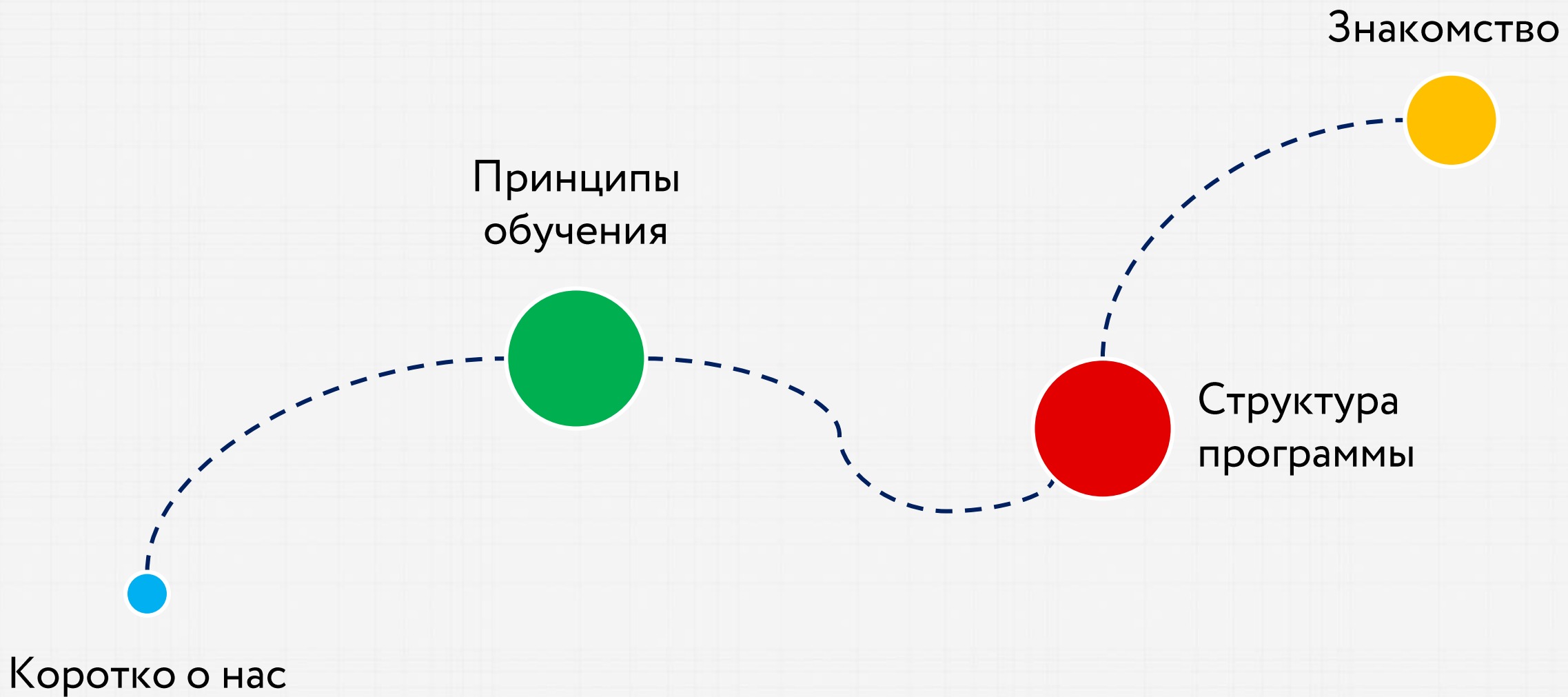




# Специалист по большим данным

МАРАФОН ДЛИНОЙ В 12 НЕДЕЛЬ

# План





# Коротко о нас



# Развитие

Выпускники:

- 280 с открытых программ
- 260 с корпоративных



**СИБУР**

Корп. клиенты

Data-driven product



1 Data Engineer

2 Data Engineer

3 BD for Executives

Y BD for Executives

1 Deep Learning

3 Deep Learning

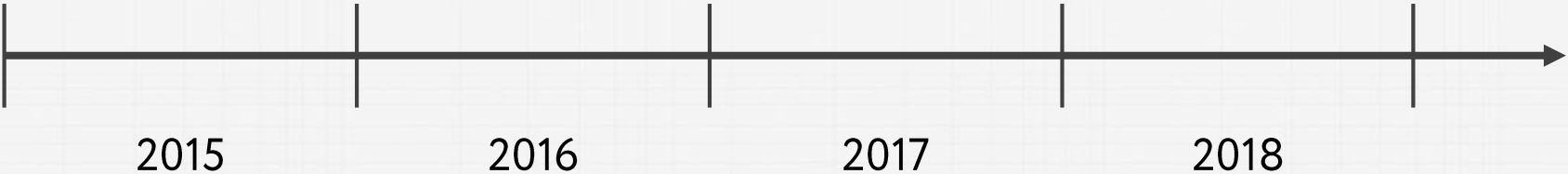
X Deep Learning

3 Big Data

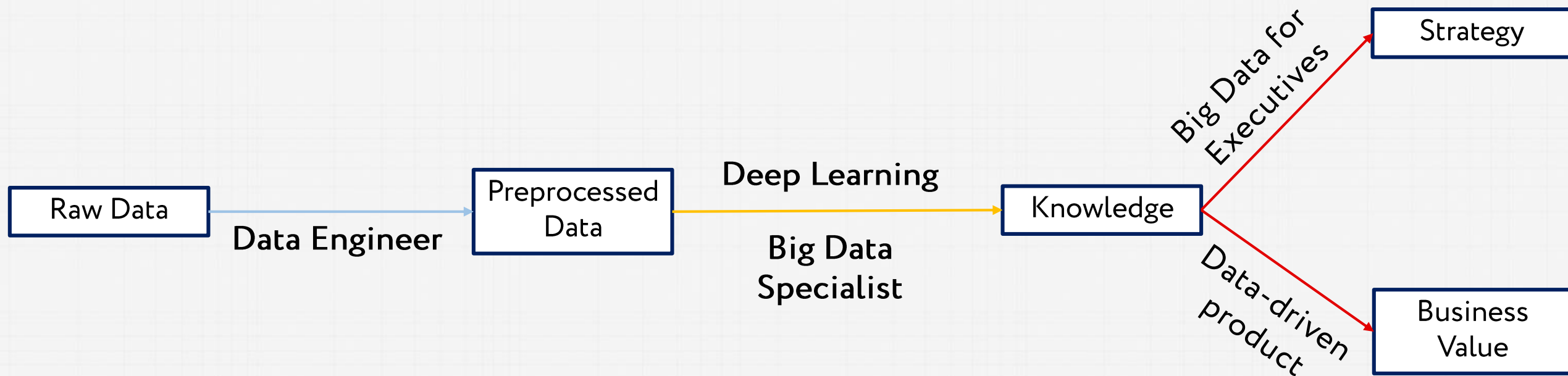
2 Big Data

2 Big Data

2 Big Data



# Программы





# Программы



# Принципы обучения



Взрослые отличаются от детей



- { 1. Материал должен быть ориентирован на конкретные задачи }



# Наша программа

Построение DMP-  
системы

Разработка  
рекомендательной  
системы

{ 2. Слушатели должны иметь возможность  
сразу же применить знания в работе или в  
жизни }



# Наша программа

10+

лабораторных работ

2

проекта



# Модуль 1



1. Развернуть кластер в облаке, установить HortonWorks, запустить MapReduce job.
2. Отфильтровать данные по пользователям и положить их в HBase, построить топ-350 url.
3. Классифицировать пользователей по их логам, определить наиболее релевантные домены для одной из групп с использованием Hive.
4. Прогнозирование оттока клиентов банка.
5. Классификация отзывов в интернете, определение схожести текстов вакансий.

**Core project.** Прогнозирование пола и возрастной категории пользователей по логам.



## Модуль 2



1. Построение неперсонализированных рекомендаций фильмов.
  2. Content-based рекомендательная система онлайн-курсов.
  3. Коллаборативная фильтрация для рекомендации фильмов.
  4. Соревнование по построению наилучшей рекомендательной системы фильмов.
  5. Рекомендательная система видео-контента (ТВ-программы и фильмы).
- Core project. Построение рекомендательной системы для онлайн-магазина.**

{ 3. В основе обучения должен лежать опыт обучающихся }

# Цикл Колба







# Наша программа

Лабы будут часто идти впереди занятий.

Они подготавливают контекст и правильные вопросы, которые можно задать на занятии.

{ 4. Слушатели должны обладать  
самостоятельностью }



# Наша программа

В лабах не всегда будет все разжевываться.

Придется порой самостоятельно двигаться в процессе решения, пользуясь гуглом и помощью коллег.

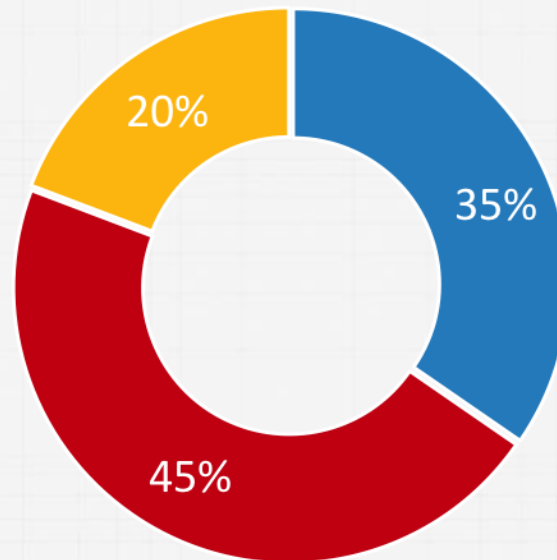
# Структура программы



# Крупными мазками

[illegible]

# Крупными мазками



■ лекция ■ семинар ■ мастер-класс



# Занятия

## Технологический трек

Linux  
MapReduce  
HDFS  
HBase  
Hive  
Spark: descriptive analysis  
Spark: exploratory analysis  
Spark: classification  
Spark: regression  
Spark: clustering & ALS  
Spark: streaming

## Алгоритмический трек

Descriptive analysis (pandas)  
Exploratory analysis (matplotlib)  
Intro to text-mining  
Parsing and cosine similarity  
Sentiment analysis  
Topic modelling  
Intro to machine learning  
ML workshop  
ML master-class (hands-on use case)  
ML in production  
Intro to Recommender System  
Non-personalized and content RS  
Collaborative filtering  
SVD, BMF  
Matrix factorizations  
Intro to Deep Learning

## Бизнес-трек

Requirement analysis in DS  
Making a story from your data  
ML master-classes  
Choosing the right metric  
(precision or recall, price of error)  
RS master-classes  
Assessing your proposal  
A/B-testing and data-driven  
organization



# Технологический



Антон Пилипенко

Big Data Engineer, Mail.ru Group



Николай Марков

Senior Data Engineer, Aligned  
Research Group



Павел Клеменков

Chief Data Scientist (marketing),  
Сбербанк





# Алгоритмический



Петр Ермаков

Head of Data & Analytics, Youla  
at Mail.Ru Group



Дмитрий Игнатов

Заместитель руководителя  
департамента анализа данных и  
искусственного интеллекта,  
ВШЭ



# Бизнес



**Александр Ульянов**

Data Science Executive Director,  
Сбербанк



**Олег Хомюк**

Руководитель R&D, Lamoda



**Александр Филатов**

Product Analytics Manager, VISA

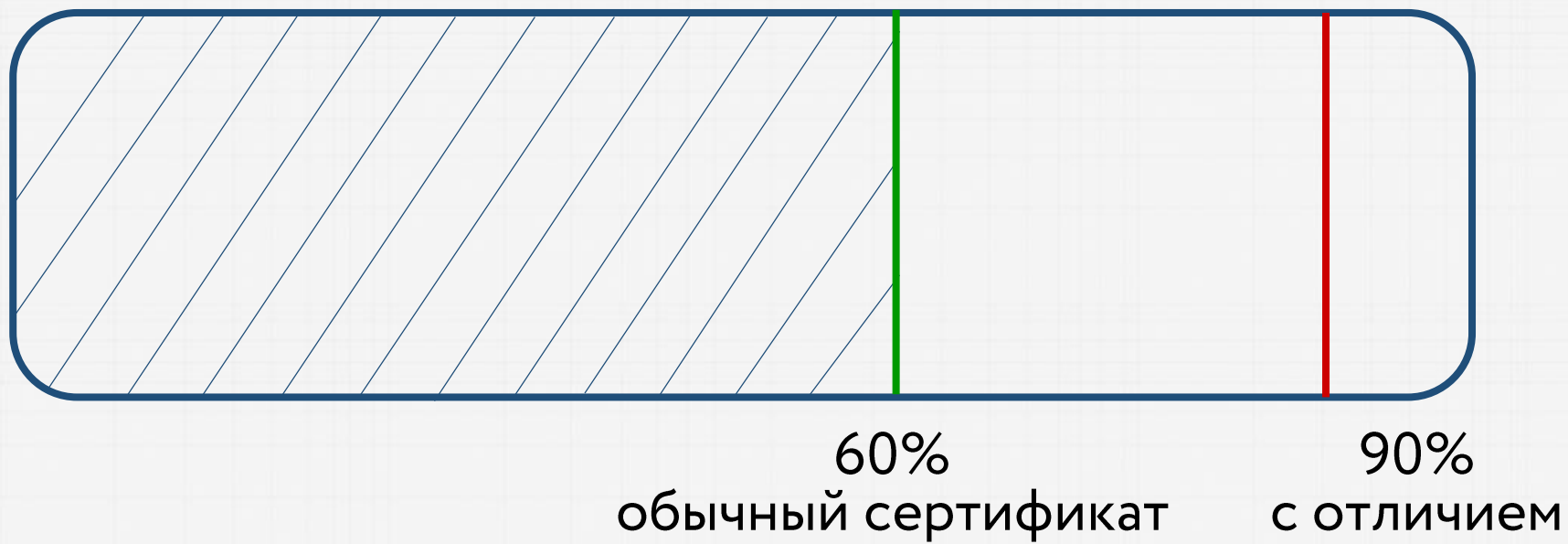


**Александр Петров**

Sr. Software Development  
Engineer, Amazon



# Сертификат



- Проект – 20 баллов
- Лаба – 10 баллов
- Суперачивка – 5 баллов
- Тест – 1 балл
- Упражнения – 0 баллов :)



# Проект

Большая задача из бизнеса. Теперь про production.  
1.5 месяца работы в фоновом режиме.  
Дает возможность командной работы.  
Попадает в ваше портфолио.



# Лаба и суперачивка

Конкретная практическая задача.

Неделя интенсивной работы.

Каждая лаба – приобретенный навык.



# Тест

Сформировать контекст занятия.

Проверить свое усвоение материала.

Вскрыть неявные вещи.

Получить дополнительные знания.

# Упражнения

Получить элементарные навыки по теме.  
Содержат в себе «подсказки» к лабам.





# Получение знаний

1. Занятия
2. Кофе-брейки
3. Slack
4. Google :)



# IT-ресурсы

1. Личный кабинет – проверка лаб, трансляция
2. Кластер – дом родной (доступ позже)
3. Slack – общение
4. GitHub – все материалы
5. Google-календарь – планирование сил




## Ближайшие шаги

23.03 – выкладываем Лабу 1

24.03 – занятие по Linux

27.03 – Дескриптивный анализ в Python

29.03 – Семинар по MapReduce

The background of the slide features a silhouette of a group of graduates standing on a hill, holding up their caps and diplomas in celebration against a sunset sky. A large, semi-transparent red rectangle is overlaid on the center of the image, serving as a background for the title text.

# Напутствия и лайфхаки от выпускников

# Знакомство



# Команда



Артём Трунов

Координирует программу



Анастасия Кошель

Отвечает за трансляцию



Марина Михайлова

Отвечает за гостеприимство



Денис Шрестха

Разрабатывает и тестирует  
лабы



# Команда



Юля Патронникова

Коммуникации до и после  
программы



Александра Нестерова

Отвечает за документальное  
сопровождение



Андрей Булатов

Чтобы не было ни единого  
разрыва (СТО)



Елена Третьякова

Отвечает за всё (CEO)

# Теперь вы

1. Имя, фамилия, компания
2. Профессиональное достижение
3. Любопытный факт о себе



1 минута



# Разомнем мозги?



Ответ

```
$ ssh -i bigdata8.pem user@newprolab.com
```

```
$ aap -q jqolibi8.xmu camz@vmexzwtij.kwu
```

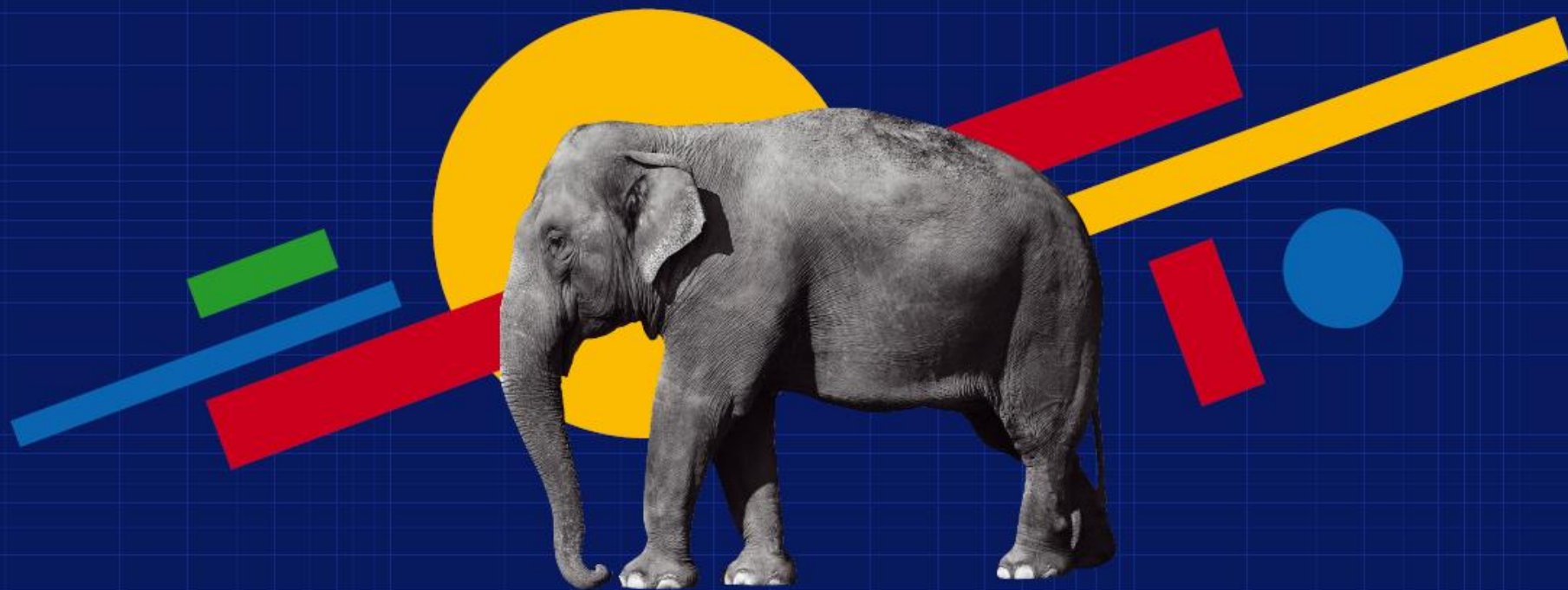
## Шифр Цезаря со смещением 8.

```
NUM_LETTERS = 26
```

```
# ord('a') = 97  
A_ASCII_SHIFT = 97
```

```
# Welcome, Newprolab BigData 8!  
SHIFT = 8
```

```
def encode(letter, shift):  
    if not letter.isalpha():  
        return letter  
  
    letter_code = ord(letter) - A_ASCII_SHIFT  
    shifted_letter_code = (letter_code + shift) % NUM_LETTERS  
  
    return chr(shifted_letter_code + A_ASCII_SHIFT)
```



# BIG DATA IS LOVE

[NEWPROLAB.COM](http://NEWPROLAB.COM)