

# Apache Spark Usage in the Open Source Ecosystem

Hossein Falaki  
@mhfalaki

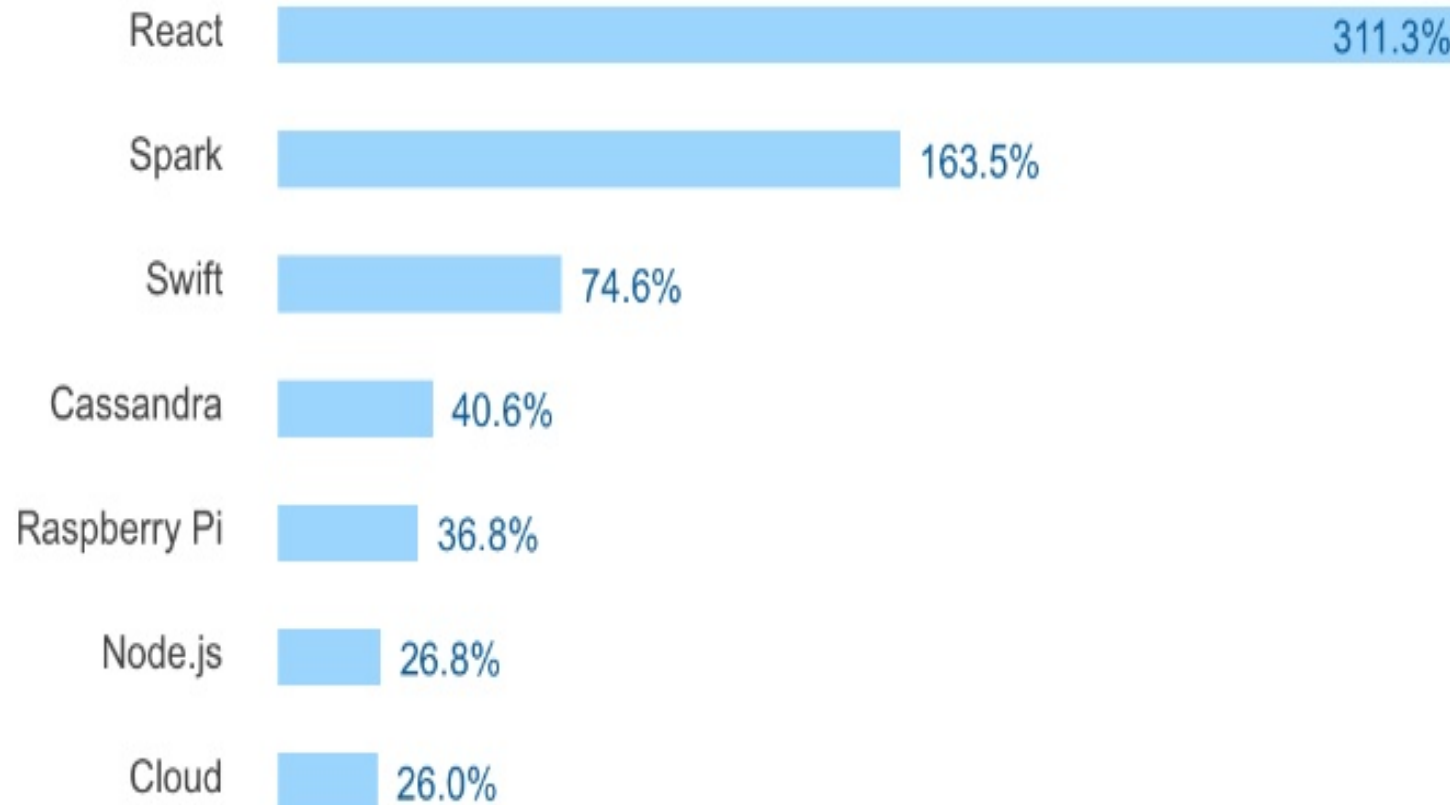
# About me

- 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

Winners

Losers



# 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





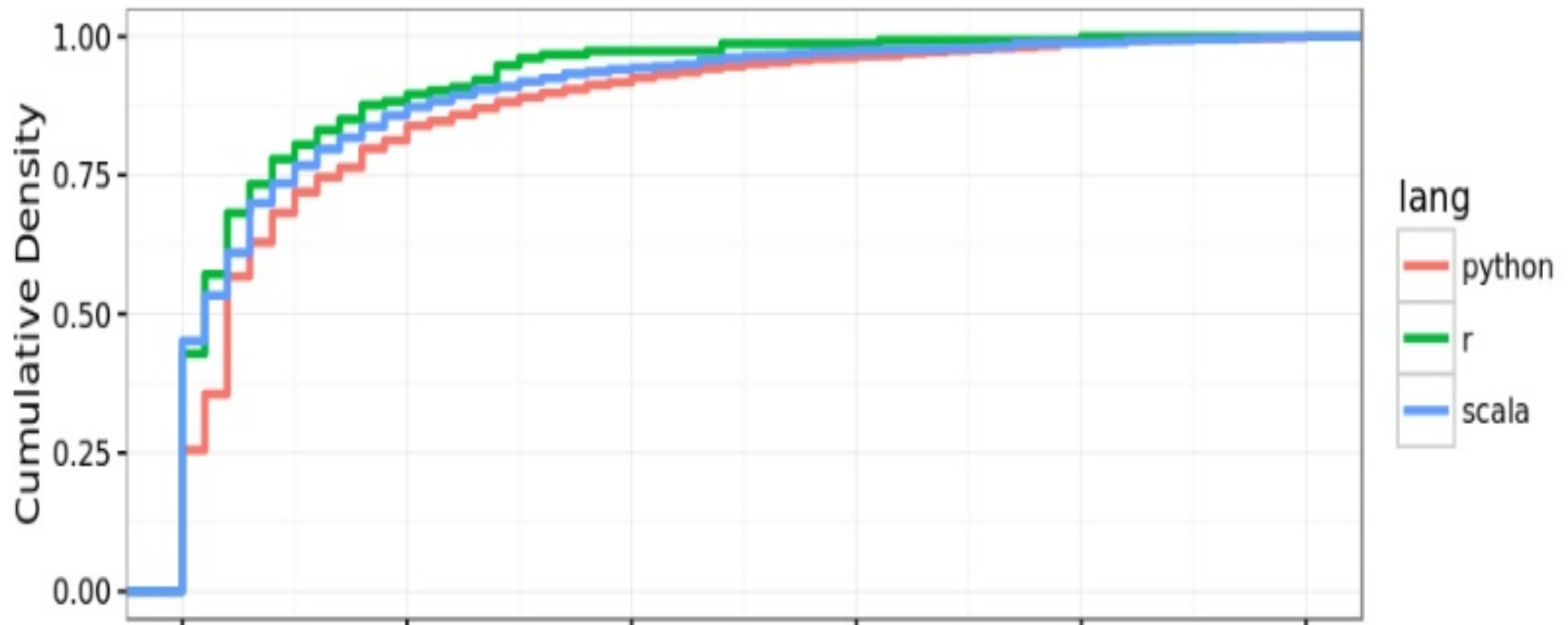
# Databricks Community Edition

- 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

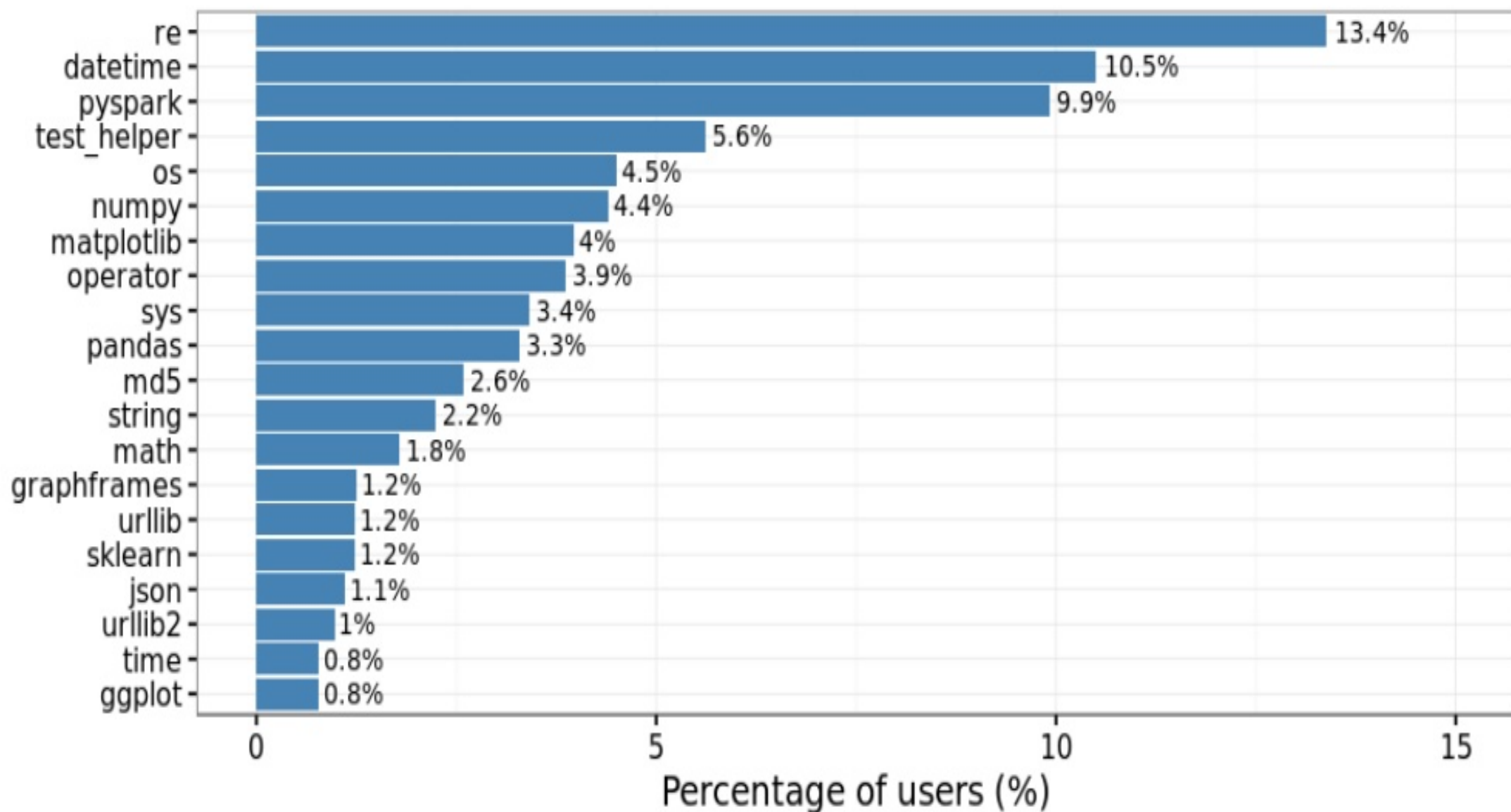
Egg File

Drop library egg here to upload

# Python Packages



# Most popular Python packages



# What is test\_helper?



» Package Index > test\_helper > 0.2

## PACKAGE INDEX >>

- [Browse packages](#)
- [Package submission](#)
- [List trove classifiers](#)
- [List packages](#)
- [RSS \(latest 40 updates\)](#)
- [RSS \(newest 40 packages\)](#)
- [Python 3 Packages](#)
- [PyPI Tutorial](#)
- [PyPI Security](#)
- [PyPI Support](#)
- [PyPI Bug Reports](#)
- [PyPI Discussion](#)
- [PyPI Developer Info](#)

## ABOUT >>

## NEWS >>

## DOCUMENTATION >>

## DOWNLOAD >>

## COMMUNITY >>

## FOUNDATION >>

## test\_helper 0.2

*A testing helper for scalable machine learning mooc*

Download  
test\_helper-0.2.tar.gz

### Not Logged In

- [Login](#)
- [Register](#)
- [Lost Login?](#)
- [Use OpenID](#)
- [Login with Google](#)

### Status

[Nothing to report](#)

File	Type	Py Version	Uploaded on	Size
<a href="#">test_helper-0.2.tar.gz</a> (md5)	Source		2015-05-11	1KB

**Author:** Daniel Liu

**Home Page:** [https://github.com/hpec/test\\_helper](https://github.com/hpec/test_helper)

**Download URL:** [https://github.com/hpec/test\\_helper/tarball/0.1](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

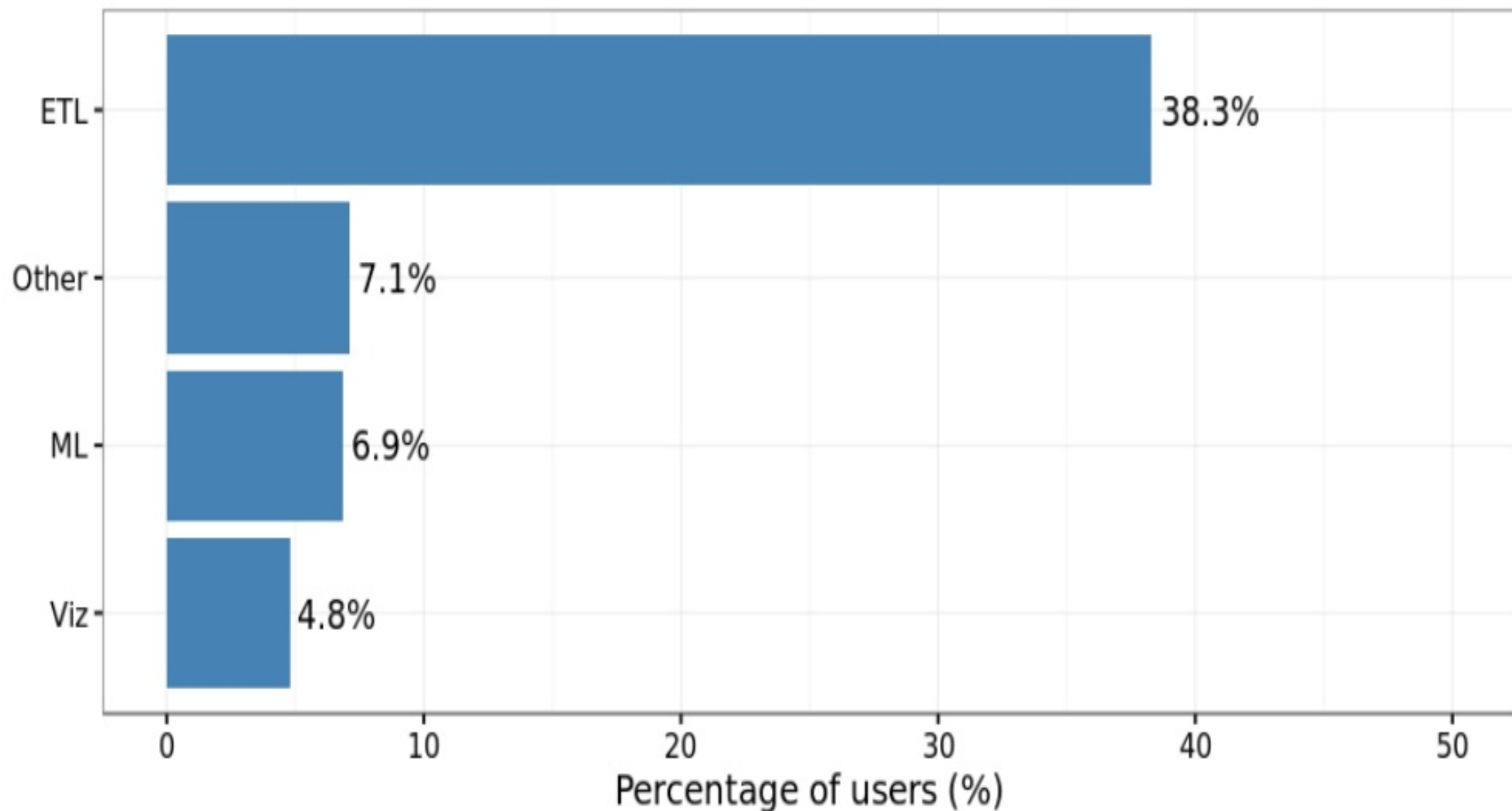
## Advanced analytics

- numpy
- sklearn
- graphframes
- tensorflow
- scipy

## Other

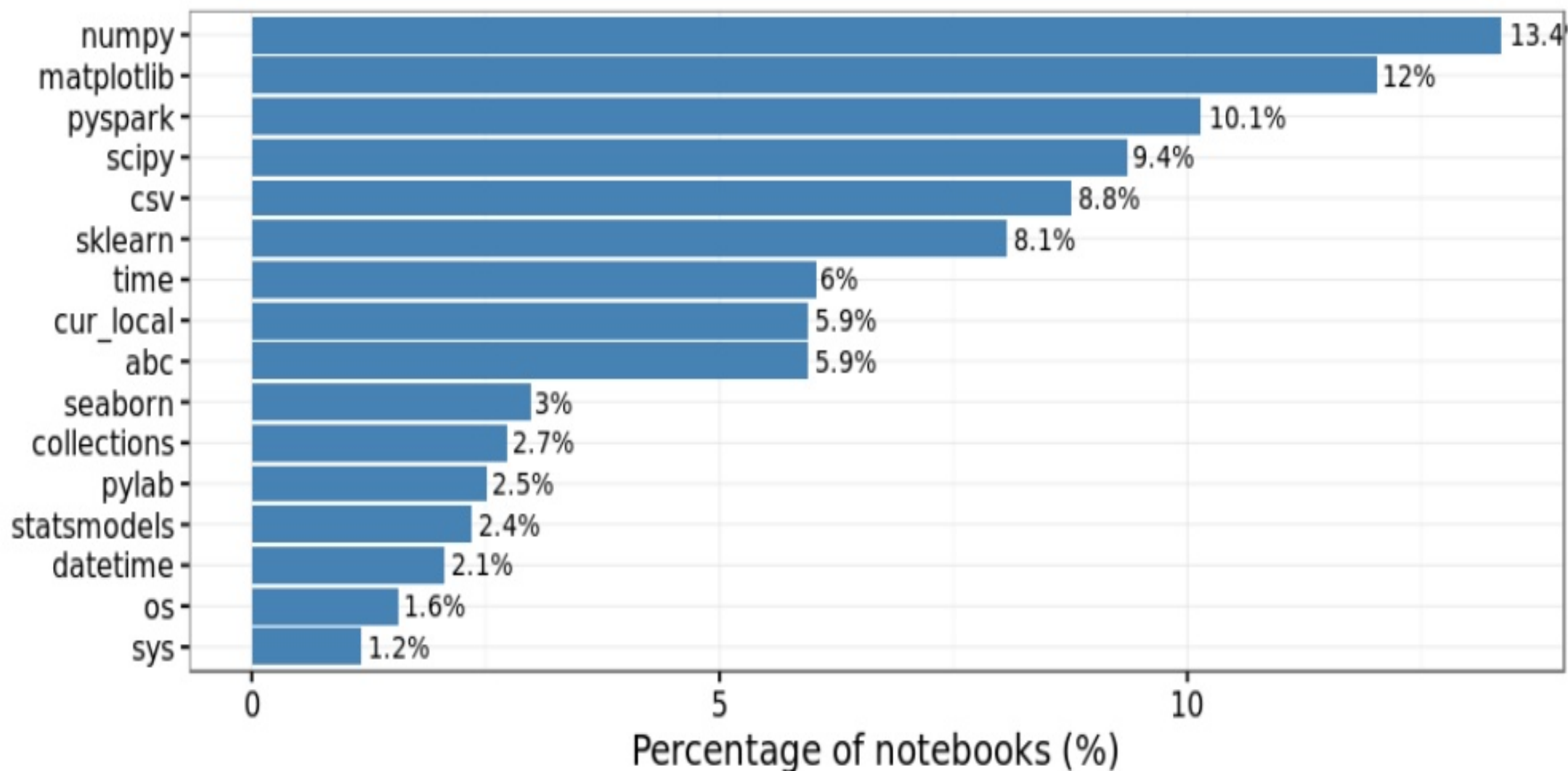
- test\_helper
- os
- md5

# Python package categories



# What packages go together?

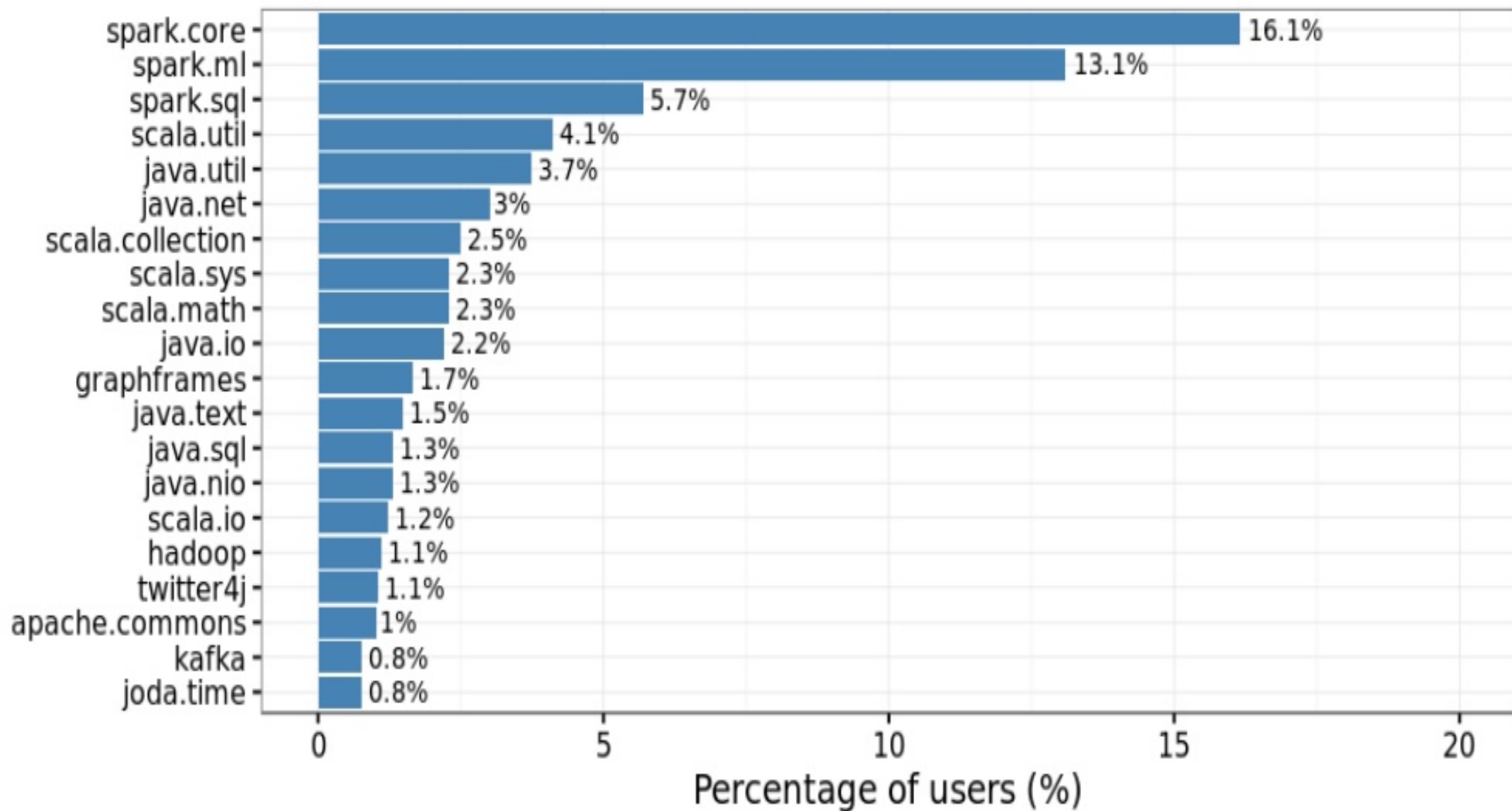
Notebooks that used pandas also used



# Scala Packages



# Most popular Scala libraries



# What are these?

## ETL

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

## Visualization

- ?

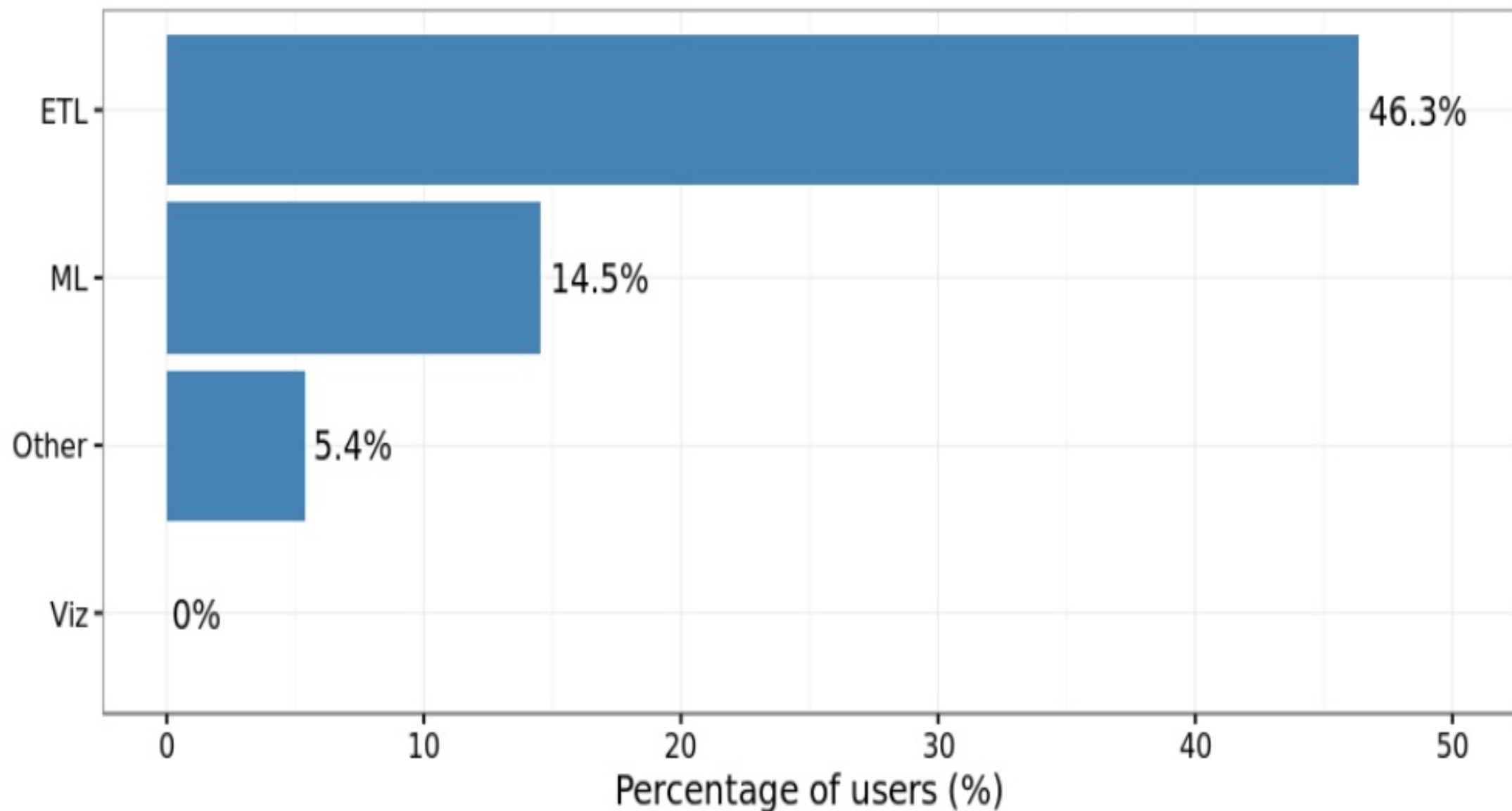
## Advanced\_analytics

- spark.ml
- graphframes

## Other

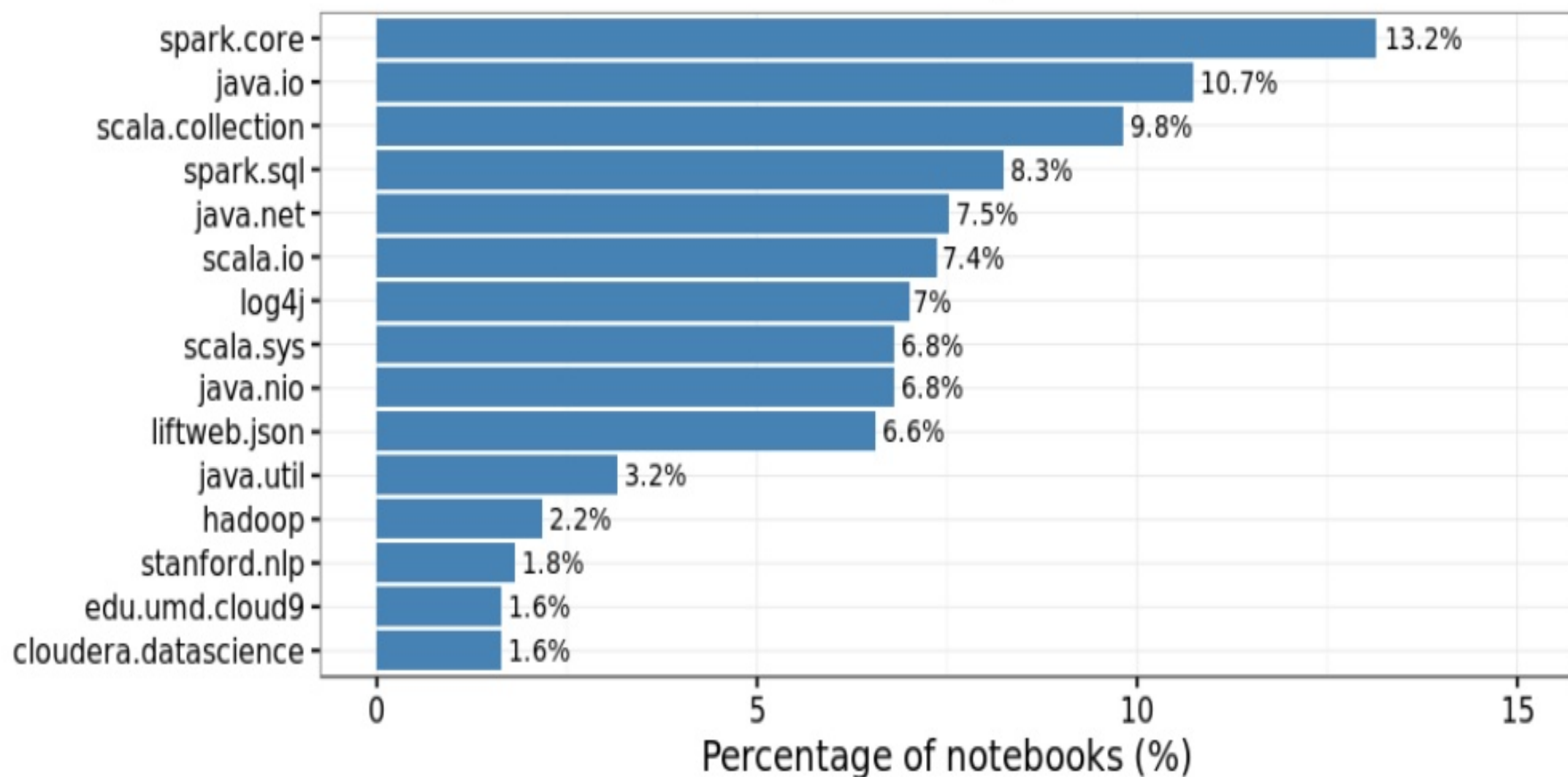
- java.net
- scala.sys

# Scala package categories



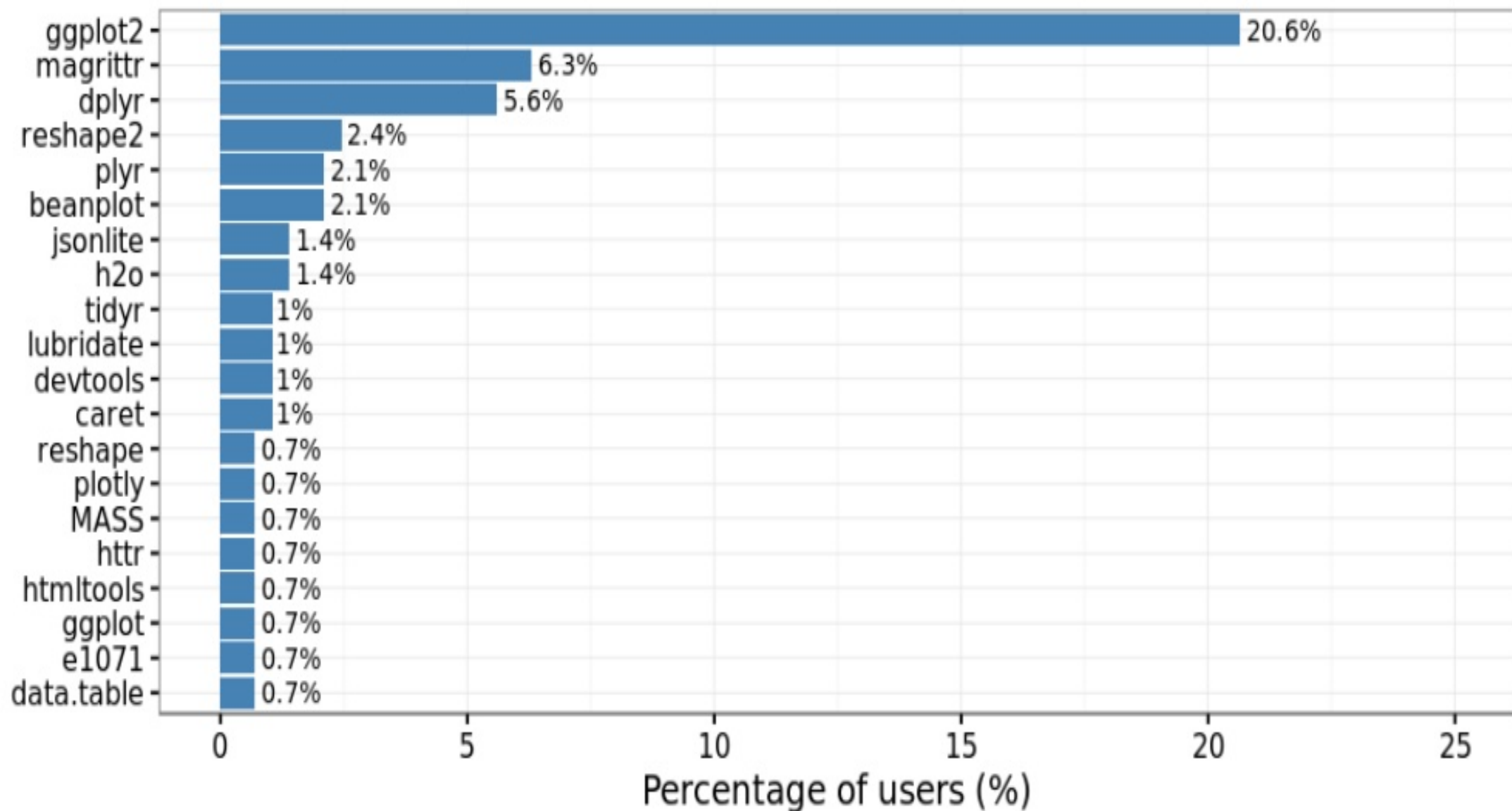
# What libraries go together?

Notebooks that used spark.ml also used



# R Packages

# Most popular R packages





# What are these?

## ETL

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

## Visualization

- ggplot2
- beanplot
- plotly
- ...

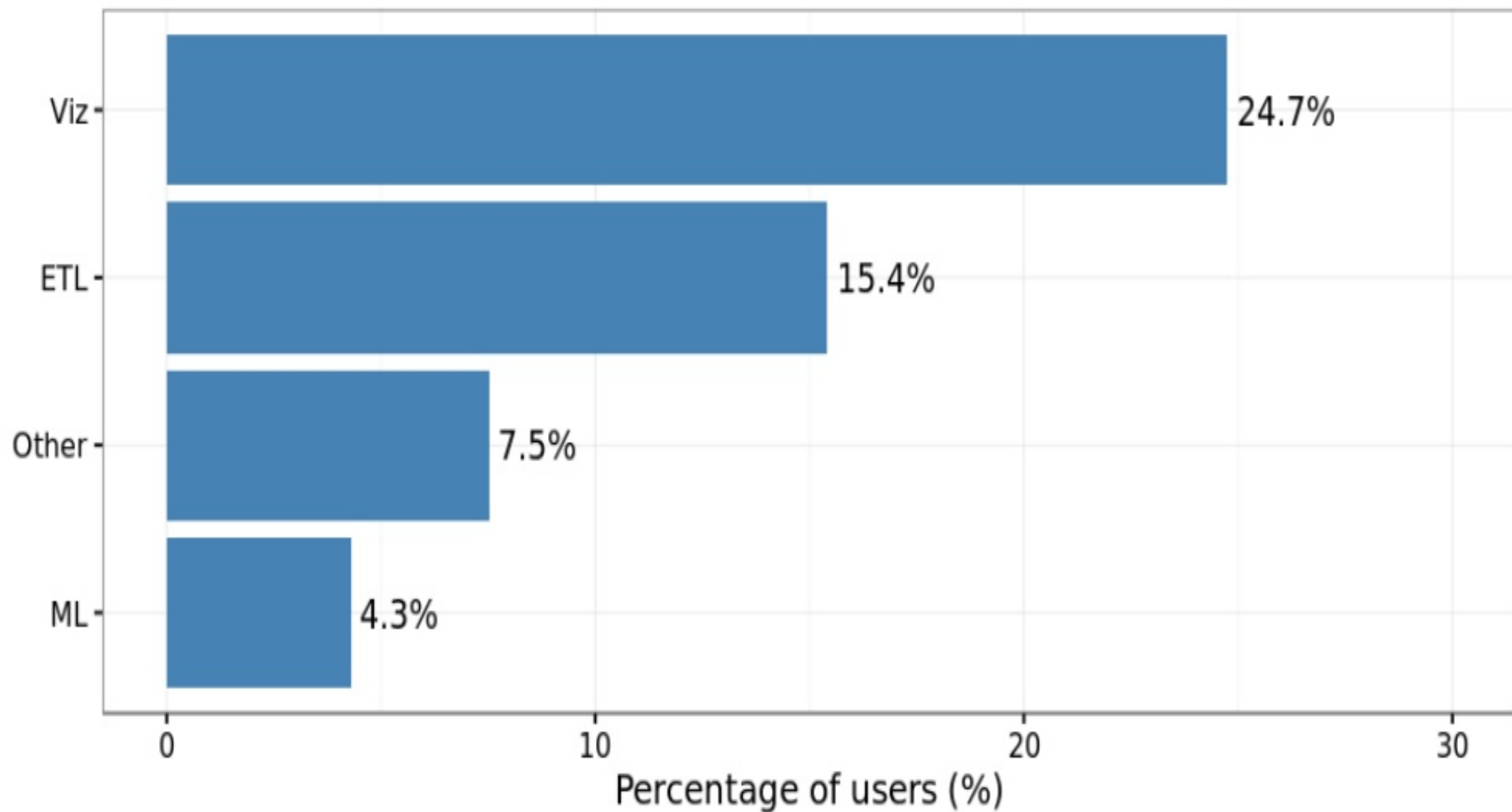
## Advanced analytics

- sparkr
- h2o
- caret
- e1071

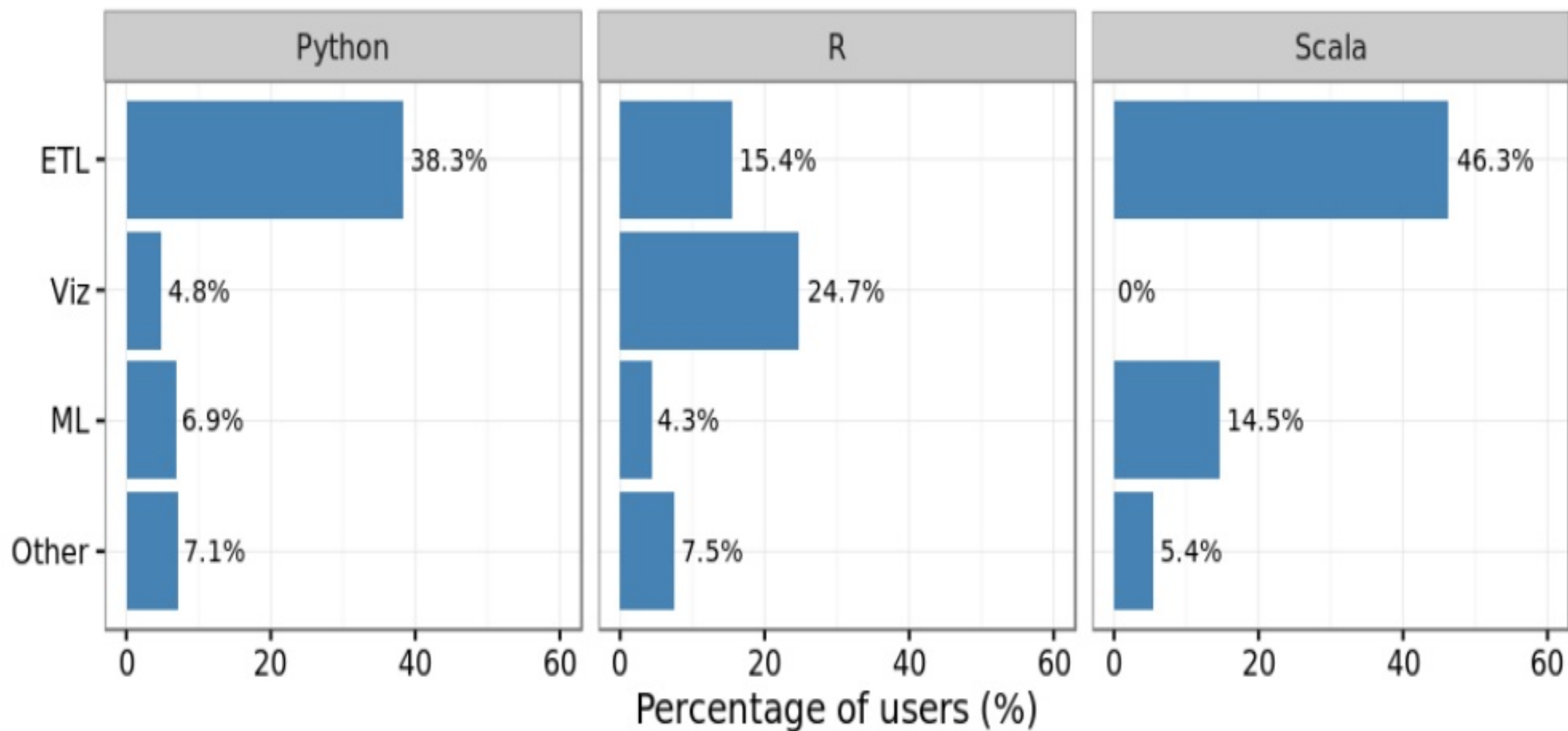
## Other

- devtools
- magrittr

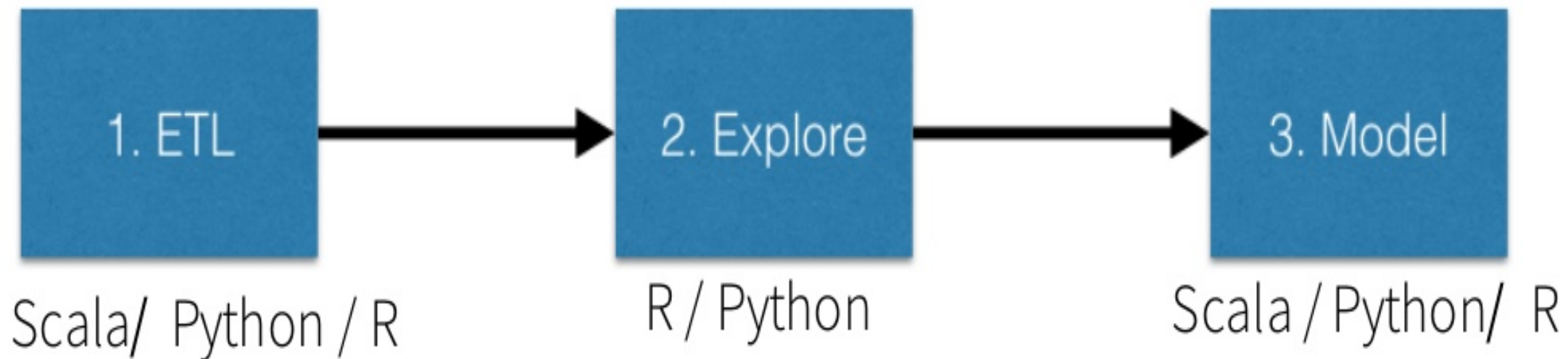
# R package categories



# Comparing Python, Scala & R



# Languages have unique features



- 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

Cancel

Create 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)

I automatically t  
de your Databricks





Thank you!

# What packages are used together?

Notebooks that used sparkr also used

