What is Data Science?

- Fundamentally an **interdisciplinary** subject.
- Data science comprises **three distinct and overlapping areas:**
  - the skills of a **statistician** who knows how to model and summarize datasets (which are growing ever larger);
  - the skills of a **computer scientist** who can design and use algorithms to efficiently store, process, and visualize this data;
  - and the **domain expertise**—what we might think of as "classical" training in a subject—necessary both to formulate the right questions and to put their answers in context.

Think of **data science** not as a new domain of knowledge to learn.

But a new **set of skills** that you can apply within your current area of expertise. Whether you are:

- reporting election results,
- forecasting stock returns,
- optimizing online ad clicks,
- identifying microorganisms in microscope photos,
- seeking new classes of astronomical objects,
- or working with data in any other field.

Why Python?

- Python has emerged as a first-class tool for **scientific computing tasks**, including the **analysis and visualization of large datasets**.
- The language itself was **not specifically designed** with data analysis or scientific computing in mind.

- The usefulness of Python for data science stems from the **large and active ecosystem of third-party packages:**
- **NumPy** for manipulation of homogeneous array-based data,
- **Pandas** for manipulation of heterogeneous and labeled data,
- **SciPy** for common scientific computing tasks,
- **Matplotlib** for publication-quality visualizations,
- **IPython** for interactive execution and sharing of code,
- **Scikit-Learn** for machine learning,
- and **many more** tools.

IPython and Jupyter

- There are **many options for development environments** for Python.
- Here we use **IPython,** and in particular **Jupyter** which is based on IPython.

- If Python is the **engine** of our data science task, you might think of Jupyter as the **interactive control panel.**
- Perhaps the **most familiar interface** provided by the Jupyter project is the Jupyter Notebook, a **browser-based environment** that is useful for development, collaboration, sharing, and even publication of data science results.

> ***NOTE:*** Study Chapter 1, if you are not familiar with IPython and/or Jupyter.