# pandas!!!

# Two Months Salary

# What’s so special about diamonds?

Each one is said to be unique.

The Four C’s:

- Carat
- Color
- Clarity
- Cut

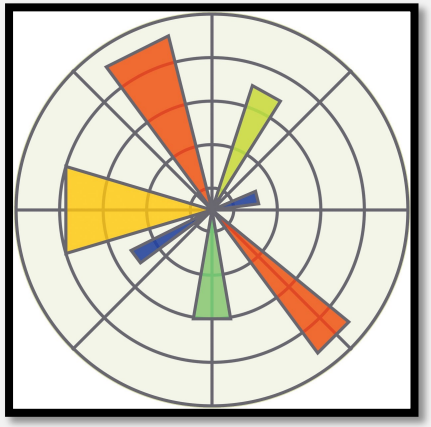
# What’s so special about diamonds?

The perfect gift to represent engagement.

Traditionally should cost two months salary.

# Data Analysis in Python

pandas and matplotlib



# Data Analysis in Python

Rarely conducted from scratch.

Generally faster and easier if done using open source packages.

# What is open source?

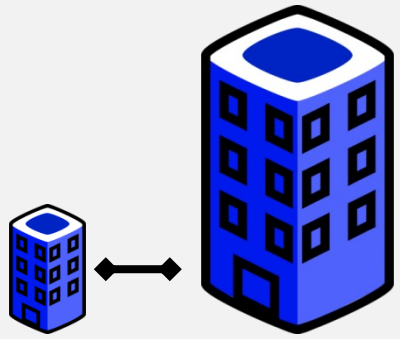
https://opensource.org/

# The Great Divide

**Pre-Business Intelligence**

Data was disorganized and rarely taken advantage of.

Lack of computer science technology was a major barrier.

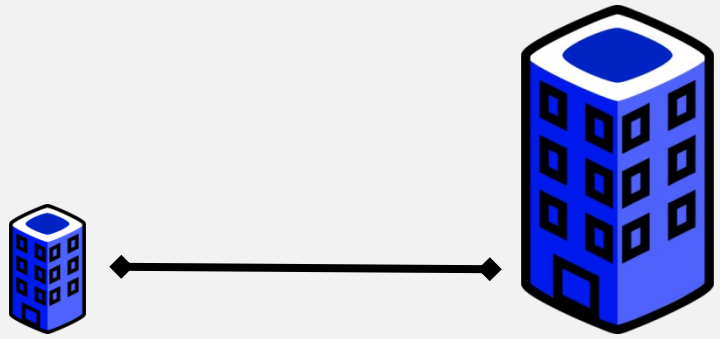


# The Great Divide

**Business Intelligence**

Large companies could afford expensive technological solutions.

Data became organized and adopters had a better picture as to what was going on.

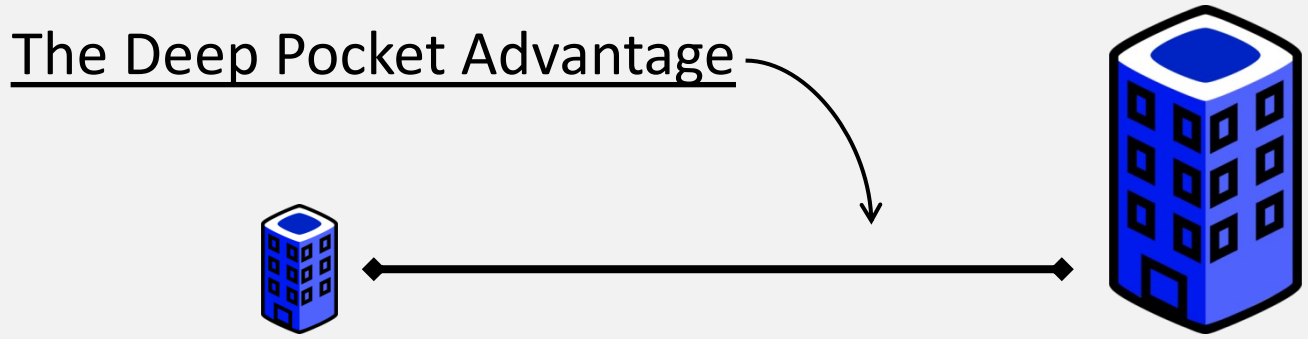


# The Great Divide

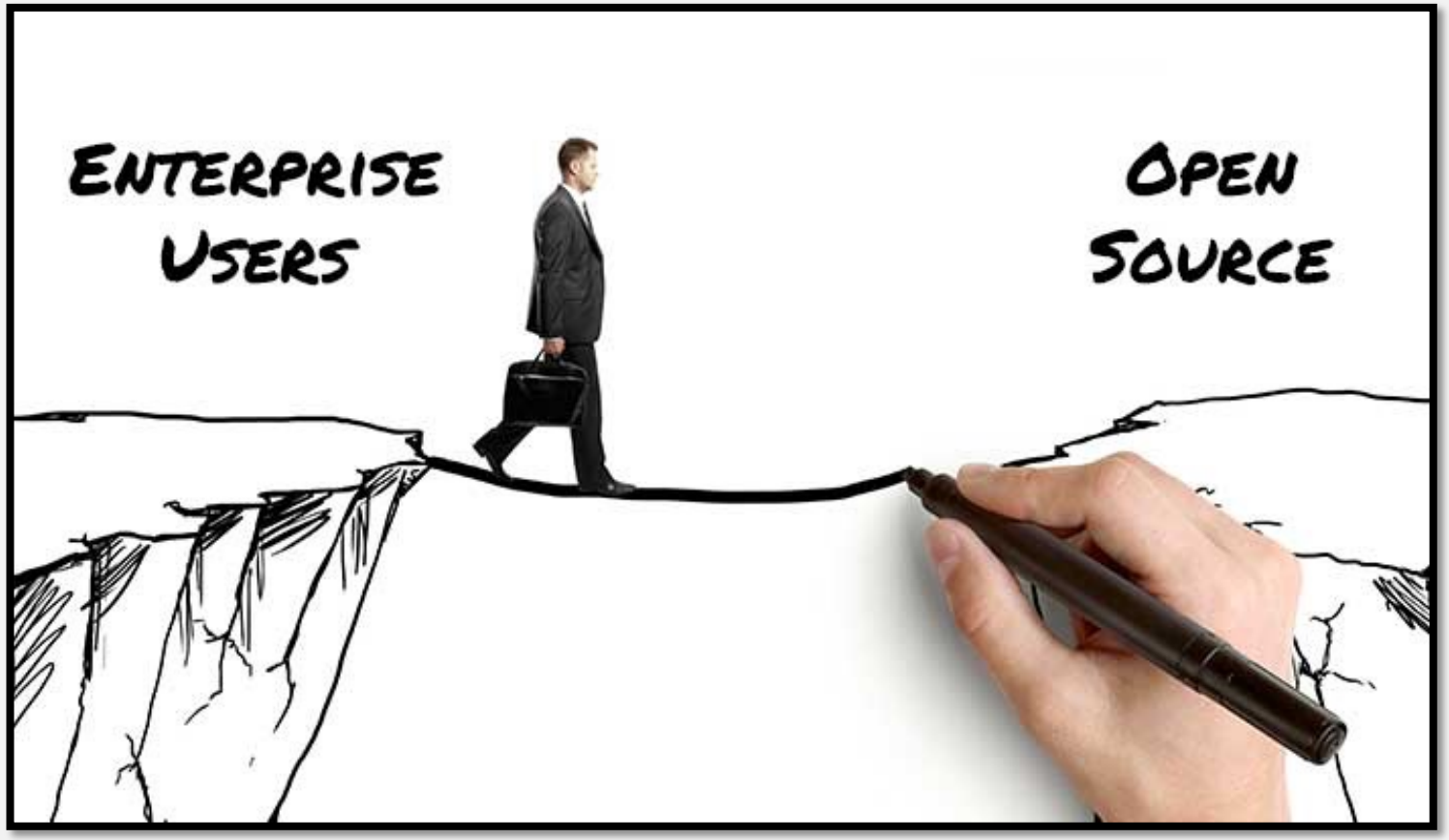
## Analytics

Data was disorganized and rarely taken advantage of.

Lack of computer science technology was a major barrier



# The Death of the Deep Pocket Advantage



# pandas

Data Science Essentials

# What is pandas?

World class open source project

Addresses Python’s need for data analysis

Home to the DataFrame, Python’s version of a spreadsheet

# pandas

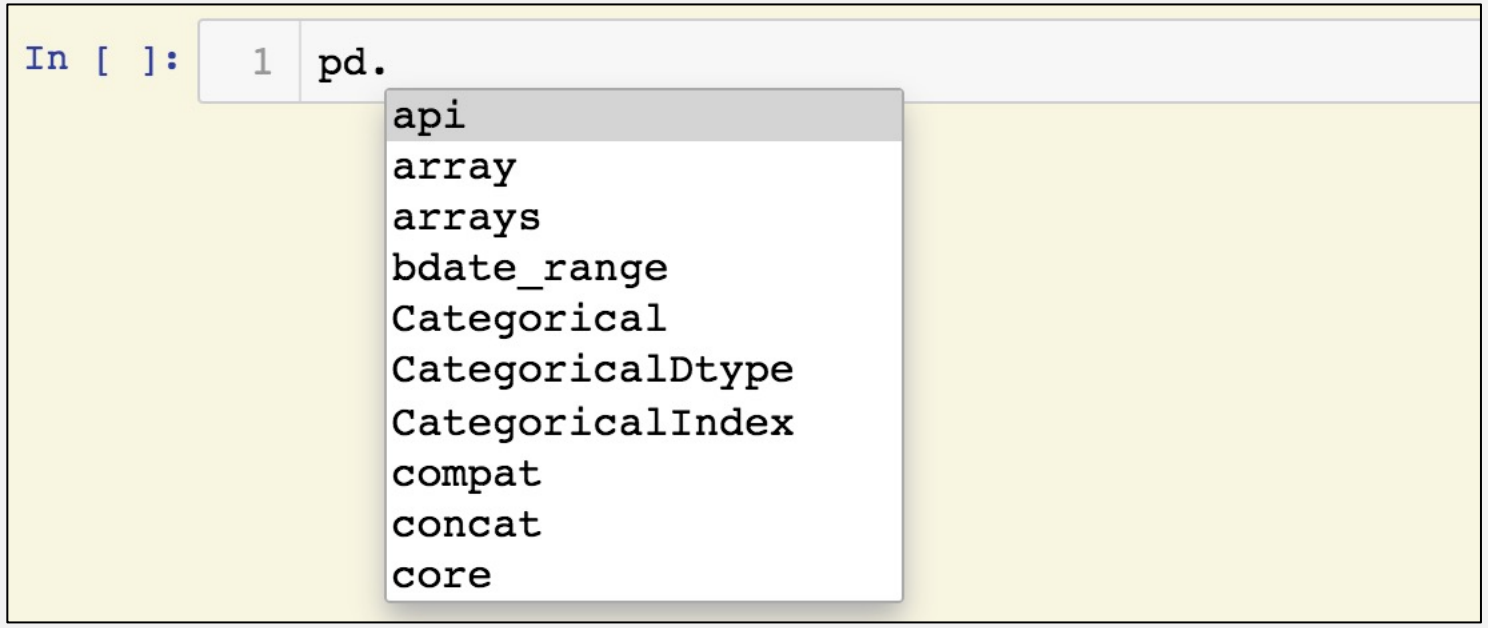
Generally imported as pd

```
In [ ]:  1  import pandas as pd
```

This makes it easier to reference pandas functions and sub-packages

# pandas

Jupyter Notebook will list available functions in ***most*** packages.



# pandas Practice

Accessing Packages and Optional Arguments

# Useful Techniques (1 of 2)

df (pd.DataFrame) is a generic way to say dataset

| Command | Description |
|---|---|
| df | accesses several rows and columns of a dataset |
| df.head() | accesses the first *n* rows in a dataset |
| df.tail() | accesses the last *n* rows in a dataset |
| df.shape() | accesses the dimensions of a dataset |
| df.columns | acesses the column names in a dataset |
| df.count() | counts the number of non-missing values in a dataset |

# Useful Techniques (2 of 2)

df (pd.DataFrame) is a generic way to say dataset

| Command | Description |
|---|---|
| df[df[col] > x] | accesses rows of a dataset that meet specified criteria |
| df. .sort_values([cols], ascending= True) | sorts the values of a given dataset |
| df.info() | accesses column names, number of non-missing values, and variable types |
| df.describe() | accesses the mean and quantiles of a dataset |
| df.loc[rows, cols] | accesses specified rows and columns of a dataset (numpy required) |
| df.iloc[row #’s, col #’s] | accesses specified rows and columns of a dataset (numpy required) |