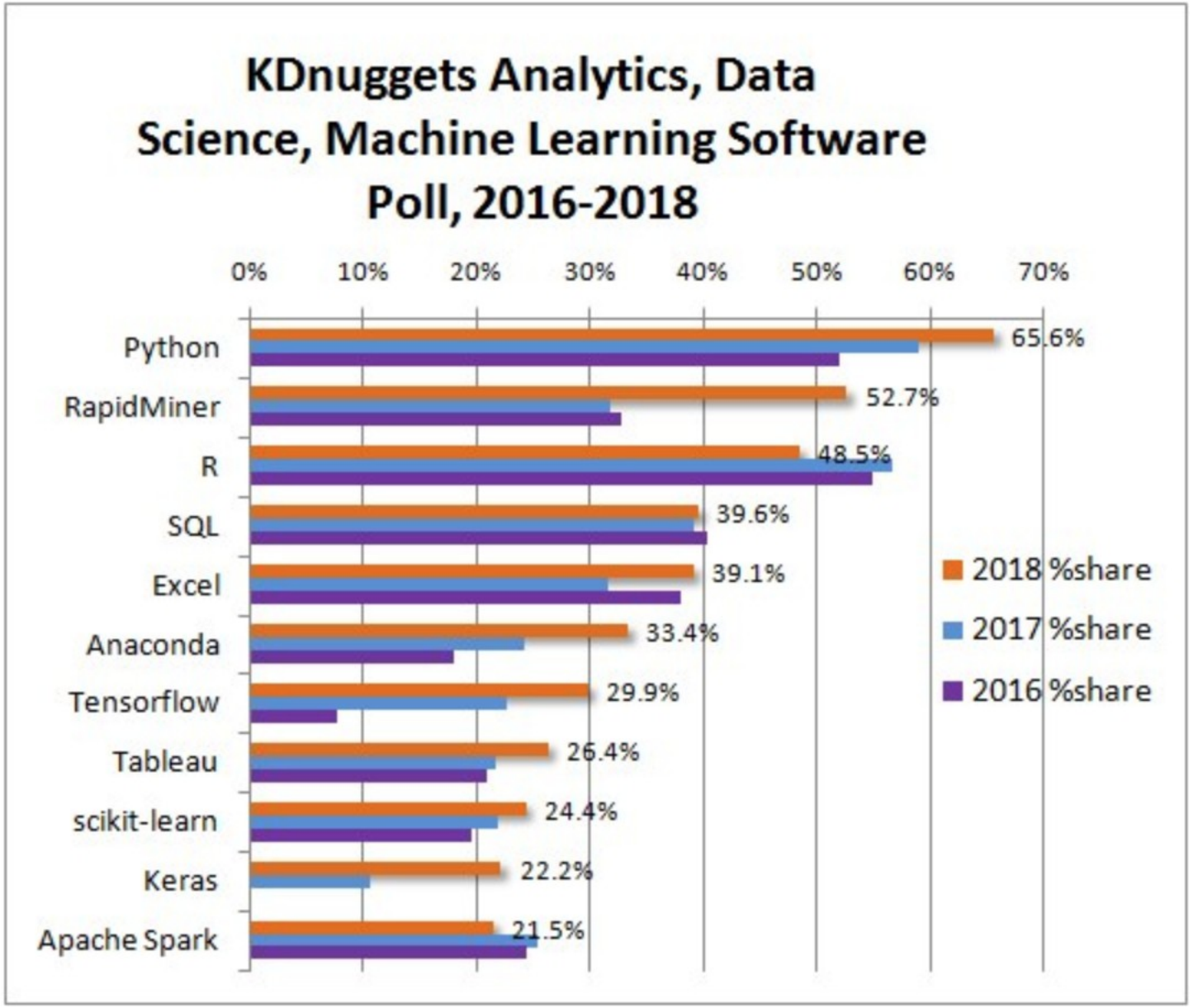

# Pandas and Jupyter Notebooks

Data Science in Python

# The 19th annual KDnuggets Software Poll

- KDnuggets is a leading site on AI, Analytics, Big Data, Data Mining, Data Science, and Machine Learning (according to KDnuggets)
- Participants on average chose ~7 tools
- Python has grown over the last few years, while R seemed to decline from 2017 to 2018

What makes Python such a great tool for data science?



# pandas

*“pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.”*

## Pandas DataFrame

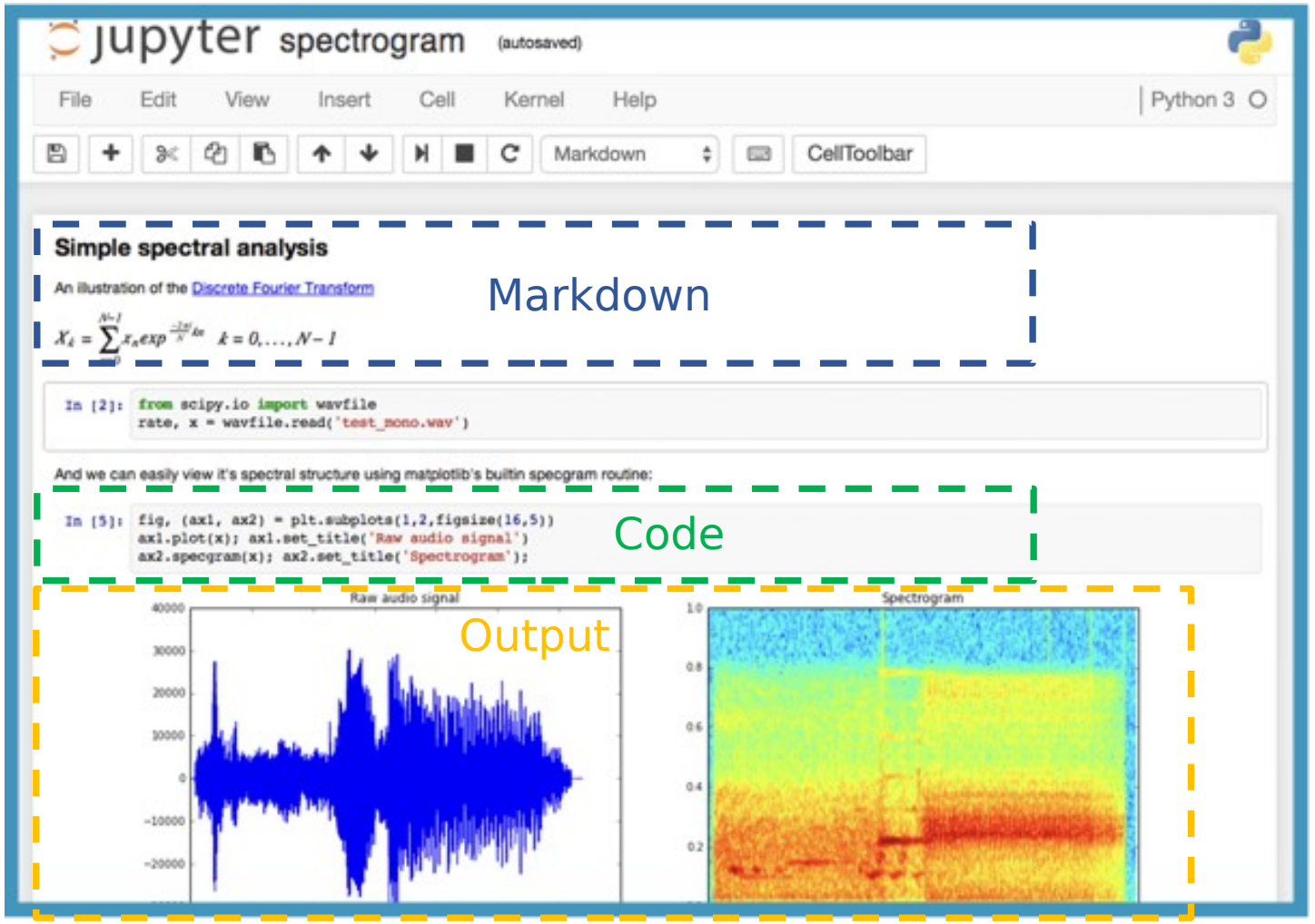
| | account | campaign | date | successes | trials | rate |
|---|---|---|---|---|---|---|
| **455** | 1 | Campaign #76 | 2012-08-14 11:56:20 -0400 | 2 | 2 | 1.000000 |
| **449** | 1 | Campaign #78 | 2012-08-14 12:06:20 -0400 | 2 | 2 | 1.000000 |
| **438** | 1 | Campaign #87 | 2012-08-14 18:06:30 -0400 | 27 | 118 | 0.228814 |
| **431** | 1 | Campaign #95 | 2012-08-15 00:07:42 -0400 | 22 | 118 | 0.186441 |
| **422** | 1 | Campaign #99 | 2012-08-15 01:27:48 -0400 | 25 | 120 | 0.208333 |

## Functionality

- Read and write data sets of common types: CSV, text files, Microsoft Excel, and SQL databases.
- Merging and joining of datasets
- Slicing, indexing and subsetting large sets of data
- Groupby engine for aggregating and transforming data
- Optimized for performance, critical code paths written in Cython or C

# Jupyter Notebooks

- Open source web application that allows you to create and share documents that contain code, equations, visualizations, and explainatory text.
- REPL (A Read–Eval–Print Loop) is an interactive and efficient way to find errors in your code and get descriptive feedback fast in Python
- Instantaneous feedback allows for high-speed development
- Robust data visualization



# Live Demo