

ANACONDA WHITEPAPER

HARNESSING OPEN DATA SCIENCE FOR PREDICTIVE ANALYTICS

By: Christine Doig, Data Scientist and Product Marketing Manager
January 2016



IN THIS WHITEPAPER

Open Data Science is an inclusive movement that makes the open source tools of data science – data, analytics and computation – easily work together as a connected ecosystem. However, not all open source data science tools and platforms fully embrace the open connectedness of this movement. Many existing open source technologies are isolated and lack the interoperability with the underlying operational infrastructure foundation, but that interoperability is key to success in building, sharing and deploying predictive analytics in enterprises.

Anaconda, a modern open source analytics platform powered by Python, is the leading full-stack Open Data Science platform. This means Anaconda addresses difficult and complex operational problems while providing Data Science teams with the power of the latest innovations in open source analytics. Anaconda also makes it easy to collaborate across the entire Data Science team, no matter where in the world members may be located.

In this paper, you'll learn how Anaconda enables powerful predictive analytic solutions for enterprises, including:

- Managing operational issues easily
- Creating predictive analytic models with Python, R and Jupyter Notebooks
- Integrating your predictive analytic models into intelligent web apps, interactive visualizations or embedding them into any existing operational process

GETTING STARTED AND TAMING THE ANARCHY

While one of the most powerful values of open source is that innovation is constant, it can be a tough challenge for enterprises. IT organizations were established when proprietary vendors produced new versions every year. But in today's open source world, changes occur continuously. So, how can organizations harness the power of open source innovation, while taming the fluidity of that continuous change?

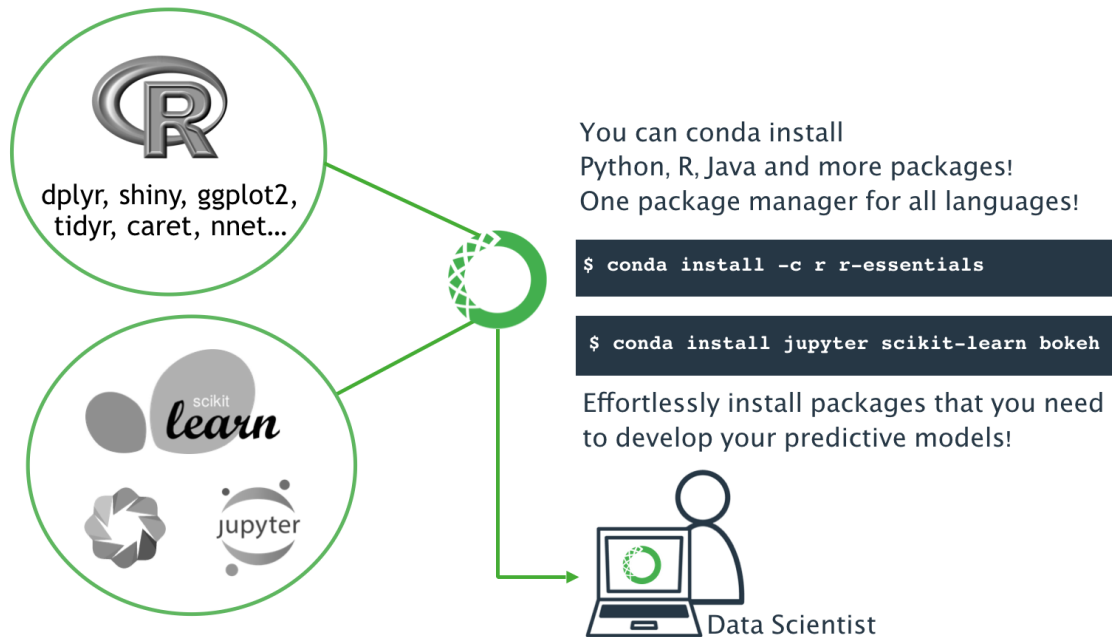
Enter conda, the powerful Anaconda package dependency and environment manager. Conda is a one stop shop to manage all your open source software:

- Across platforms – Windows, Linux, OS X
- Across languages – R, Python, Scala, Java, C/C++, Fortran and more

Conda makes it easy for Data Scientists, Developers and DevOps to install packages with a single command. Conda automatically identifies all the package dependencies and installs them. This happens in minutes rather than in the hours or days that it typically takes to install packages.

While Anaconda comes pre-loaded with 100+ commonly used Python packages, it can also access a growing repository containing more than 700 additional Python and R packages certified to work with Anaconda. By default, Anaconda pulls these packages from a cloud repository, but that repository can also be

mirrored behind your firewall. Your own proprietary or third party packages can also be added to the repository. For Data Scientists, Developers and DevOps wanting lighter weight installations, we offer Miniconda, which just installs Python and conda, allowing full customization of the packages in your environment.

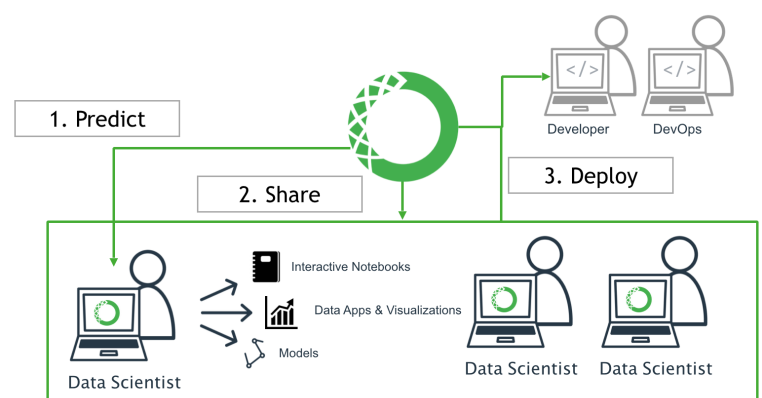


EXPLOITING THE POWER OF OPEN DATA SCIENCE FOR PREDICTIVE ANALYTICS

Data Scientists use various methods to create predictive analytic models. Some choose to write exclusively in a programming language such as Python, R or C/C++ via the command line or IDE (interactive development environment). Some prefer to use a visual-based interface to drag and drop to create analytic pipelines. Others today prefer to use Notebooks, which are combinations of code, narrative and visualizations. By definition, Open Data Science does not prescribe how to build a predictive model, but few platforms fully embrace this inclusiveness. Anaconda supports Data Scientists of all stripes – from the most seasoned to the newcomer – allowing them to select their tool of choice based on their organization, team or individual preference. Anaconda supports creating predictive models using Jupyter Notebooks, R Studio, Microsoft Excel, Orange Data Mining, the command line and any other Data Science workbench that supports Python. This range of diversity to build predictive analytics is absolutely

necessary not only today, but also into the future as the analytic model building tools continue to evolve. Anaconda, as the leader in Open Data Science, will continue to evolve with them.

Over the past few years, analytics innovation has moved into the fast lane. Researchers and academics now immediately



open source their algorithms. This acceleration of innovation, available only in open source, empowers Data Scientists to use any combination of tried and true methods along with the latest innovations to achieve and, oftentimes, supersede the goals of their analysis. With Anaconda, Data Scientists can leverage legacy algorithms written in Fortran and C/C++, along with latest innovations in Python, R, Scala and Java to create valuable predictive models for their enterprises. By standing on the shoulders of giants that came before them, Data Scientists today can create more powerful and valuable predictive models for their enterprises.

While Data Scientists still use hypothesis testing, they also increasingly need to use fast-to-fail techniques to quickly eliminate possibilities and narrow in on potential solutions that deliver fast time-to-value for their enterprises. In either approach, but especially crucial to fast-to-fail, collaboration is the key to sharing initial results and soliciting feedback from the entire Data Science team including the line of business, business analysts, DevOps, IT, Developers and other key stakeholders. Some teams use dashboards and standard plots to convey initial results, while Data Scientists are increasingly using powerful, browser-based interactive visualizations, often embedded into Notebooks, to publish and solicit feedback from the Data Science team. Anaconda enables the team to use any of these approaches for collaboration, including enterprise notebooks with revision control, difference identification, plus powerful tag and index searching.

Anaconda goes beyond standard plots and charts to deliver slideshows or contextually rich interactive visualizations that significantly enhance the meaning and interpretation of the data and results. This allows the entire team to explore and develop a deeper understanding until the solution meets the business goals. These intuitive and easy ways to collaborate streamline the work and enhance the productivity of the entire Data Science team.








ENCAPSULATING MAKES TESTING AND DEPLOYMENT A SNAP

Once the Data Scientists are ready to deploy the solution, they no longer have to “toss it over the wall” to Developers or DevOps. Developers and DevOps have already been involved in the collaboration process and the predictive model is easily packaged using conda. Conda encapsulates all the work necessary to reproduce the predictive model from the development phase into the testing and deployment stages of the lifecycle. This bundle includes all the predictive model dependencies such as open source packages, proprietary packages, legacy code and third party packages. Anaconda provides automated building infrastructure to move your teams to Agile Data Science through continuous integration. This orchestration makes it simple to move the predictive model to testing and production environments.

Predictive models are often deployed into interactive visualizations, intelligent web apps or embedded into operational processes. While Anaconda includes capabilities to create and deploy interactive visualization and intelligent web apps, it also includes an automatic API generator that makes it easy to embed your predictive model into your tools of choice or into any new or existing operational process.



DevOps

-  Interactive Notebooks
-  Interactive Visualizations
-  Models
-  Slides/ Presentations
-  Intelligent Web Apps
-  RESTful Service for Embedding
-  Dashboards

SUMMARY

Open Data Science is the foundation to modernizing predictive analytics. The key principle in this new foundation is leveraging all the innovation available in open source without experiencing the chaos typically experienced with open source.

Anaconda is a modern open source analytics platform powered by Python that is the leading full-stack Open Data Science platform where Data Science teams can efficiently and effectively:

- Manage packages, dependencies and environments for the entire Open Data Science stack
- Collaborate with the entire Data Science team across the globe
- Deploy intelligent interactive visualizations, web apps or embedded processes
- Iterate quickly to build and evaluate high value predictive models

Anaconda gives superpowers to your data science team to accelerate time-to-value, connect the dots in your data and empower your entire team.

ABOUT CONTINUUM ANALYTICS

Continuum Analytics is the creator and driving force behind Anaconda, the leading modern open source analytics platform powered by Python. We put superpowers into the hands of people who are changing the world.

With more than 2M downloads annually and growing, Anaconda is trusted by the world's leading businesses across industries – financial services, government, health & life sciences, technology, retail & CPG, oil & gas – to solve the world's most challenging problems. Anaconda does this by helping everyone in the Data Science team discover, analyze and collaborate by connecting their curiosity and experience with data. With Anaconda, teams manage their Open Data Science environments without any hassles to harness the power of the latest open source analytic and technology innovations.

Our community loves Anaconda because it empowers the entire Data Science team – Data Scientists, Developers, DevOps, Architects, and Business Analysts – to connect the dots in their data and accelerate the time-to-value that is required in today's world. To ensure our customers are successful, we offer comprehensive support, training and professional services.

Continuum Analytics' Founders and Developers have created or contribute to some of the most popular Open Data Science technologies, including NumPy, SciPy, Matplotlib, Pandas, Jupyter/IPython, Bokeh, Numba and many others. Continuum Analytics is venture-backed by General Catalyst and BuildGroup.

To learn more, visit <http://www.continuum.io>