СБЕРБАНК

APACHE SPARK
ДЛЯ ДАТА ИНЖИНИРИНГА

КОРОТКО О НАС

Newprolab

- Занимаемся обучением работе с данными с 2015 года.
- Широкая линейка программ про данные на рынке.
- Используем андрагогику и интенсивные практикоориентированные программы.

История

Выпускники:

- 800+ с открытых программ
- 850+ с корпоративных

3 Big Data

1 Deep Learning

2 Big Data

1 Data Engineer

3 for Executives

3 Deep Learning

2 Big Data

2 Data Engineer

2 for Executives

1 Deep Learning

2 Big Data

2 Scala

2 Data Engineer

3 for Executives

1 Deep Learning

2 Big Data

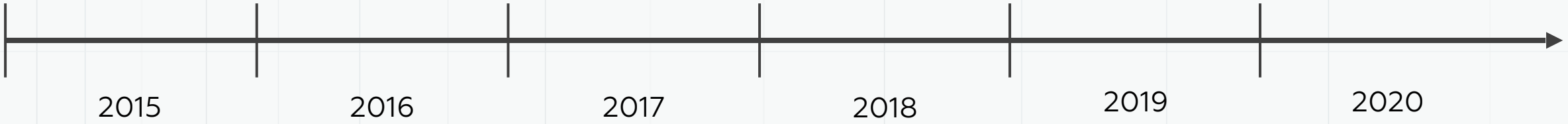
1 Scala

2 Data Engineer

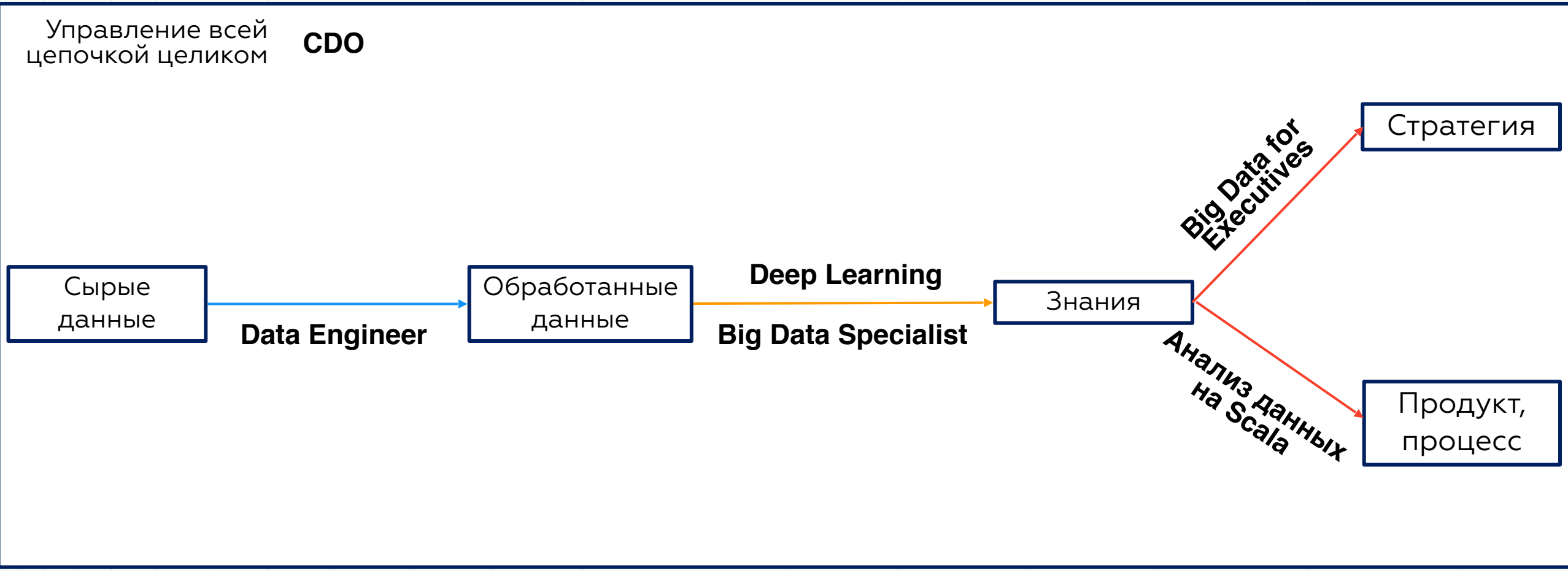
Клуб **CDO**

2 Deep Learning

2 Big Data



Линейка программ



О ПРОГРАММЕ

Цели

Научить работе со Spark для различных задач по дата инжинирингу: от их предобработки и формирования витрин до построения задач по мониторингу и дообучению моделей машинного обучения.

Формат

- 2 занятия в неделю
- Каждое занятие: 3 часа.
- 10 лаб (практических домашних задач).
- 3 теста.

Занятия

1. Hadoop для Spark-пользователя.
2. Введение в Scala.
3. Введение в Spark.
4. Spark Dataframes I, II, III.
5. Spark Structured Streaming I, II.
6. Разборы лаб и Q&A 1, 2, 3.
7. Spark Structured Streaming III, IV.
8. Разборы лаб и Q&A 4, 5, 6.
9. Spark ML-инжиниринг.
10. Мониторинг и оптимизация Spark.
11. Разборы лаб и Q&A 7, 8.

Лабы

1. Дескриптивный анализ рейтингов фильмов на Scala.
2. Подбор топ-350 релевантных URL для автолюбителей.
3. Построение витрины данных из разных внешних источников (Cassandra, PostgreSQL, Elasticsearch, HDFS).
4. Сохранение логов в Spark из Kafka по расписанию*.
5. Подготовка матрицы users x items по логам.
6. Подготовка расширенной матрицы фичей.
7. Обучение и дообучение ML-модели по расписанию*.
8. Мониторинг качества работы модели.

* есть доп. задание со звездочкой

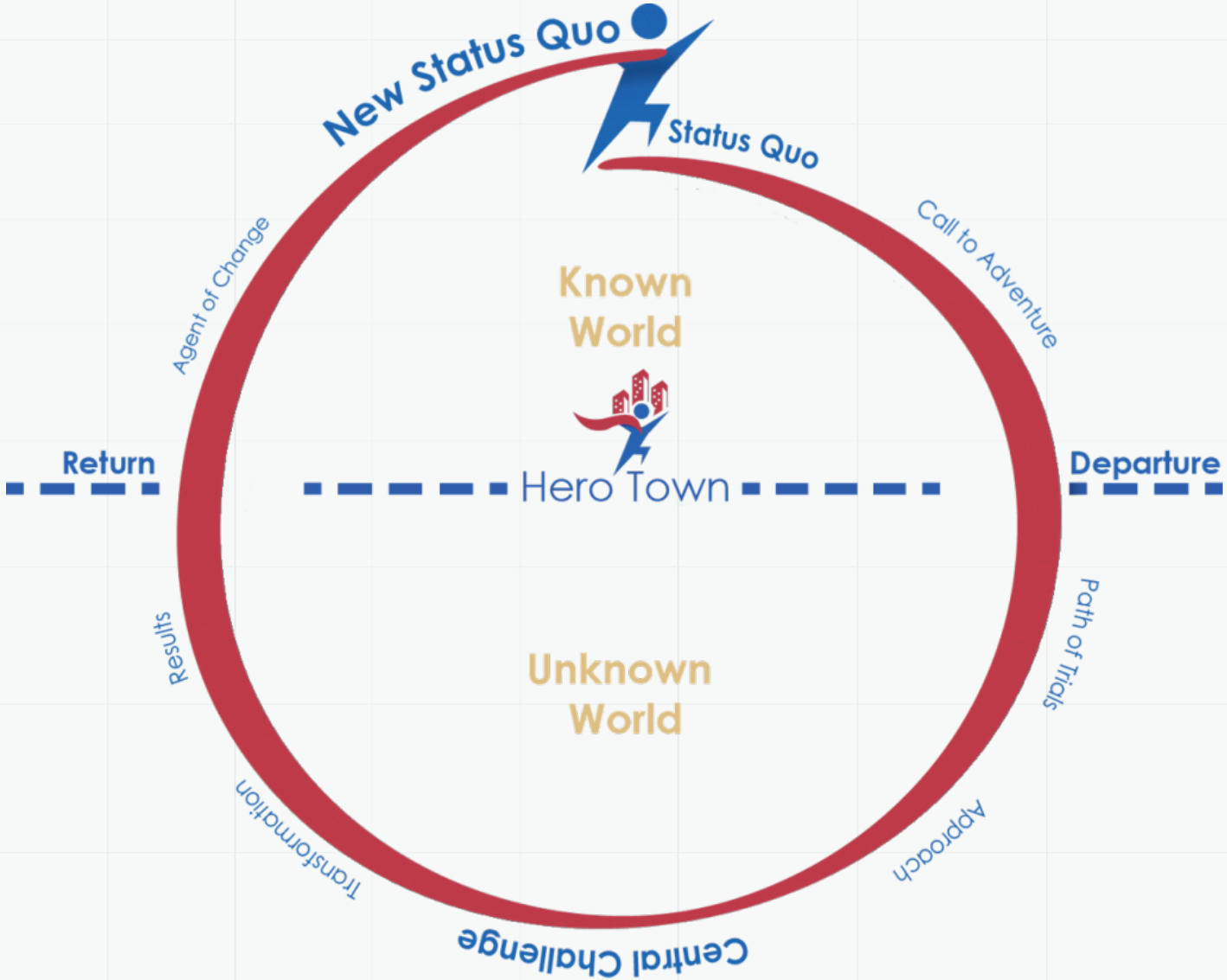
Расписание

Занятия будут проходить в Zoom.

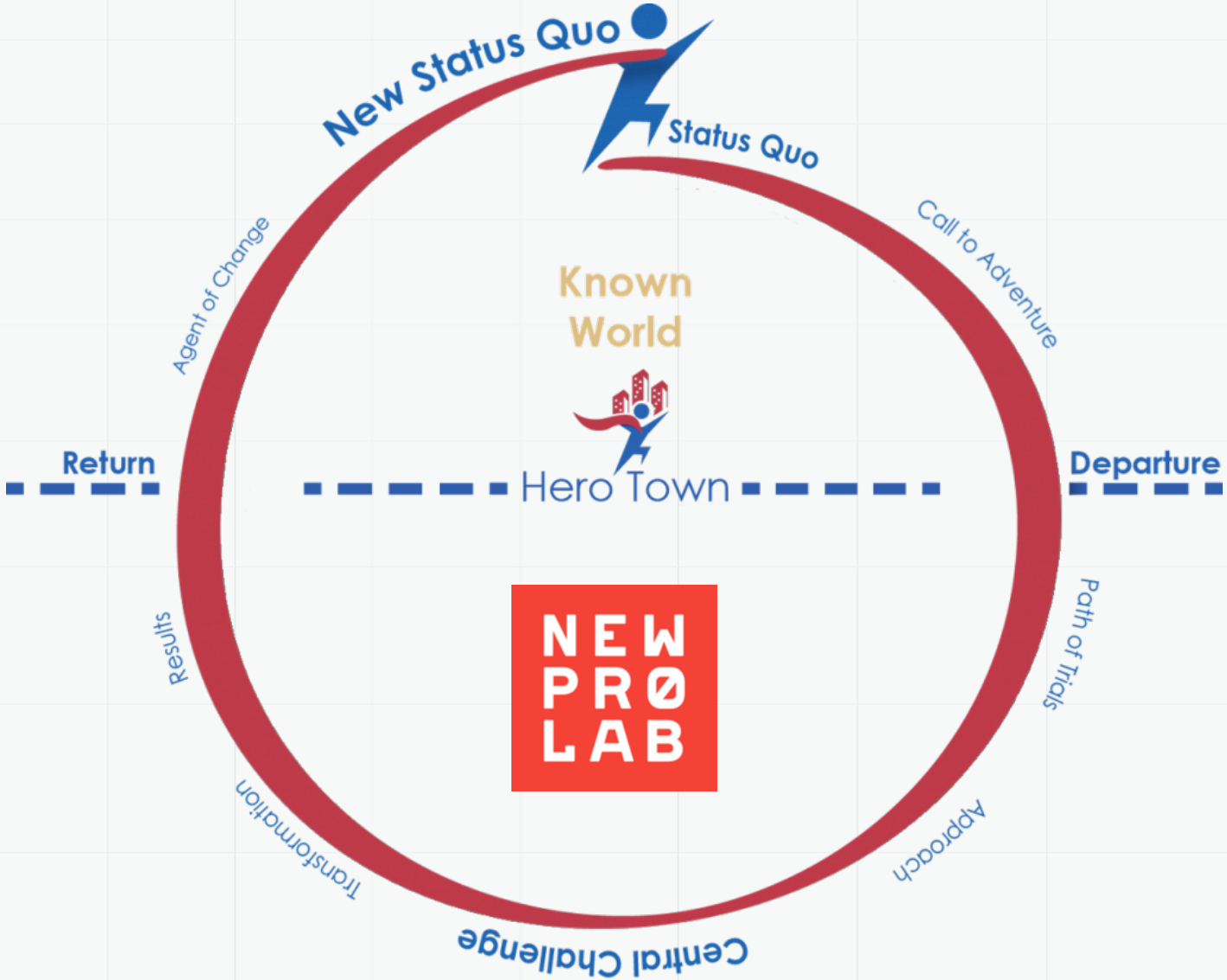
19:00 – 22:00.

ПРИНЦИПЫ

Hero's journey



Hero's journey



Главная идея

Баланс между самостоятельностью и поддержкой.

Другие принципы

- Занятия: задавайте вопросы (не существует глупых вопросов).
- Лабы: просите помощи у сокурсников и координаторов, но вначале попробуйте решить сами.
- Делайте лабы заранее. В последний момент можно не успеть.

ИНФРАСТРУКТУРА

Ресурсы

1. Общий на всех кластер со Spark 2.4.7. Конфигурация: 18 нод по 16 CPU, 80GB RAM. 2 мастера с 32 ядрами и 256GB RAM.
2. Доступ к кластеру по SSH и через JupyterHub.
3. Личный кабинет с календарем занятий и чекерами для лаб.
4. GitHub (**доступ!**).
5. Slack.

КОМАНДА ПРОГРАММЫ

Координаторы

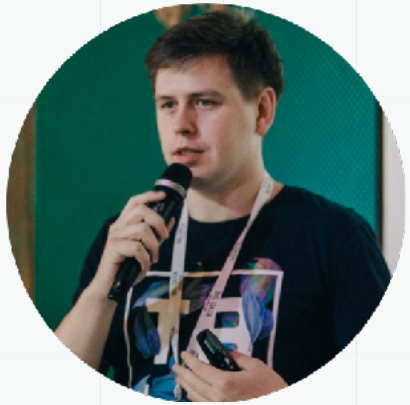
Виталий Монастырев

Senior Data Engineer
Grid Dynamics

Андрей Качетов

Head of ML operations
Альфа-Банк

Преподаватели

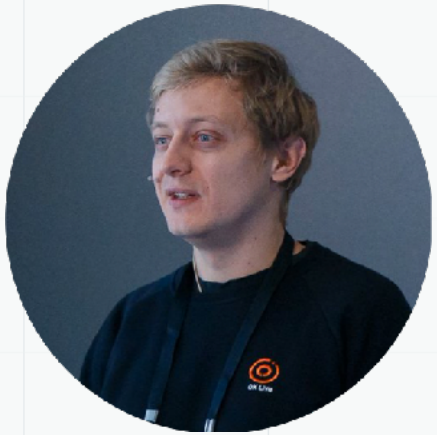


Егор Матешук

Технический директор
ГПМ Дата

Андрей Титов

Senior Spark Engineer,
NVIDIA



Дмитрий Бугайченко

Управляющий директор,
Сбербанк

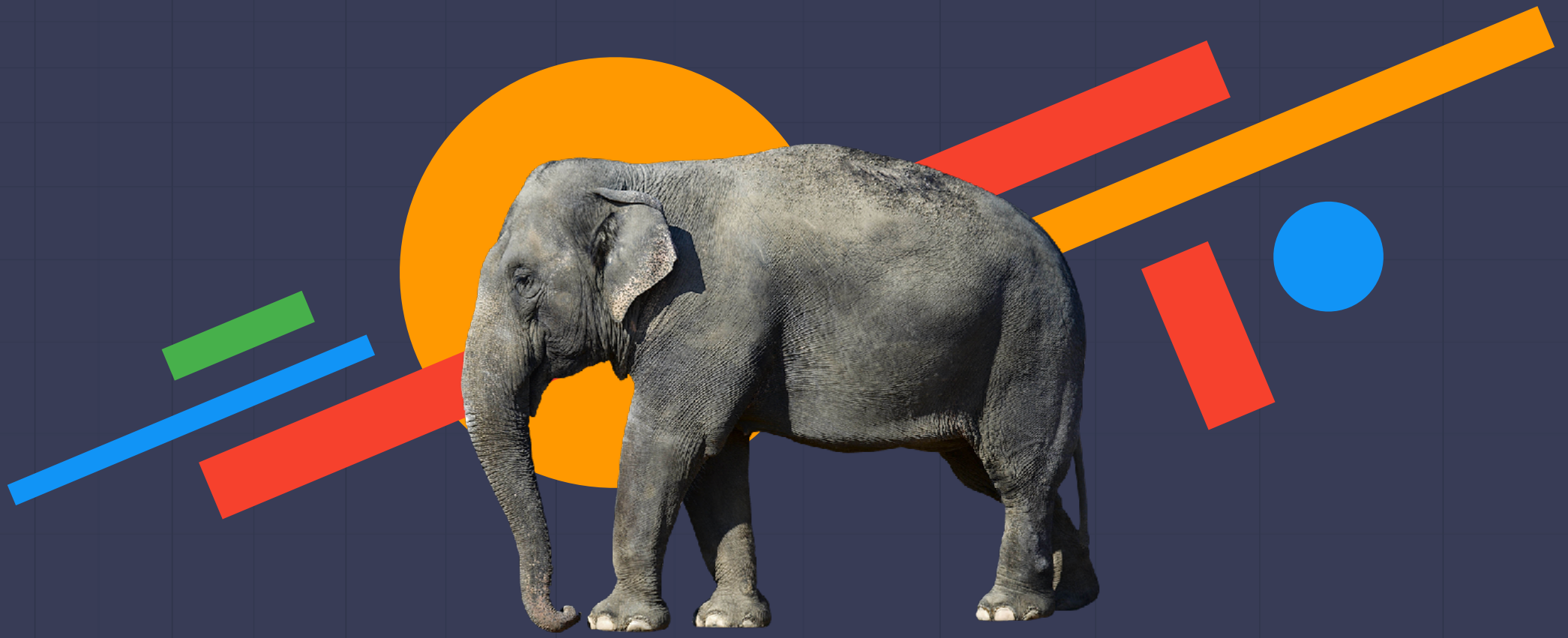
УСПЕШНОСТЬ ПРОХОЖДЕНИЯ

Сертификат

1. 6 из 10 лаб сданы успешно и в срок.
2. Всего 3 теста по 10 вопросов. Нужно ответить правильно суммарно минимум на 15.

С отличием

1. 8 из 10 лаб сданы успешно и в срок.
2. Нужно ответить правильно суммарно минимум на 23.



Big Data is Love

NEWPROLAB.COM