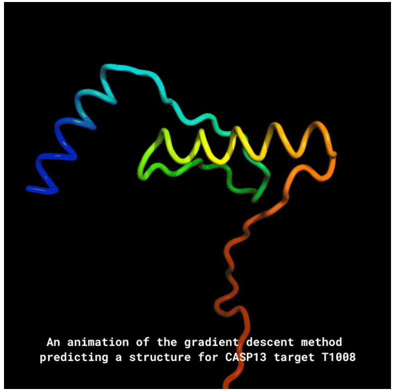


生物信息学编程

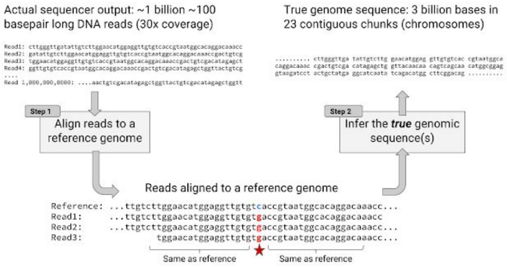
这些生物信息学工具是用 Python 编写/封装的：

- hisat2, PyMOL,Galaxy, HiC-Pro, higlass ...

机器学习、深度学习技术在生物数据上的应用是一个很有前景的发展方向。


An animation of the gradient descent method
predicting a structure for CASP13 target T1008

2019 AlphaFold


Actual sequencer output: ~1 billion ~100 basepair long DNA reads (30x coverage)
Read1: cttgggttgatattgtcttggaacatggaggttgtgtcaccgtaatggcacaggacaaacc
Read2: gatattgtcttggaacatggaggttgtgtcaccgtaatggcacaggacaaaccgactgtcg
Read3: tggaacatggaggttgtgtcaccgtaatggcacaggacaaaccgactgtcgacatagagct
Read4: ggttgtgtcaccgtaatggcacaggacaaaccgactgtcgacatagagctggttactgtcg
....
Read 1,000,000,0000: ....aactgtcgacatagagctggttactgtcgacatagagctggtt
Step 1
Align reads to a reference genome
Reads aligned to a reference genome
Reference: ...ttgtcttggaacatggaggttgtgtcaccgtaatggcacaggacaaacc...
Read1: ...ttgtcttggaacatggaggttgtgtgaccgtaatggcacaggacaaacc
Read2: ...ttgtcttggaacatggaggttgtgtgaccgtaatggcacaggacaaacc...
Read3: tggaacatggaggttgtgtgaccgtaatggcacaggacaaacc...
Same as reference
Same as reference
Step 2
Infer the true genomic sequence(s)
True genome sequence: 3 billion bases in 23 contiguous chunks (chromosomes)
.... cttgggttga tattgtcttg gaacatggag gttgtgtcac cgtaatggca
caggacaaac cgactgtcga catagagctg gttacaacaa cagtcagcaa catggcggag
gtaagatcct actgctatga ggcatcaata tcagacatgg cttcggacag ....

2017 DeepVarient

DeepSea, DeepBind, Basset, SPEID, ...

推荐阅读这篇文献：
https://arxiv.org/abs/1903.00342

如何学习?

-> 如何掌握基础知识？

-> 如何填平基础知识与实践活动之间的鸿沟？

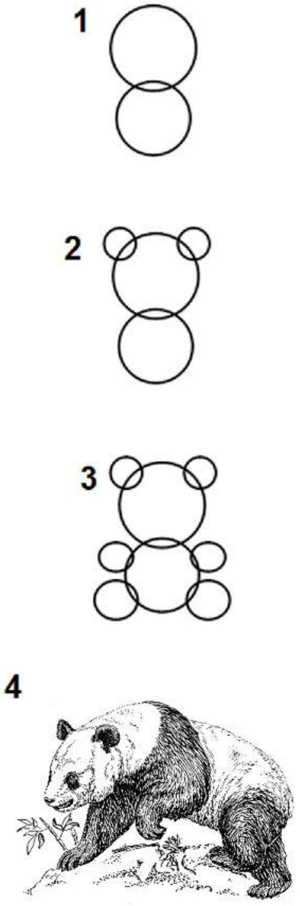
代码练习:

工作之中的练习: 尝试运用 Python 解决实际问题与简化工作流程

工作之外的练习: codewars、leetcode

阅读优质代码:

GitHub 上寻找优质的项目，阅读它。



遇到问题怎么办？

1. 搜索引擎是第一选择，善用搜索引擎。

2. 提问：

网上提问：

a. stackoverflow
b. stackexchange
c. reddit
d. GitHub issue

3. 寻找并阅读与你的目的相似的代码，GitHub是一个好去处。

需要注意，网上提问不是一件随随便便的事，关于网上提问，有很多需要注意的技巧与礼仪：
可以阅读学习一下这篇《提问的智慧》，有助于你提出一个好文题，节省自己与他人的时间。