Apache Spark Usage in the Open Source Ecosystem

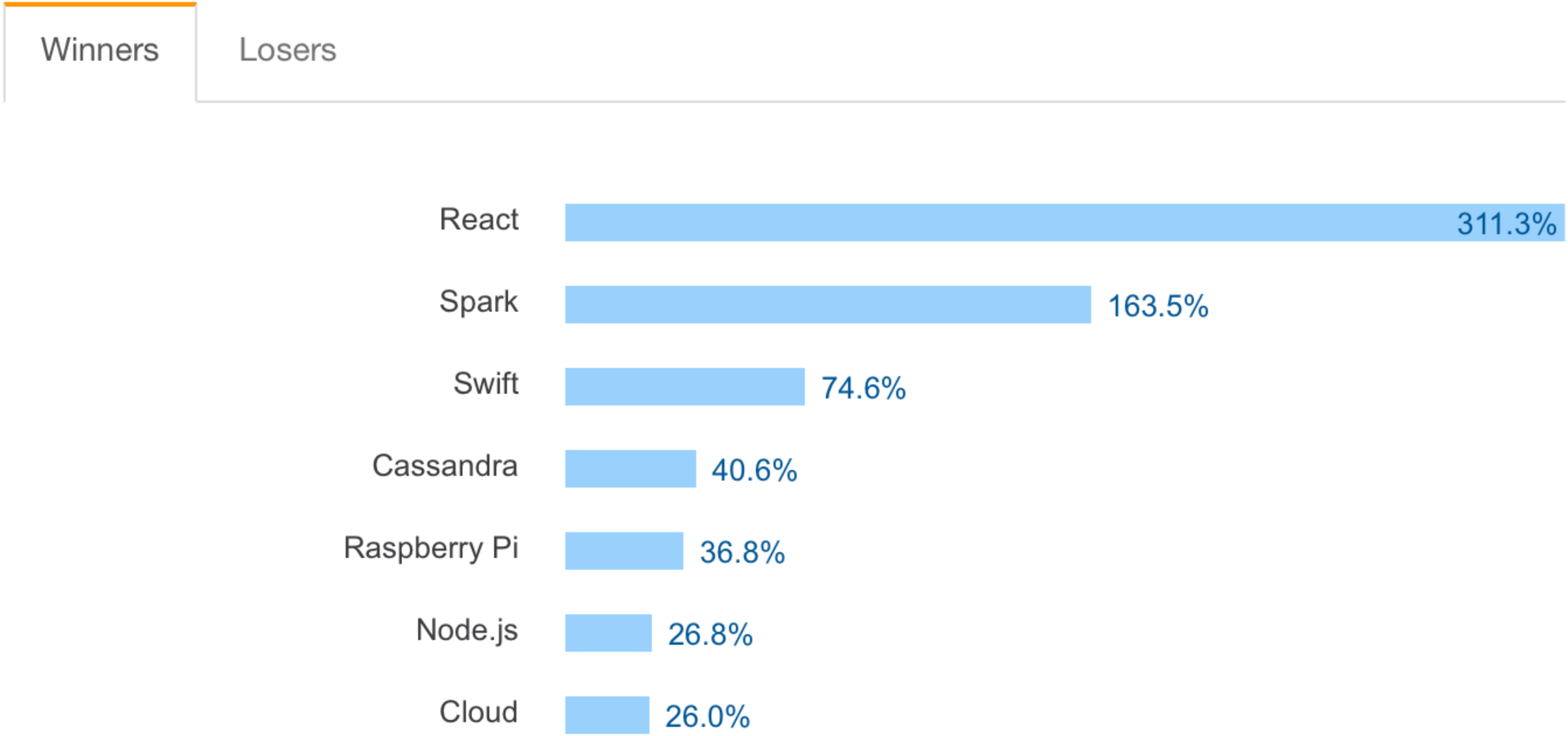
Hossein Falaki
@mhfalaki

databricks™

About me

- Software Engineer / part-time Data Scientist at Databricks
- I started using Apache Spark since version 0.6
- Developed first version of Apache Spark CSV data source
- Worked on SparkR and R notebooks at Databricks

Stackoverflow 2016 trending tech

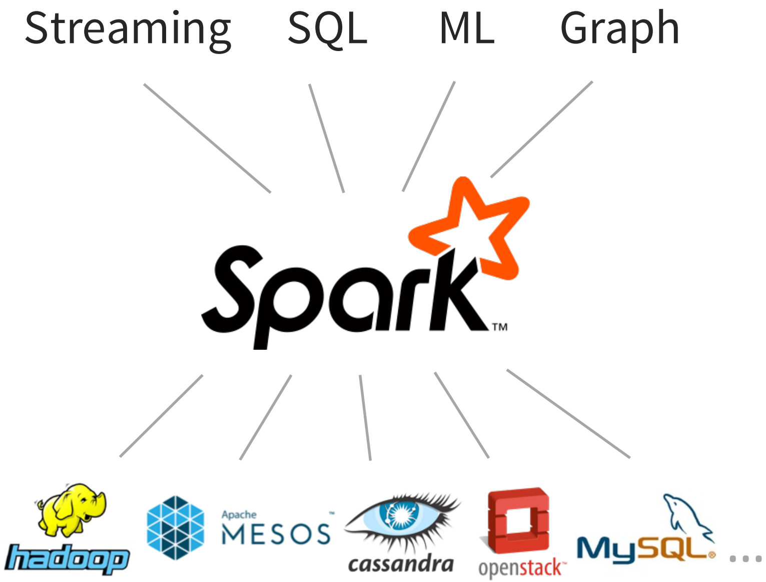


Apache Spark Philosophy

1. Unified engine
   Support end-to-end applications

2. High-level APIs
   Easy to use, rich optimizations

3. Integrate broadly
   Storage systems, libraries, etc



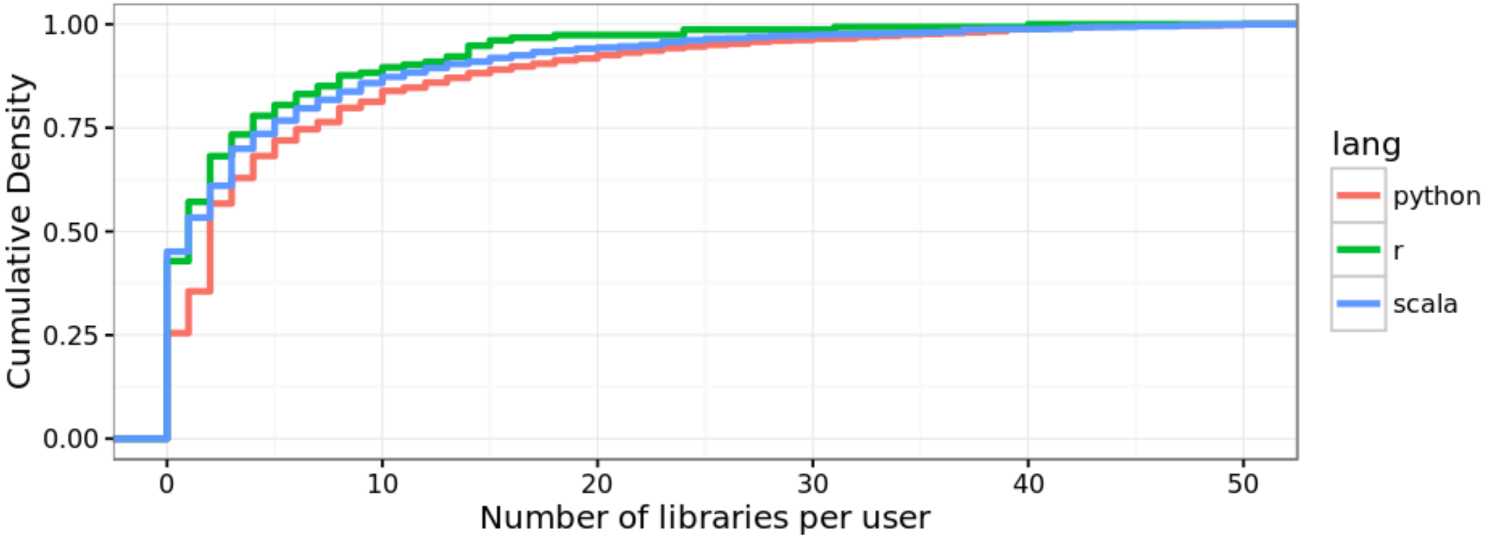
Streaming SQL ML Graph

Databricks Community Edition

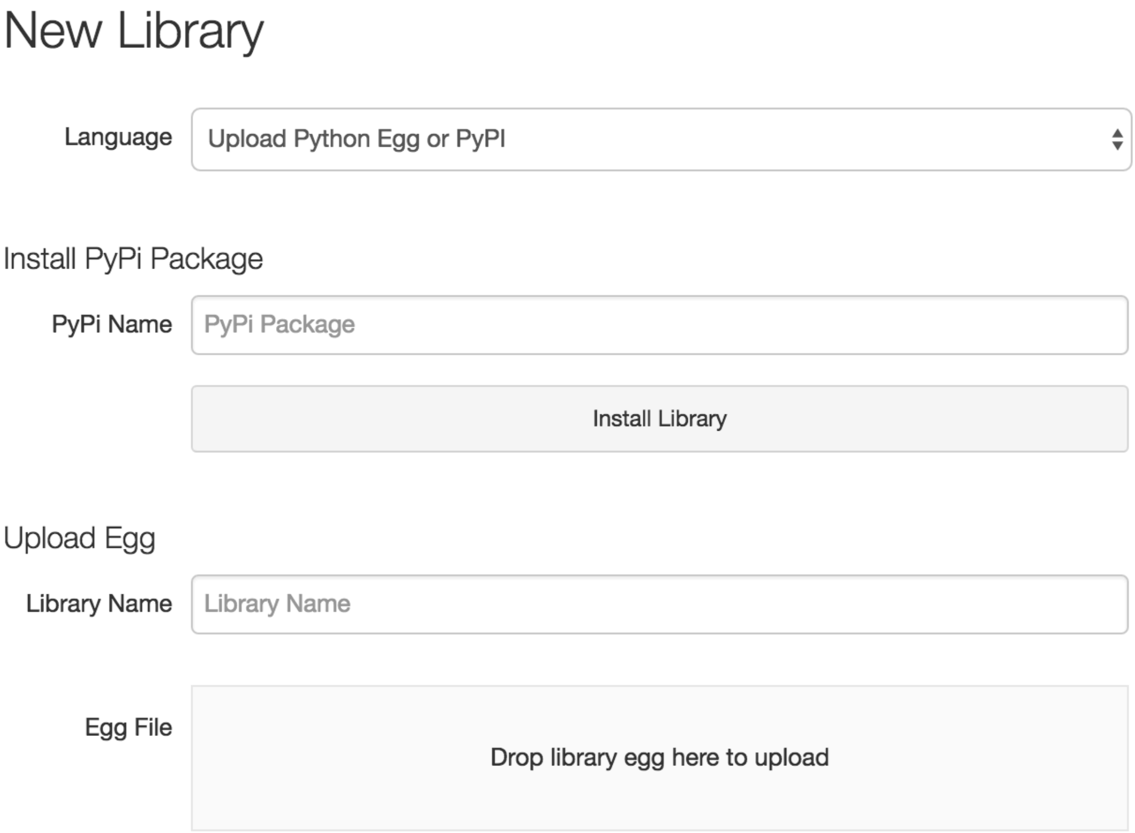
- In February Databricks launched a free version of its cloud based platform in beta
- Since then more than 8,000 users registered
- Users created over 61,000 notebooks in different languages
- This is an analysis of third party libraries that our beta users imported to complement Apache Spark in Scala, Python, and R

What % of users use other libraries

Language	% users importing external libs	Average # libs	Median # libs
Python	75 %	9	2
Scala	55 %	3	1
R	57 %	6	1



Installing libraries is easy



New Library

Language: Upload Python Egg or PyPI

Install PyPi Package

PyPi Name: PyPi Package

Install Library

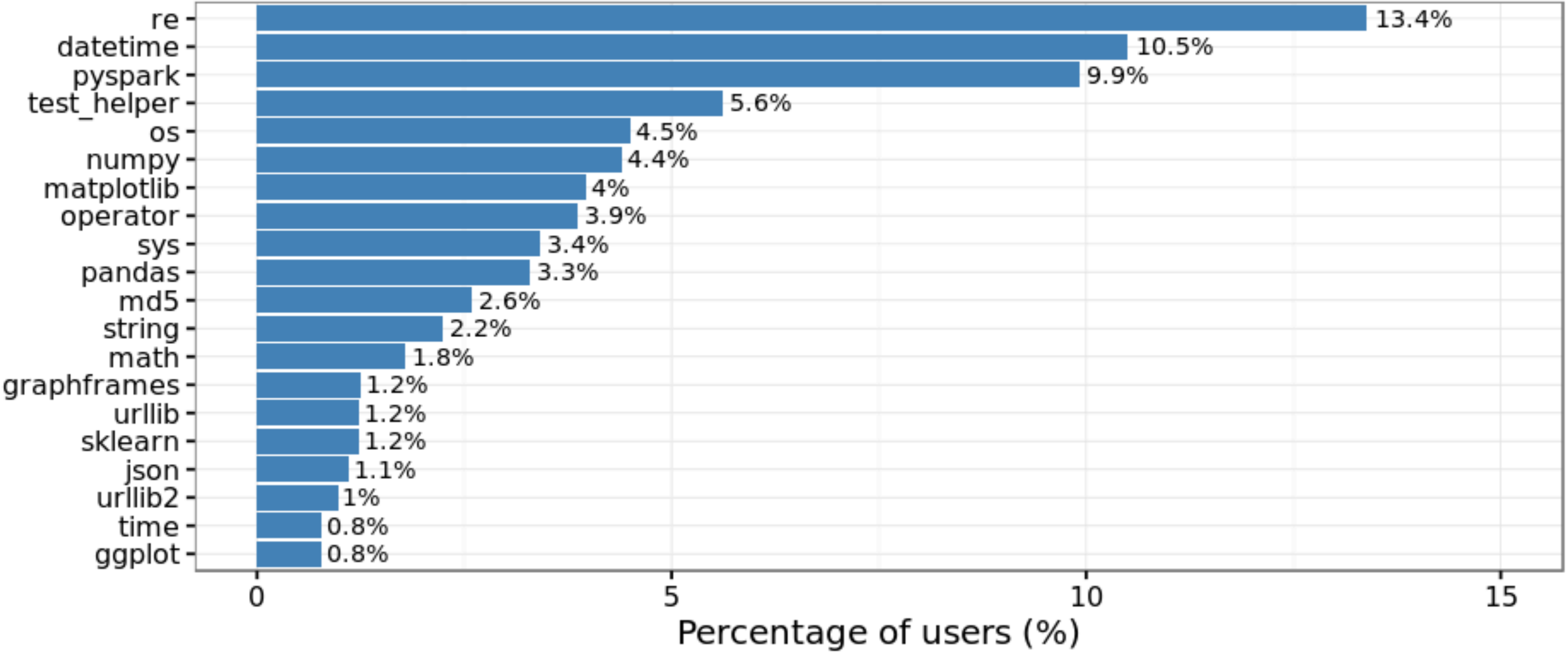
Upload Egg

Library Name: Library Name

Egg File: Drop library egg here to upload

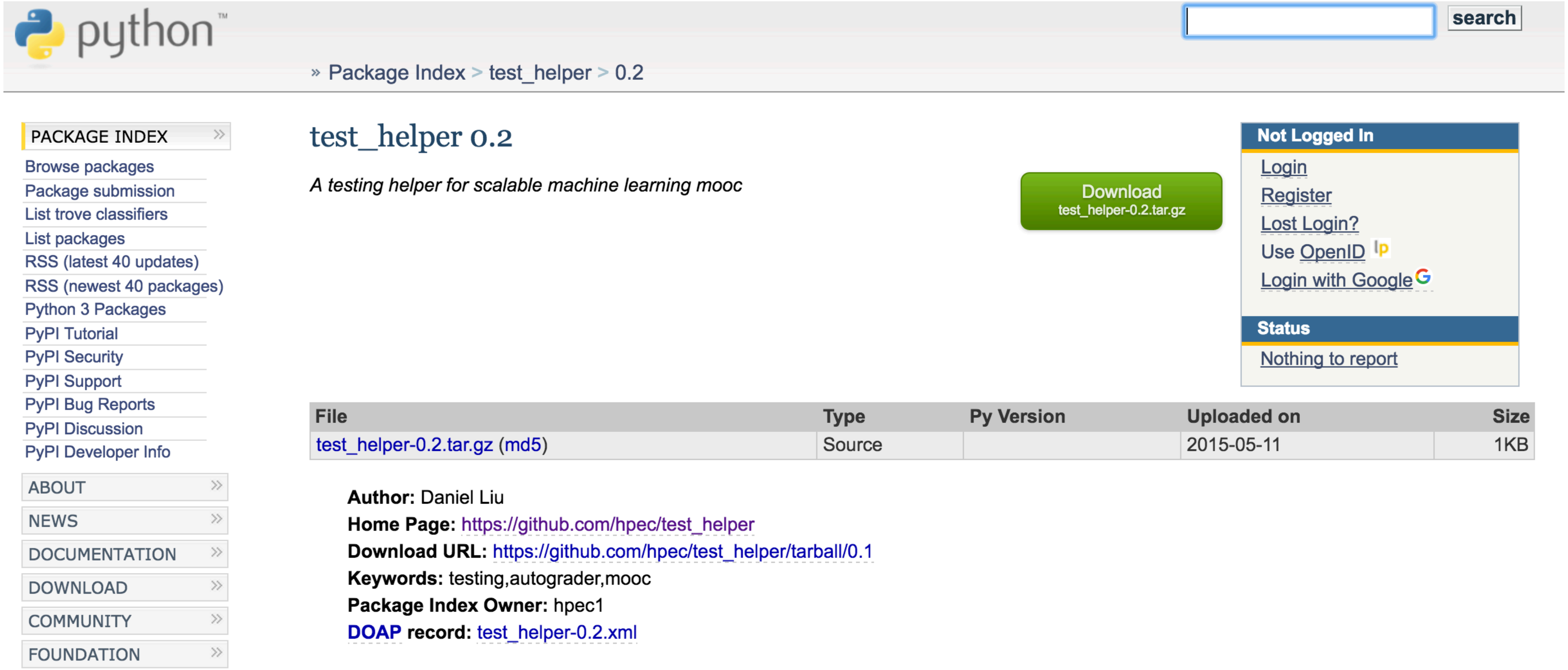
Python Packages

Most popular Python packages



databricks

What is test_helper?



test_helper 0.2

A testing helper for scalable machine learning mooc

File	Type	Py Version	Uploaded on	Size
test_helper-0.2.tar.gz (md5)	Source		2015-05-11	1KB

Author: Daniel Liu
Home Page: https://github.com/hpec/test_helper
Download URL: https://github.com/hpec/test_helper/tarball/0.1
Keywords: testing,autograder,mooc
Package Index Owner: hpec1
DOAP record: test_helper-0.2.xml

What are these?

ETL

- re
- datetime
- pandas
- json
- csv
- string
- math / operator
- urllib / urllib2

Visualization

- matplotlib
- ggplot
- seaborn

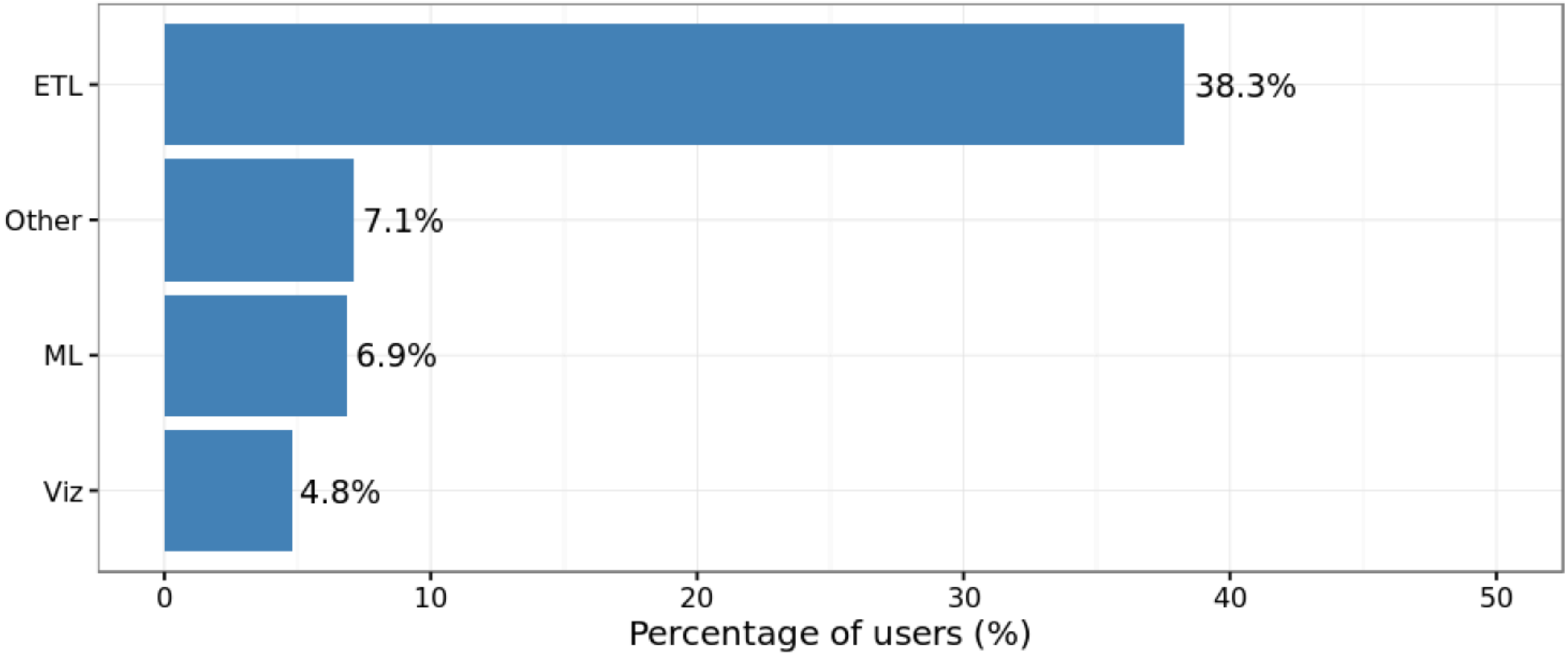
Advanced analytics

- numpy
- sklearn
- graphframes
- tensorflow
- scipy

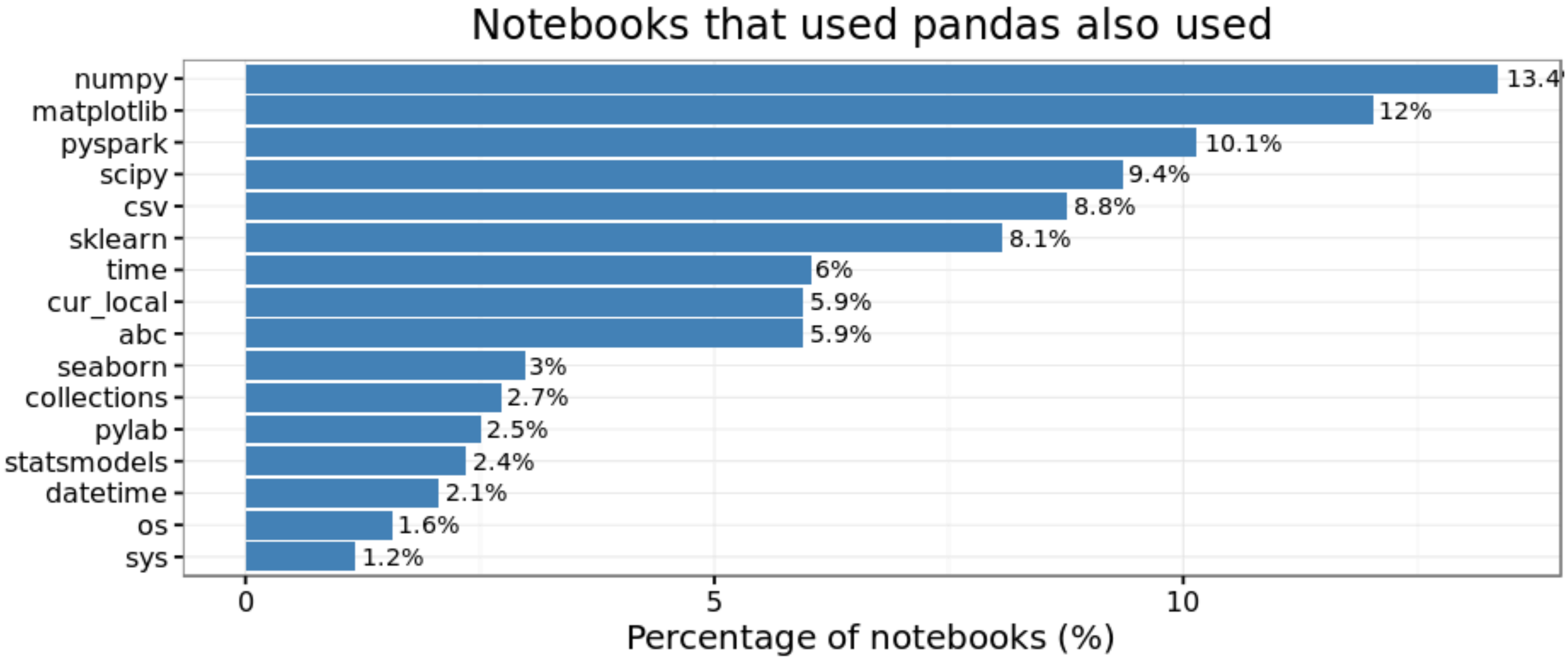
Other

- test_helper
- os
- md5

Python package categories



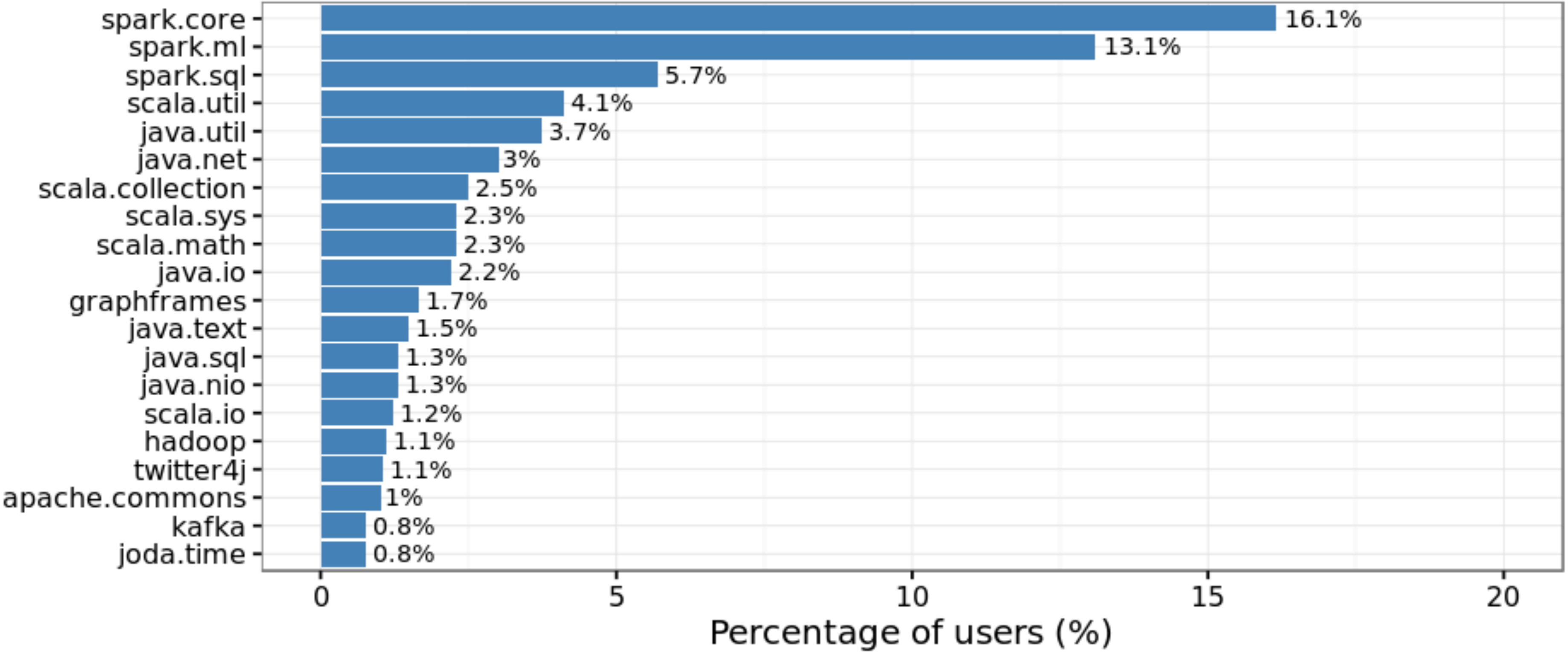
What packages go together?



databricks

Scala Packages

Most popular Scala libraries



databricks

What are these?

ETL

- java/scala util
- scala.collection
- scala.math
- java.{io, nio}
- java.text
- o.a.commons
- kafka
- twitter4j

Visualization

- ?

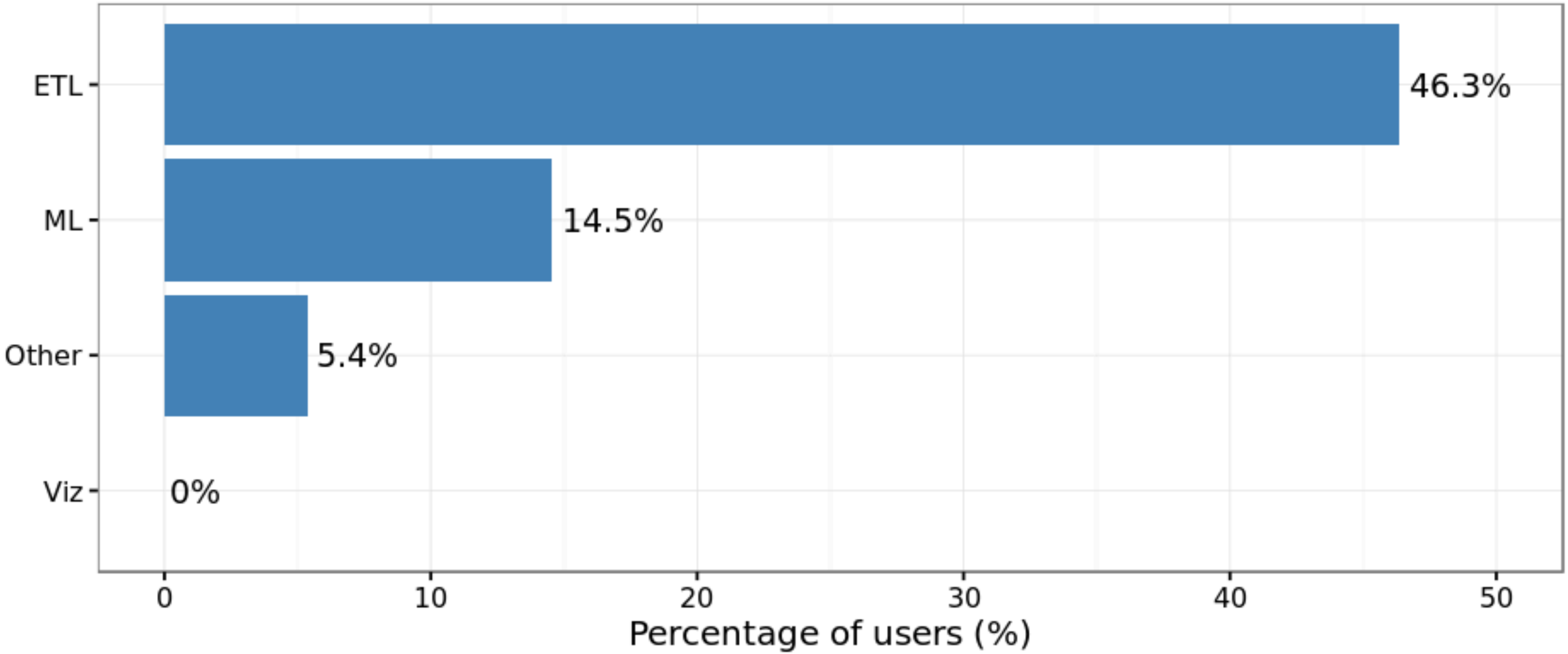
Advanced_analytics

- spark.ml
- graphframes

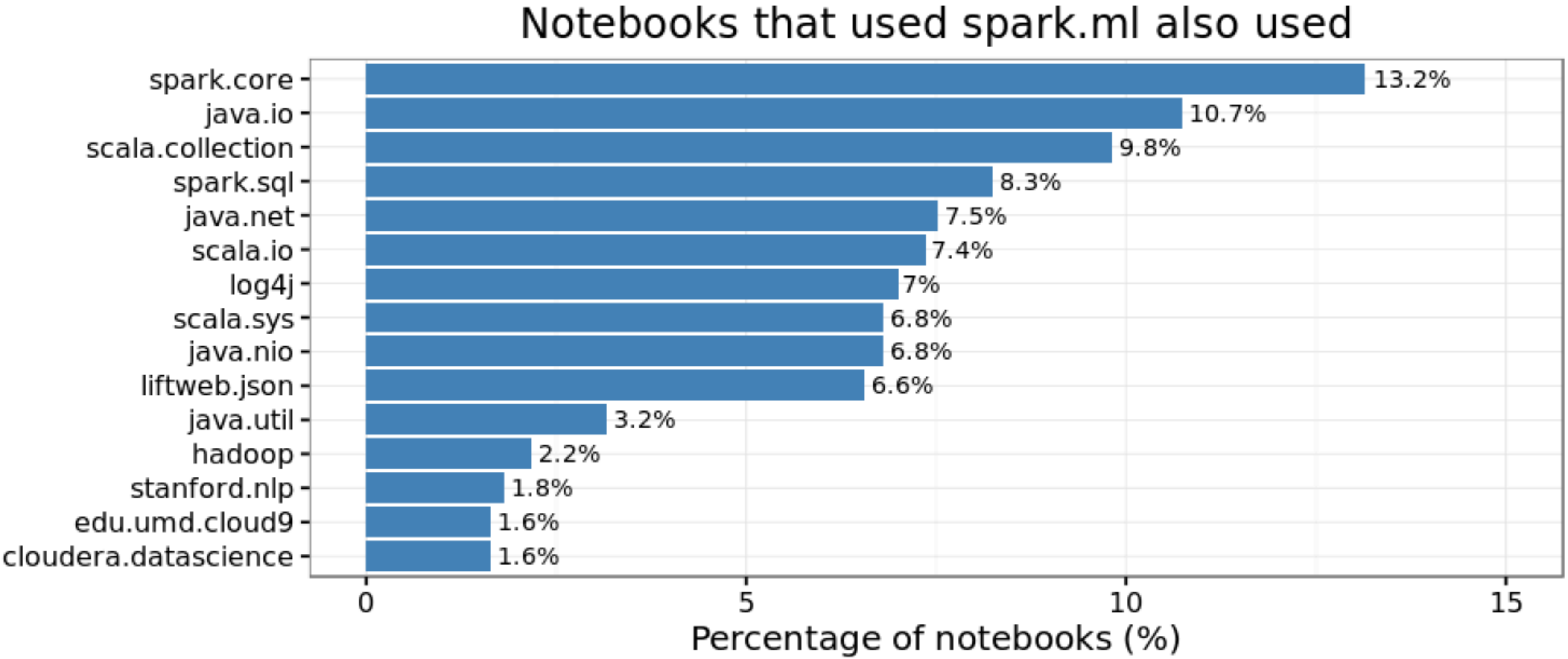
Other

- java.net
- scala.sys

Scala package categories

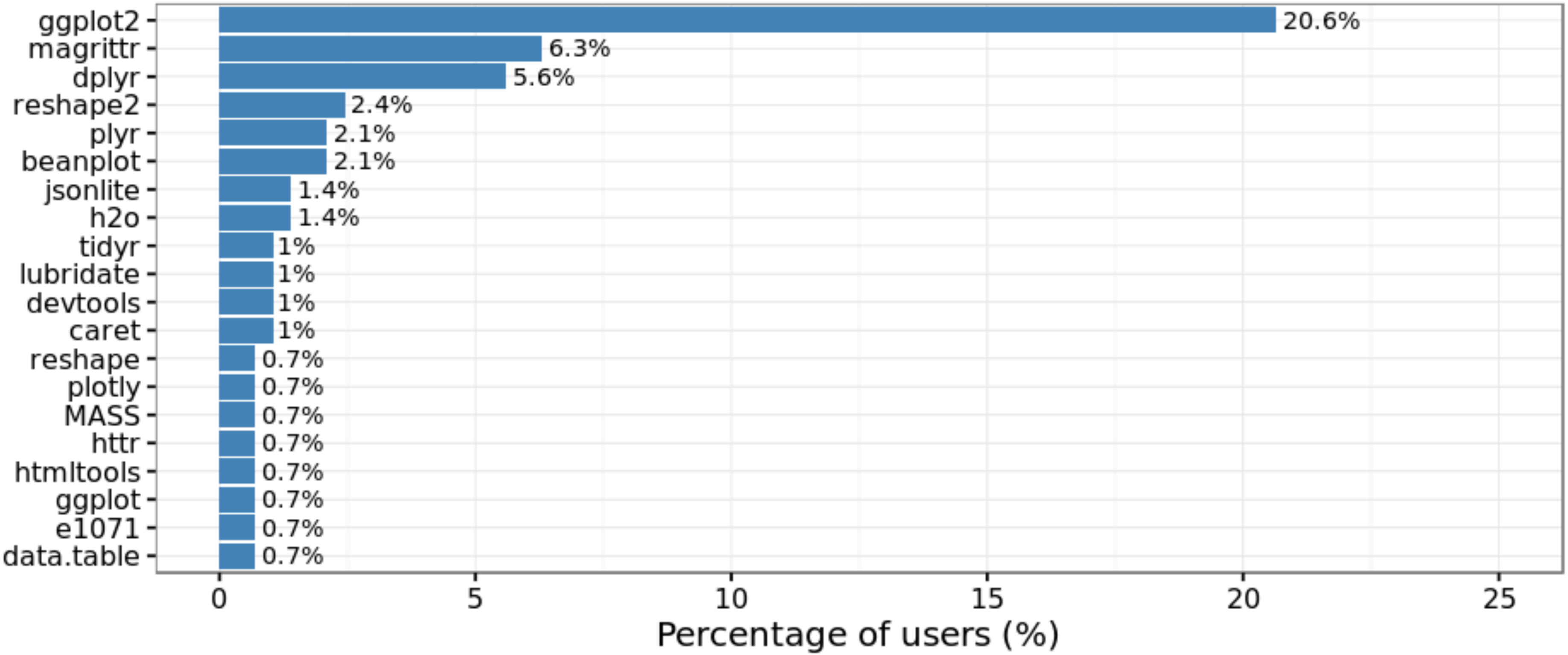


What libraries go together?



R Packages

Most popular R packages



What are these?

ETL

- dplyr
- plyr
- reshape2
- jsonlite
- tidyr
- lubridate
- httr
- data.table

Visualization

- ggplot2
- beanplot
- plotly
- …

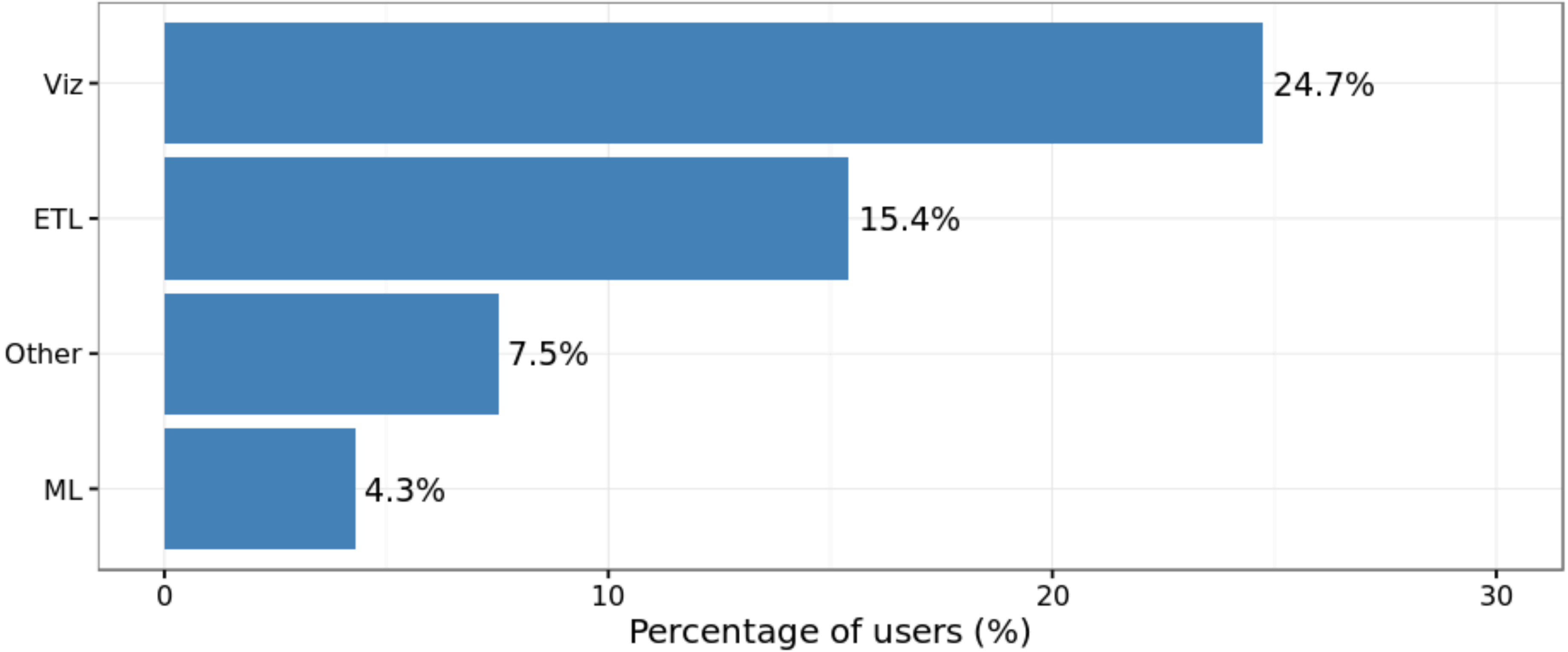
Advanced analytics

- sparkr
- h2o
- caret
- e1071

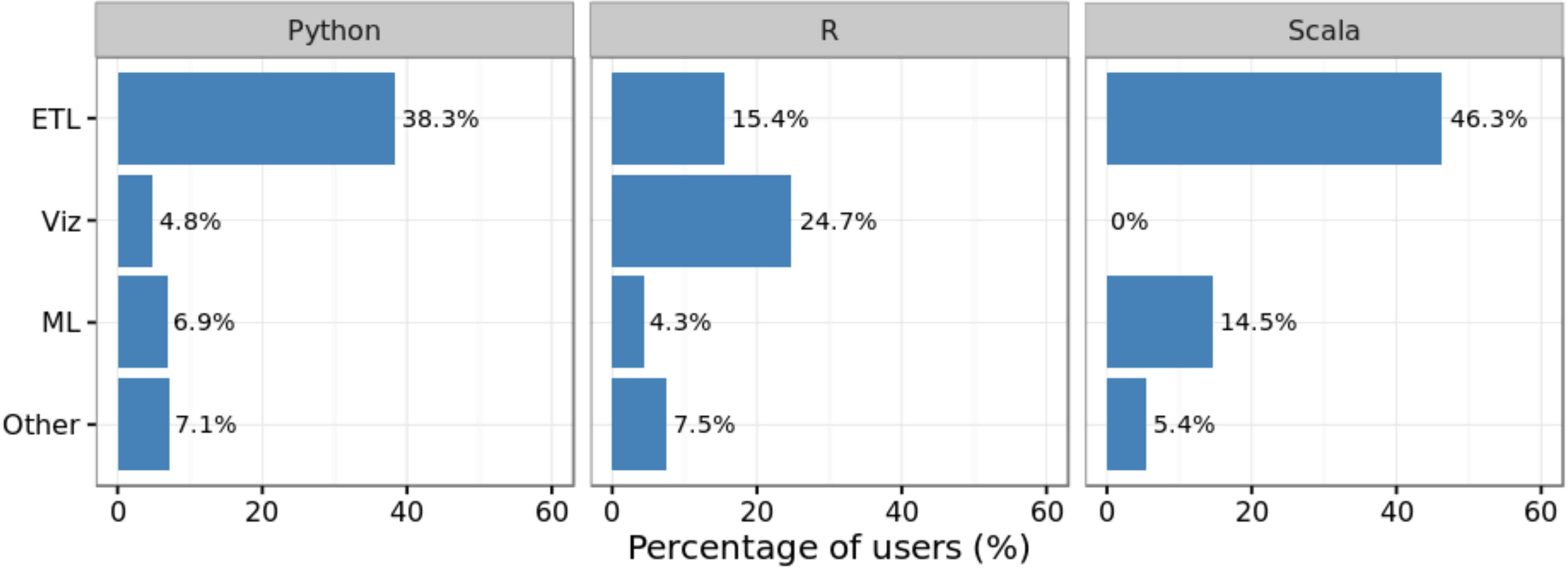
Other

- devtools
- magrittr

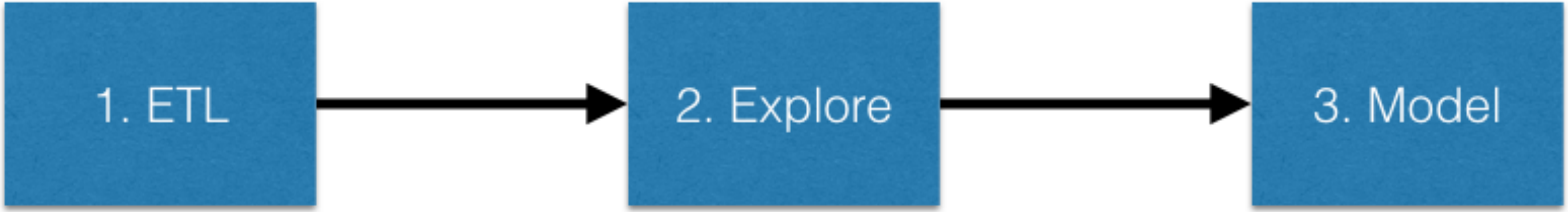
R package categories



Comparing Python, Scala & R



Languages have unique features



Scala/ Python / R

R / Python

Scala / Python/ R

- 25 % of users, use multiple languages
- 3% of notebooks mix different languages

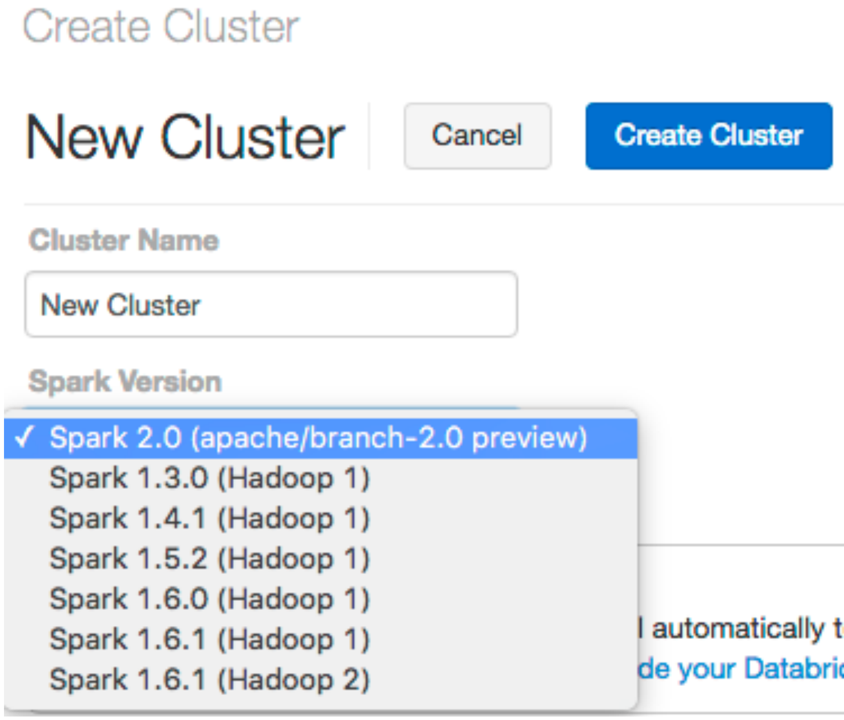
Summary

- Spark users extensively mix it with other packages in different languages
  - One of goals of Spark project is working well with other projects
- ETL related libraries are the most popular category
  - Opportunities for new data sources
- Notebooks are being used for “small data” as well as “big data.”
- Languages and their ecosystems have diverse capabilities. Users seem to be mixing languages to their advantage
  - Scala is missing visualization libraries

Try your favorite library in Databricks

Try latest version of Apache Spark and preview of Spark 2.0

http://databricks.com/ce



Create Cluster

New Cluster

Cluster Name

New Cluster

Spark Version

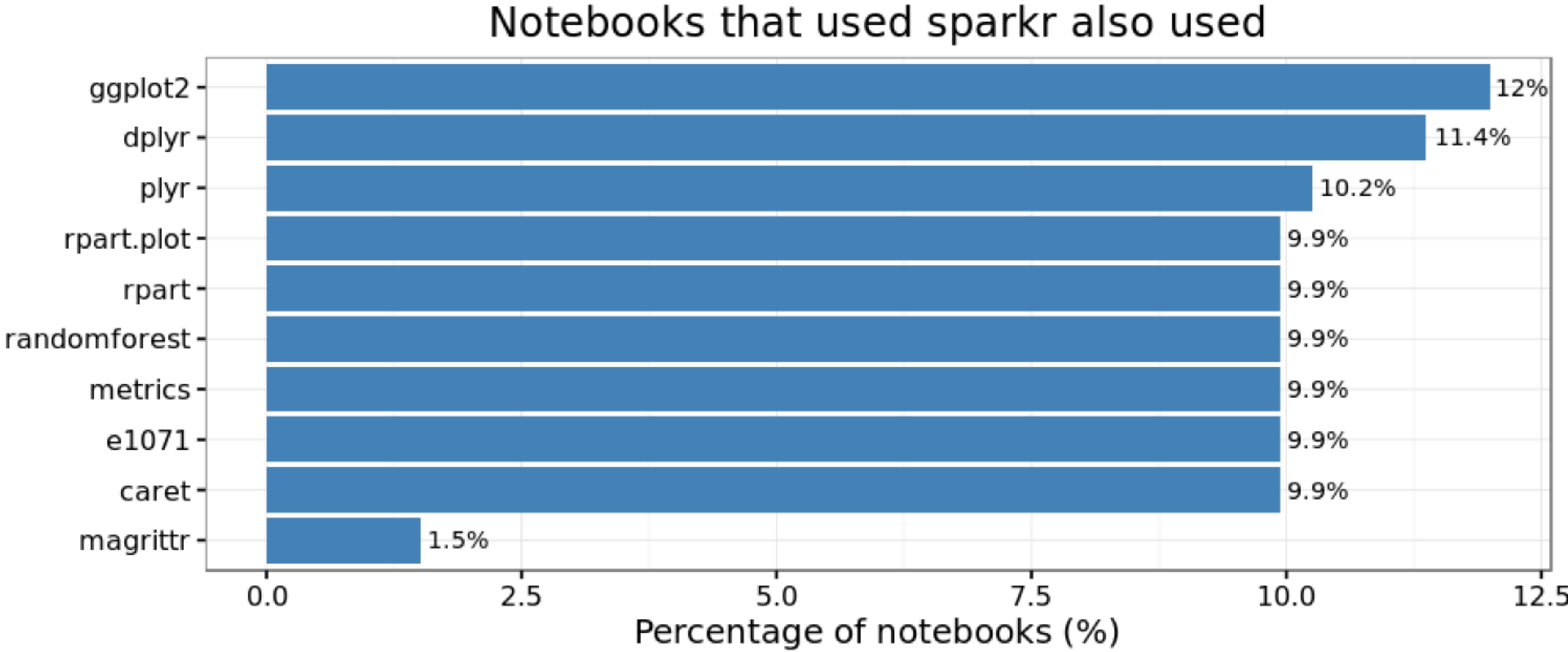
- ✓ Spark 2.0 (apache/branch-2.0 preview)
- Spark 1.3.0 (Hadoop 1)
- Spark 1.4.1 (Hadoop 1)
- Spark 1.5.2 (Hadoop 1)
- Spark 1.6.0 (Hadoop 1)
- Spark 1.6.1 (Hadoop 1)
- Spark 1.6.1 (Hadoop 2)

l automatically t
de your Databric

Thank you!

databricks™

What packages are used together?



databricks