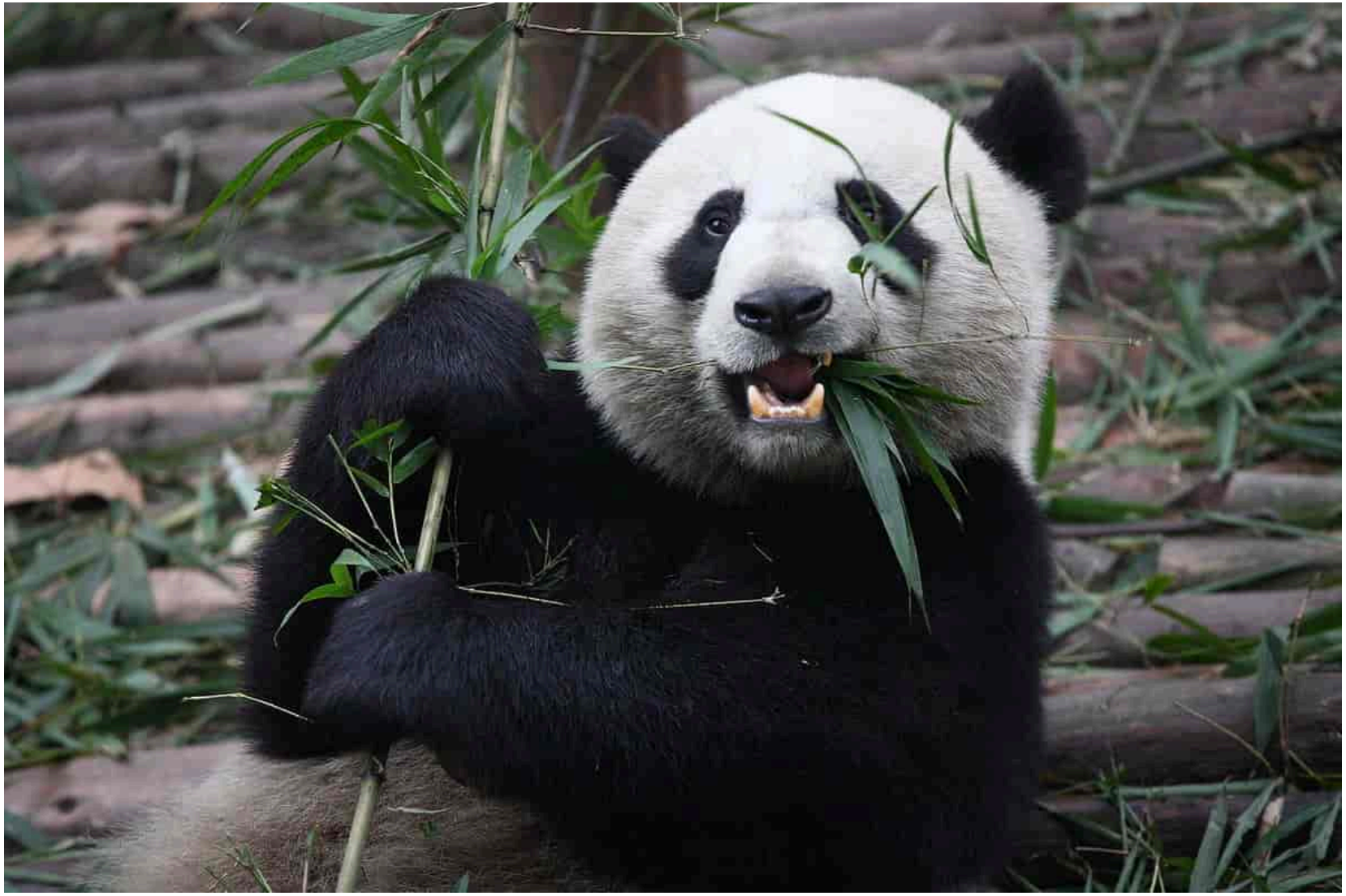
Session 2 - Pandas



What we will learn

- What is pandas?
- Pandas basics
  - Series, DataFrames
  - Accessing data
  - Data I/O
  - Changing the shape of data
  - Sorting
  - Combining data
- Exercises

Pandas

- Pandas is *the* Python package for data analysis
- High performance (-ish)
- Single machine (i.e. data sets should fit into RAM)
  - Alternative: Spark
- Supports many data formats
- Under active development
- We won't cover everything here
  - Notably: `Categorical`, `nullable int` data types.
  - Some deeper dives into e.g. `MultiIndex`
  - Some fine-print

Basic building blocks - Series

A `Series` is a collection of ‘measurements’ or data.

Temperature (°C)
21
25
19

Data in a series can be used in calculations, summarized, plotted etc. `Sereis` objects have an `Index` that determines how data can be accessed.

Basic building blocks - Data Frames

A `DataFrame` is a collection of `Series` (approximately).

Temperature (°C)	Cloud cover (%)
21	20
25	12
19	40

Installing Pandas

You should be fine with

```
pip install pandas
pip install matplotlib   # for visualization
```