Все, что вы хотели знать об инструментах для Data Science, но боялись спросить

Data Science в России и мире

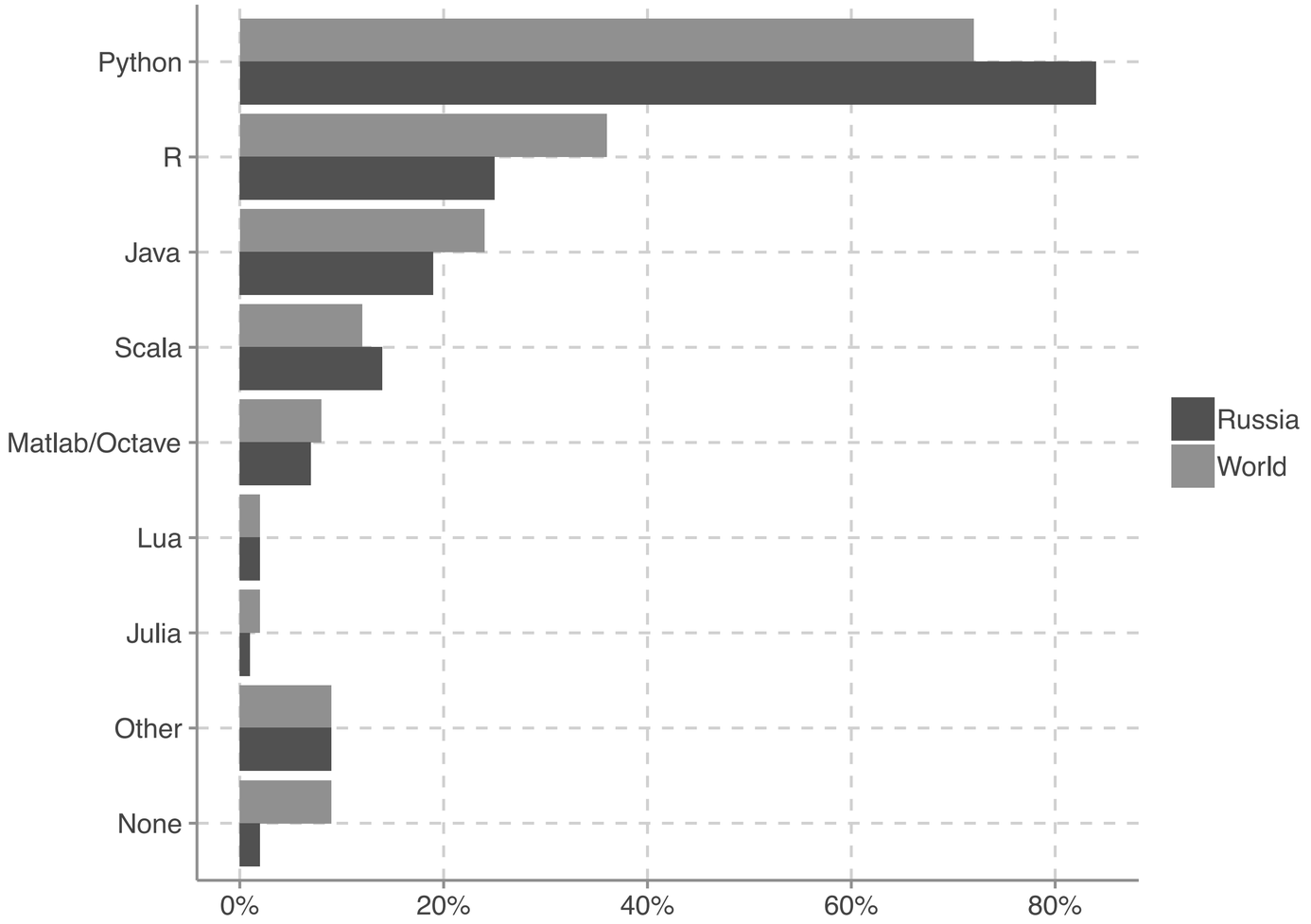
Виталий Худобахшов

1. Языки программирования

Доля специальных языков программирования, используемых в Data Science, стремительно сокращается

Распределение языков в Data Science

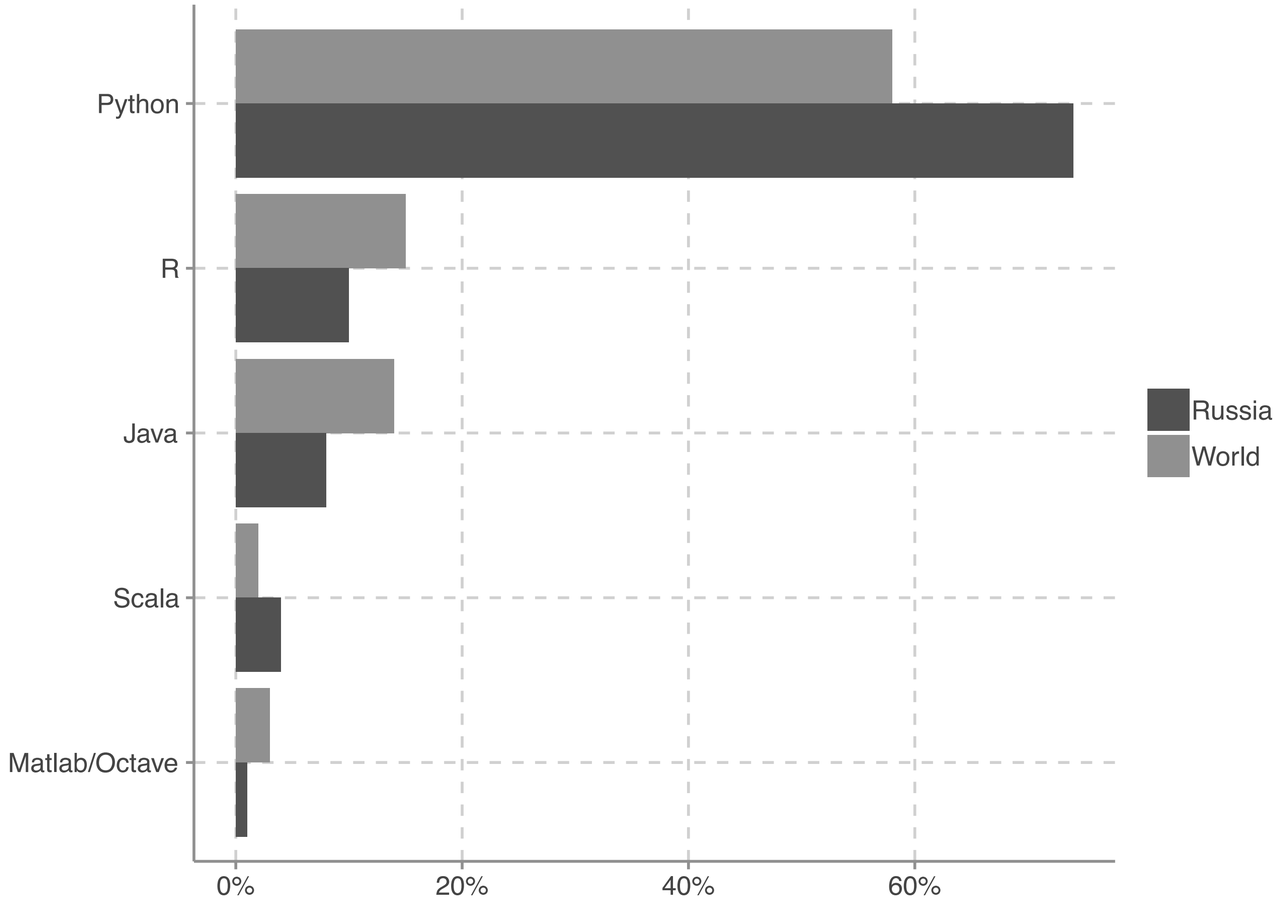
- Язык **R** в мире (**36%**) куда популярнее, чем в России (**25%**).
- **Java** занимает 3-ю позицию.
- Доля тех, кто не использует языки программирования, в мире выше.



Основной язык программирования

—

- Среди респондентов в мире доля тех, кто в качестве основного языка использует **R** выше, чем тех, кто использует **Java.**
- **Python,** как основной язык, менее популярен в мире чем в России.
- За пределами России доля **Matlab/Octave** выше, чем **Scala.**



Студенты

98% опрошенных студентов в России пишут на **Python**

2. Библиотеки для анализа данных

—

R, python, shiny, dplyr, purrr, ditto, ggplot2, spark, sawk, pyspark, sparkr, jupyter, vulpix, git, numpy, pandas, feebas, scikit-learn, h2o.ai, sparklin-water, tensorflow, keras, onyx, ekans, hadoop, scala, metapod

Что из этого фреймворки, а что покемоны?

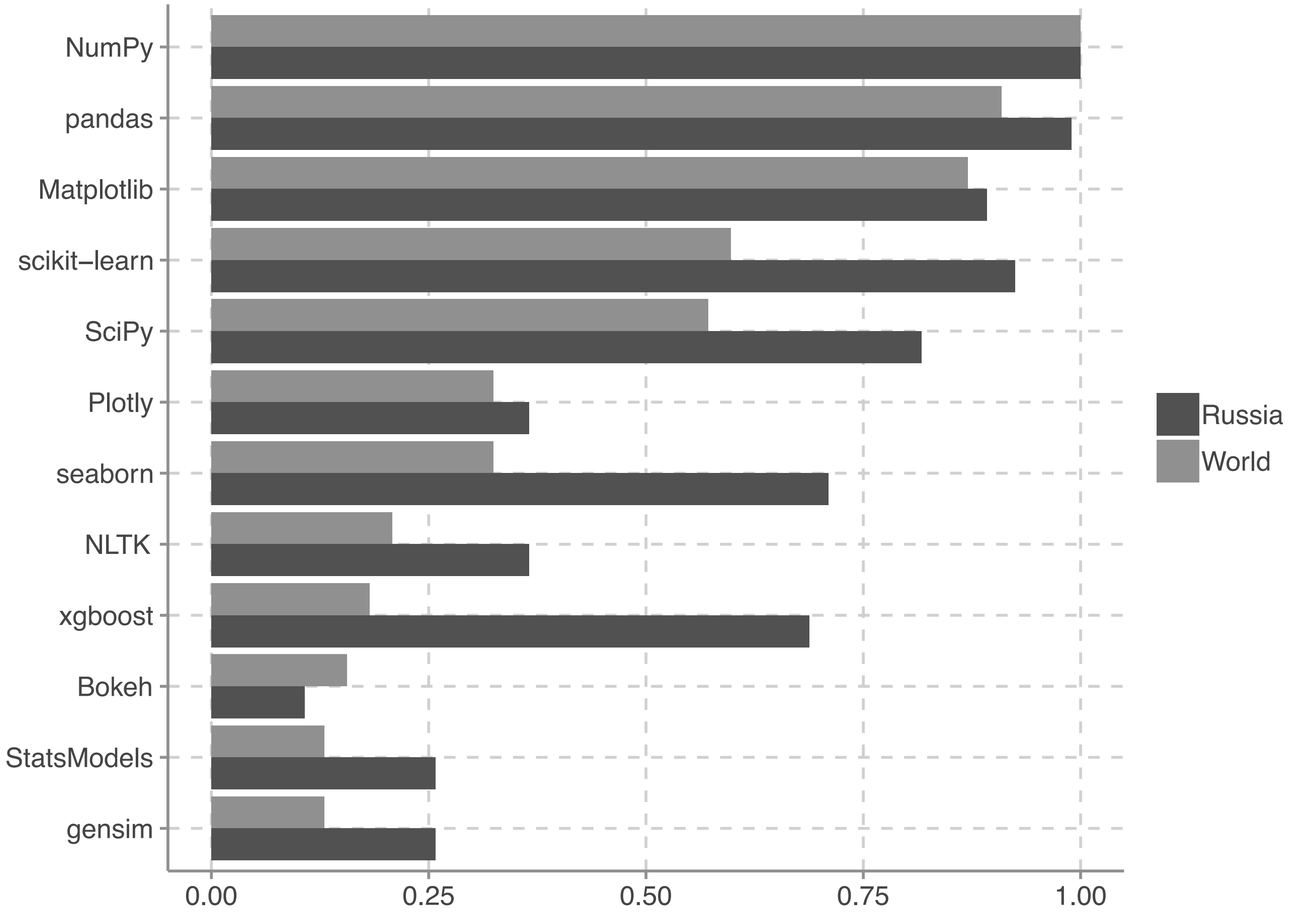
Теорема о покемонах

—

Теорема (Худобахшов В. А.)**.** *Для любого покемона рано или поздно найдется хотя бы один репозиторий в GitHub с таким именем.*

Библиотеки для Python

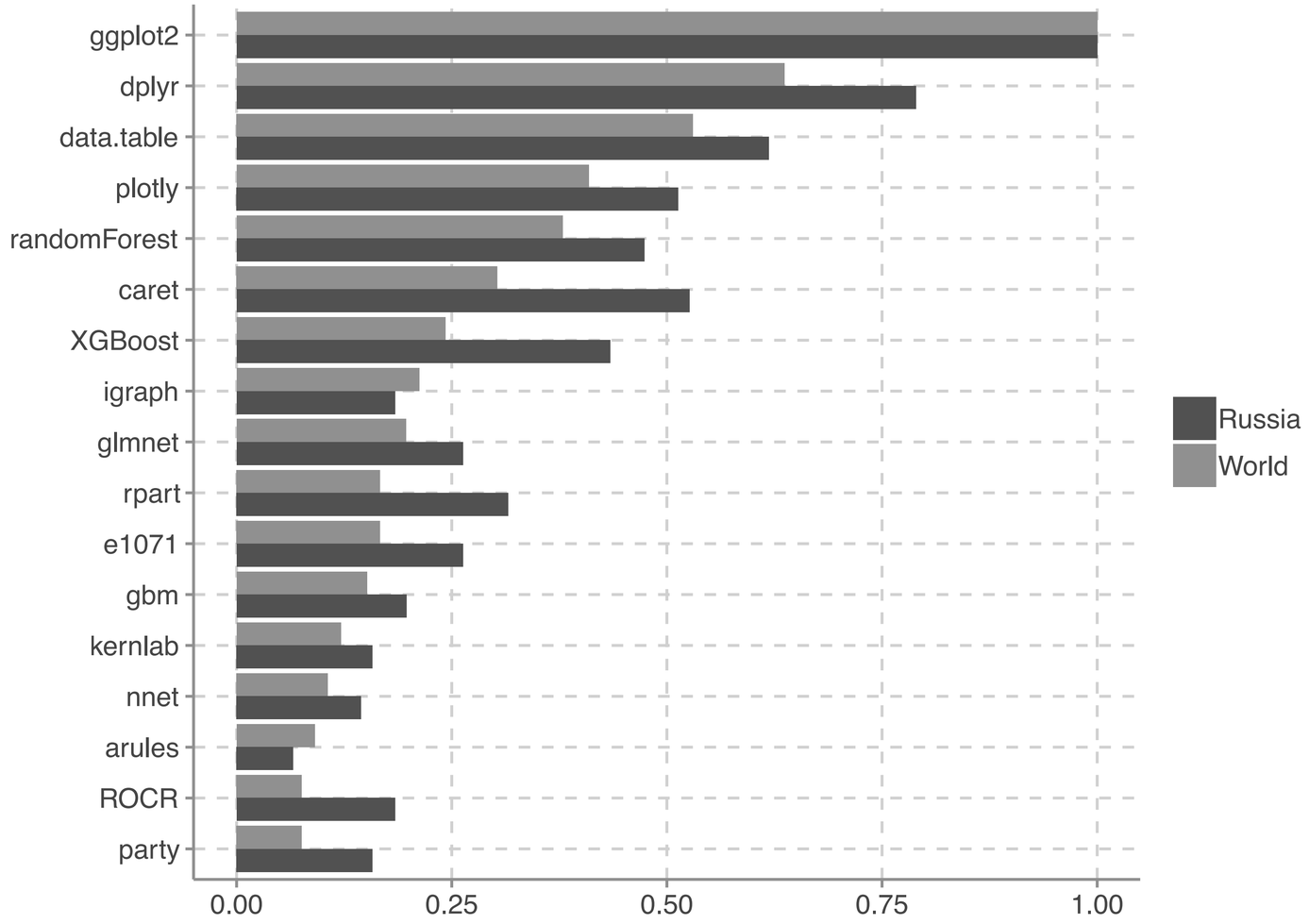
- Популярность **XGBoost** и **Seaborn** - это чисто российская специфика.
- **В России люди используют почти все подряд (или так говорят).**



Библиотеки для R

—

- Все указывает на то, что люди, пишущие на **R**, в среднем более «добросовестны».
- **ggplot2** лидирует по популярности со значительным отрывом.



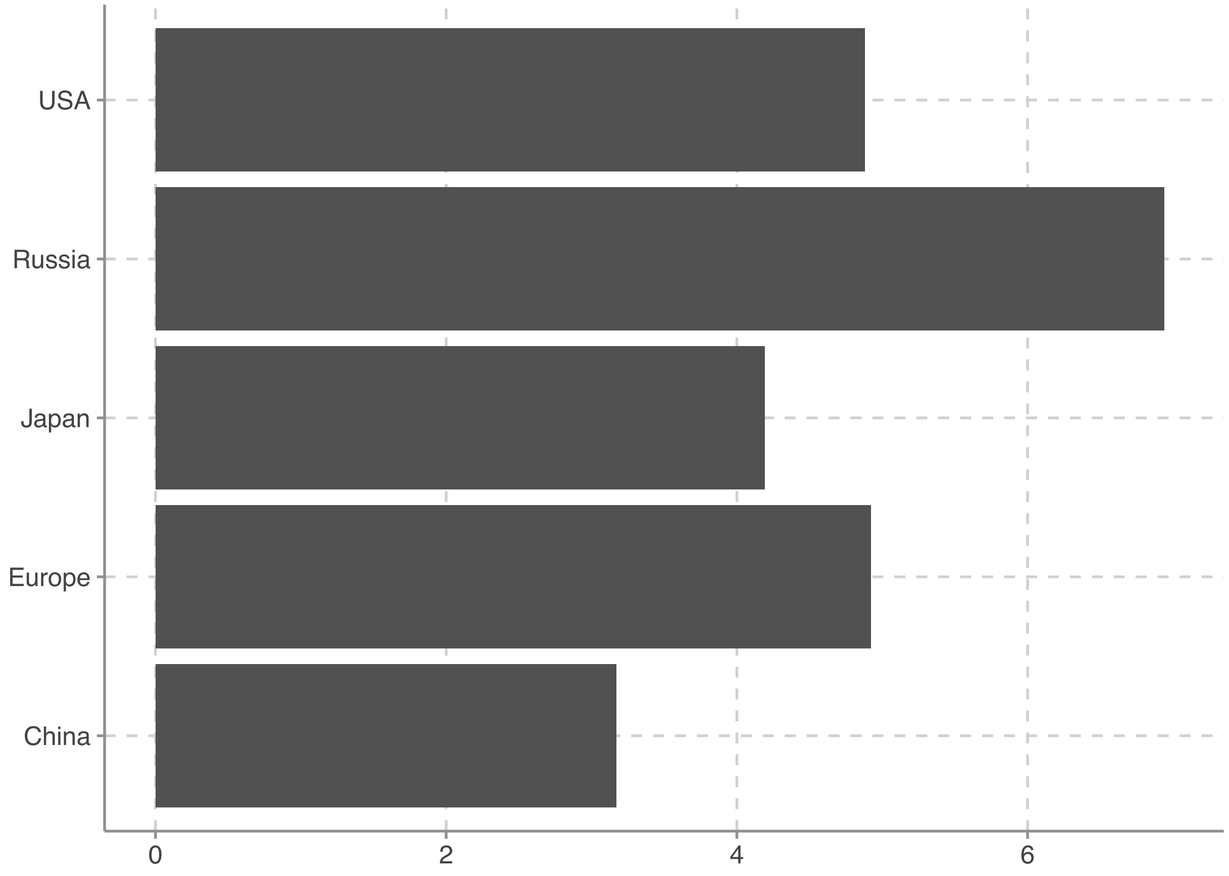
Среднее количество Python библиотек на человека

4.38

в мире

6.94

в России



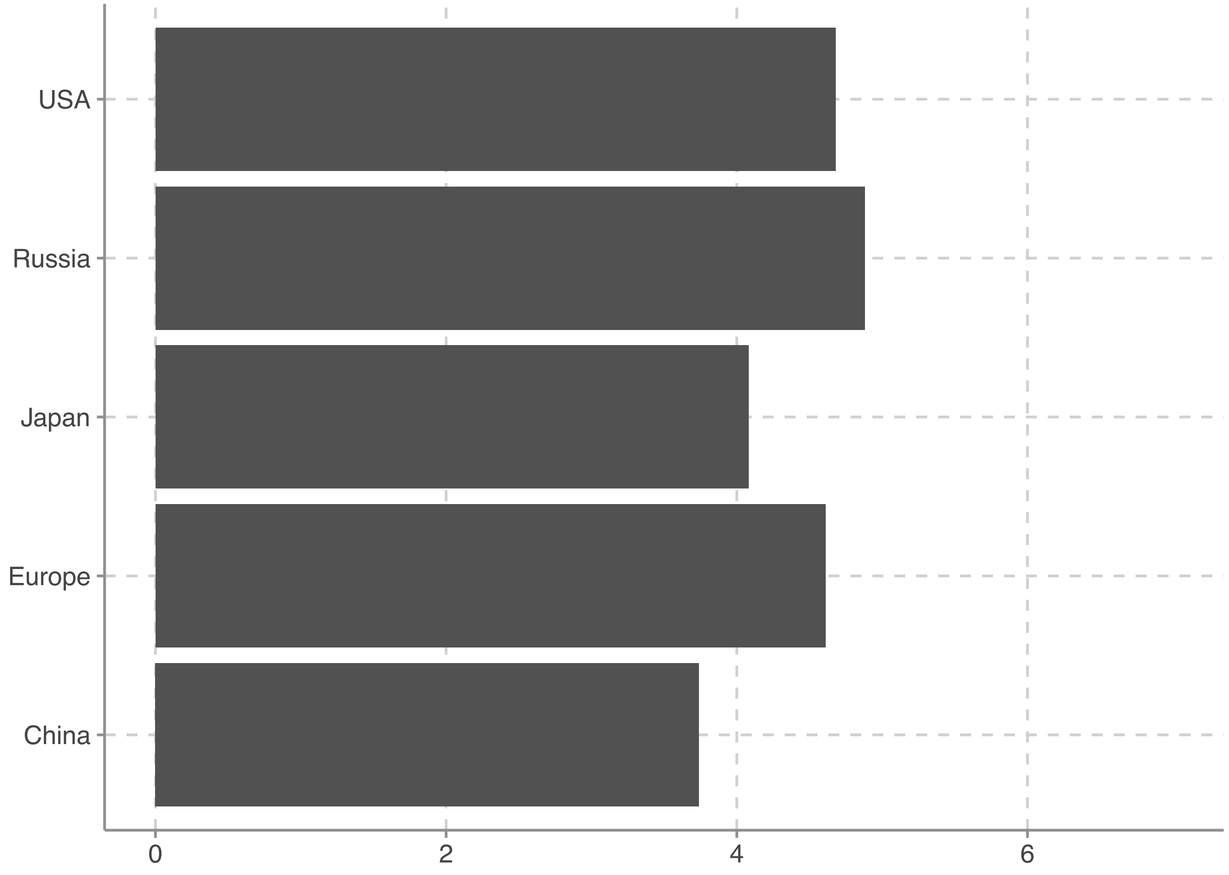
Среднее количество R библиотек на человека

4.38

в мире

4.88

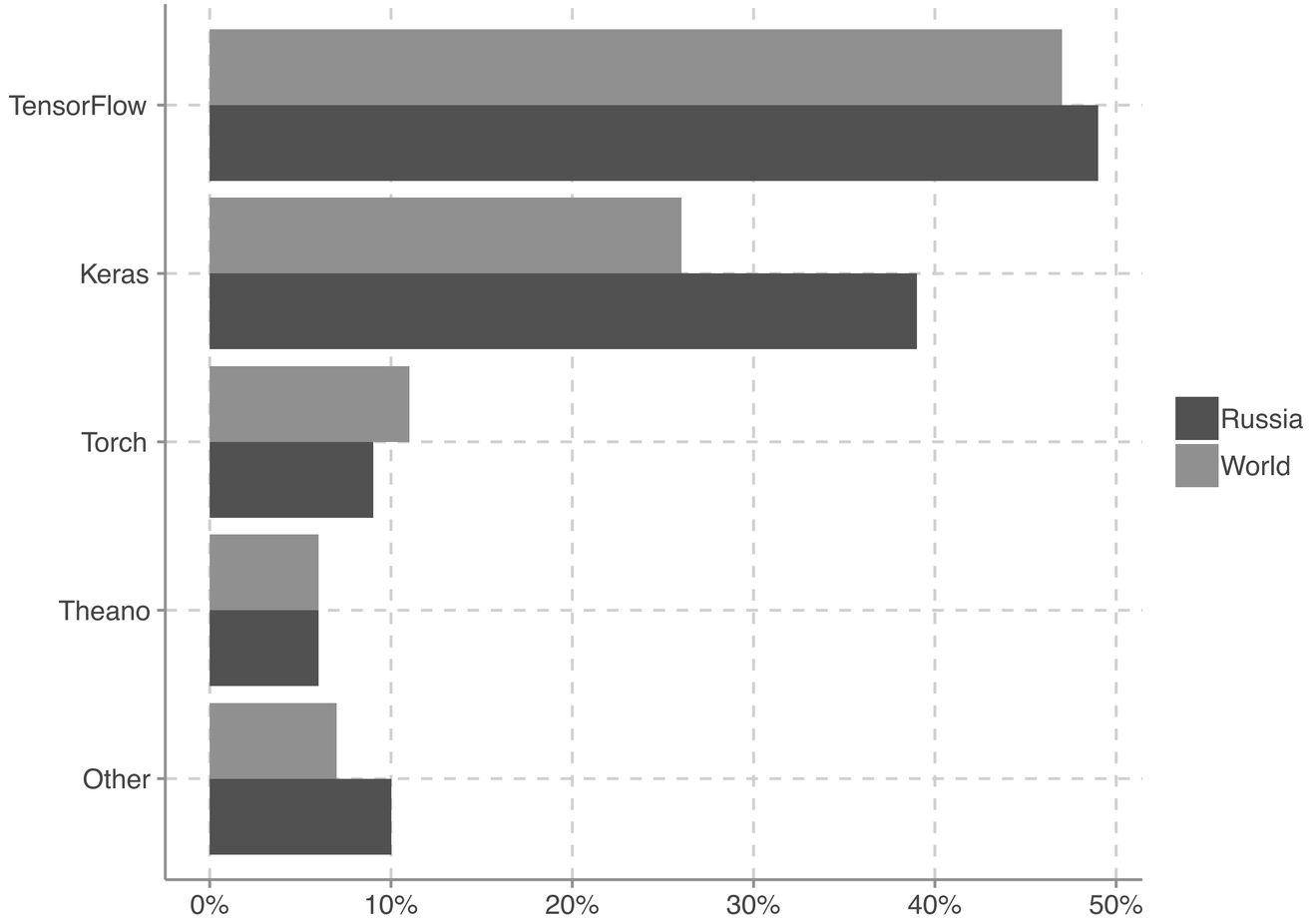
в России



Библиотеки для глубокого обучения

60%

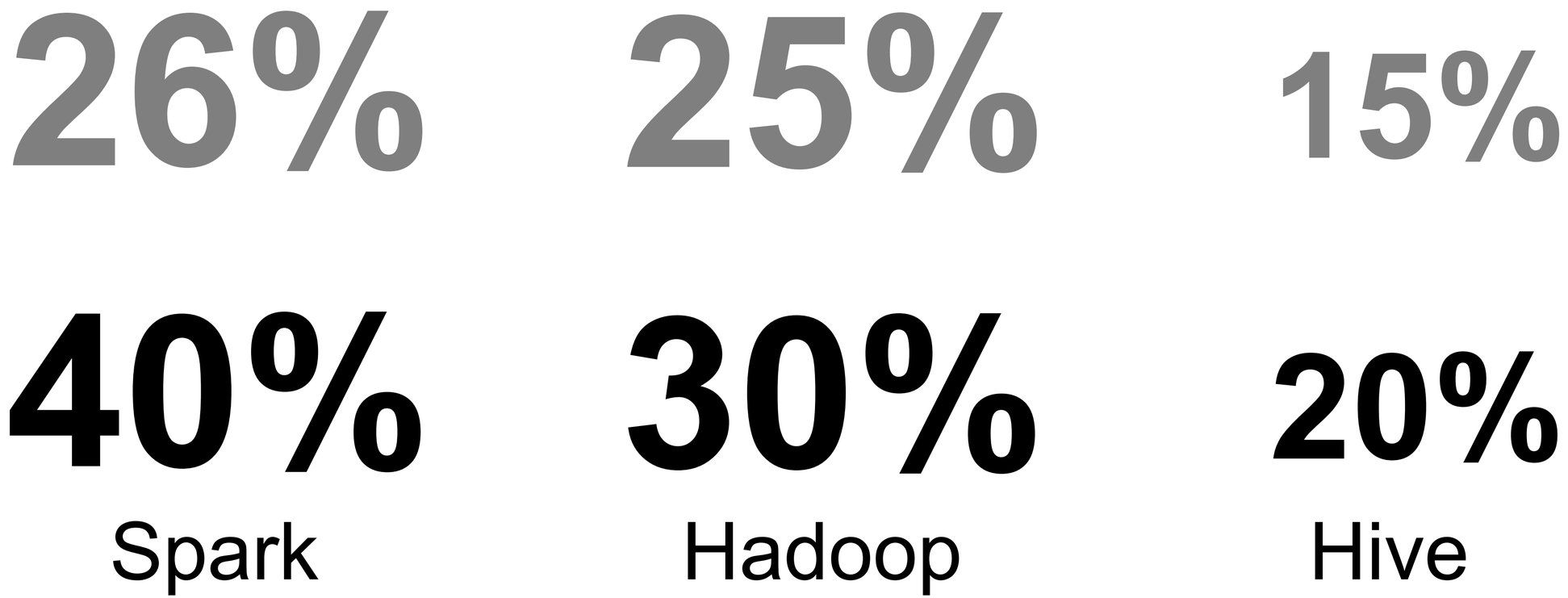
респондентов используют фреймворки для глубокого обучения



3. Мои большие данные больше чем твои!

50% респондентов так или иначе занимаются обработкой больших данных

Самые популярные технологии в больших данных



Аутсайдеры

Apache Pig, Apache Beam, Apache Flink, Apache Tez, Apache Samza и Dask находятся ниже отметки

5%

Использование Apache Spark вне JVM

Люди готовы пойти на любые мучения, лишь бы не учить **Scala**

19%

PySpark

13%

SparkR

Вычислительные ресурсы

30% респондентов используют облачные сервисы для анализа данных

Вычислительные ресурсы

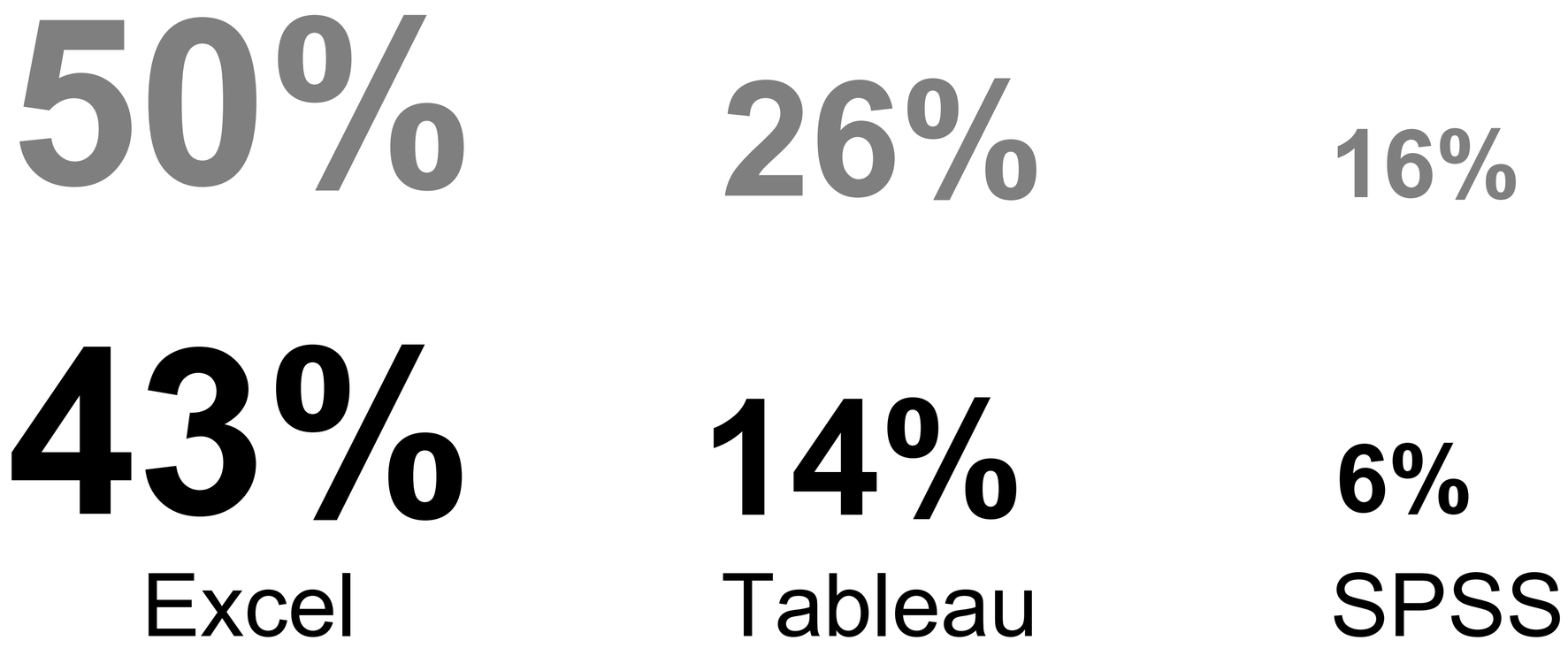
>20% респондентов используют кластер из **4 – 10** узлов

4. Визуальные средства анализа данных

—

В эпоху тотального неумения программировать важнейшими инструментами анализа данных являются **Excel** и **Tableau**

Визуализация данных



Визуальные средства анализа данных

—

Лидером рынка является **Azure ML** от Microsoft с долей **19%** в мире и **5%** в России

59%

в мире

20%

в России

5. Люди и деньги

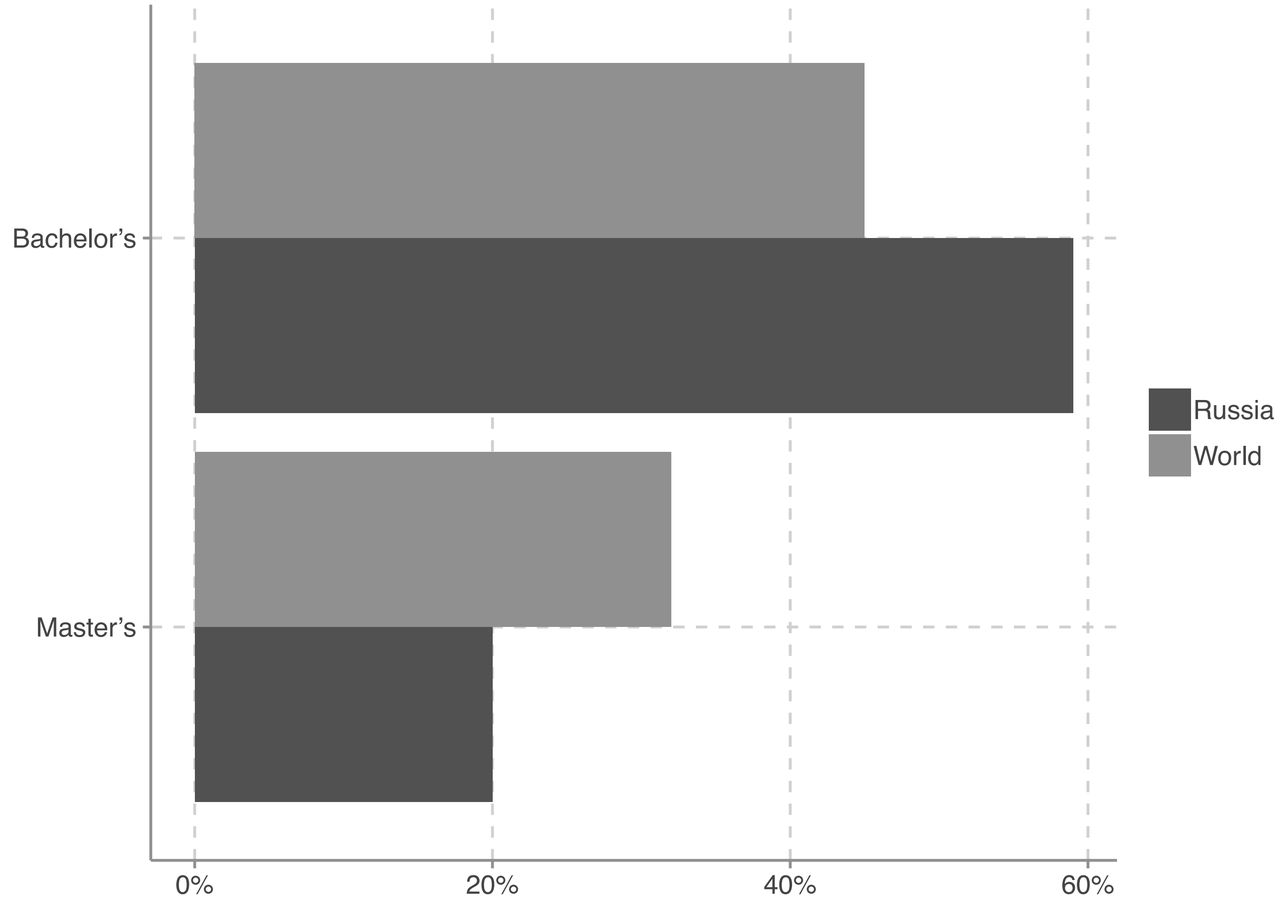
—

Только лишь от 40% до 50% респондентов занимаются анализом данных как **основной деятельностью**

Образование

—

- Уровень образование респондентов за границей несколько выше, чем в России.
- Но только **6%** за границей имеют степень выше магистерской.
- В России **7%** специалистов и **10%** аспирантов.



Зарплаты в России

—

Средняя зарплата специалиста по анализу данных на **R** и **Python**

140 т.р.

Зарплаты в России

25% респондентов получают больше **180 т.р.**

Спасибо за внимание

—

jetbrains.com