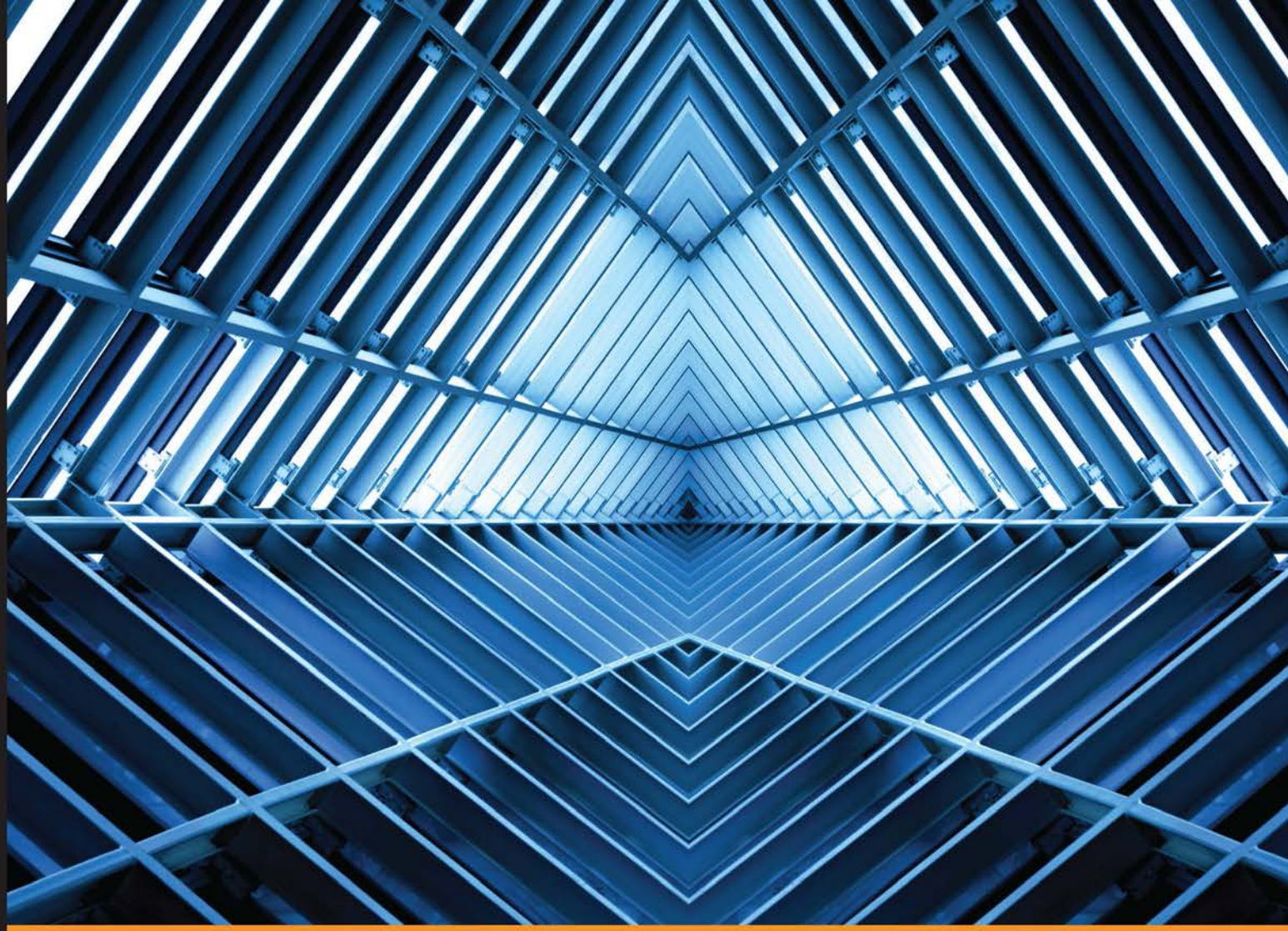

Community Experience Distilled

Learning IPython for Interactive Computing and Data Visualization

Second Edition

Get started with Python for data analysis and numerical computing in the Jupyter notebook

Cyrille Rossant

[PACKT] PUBLISHING open source* community experience distilled

Learning IPython for Interactive Computing and Data Visualization

Second Edition

Get started with Python for data analysis and numerical computing in the Jupyter notebook

Cyrille Rossant

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

Learning IPython for Interactive Computing and Data Visualization

Second Edition

First published: April 2013

Second edition: October 2015

Production reference: 1151015

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-698-9

www.packtpub.com

Credits

Author
Cyrille Rossant

Reviewers
Damián Avila
Nicola Rainiero
G Scott Stukey

Commissioning Editor
Kartikey Pandey

Acquisition Editors
Kartikey Pandey
Richard Brookes-Bland

Content Development Editor
Arun Nadar

Technical Editor
Pranil Pathare

Copy Editor
Stephen Copestake

Project Coordinator
Shweta H Birwatkar

Proofreader
Safis Editing

Indexer
Monica Ajmera Mehta

Production Coordinator
Conidon Miranda

Cover Work
Conidon Miranda

About the Author

Cyrille Rossant is a researcher in neuroinformatics, and is a graduate of Ecole Normale Superieure, Paris, where he studied mathematics and computer science. He has worked at Princeton University, University College London, and College de France. As part of his data science and software engineering projects, he gained experience in machine learning, high-performance computing, parallel computing, and big data visualization.

He is one of the main developers of VisPy, a high-performance visualization package in Python. He is the author of the *IPython Interactive Computing and Visualization Cookbook*, *Packt Publishing*, an advanced-level guide to data science and numerical computing with Python, and the sequel of this book.

I am grateful to Nick Fiorentini for his help during the revision of the book. I would also like to thank my family and notably my wife Claire for their support.

About the Reviewers

Damián Avila is a software developer and data scientist (formerly a biochemist) from Córdoba, Argentina.

His main focus of interest is data science, visualization, finance, and IPython/Jupyter-related projects.

In the open source area, he is a core developer for several interesting and popular projects, such as IPython/Jupyter, Bokeh, and Nikola. He has also started his own projects, being RISE, an extension to enable amazing live slides in the Jupyter notebook, the most popular one. He has also written several tutorials about the Scientific Python tools (available at Github) and presented several talks at international conferences.

Currently, he is working at Continuum Analytics.

Nicola Rainiero is a civil geotechnical engineer with a background in the construction industry as a self-employed designer engineer. He is also specialized in the renewable energy field and has collaborated with the Sant'Anna University of Pisa for two European projects, REGEOCITIES and PRISCA, using qualitative and quantitative data analysis techniques.

He has an ambition to simplify his work with open software and use and develop new ones; sometimes obtaining good results, at other times, negative. You can reach Nicola on his website at `http://rainnic.altervista.org`.

A special thanks to Packt Publishing for this opportunity to participate in the reviewing of this book. I thank my family, especially my parents, for their physical and moral support.

---

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

Data analysis skills are now essential in scientific research, engineering, finance, economics, journalism, and many other domains. With its high accessibility and vibrant ecosystem, Python is one of the most appreciated open source languages for data science.

This book is a beginner-friendly introduction to the Python data analysis platform, focusing on IPython (Interactive Python) and its Notebook. While IPython is an enhanced interactive Python terminal specifically designed for scientific computing and data analysis, the Notebook is a graphical interface that combines code, text, equations, and plots in a unified interactive environment.

The first edition of *Learning IPython for Interactive Computing and Data Visualization* was published in April 2013, several months before the release of IPython 1.0. This new edition targets IPython 4.0, released in August 2015. In addition to reflecting the novelties of this new version of IPython, the present book is also more accessible to non-programmer beginners. The first chapter contains a brand new crash course on Python programming, as well as detailed installation instructions.

Since the first edition of this book, IPython's popularity has grown significantly, with an estimated user base of several millions of people and ongoing collaborations with large companies like Microsoft, Google, IBM, and others. The project itself has been subject to important changes, with a refactoring into a language-independent interface called the Jupyter Notebook, and a set of backend kernels in various languages. The Notebook is no longer reserved to Python; it can now also be used with R, Julia, Ruby, Haskell, and many more languages (50 at the time of this writing!).

The Jupyter project has received significant funding in 2015 from the Leona M. and Harry B. Helmsley Charitable Trust, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation, which will allow the developers to focus on the growth and maturity of the project in the years to come.

Here are a few references:

- Home page for the Jupyter project at `http://jupyter.org/`
- Announcement of the funding for Jupyter at `https://blog.jupyter.org/2015/07/07/jupyter-funding-2015/`
- Detail of the project's grant at `https://blog.jupyter.org/2015/07/07/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science/`

What this book covers

Chapter 1, *Getting Started with IPython*, is a thorough and beginner-friendly introduction to Anaconda (a popular Python distribution), the Python language, the Jupyter Notebook, and IPython.

Chapter 2, *Interactive Data Analysis with pandas*, is a hands-on introduction to interactive data analysis and visualization in the Notebook with pandas, matplotlib, and seaborn.

Chapter 3, *Numerical Computing with NumPy*, details how to use NumPy for efficient computing on multidimensional numerical arrays.

Chapter 4, *Interactive Plotting and Graphical Interfaces*, explores many capabilities of Python for interactive plotting, graphics, image processing, and interactive graphical interfaces in the Jupyter Notebook.

Chapter 5, *High-Performance and Parallel Computing*, introduces the various techniques you can employ to accelerate your numerical computing code, namely parallel computing and compilation of Python code.

Chapter 6, *Customizing IPython*, shows how IPython and the Jupyter Notebook can be extended for customized use-cases.

What you need for this book

The following software is required for the book:

- Anaconda with Python 3
- Windows, Linux, or OS X can be used as a platform

Who this book is for

This book targets anyone who wants to analyze data or perform numerical simulations of mathematical models.

Since our world is becoming more and more data-driven, knowing how to analyze data effectively is an essential skill to learn. If you're used to spreadsheet programs like Microsoft Excel, you will appreciate Python for its much larger range of analysis and visualization possibilities. Knowing this general-purpose language will also let you share your data and analysis with other programs and libraries.

In conclusion, this book will be useful to students, scientists, engineers, analysts, journalists, statisticians, economists, hobbyists, and all data enthusiasts.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "Run it with a command like `bash Anaconda3-2.3.0-Linux-x86_64.sh` (if necessary, replace the filename by the one you downloaded)."

A block of code is set as follows:

```
def load_ipython_extension(ipython):
    """This function is called when the extension is loaded.
    It accepts an IPython InteractiveShell instance.
    We can register the magic with the `register_magic_function`
    method of the shell instance."""
    ipython.register_magic_function(cpp, 'cell')
```

Any command-line input or output is written as follows:

```
$ python
Python 3.4.3 |Anaconda 2.3.0 (64-bit)| (default, Jun  4 2015, 15:29:08)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-1)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "To create a new notebook, click on the **New** button, and select **Notebook (Python 3)**."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.
You can also report any issues at `https://github.com/ipython-books/minibook-2nd-code/issues`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you. You will also find the book's code on this GitHub repository: `https://github.com/ipython-books/minibook-2nd-code`.

Downloading the color images of this book

We also provide you with a PDF file that has color images of the screenshots/diagrams used in this book. The color images will help you better understand the changes in the output. You can download this file from `https://www.packtpub.com/sites/default/files/downloads/6989OS_ColouredImages.pdf`.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.

1
Getting Started with IPython

In this chapter, we will cover the following topics:

- What are Python, IPython, and Jupyter?
- Installing Python with Anaconda
- Introducing the Notebook
- A crash course on Python
- Ten Jupyter/IPython essentials

What are Python, IPython, and Jupyter?

Python is an open source general-purpose language created by Guido van Rossum in the late 1980s. It is widely-used by system administrators and developers for many purposes: for example, automating routine tasks or creating a web server. Python is a flexible and powerful language, yet it is sufficiently simple to be taught to school children with great success.

In the past few years, Python has also emerged as one of the leading open platforms for data science and high-performance numerical computing. This might seem surprising as Python was not originally designed for scientific computing. Python's interpreted nature makes it much slower than lower-level languages like C or Fortran, which are more amenable to number crunching and the efficient implementation of complex mathematical algorithms.

However, the performance of these low-level languages comes at a cost: they are hard to use and they require advanced knowledge of how computers work. In the late 1990s, several scientists began investigating the possibility of using Python for numerical computing by interoperating it with mainstream C/Fortran scientific libraries. This would bring together the ease-of-use of Python with the performance of C/Fortran: the dream of any scientist!

Consequently, the past 15 years have seen the development of widely-used libraries such as NumPy (providing a practical array data structure), SciPy (scientific computing), matplotlib (graphical plotting), pandas (data analysis and statistics), scikit-learn (machine learning), SymPy (symbolic computing), and Jupyter/IPython (efficient interfaces for interactive computing). Python, along with this set of libraries, is sometimes referred to as the SciPy stack or PyData platform.

Competing platforms

Python has several competitors. For example, MATLAB (by Mathworks) is a commercial software focusing on numerical computing that is widely-used in scientific research and engineering. SPSS (by IBM) is a commercial software for statistical analysis. Python, however, is free and open source, and that's one of its greatest strengths. Alternative open source platforms include R (specialized in statistics) and Julia (a young language for high-performance numerical computing).

More recently, this platform has gained popularity in other non-academic communities such as finance, engineering, statistics, data science, and others.

This book provides a solid introduction to the whole platform by focusing on one of its main components: Jupyter/IPython.

Jupyter and IPython

IPython was created in 2001 by Fernando Perez (the *I* in *IPython* stands for "interactive"). It was originally meant to be a convenient command-line interface to the scientific Python platform. In scientific computing, trial and error is the rule rather than the exception, and this requires an efficient interface that allows for *interactive* exploration of algorithms, data, and graphs.

In 2011, IPython introduced the interactive **Notebook**. Inspired by commercial software such as Maple (by Maplesoft) or Mathematica (by Wolfram Research), the Notebook runs in a browser and provides a unified web interface where code, text, mathematical equations, plots, graphics, and interactive graphical controls can be combined into a single document. This is an ideal interface for scientific computing. Here is a screenshot of a notebook: