End-to-End Machine Learning: The Python ML Ecosystem

Glenn Bruns

CSUMB

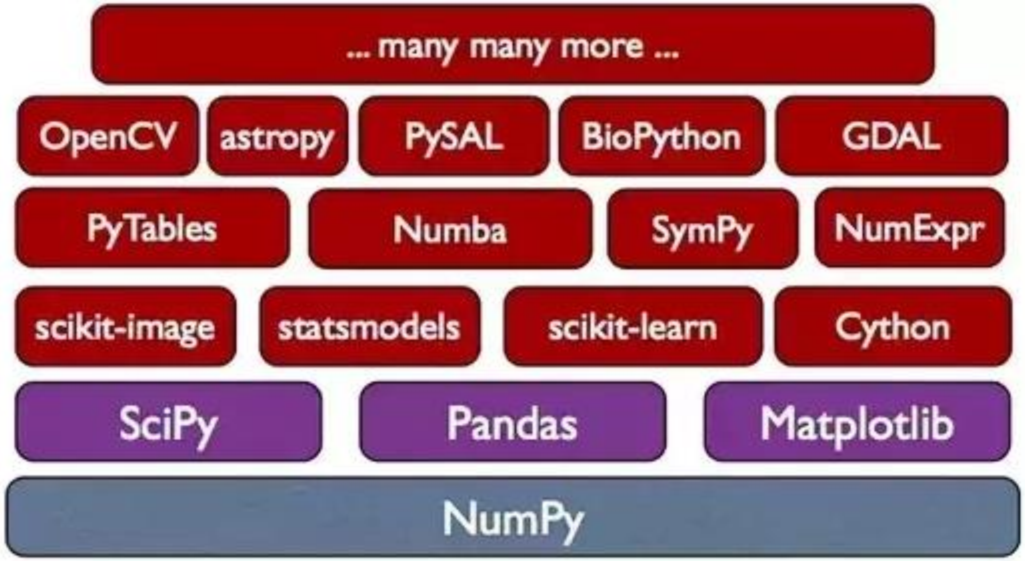
Learning outcomes

After this lecture you should be able to:

1. List the packages and tools we’ll use for machine learning with Python
2. Explain the role of each package/tool
3. Load csv data using Pandas
4. Do basic data exploration with NumPy
5. Edit and run Python programs using Spyder

The Python machine learning ecosystem

- □ NumPy
- □ SciPy
- □ Scikit-Learn
- □ Pandas
- □ Matplotlib



source: quora.com/What-is-the-relationship-among-NumPy-SciPy-Pandas-and-Scikit-learn-and-when-should-I-use-each-one-of-them

What the packages do

NumPy: supports large, multi-dimensional arrays & matrices, plus functions on them

SciPy: builds on NumPy; adds algorithms and convenience functions

Scikit-Learn: builds on SciPy; adds machine learning

Pandas: data manipulation and analysis

- Series – labelled, 1D array
- DataFrame – labelled, 2D data

Matplotlib: plotting

Reminder: data science process

- Define goal
- Data acquisition and cleaning
- Data exploration
- Machine learning
- Present results

Reading a CSV file

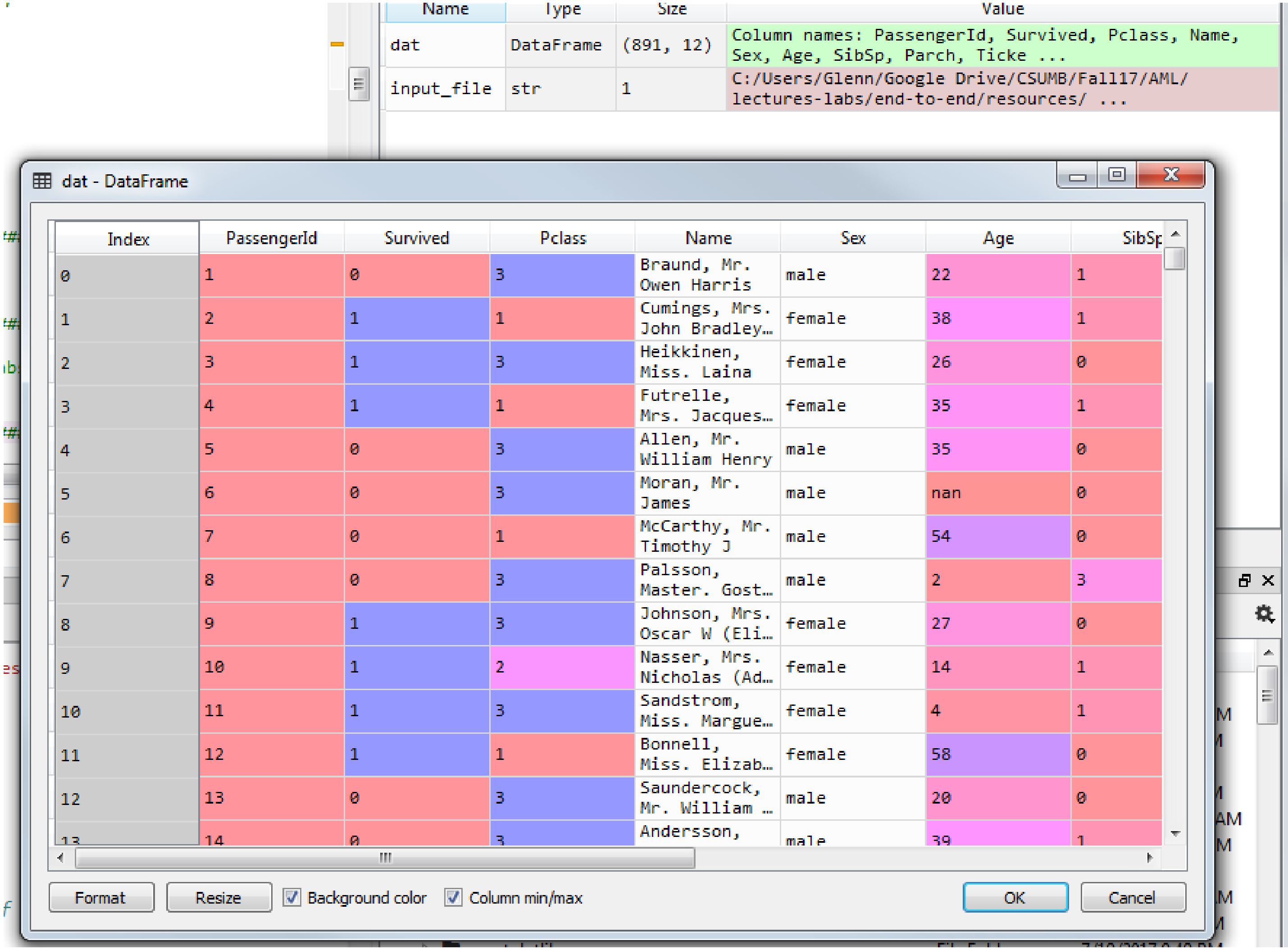
One way, using Pandas:

```
import numpy as np
import pandas as pd
input_file = "C:/Users/…/titanic-data/train.csv"
dat = pd.read_csv(input_file)
```

dat is a pandas DataFrame

(see pandas.pydata.org/pandas-docs/stable/dsintro.html)

Examining the data in Spyder



Name	Type	Size	Value
dat	DataFrame	(891, 12)	Column names: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticke ...
input_file	str	1	C:/Users/Glenn/Google Drive/CSUMB/Fall17/AML/lectures-labs/end-to-end/resources/ ...

dat - DataFrame

Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
0	1	0	3	Braund, Mr. Owen Harris	male	22	1
1	2	1	1	Cumings, Mrs. John Bradley…	female	38	1
2	3	1	3	Heikkinen, Miss. Laina	female	26	0
3	4	1	1	Futrelle, Mrs. Jacques…	female	35	1
4	5	0	3	Allen, Mr. William Henry	male	35	0
5	6	0	3	Moran, Mr. James	male	nan	0
6	7	0	1	McCarthy, Mr. Timothy J	male	54	0
7	8	0	3	Palsson, Master. Gost…	male	2	3
8	9	1	3	Johnson, Mrs. Oscar W (Eli…	female	27	0
9	10	1	2	Nasser, Mrs. Nicholas (Ad…	female	14	1
10	11	1	3	Sandstrom, Miss. Margue…	female	4	1
11	12	1	1	Bonnell, Miss. Elizab…	female	58	0
12	13	0	3	Saundercock, Mr. William …	male	20	0
13	14	0	3	Andersson,	male	39	1

☑ Background color ☑ Column min/max

Basic data exploration

```
# similar to R’s ‘summary’
dat.info()

# number of elements
dat.size

# dimensions
dat.shape

# number of rows
len(dat)
dat.shape[0]    # alternative way

# beginning of the data frame
dat.head()
```

Running code in Spyder

- □ As in R, highlight code and hit ctrl-Enter
- □ To clear your environment, enter `%reset` in the console

Running dat.info()

```
In [227]: dat.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

Summary

□ Packages in the Python ML ecosystem

□ Reading CSV data and basic data exploration

□ Edit and run Python programs using Spyder