Computational Statistics and Data Analytics

November 14, 2019

CDC Short Course

Xiaoming Huo, Georgia Tech

CREATING THE NEXT®

Schedule

- 8:30 - 9:30 **Lect. 1: Introduction**
- 9:45 - 10:45 Lect. 2: Classification
- 11:00 - 12:00 Lect. 3: Clustering
- 12:00 - 1:15 *Lunch*
- 1:15 - 2:15 Lect. 4: Tree-based Methods
- 2:30 - 3:30 Lect. 5: Principal Component Analysis
- 3:45 - 4:15 Lect. 6: Summary
- 4:15 *Adjourn*
- Online *resources*: tinyurl.com/StatComCDC

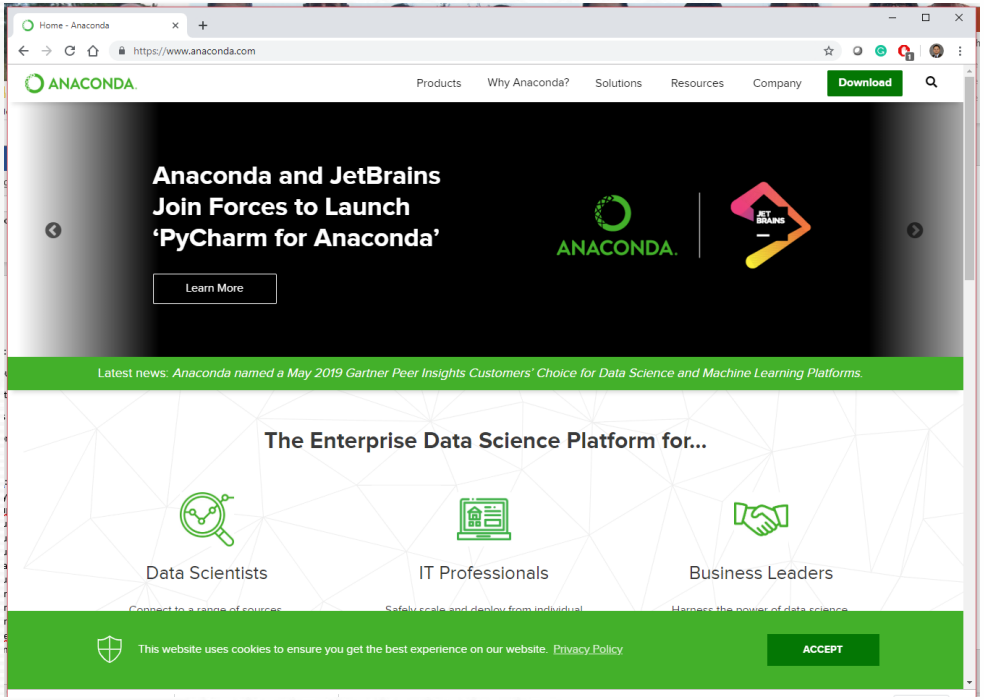
courses.d2l.ai/berkeley-stat-157

Agenda

- **Anaconda**
- Jupyter Notebooks,
- Reproducible Research
- GitHub – version control

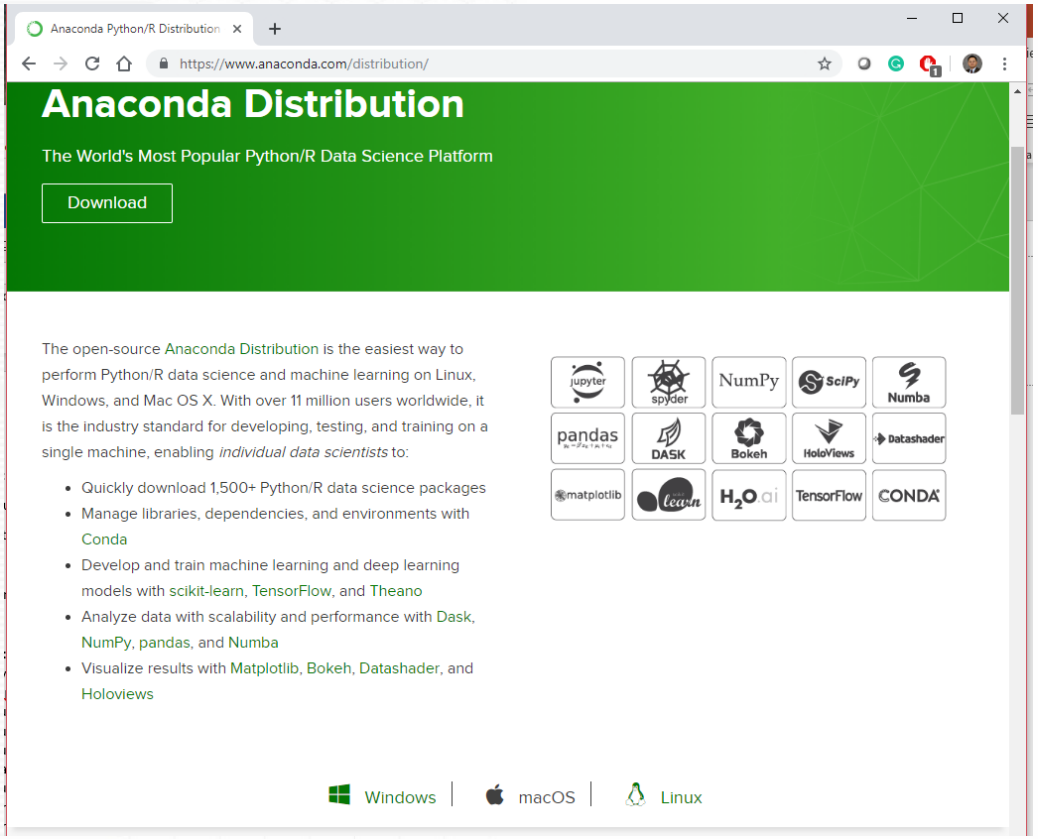
Anaconda

- https://www.anaconda.com/

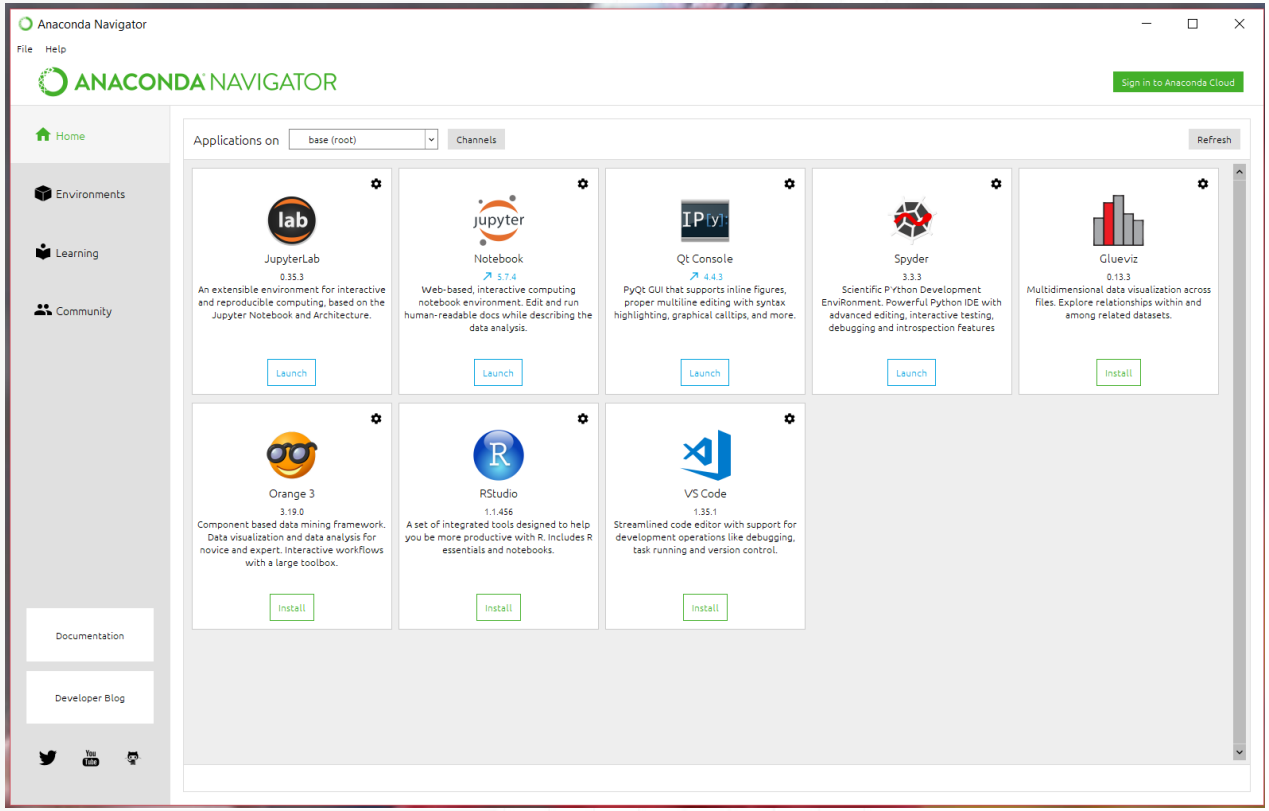


Download

https://www.anaconda.com/distribution/



After installation



GitHub


GitHub Inc.
Company
github.com
GitHub is an American company that provides hosting for software development version control using Git. It is a subsidiary of Microsoft, which acquired the company in 2018 for $7.5 billion. Wikipedia
Founded: 2008
Headquarters: San Francisco, CA
CEO: Nat Friedman (Oct 29, 2018–)
Parent organization: Microsoft Corporation
Founders: Tom Preston-Werner, Chris Wanstrath, P. J. Hyett, Scott Chacon
Subsidiary: Easel Inc.


ndarray
courses.d2l.ai/berkeley-stat-157

Georgia Tech

N-dimensional Array Examples

- N-dimensional array, short for ndarray, is the main data structure for machine learning and neural networks

0-d (scalar)

```
1.0
```

A class label

1-d (vector)

```
[1.0, 2.7, 3.4]
```

A feature vector

2-d (matrix)

```
[[1.0, 2.7, 3.4]
 [5.0, 0.2, 4.6]
 [4.3, 8.5, 0.2]]
```

A example-by-feature matrix

ND Array Examples, cont

3-d

```
[[[0.1, 2.7, 3.4]
  [5.0, 0.2, 4.6]
  [4.3, 8.5, 0.2]]
 [[3.2, 5.7, 3.4]
  [5.4, 6.2, 3.2]
  [4.1, 3.5, 6.2]]]
```

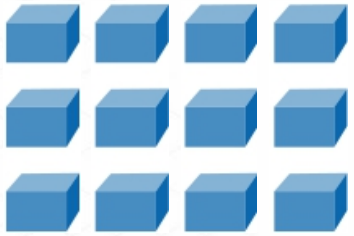
A RGB image (width x height x channels)

4-d

```
[[[[. . .
   . . .
   . . .]]]
```

A batch of RGB images (batch-size x width x height x channels)

5-d



```
[[[[. . .
   . . .
   . . .]]]
```

A batch of videos (batch-size x time x width x height x channels)

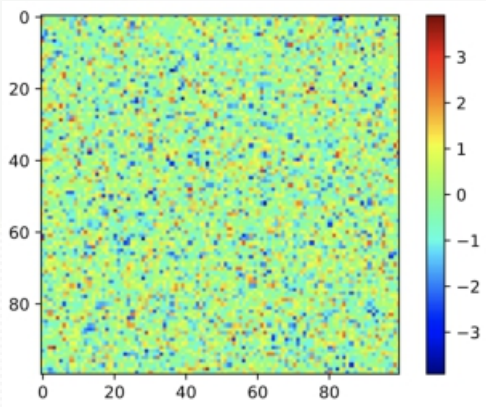
Georgia Tech

Create Arrays

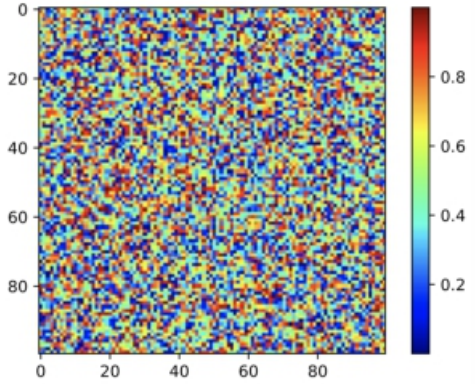
- Create arrays with
  - A shape, e.g. 3-by-4 matrix
  - Data type for each element, e.g. float
  - Element values, e.g all 0s, or random values

100-by-100 matrix with elements generated from

A normal distribution



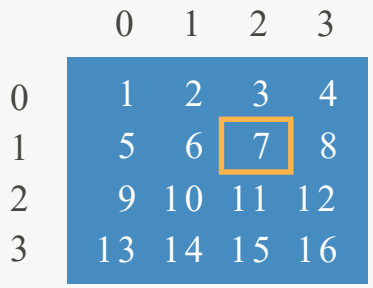
A uniform distribution



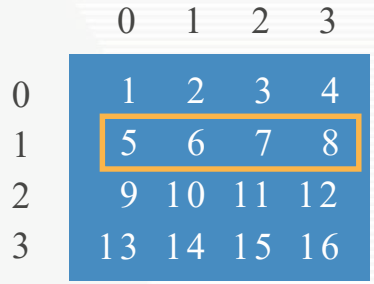
Georgia Tech

Access Elements

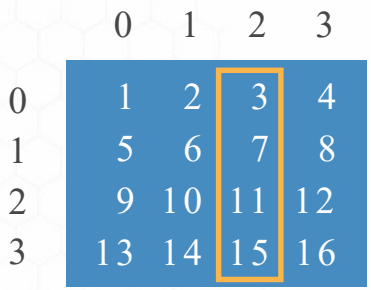
An element: [1, 2]

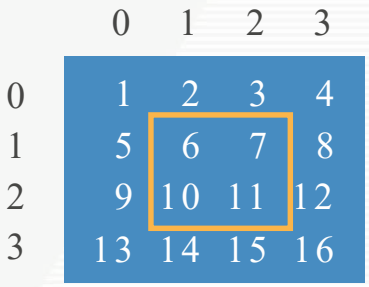


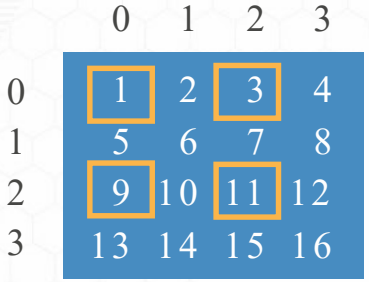
A row: [1, :]



A column: [:, 2]







Georgia Tech

Pointers

- d2l.ai/chapter_crashcourse/ndarray.html
- beta.mxnet.io/guide/crash-course/1-ndarray.html
- Gilbert Strang's Linear Algebra course github.com/juanklopper/MIT_OCW_Linear_Algebra_18_06
- Berkeley HPC course (GPU sections) sites.google.com/lbl.gov/cs267-spr2018/

Georgia Tech

Summary

- **Links for online resources**
  Course materials, URLs
- **Software**
  Installation, cloud
- **Linear Algebra**
  Basic notation
- **NDArray**

Schedule

- 8:30 - 9:30 Lect. 1: Introduction
- 9:45 - 10:45 Lect. 2: Classification
- 11:00 - 12:00 Lect. 3: Clustering
- 12:00 - 1:15 *Lunch*
- 1:15 - 2:15 Lect. 4: Tree-based Methods
- 2:30 - 3:30 Lect. 5: Principal Component Analysis
- 3:45 - 4:15 Lect. 6: Summary
- 4:15 *Adjourn*
- Online *resources*: tinyurl.com/StatComCDC