

- [Ivan Chen](#)
- [Sign Out](#)

[Subscribe to DSC Newsletter](#)

- All Blog Posts
- My Blog
- Edit Blog Posts
- Add



Top 20 Python libraries for data science in 2018

- Posted by Igor Bobriakov on June 13, 2018 at 2:30am
- [Send Message](#) [View Blog](#)

Python continues to take leading positions in solving data science tasks and challenges. Last year we made a blog post overviewing the Python's libraries that proved to be the most helpful at that moment. This year, we expanded our list with new libraries and gave a fresh look to the ones we already talked about, focusing on the updates that have been made during the year.

Our selection actually contains more than 20 libraries, as some of them are alternatives to each other and solve the same problem. Therefore we have grouped them as it's difficult to distinguish one particular leader at the moment.

Core Libraries & Statistics

1. NumPy (Commits: 17911, Contributors: 641)

Traditionally, we start our list with the libraries for scientific applications, and NumPy is one of the principal packages in this area. It is intended for processing large multidimensional arrays and matrices, and an extensive collection of high-level mathematical functions and implemented methods makes it possible to perform various operations with these objects.

During the year, a large number of improvements have been made to the library. In addition to bug fixes and compatibility issues, the crucial changes regard styling possibilities, namely the printing format of NumPy objects. Also, some functions can now handle files of any encoding that is available in Python.

2. SciPy (Commits: 19150, Contributors: 608)

Another core library for scientific computing is SciPy. It is based on NumPy and therefore extends its capabilities. SciPy main data structure is again a multidimensional array, implemented by Numpy. The package contains tools that help with solving linear algebra, probability theory, integral calculus and many more tasks.

SciPy faced major build improvements in the form of continuous integration into different operating systems, new functions and methods and, what is especially important - the updated optimizers. Also, many new BLAS and LAPACK functions were wrapped.

3. Pandas (Commits: 17144, Contributors: 1165)

Pandas is a Python library that provides high-level data structures and a vast variety of tools for analysis. The great feature of this package is the ability to translate rather complex operations with data into one or two commands. Pandas contain many built-in methods for grouping, filtering, and combining data, as well as the time-series functionality. All of this is followed by impressive speed indicators.

There have been a few new releases of the pandas library, including hundreds of new features, enhancements, bug fixes, and API changes. The improvements regard pandas abilities for grouping and sorting data, more suitable output for the *apply* method, and the support in performing custom types operations.

4. StatsModels (Commits: 10067, Contributors: 153)

Statsmodels is a Python module that provides many opportunities for statistical data analysis, such as statistical models estimation, performing statistical tests, etc. With

Embrace the Modern Analytics Lifecycle - June 26

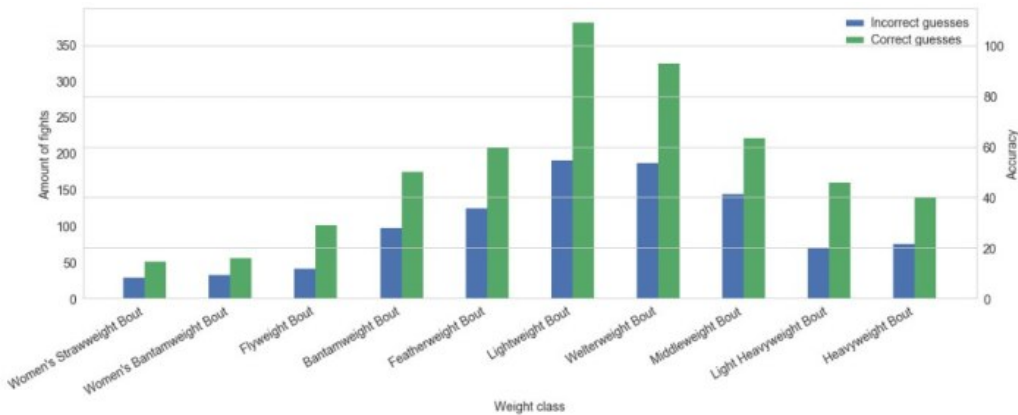
Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)

5. Matplotlib (Commits: 25747, Contributors: 725)

Matplotlib is a low-level library for creating two-dimensional diagrams and graphs. With its help, you can build diverse charts, from histograms and scatterplots to non-Cartesian coordinates graphs. Moreover, many popular plotting libraries are designed to work in conjunction with matplotlib.

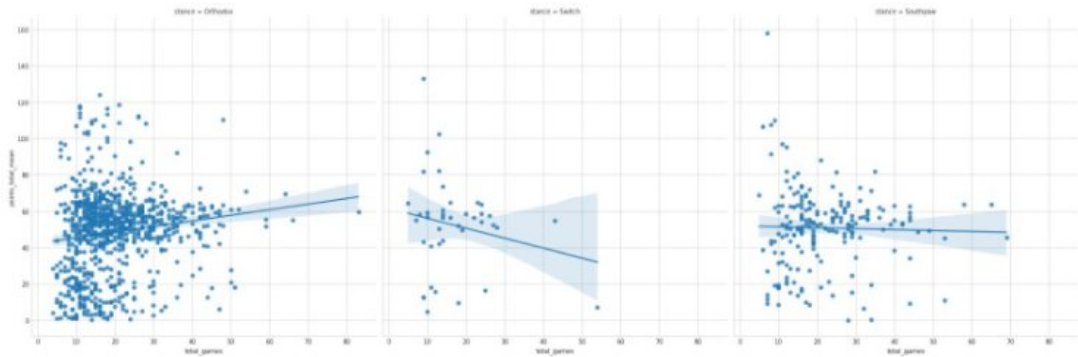
There have been style changes in colors, sizes, fonts, legends, etc. As an example of an appearance improvements are an automatic alignment of axes legends and among significant colors improvements is a new colorblind-friendly color cycle.



6. Seaborn (Commits: 2044, Contributors: 83)

Seaborn is essentially a higher-level API based on the matplotlib library. It contains more suitable default settings for processing charts. Also, there is a rich gallery of visualizations including some complex types like time series, jointplots, and violin diagrams.

The seaborn updates mostly cover bug fixes. However, there were improvements in compatibility between FacetGrid or PairGrid and enhanced interactive matplotlib backends, adding parameters and options to visualizations.



7. Plotly (Commits: 2906, Contributors: 48)

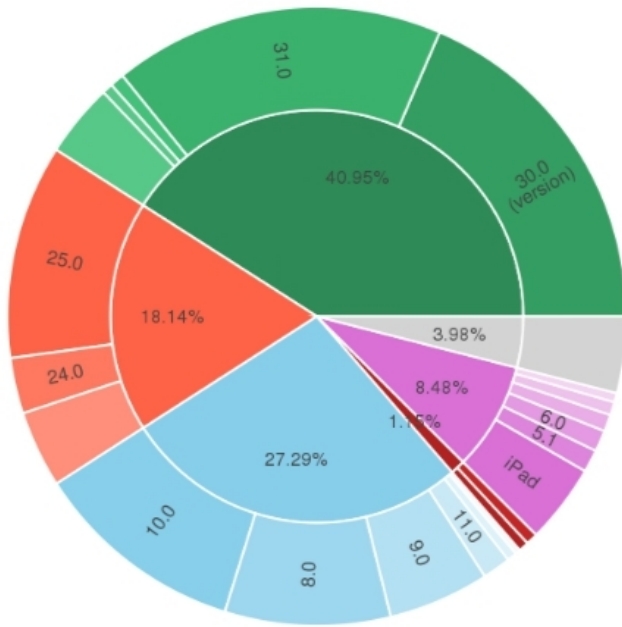
Plotly is a popular library that allows you to build sophisticated graphics easily. The package is adapted to work in interactive web applications. Among its remarkable visualizations are contour graphics, ternary plots, and 3D charts.

The continuous enhancements of the library with new graphics and features brought the support for "multiple linked views" as well as animation, and crosstalk integration.

8. Bokeh (Commits: 16983, Contributors: 294)

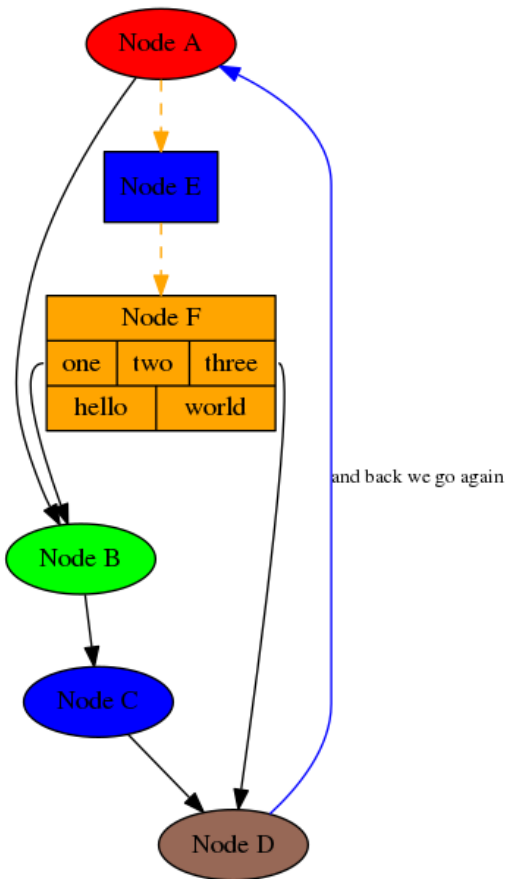
The Bokeh library creates interactive and scalable visualizations in a browser using JavaScript widgets. The library provides a versatile collection of graphs, styling possibilities, interaction abilities in the form of linking plots, adding widgets, and defining callbacks, and many more useful features.

Bokeh can boast with improved interactive abilities, like a rotation of categorical tick labels, as well as small zoom tool and customized tooltip fields enhancements.



9. Pydot (Commits: 169, Contributors: 12)

Pydot is a library for generating complex oriented and non-oriented graphs. It is an interface to Graphviz, written in pure Python. With its help, it is possible to show the structure of graphs, which are very often needed when building neural networks and decision trees based algorithms.



Machine Learning

10. Scikit-learn (Commits: 22753, Contributors: 1084)

This Python module based on NumPy and SciPy is one of the best libraries for working with data. It provides algorithms for many standard machine learning and data mining tasks such as clustering, regression, classification, dimensionality reduction, and model selection.

Embrace the Modern Analytics Lifecycle - June 26

Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)

11. XGBoost / LightGBM / CatBoost (Commits: 3277 / 1083 / 1509, Contributors: 280 / 79 / 61)

Gradient boosting is one of the most popular machine learning algorithms, which lies in building an ensemble of successively refined elementary models, namely decision trees. Therefore, there are special libraries designed for fast and convenient implementation of this method. Namely, we think that XGBoost, LightGBM, and CatBoost deserve special attention. They are all competitors that solve a common problem and are used in almost the same way. These libraries provide highly optimized, scalable and fast implementations of gradient boosting, which makes them extremely popular among data scientists and Kaggle competitors, as many contests were won with the help of these algorithms.

12. Eli5 (Commits: 922, Contributors: 6)

Often the results of machine learning models predictions are not entirely clear, and this is the challenge that eli5 library helps to deal with. It is a package for visualization and debugging machine learning models and tracking the work of an algorithm step by step. It provides support for scikit-learn, XGBoost, LightGBM, lightning, and sklearn-crfsuite libraries and performs the different tasks for each of them.

Deep Learning

13. TensorFlow (Commits: 33339, Contributors: 1469)

TensorFlow is a popular framework for deep and machine learning, developed in Google Brain. It provides abilities to work with artificial neural networks with multiple data sets. Among the most popular TensorFlow applications are object identification, speech recognition, and more. There are also different layer-helpers on top of regular TensorFlow, such as tflearn, tf-slim, skflow, etc.

This library is quick in new releases, introducing new and new features. Among the latest are fixes in potential security vulnerability and improved TensorFlow and GPU integration, such as you can run an Estimator model on multiple GPUs on one machine.

14. PyTorch (Commits: 11306, Contributors: 635)

PyTorch is a large framework that allows you to perform tensor computations with GPU acceleration, create dynamic computational graphs and automatically calculate gradients. Above this, PyTorch offers a rich API for solving applications related to neural networks.

The library is based on Torch, which is an open source deep learning library implemented in C with a wrapper in Lua. The Python API was introduced in 2017 and from that point on, the framework is gaining popularity and attracting an increasing number of data scientists.

15. Keras (Commits: 4539, Contributors: 671)

Keras is a high-level library for working with neural networks, running on top of TensorFlow, Theano, and now as a result of the new releases, it is also possible to use CNTK and MxNet as the backends. It simplifies many specific tasks and greatly reduces the amount of monotonous code. However, it may not be suitable for some complicated things.

This library faced performance, usability, documentation, and API improvements. Some of the new features are Conv3DTranspose layer, new MobileNet application, and self-normalizing networks.

Distributed Deep Learning

16. Dist-keras / elephas / spark-deep-learning (Commits: 1125 / 170 / 67, Contributors: 5 / 13 / 11)

Deep learning problems are becoming crucial nowadays since more and more use cases require considerable effort and time. However, processing such an amount of data is much easier with the use of distributed computing systems like Apache Spark which again expands the possibilities for deep learning. Therefore, dist-keras, elephas, and spark-deep-learning are gaining popularity and developing rapidly, and it is very difficult to single out one of the libraries since they are all designed to solve a common task. These packages allow you to train neural networks based on the Keras library directly with the help of Apache Spark. Spark-deep-learning also provides tools to create a pipeline with Python neural networks.

Natural Language Processing

17. NLTK (Commits: 13041, Contributors: 236)

NLTK is a set of libraries, a whole platform for natural language processing. With the help of NLTK, you can process and analyze text in a variety of ways, tokenize and tag it, extract information, etc. NLTK is also used for prototyping and building research systems.

The enchantments to this library cover minor changes in APIs and compatibility and a new interface to CoreNLP.

18. SpaCy (Commits: 8623, Contributors: 215)

SpaCy is a natural language processing library with excellent examples, API documentation, and demo applications. The library is written in the Cython language which is C extension of Python. It supports almost 30 languages, provides easy deep learning integration and promises robustness and high accuracy. Another great feature of spaCy is an architecture designed for entire documents processing, without breaking the document into phrases.

19. Gensim (Commits: 3603, Contributors: 273)

Gensim is a Python library for robust semantic analysis, topic modeling and vector-space modeling, and is built upon Numpy and Scipy. It provides an implementation of popular NLP algorithms, such as word2vec. Although gensim has its own models.wrappers.fasttext implementation, the fasttext library can also be used for efficient learning of word representations.

Data Scraping

20. Scrapy (Commits: 6625, Contributors: 281)

Scrapy is a library used to create spiders bots that scan website pages and collect structured data. In addition, Scrapy can extract data from the API. The library happens to be very handy due to its extensibility and portability.

Embrace the Modern Analytics Lifecycle - June 26














Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

Register Now



This is our enriched collection of Python libraries for data science in 2018. Comparing to the previous year, some new modern libraries are gaining popularity while the ones that have become classical for data scientific tasks are continuously improving.

Again, there is a table that shows detailed statistics of github activities.

| Github data Python 2018 | | | | | | | | | | |
|---|---------------------------|---------|--------------|----------|-------|--------|--------|---|----------------------|-----------------------|
| Library Name | Type | Commits | Contributors | Releases | Watch | Star | Fork | Commits/ Contributors | Commits/ Releases | Star/ Contributors |
|  matplotlib | Visualization | 25 747 | 725 | 70 | 498 | 7 292 | 398 | 36 | 368 | 10 |
|  Bokeh | Visualization | 16 983 | 294 | 58 | 363 | 7 615 | 2 000 | 58 | 293 | 26 |
|  plotly | Visualization | 2 906 | 48 | 8 | 198 | 3 444 | 850 | 61 | 363 | 72 |
| Seaborn | Visualization | 2 044 | 83 | 13 | 205 | 4 856 | 752 | 25 | 157 | 59 |
| <i>pydot</i> | Visualization | 169 | 12 | 12 | 17 | 193 | 80 | 14 | 14 | 16 |
|  learn | Machine learning | 22 753 | 1 084 | 86 | 2 114 | 28 098 | 14 005 | 21 | 265 | 26 |
| XGBoost LightGBM CatBoost | Machine learning | 3277 | 280 | 9 | 868 | 11 991 | 5 425 | 12 | 364 | 43 |
| | Machine learning | 1083 | 79 | 14 | 363 | 5 488 | 1 467 | 14 | 77 | 69 |
| | Machine learning | 1509 | 61 | 20 | 157 | 2 780 | 369 | 25 | 75 | 46 |
| eli5 | Machine learning | 922 | 6 | 22 | 39 | 672 | 89 | 154 | 42 | 112 |
|  SciPy | Data wrangling | 19 150 | 608 | 99 | 301 | 4 447 | 2 318 | 31 | 193 | 7 |
|  NumPy | Data wrangling | 17 911 | 641 | 136 | 390 | 7 215 | 2 766 | 28 | 132 | 11 |
| pandas  | Data wrangling | 17 144 | 1 165 | 93 | 858 | 14 294 | 5 788 | 15 | 184 | 12 |
|  Statsmodels <small>Statistics in Python</small> | Statistics | 10 067 | 153 | 21 | 234 | 2 868 | 1 240 | 66 | 479 | 19 |
|  TensorFlow | Deep learning | 33 339 | 1 469 | 58 | 7 968 | 99 664 | 62 952 | 23 | 575 | 68 |
| PYTORCH | Deep learning | 11 306 | 635 | 16 | 816 | 15 512 | 3 483 | 18 | 707 | 24 |
|  Keras | Deep learning | 4 539 | 671 | 41 | 1 673 | 29 444 | 10 964 | 7 | 1111 | 44 |
| dist-keras elephas spark-deep-learning | Distributed deep learning | 1125 | 5 | 7 | 41 | 431 | 106 | 225 | 161 | 86 |
| | | 170 | 13 | 5 | 97 | 913 | 189 | 13 | 34 | 70 |
| | | 67 | 11 | 3 | 116 | 920 | 206 | 6 | 22 | 84 |
|  NLTK <small>Natural Language Toolkit</small> | NLP | 13 041 | 236 | 24 | 467 | 6 405 | 1 804 | 55 | 543 | 27 |
| spaCy | NLP | 8 623 | 215 | 56 | 425 | 9 258 | 1 446 | 40 | 154 | 43 |
| gensim | NLP | 3 603 | 273 | 52 | 415 | 6 995 | 2 689 | 13 | 69 | 26 |
|  Scrapy | Data scraping | 6 625 | 281 | 81 | 1 723 | 27 277 | 6 469 | 24 | 82 | 97 |
| Last reviewed: 13.02.2018 | | | | | | | | Created by  ActiveWizards | | |

Even though we have extended our list this year, it still may not cover some other great and useful libraries that deserve to be looked at. So, share your favorites in the comment section below, as well as any ideas about the packages that we mentioned.

Thank you for your attention!

Originally posted here.


Views: 8217

Like

9 members like this

Share

Tweet



Facebook

Like 122

•

< Previous Post



Visual Mode HTML Editor



10pt

B*I*~~S~~U

Follow – Email me when people comment

Add Comment



Comment by Mohammad on Saturday

Great post for any data science practitioner! Thank you for creating this list!



Comment by Jessica Tanon on Thursday

Thanks for taking the time to put this list together !

RSS

IVAN CHEN

Sign Out

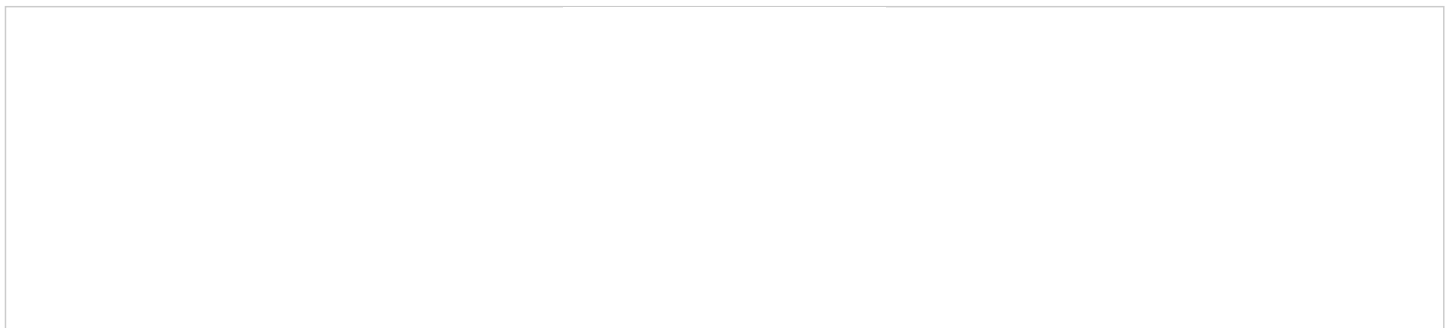
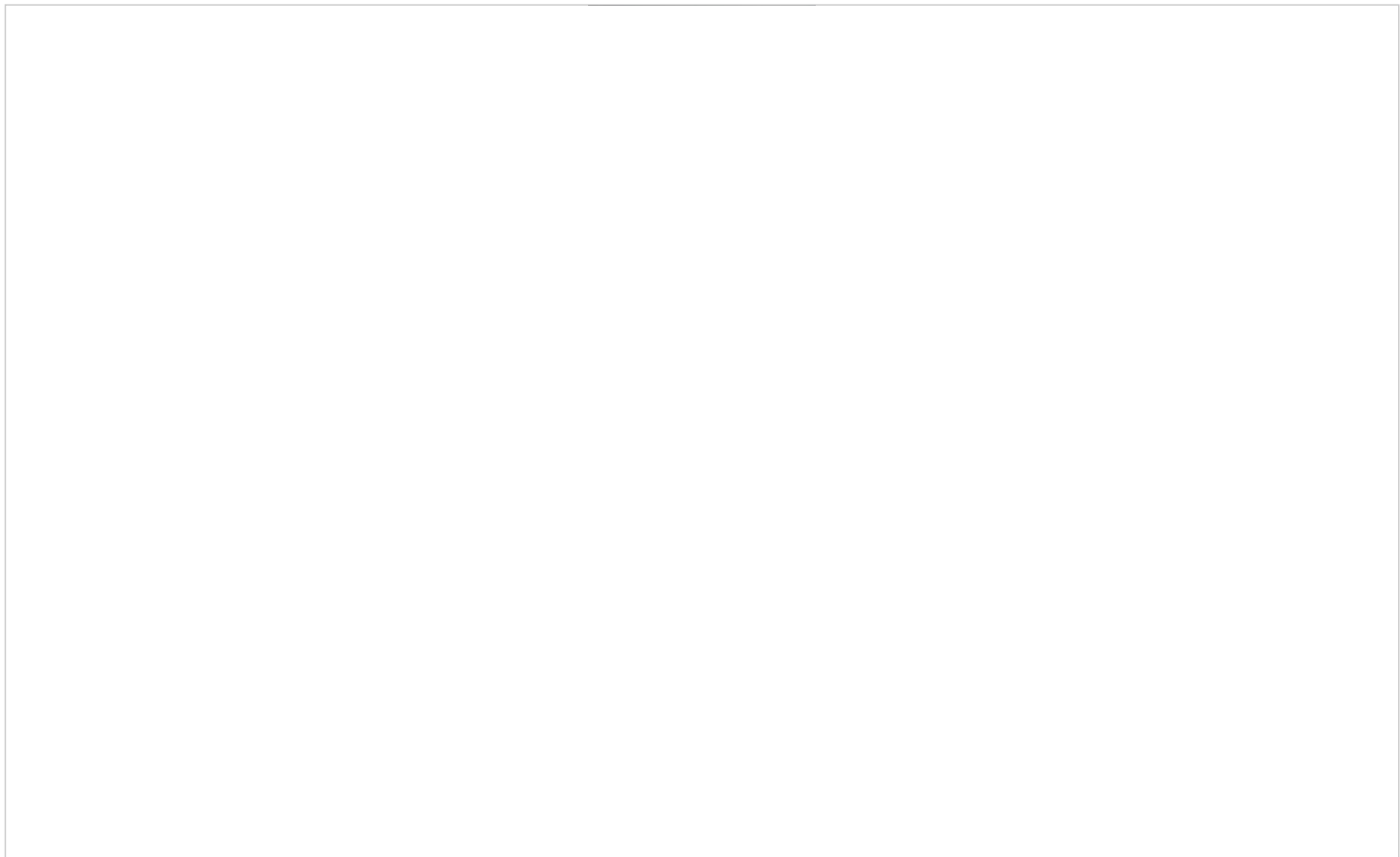
- [Inbox](#)
- [Alerts](#)
- [Friends – Invite](#)
- [Settings](#)

Embrace the Modern Analytics Lifecycle - June 26

Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)





FOLLOW US

@DataScienceCtrl | [RSS Feeds](#)

TOP CONTENT



1 [Top 20 Python libraries for data science in 2018](#)



2 [Yet another introduction to Neural Networks](#)



3 [State-of-the-Art Machine Learning Automation with HDT](#)

Embrace the Modern Analytics Lifecycle - June 26



Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)



5

Data + Art STEAM Project: Initial Results



6

The Most Popular Language For Machine Learning Is ...

- [RSS](#)
- [View All](#)

ANNOUNCEMENTS

MS in Applied Business Analytics at JU

7 Ways to Fast Track Analytics Initiatives

Online Master's in Data Science From SMU

Building Better Predictive Models with Anaconda

Building a better map

Strata NY—Dive deep into the business of data

Customer Analytics with Wharton Executive Education

Whitepaper: Delivering data-informed insights

Analytics Reimagined – Virtual Conference

Join NYU Stern's MS in Business Analytics Program

VIDEOS



•

DSC Webinar Series: Life as a Data Scientist and How to Become One

Added by Tim Matteson 0 Comments 1 Like +1



•

DSC Webinar Series: Modern Data Science Workflows

Added by Tim Matteson 0 Comments 0 Likes +1



•

DSC Webinar Series: An Expert's Guide to Apache Spark™

Added by Tim Matteson 1 Comment 0 Likes +1

- [Add Videos](#)
- [View All](#)

RESOURCES

Embrace the Modern Analytics Lifecycle - June 26



Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)

- [Selection of best articles from our past weekly digests](#)
- [Statistical Analysis Advisor Chart](#)
- [Selection of best articles from our past weekly digests](#)
- [Free Online Book: Forecasting, Principles and Practice](#)
- [38 Seminal Articles Every Data Scientist Should Read](#)
- [Black-box Confidence Intervals: Excel and Perl Implementation](#)

TOP CATEGORIES

[Machine Learning](#)

[R Programming](#)

[Python for Data Science](#)

[Visualization, Dashboards](#)

[NoSQL and NewSQL](#)

[Big Data](#)

[Cheat Sheets](#)

[Internet of Things](#)

[Excel](#)

© 2018 Data Science Central™ Powered by **NING**

[Badges](#) | [Report an Issue](#) | [Privacy Policy](#) | [Terms of Service](#)



Embrace the Modern Analytics Lifecycle - June 26



Join Radiant Advisors for a discussion surrounding new research that explores solutions and reference architectures for data science platforms on Azure – all in the context of a modern analytics lifecycle.

[Register Now](#)