Northeastern University

DS 5110 – Lecture 1
Introduction

Roi Yehoshua

Agenda

- Course objectives
- The data science pipeline

DS 5110 Course Objectives

Northeastern University

- Learn to work with the data science libraries in Python (NumPy, Matplotlib, Pandas)
- Learn to work with different types of files text, CSV and JSON files
- Analyze data using Pandas
- Learn how to do exploratory data analysis (EDA)
- Understand basic database concepts
- Use ER models to design a database system
- Learn to formulate SQL queries on the data
- Integrate SQL queries within Python applications
- Describe big data tools and techniques
- Learn to build basic machine learning pipelines using Scikit-Learn

What is Data?

Northeastern University

- Collection of data objects and their attributes

Attributes

Objects

ID	*name*	*dept_name*	*salary*
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

Facets of Data

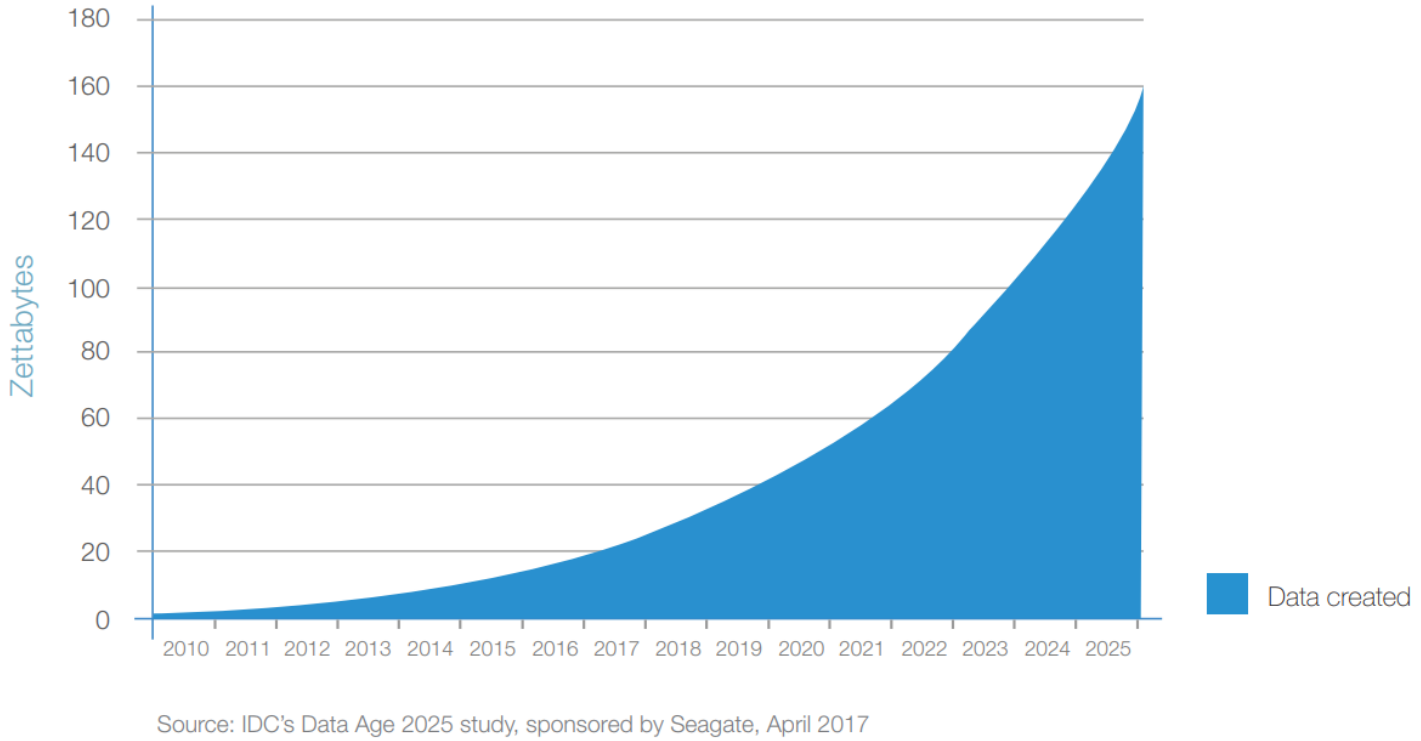
Northeastern University

- There are many different types of data
- Each requires its own tools and techniques
- Main categories of data:
  - Data matrix (e.g., spreadsheets)
  - Relational data (tables in a database)
  - Text documents (natural language)
  - Graph-based (e.g., social networks, world wide web)
  - Sequential data (e.g., genome sequences)
  - Multimedia (audio, video, images)
  - Streamed data (e.g., changing stock prices, stream of Twitter tweets)

Data Growth

Northeastern University

- The amount of raw data is increasing exponentially
- Data is eating the world: prediction is that 163 ZB will be created in 2025

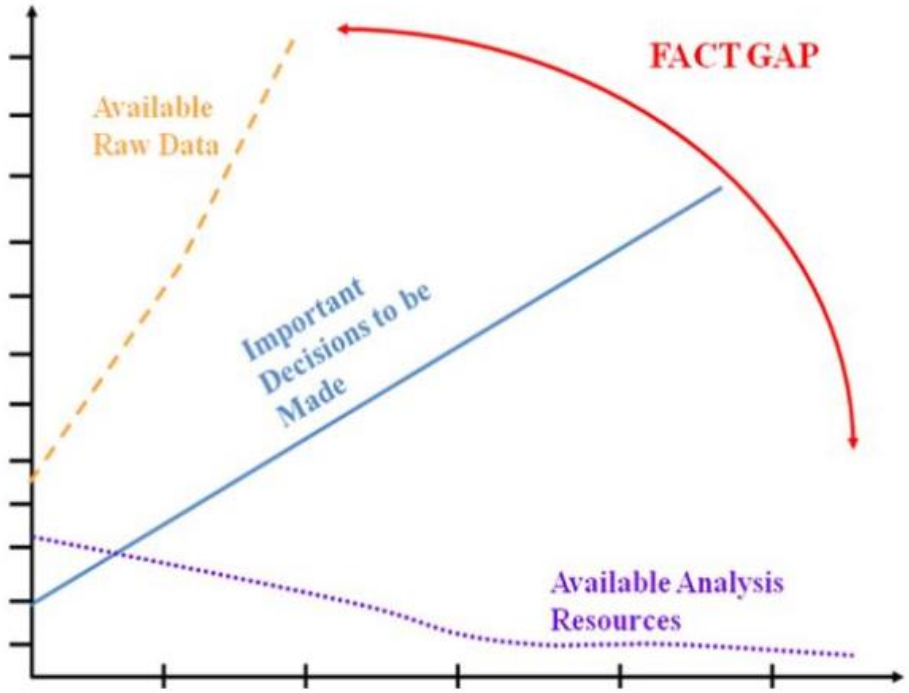


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

1 KB (Kilobyte)	10^3 bytes
1 MB (Megabyte)	10^6 bytes
1 GB (Gigabyte)	10^9 bytes
1 TB (Terabyte)	10^{12} bytes
1 PB (Petabyte)	10^{15} bytes
1 EB (Exabyte)	10^{18} bytes
1 ZB (Zettabyte)	10^{21} bytes
1 YB (Yottabyte)	10^{24} bytes

The Fact Gap

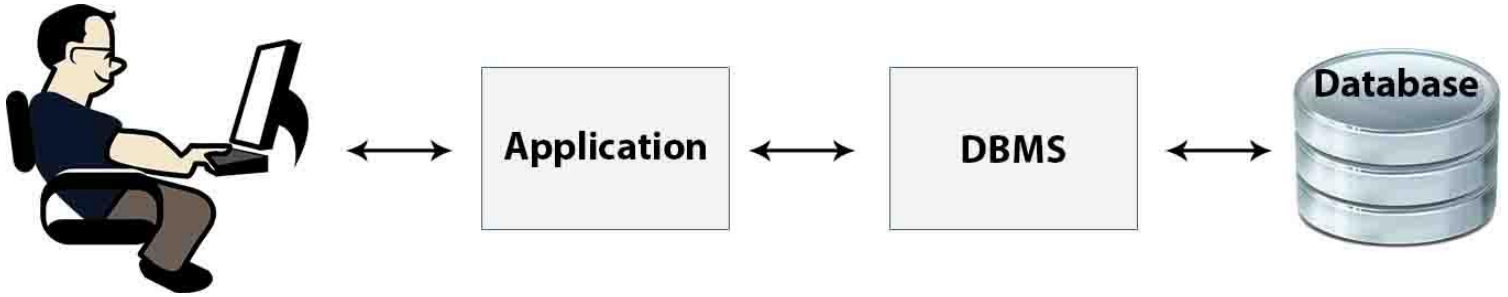
- The amount of raw data is increasing exponentially
- Growing number of decisions requiring the data
- But slower growth in resources available to analyze the data



Database and DBMS

Northeastern University

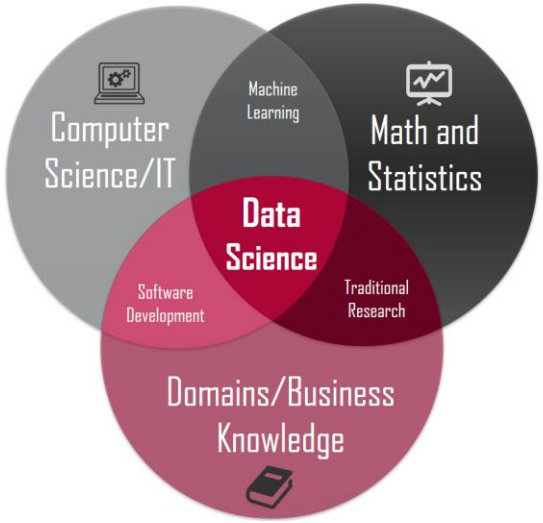
- A **database** is a an organized collection of structured data stored in a computer
- A **database management system** (DBMS) is software designed to store, retrieve, define, and manage data in a database
- DBMS provide users with an abstract view of the data
  - By hiding certain details of how the data are stored and maintained
- A **database system** includes the database, DBMS and associated applications



What is Data Science?

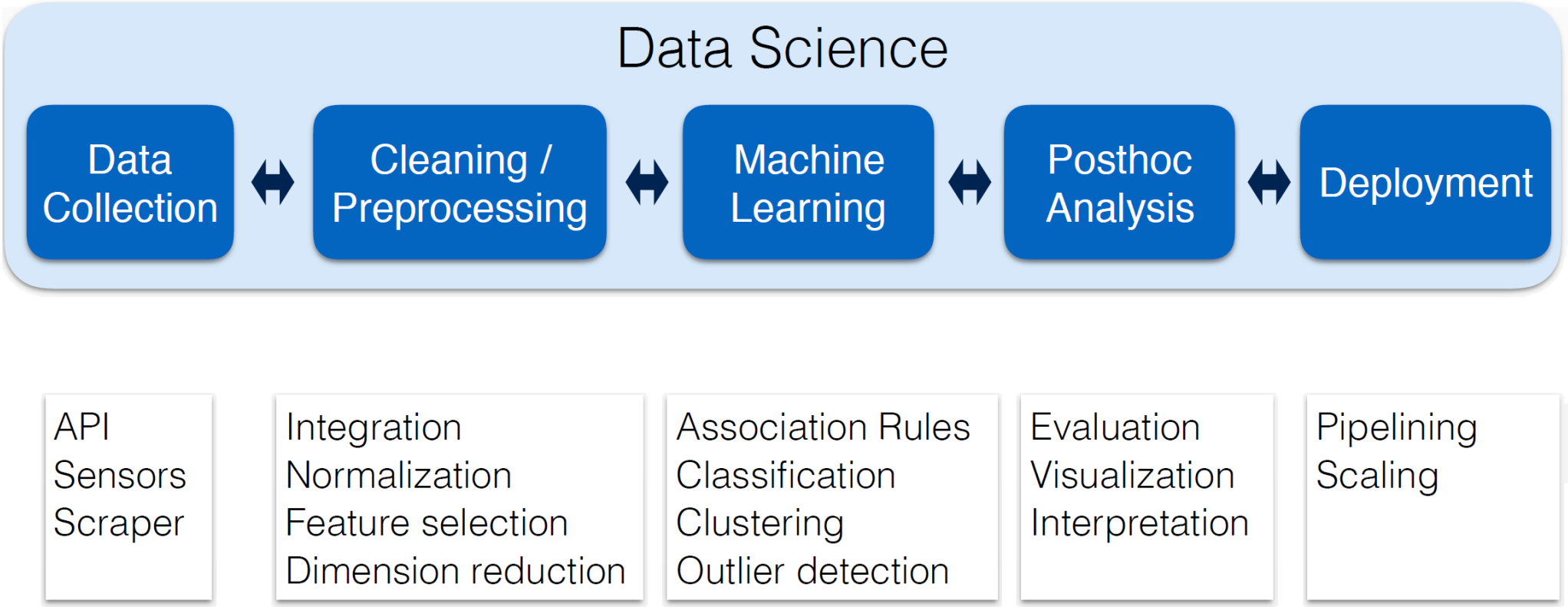
Northeastern University

- Data science is the discipline of the extraction of knowledge from data
- Data science relies on:
  - Computer science (for data structures, algorithms, visualization, big data support, and general programming)
  - Statistics (for regressions and inference)
  - Domain knowledge (for asking questions and interpreting results)



The Data Science Pipeline

Northeastern University



Data Collection

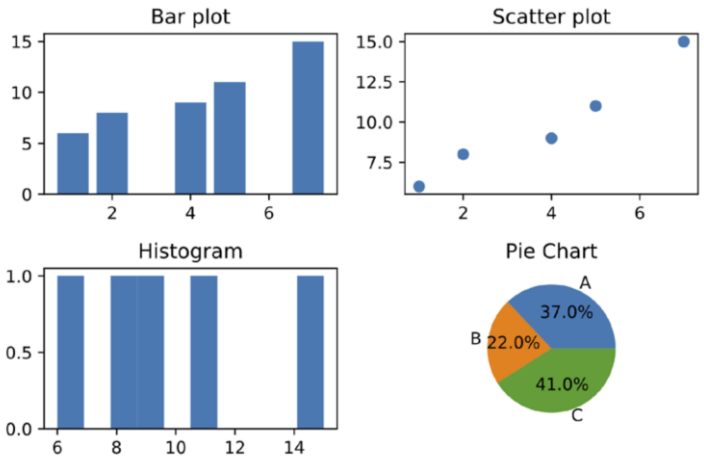
Northeastern University

- This step focuses on getting high-quality data
- The data may reside in a centralized repository or distributed across multiple sites
- The data may be owned by more than one organization
- The data may be stored in variety of formats (e.g., files, spreadsheets, relational DB)
- Data security and privacy might be an issue

Exploratory Data Analysis (EDA)

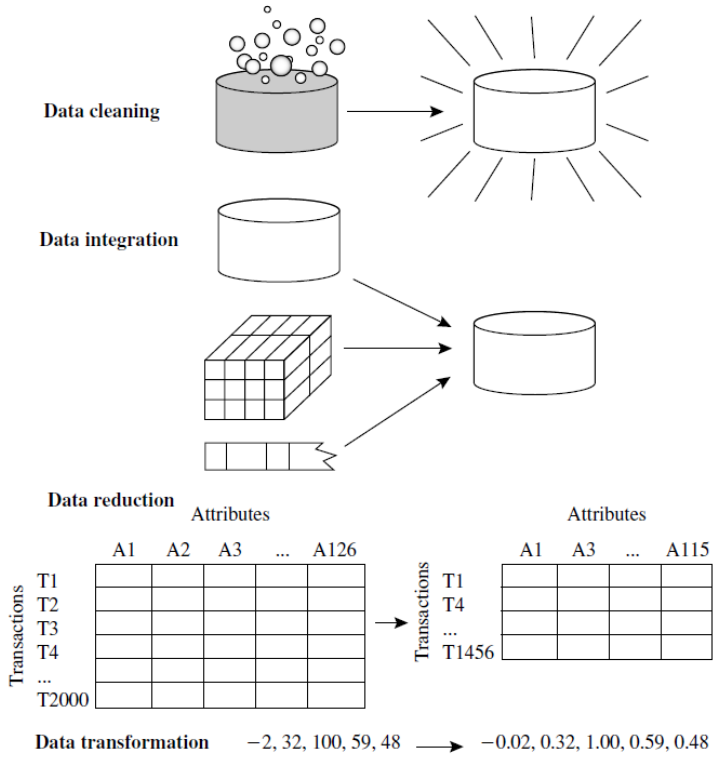
Northeastern University

- A preliminary exploration of the data is used to better understand its characteristics
- During this phase we look for patterns, correlations, and deviations based on visual and descriptive techniques
- Key motivations for data exploration include:
  - Helping to select the right tools for preprocessing and data analysis
  - Making use of humans' abilities to visually recognize patterns
- Techniques used in this phase:
  - Summary statistics
  - Visualization



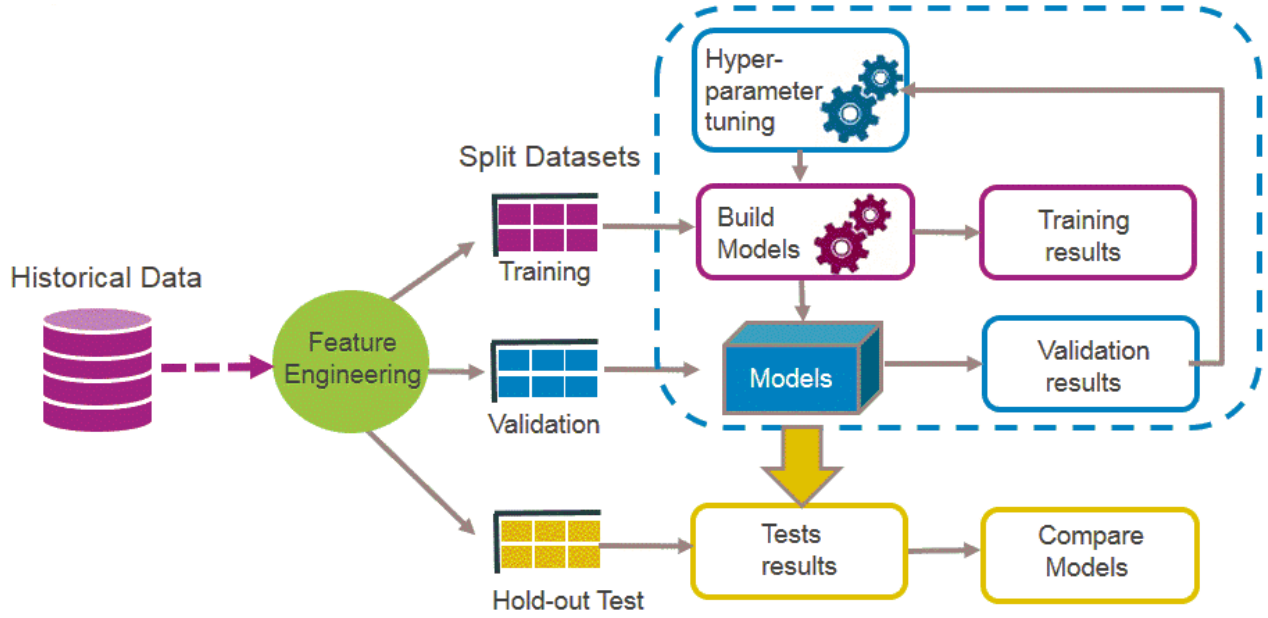
Data Preparation

- In data science there's a well-known saying: *Garbage in equals garbage out*
- Data preparation involves many steps to get the raw data ready for analysis:
  - Imputing missing values
  - Handling noisy data and outliers
  - Attribute transformation
  - Discretization
  - Normalization
  - Sampling
  - Feature selection
  - Dimensionality reduction
  - Feature extraction



Building the Model

- With clean data in place you're ready to build your data analysis model
- In this stage you'll use techniques from machine learning and statistics depending on the task at hand (classification, clustering, association discovery, etc.)
- A typical flow of building a model:



Interpretation and Evaluation

Northeastern University

- The final step integrates the data mining results into the decision support systems
- Visualization of the results allows analysts and decision makers to explore the data mining results from different view points
- Statistical measures or hypothesis testing methods are used to eliminate spurious data mining results

Challenges in Data Science

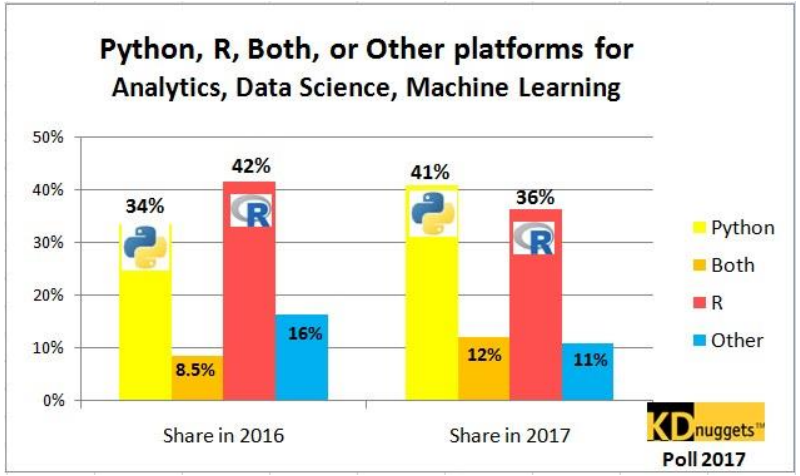
Northeastern University

- Large and high-dimensional data
  - Large number of variables (features/dimensions), number of instances
  - Multi terabyte databases
  - Need efficient algorithms, parallel / distributed computing
- Data quality
  - Missing and noisy data
- Complex and heterogeneous data
- Changing data
- Overfitting
  - Models adapt to the noise in training data, instead of finding the general patterns
- Privacy preservation
- Use of domain knowledge

What is Python?

Northeastern University

- Python is a powerful high-level, interpreted, object-oriented programming language
- It has a simple and easy-to-use syntax
- Portable across different platforms and operating systems
- Has a rich variety of native data structures such as lists and dictionaries
- It is one of the most popular programming languages used by data scientists



Python Code vs. C Code

Northeastern University

- Python code is typically 1/5 to 1/3 the size of equivalent C or Java code

Python example

```python
names = ["Isaac Newton", "Marie Curie", "Paul Dirac"]
for name in names:
    print(name)
```

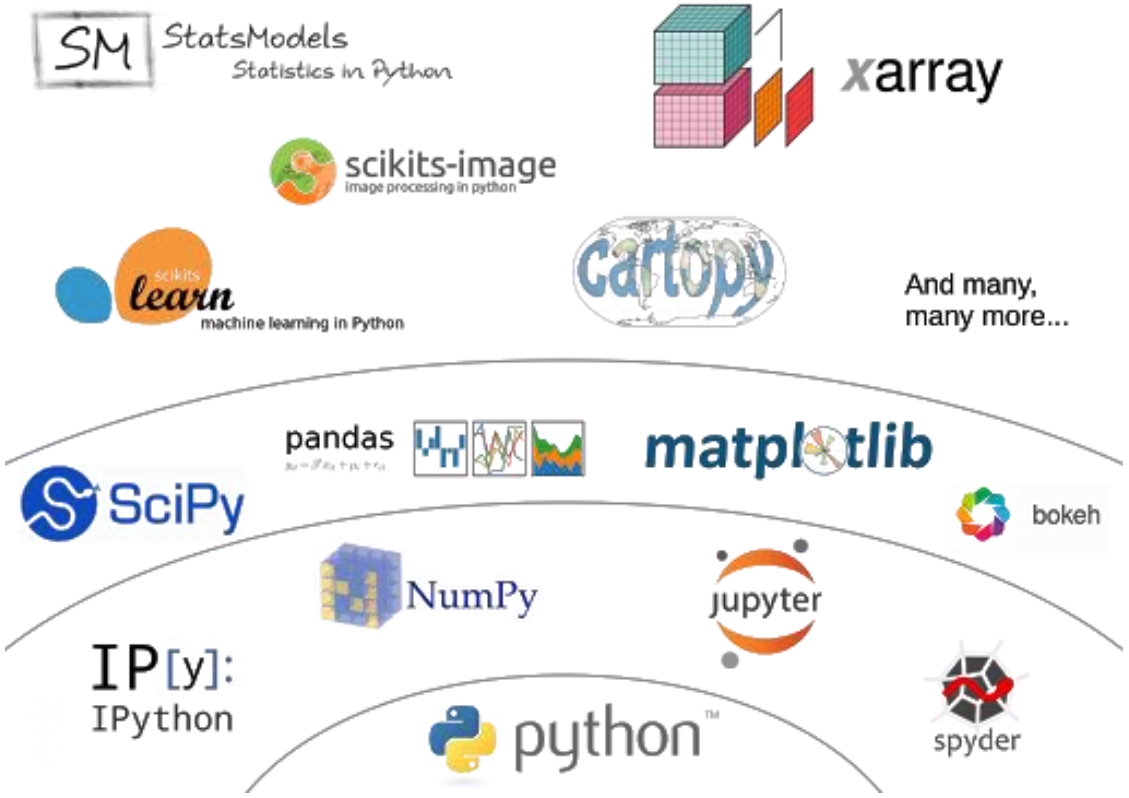
Same example in C

```c
#include <stdio.h>
#include <string.h>
#define MAX_STRING_LENGTH 20
#define NUMBER_OF_STRINGS 3

int main()
{
    char names[NUMBER_OF_STRINGS][MAX_STRING_LENGTH + 1];
    int i;

    strcpy(names[0], "Isaac Newton");
    strcpy(names[1], "Marie Curie");
    strcpy(names[2], "Paul Dirac");

    for (i = 0; i < NUMBER_OF_STRINGS; i++) {
        printf("%s\n", names[i]);
    }

    return 0;
}
```

Python for Data Science

Northeastern University

- Python provides high-performance and easy to use libraries for data science



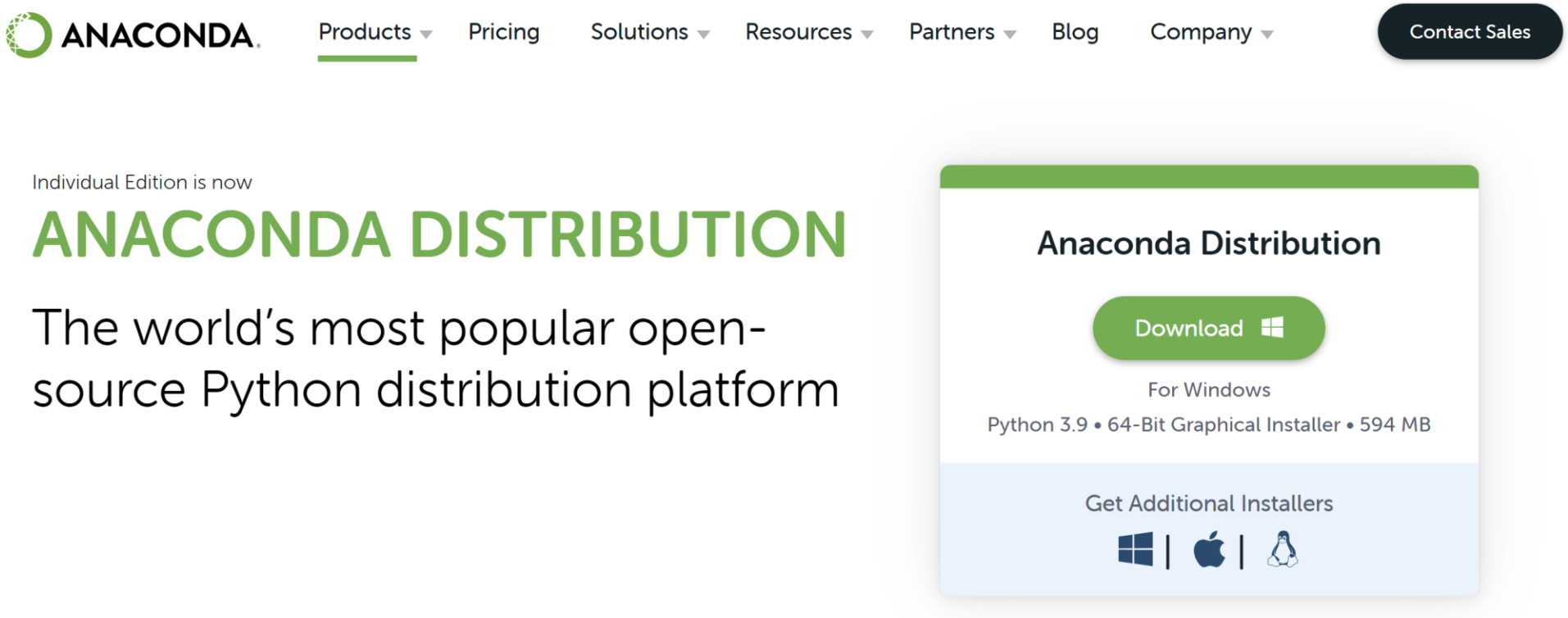
Anaconda Distribution

Northeastern University

- A **distribution of Python** is a bundle that contains an implementation of Python along with a bunch of libraries or tools
- Anaconda is the one of the most popular Python distributions
- It provides a wide variety of libraries for machine learning and data science
- Has a short and simple setup

Installing Anaconda

- Go to https://www.anaconda.com/download/
- Download the setup file

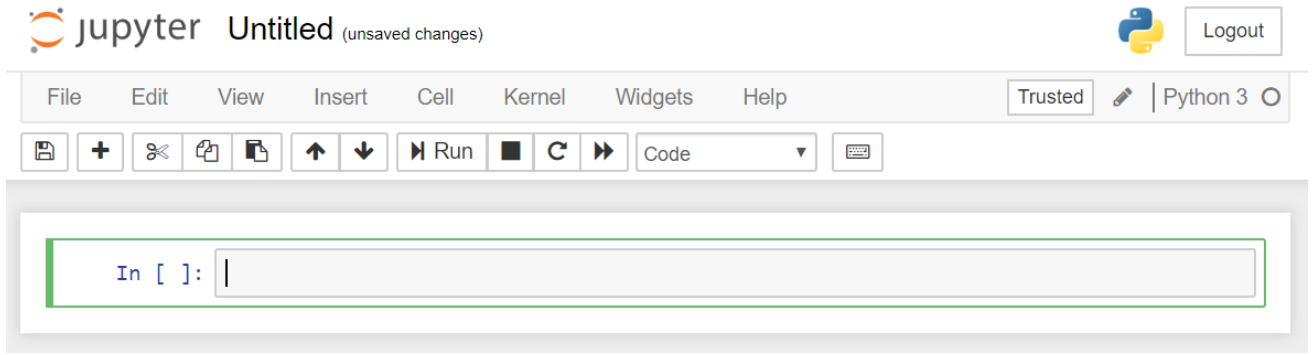


Jupyter Notebook

Northeastern University

jupyter

- An interactive environment for Python programming
- Runs within a web browser
- A notebook allows you to organize your program into a series of cells
- Each cell can contain Python code or other data such as text, plots, formulas, etc.
- Each cell may be executed individually
- Notebooks are saved in .ipynb files



Jupyter Notebook

Northeastern University

- Installation included with the Anaconda distribution
- To start Jupyter notebook type in the Command Prompt:

```
jupyter notebook
```

- How to change the Jupyter start-up folder
  - https://stackoverflow.com/questions/35254852/how-to-change-the-jupyter-start-up-folder

Python IDEs

Northeastern University

- An **IDE** (Integrated Development Environment) provides a rich set of features that make the developer's life easier, such as:
  - Debugger
  - Code profiling
  - Unit testing
  - Integration with version control systems like git
  - And many more
- Many IDEs exist for Python: PyCharm, VS Code, PyDev, Eclipse, Komodo, Spyder, …

pip

Northeastern University

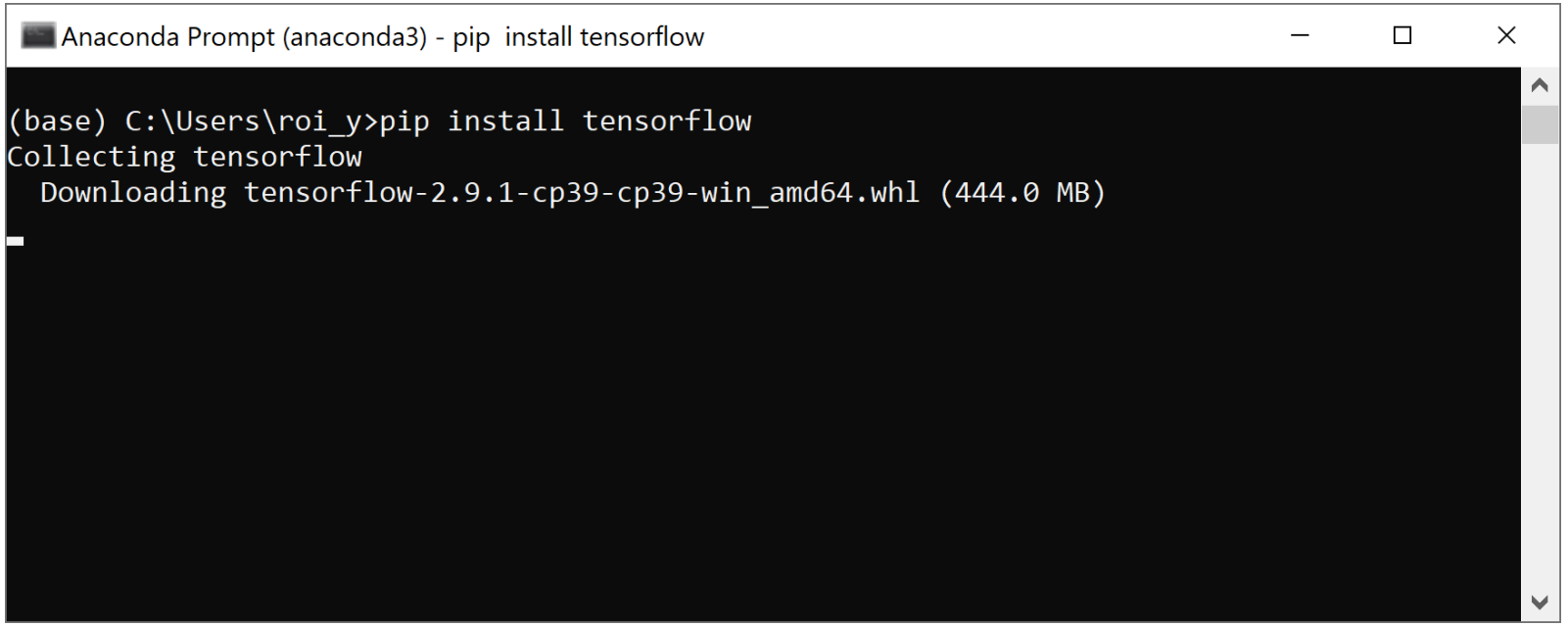
- pip is a Python package management system used to install and manage software packages written in Python
- Activated from the command prompt
- Useful pip commands:

Command	Description
pip install some-package	Install the latest version of a package
pip install some-package==1.4	Install a specific version of a package
pip install -u some-package	Upgrade an already installed package to the latest version
pip uninstall some-package	Uninstall a package
pip list	List all the packages installed
pip show some-package	Get information on an installed package

pip

Northeastern University

- For example, to install a new package use **pip install**:



PEP8

Northeastern University

- Style guide for writing Python code https://www.python.org/dev/peps/pep-0008/
- **Code is read much more often than it is written**
- The guidelines provided in PEP8 are intended to improve the readability of your code and make it consistent across the system
- Example for guidelines:
  - Surround assignment and comparison operators with a single space on either side
    - e.g., x = 10 and not x=10
  - Don’t use spaces around the = in keyword arguments
    - e.g., print(sep=‘,’) and not print(sep = ‘,’)
  - Use a maximum of 80 characters per line, if needed split long lines using \
  - Imports are always put at the top of the file

Recommended Textbooks

Northeastern University