

# Data Science UA. Лекция 1.

## Организационная

Январь 2017

# Класс

- Начальный уровень
- Студенты/люди с базовыми навыками программирования
- <20 человек

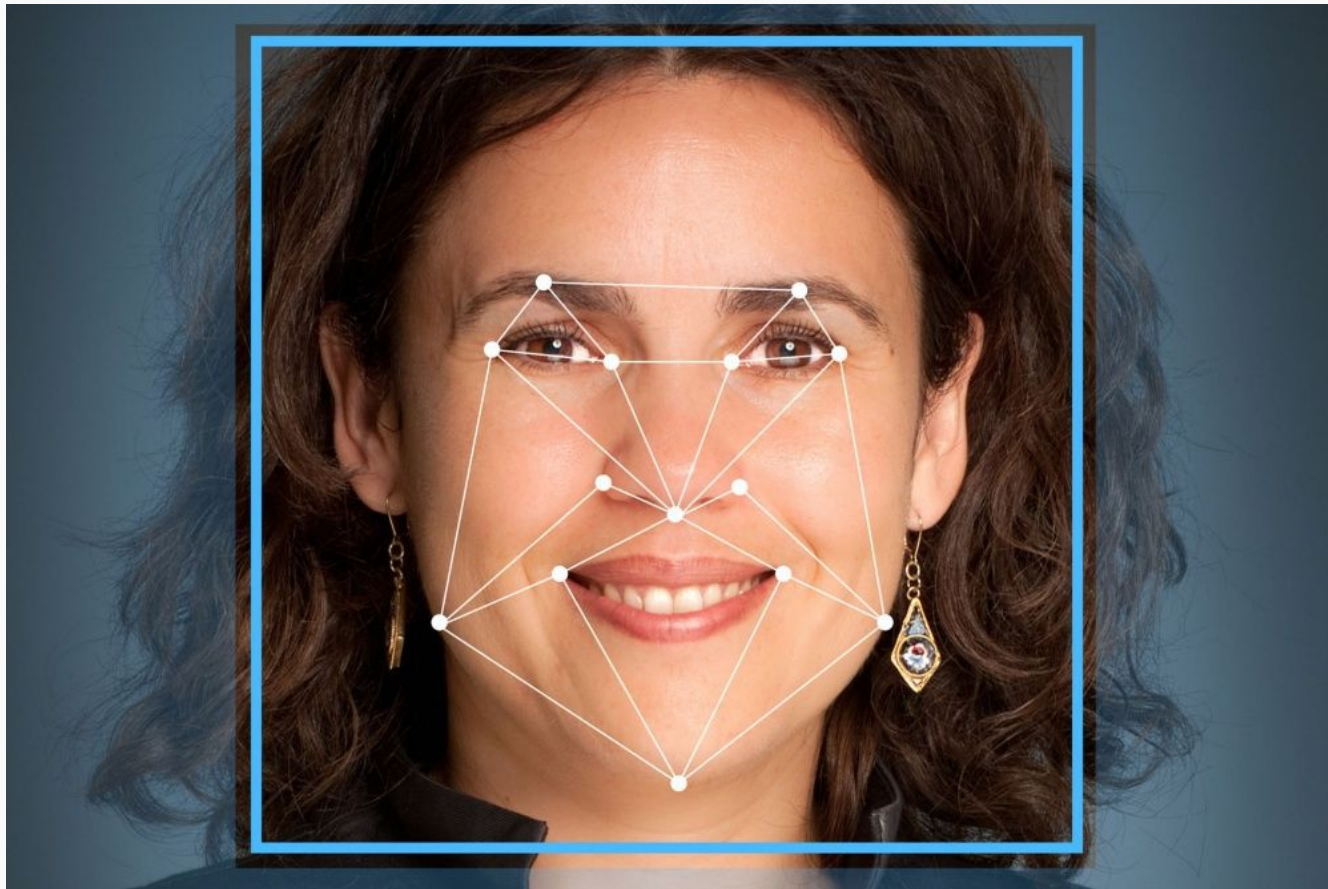
# Цель курсов

- Теоретические основы Data Science и предиктивного моделирования.
- Практические навыки по сбору, очистке и подготовке данных к использованию.
- Практические навыки моделирования и применения DS моделей в разных областях.
- Финальный проект

# Кто я?

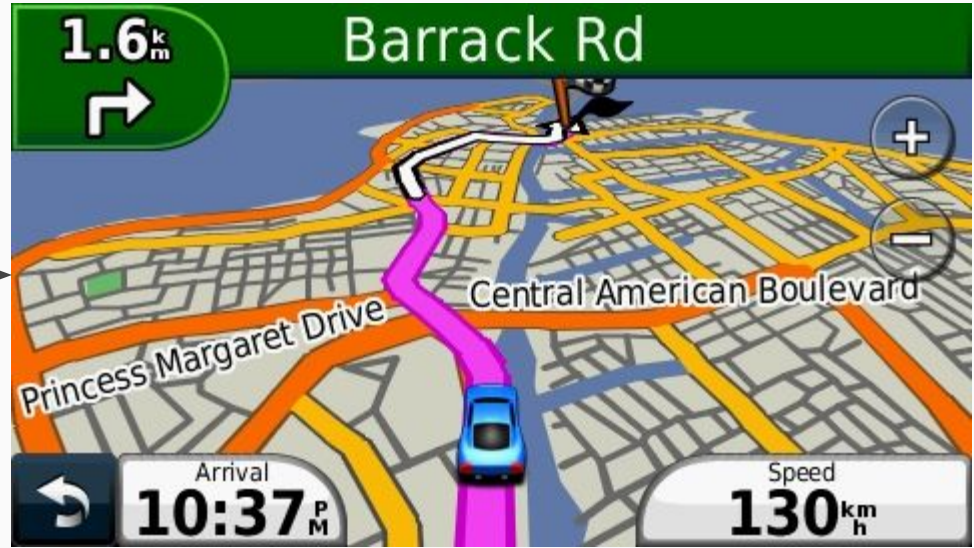
- Одинцов Михаил
- 8 лет разработки
- 3 из них - Python/DS
- Проекты связанные с распознаением лиц, топографических карт, ПО для казино и онлайн игр
- DataRobot senior software engineer

# Распознавание лиц на основе эластичных сетей и контурной информации



$$\mathbf{B} = \frac{1}{159} \begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} * \mathbf{A}.$$

# Автогенерация карт и GPS данных из топологических карт



# Программное обеспечение для казино



Король червей  
Дама червей

Вероятность стрита: 1%  
Вероятность флеша: 3%  
Вероятность 1 пары: 20%

...

# КТО ВЫ?

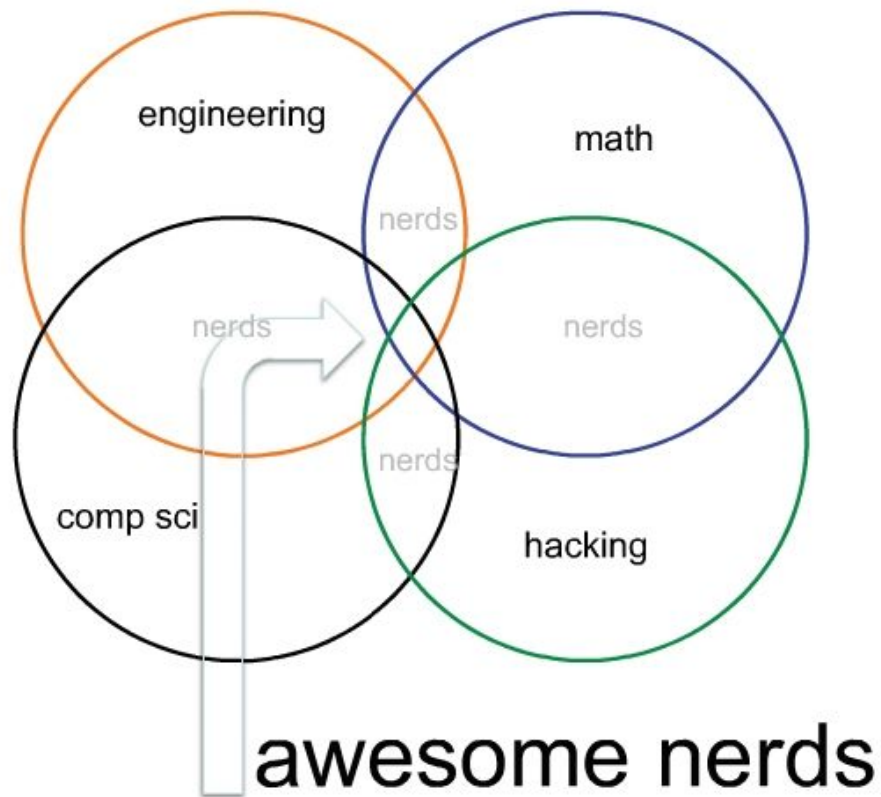
- ФИО
- Несколько слов о себе
- Навыки в программировании
- Ожидания от курса



# КТО ОНИ?

Data Scientists

# Data science?



# Питер Норвиг

- Research Director at Google
- 6 лет в NASA на должности Head of Computational Science Division
- Курс “Intro to Artificial Intelligence”



# Эндрю Ын

- Сооснователь Coursera
- Chief Scientist в Baidu
- Stanford University Associate Professor
- Основатель и ведущий разработчик Google Deep Learning Project



# Себастьян Трун

- Основатель Udacity
- Stanford University Research Professor
- Max-Planck-Research Award



# Джереми Ачин

- Сооснователь и CEO DataRobot
- Победитель множества конкурсов на Kaggle
- Сильнейшая команда DS специалистов в мире
- Мой начальник :)



# Нейт Сильвер

- Популяризатор Data Science
- Автор тематического блога “FiveThirtyEight”
- Очень точные предсказания результатов игр и выборов
- Автор “Сигнал и шум”



# Чем занимаются?

Какие применения Data Science  
находит в современном мире?

Какие проекты на стадии  
рассмотрения?

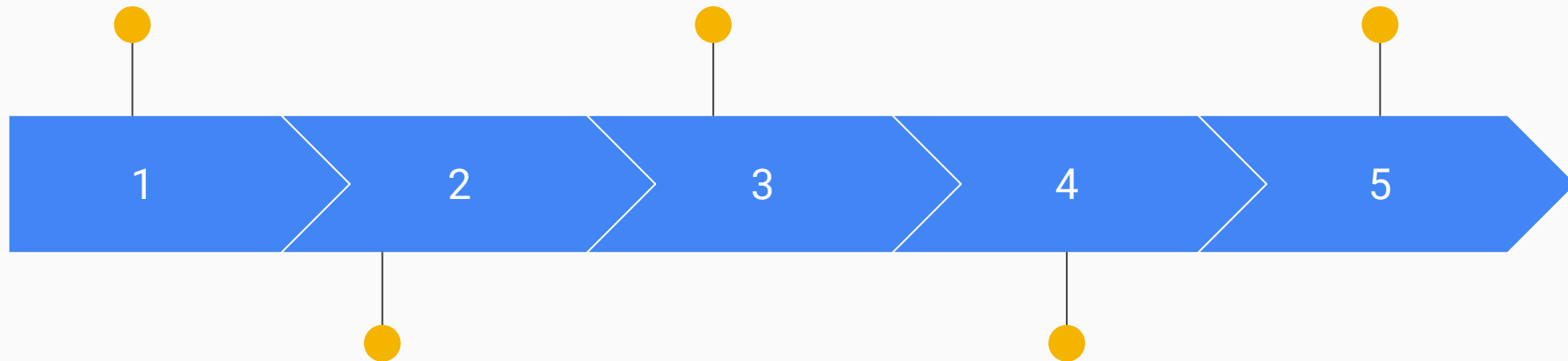
Какие возможные направления в  
будущем?



Формулировка  
задачи

Анализ данных

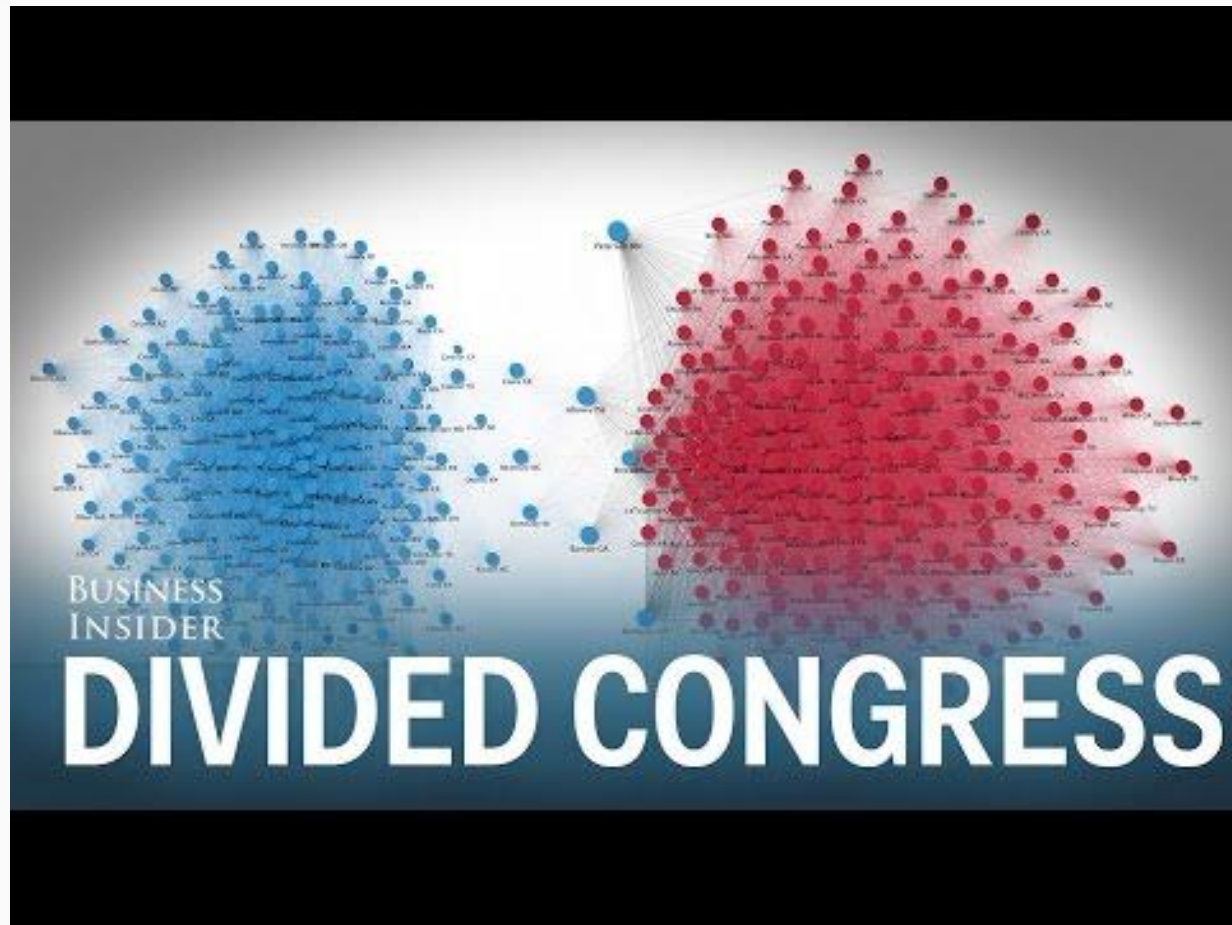
Презентация,  
использование и  
контроль модели



Сбор и очистка  
данных


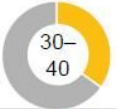
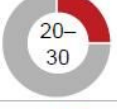
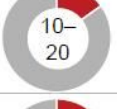

Построение модели  
данных

Хороший проект рассказывает историю



# McKinsey's 2016 Analytics. Future Of Machine Learning

There has been uneven progress in capturing value from data and analytics

	Potential impact: 2011 research	Value captured %	Major barriers
<b>Location-based data</b>	<ul style="list-style-type: none"> <li>\$100 billion+ revenues for service providers</li> <li>Up to \$700 billion value to end users</li> </ul>		<ul style="list-style-type: none"> <li>Penetration of GPS-enabled smartphones globally</li> </ul>
<b>US retail<sup>1</sup></b>	<ul style="list-style-type: none"> <li>60%+ increase in net margin</li> <li>0.5–1.0% annual productivity growth</li> </ul>		<ul style="list-style-type: none"> <li>Lack of analytical talent</li> <li>Siloed data within companies</li> </ul>
<b>Manufacturing<sup>2</sup></b>	<ul style="list-style-type: none"> <li>Up to 50% lower product development cost</li> <li>Up to 25% lower operating cost</li> <li>Up to 30% gross margin increase</li> </ul>		<ul style="list-style-type: none"> <li>Siloed data in legacy IT systems</li> <li>Leadership skeptical of impact</li> </ul>
<b>EU public sector<sup>3</sup></b>	<ul style="list-style-type: none"> <li>~€250 billion value per year</li> <li>~0.5% annual productivity growth</li> </ul>		<ul style="list-style-type: none"> <li>Lack of analytical talent</li> <li>Siloed data within different agencies</li> </ul>
<b>US health care</b>	<ul style="list-style-type: none"> <li>\$300 billion value per year</li> <li>~0.7% annual productivity growth</li> </ul>		<ul style="list-style-type: none"> <li>Need to demonstrate clinical utility to gain acceptance</li> <li>Interoperability and data sharing</li> </ul>

<sup>1</sup> Similar observations hold true for the EU retail sector.

<sup>2</sup> Manufacturing levers divided by functional application.

<sup>3</sup> Similar observations hold true for other high-income country governments.

Через 5 лет машинное обучение  
будет частью ежедневной работы  
врачей

(c) Vic Gundotra,

Microsoft and Google executive, CEO of AliveCor

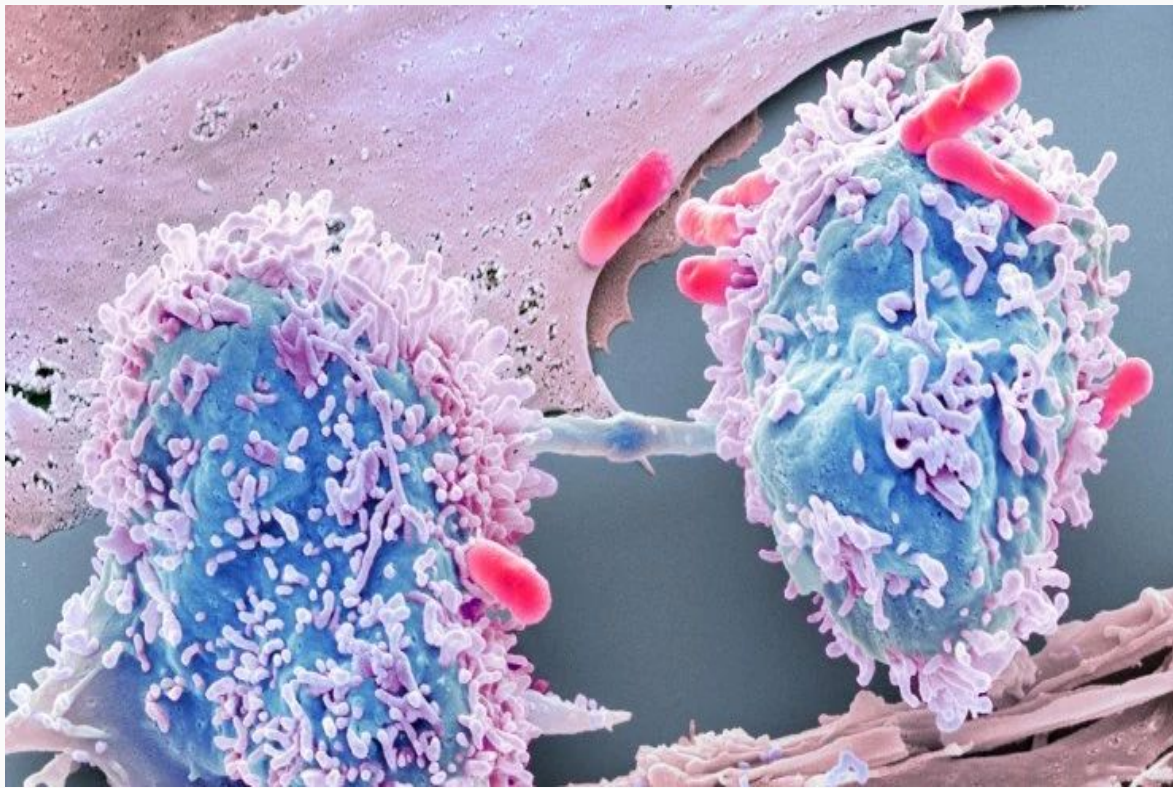
## О чем говорят летучие мыши



- Еда
- Сон
- Места для отдыха
- Нежелательные приставания



# Побеждаем рак (и другие заболевания) силой Data Science



Программы:

- U.S. Department of Veteran Affairs' Million Veteran Program
- the 100,000 Genomes Project in the U.K.
- the NIH's The Cancer Genome Atlas

Стоимость секвенирования генома упала с 10 миллионов до тысячи долларов.

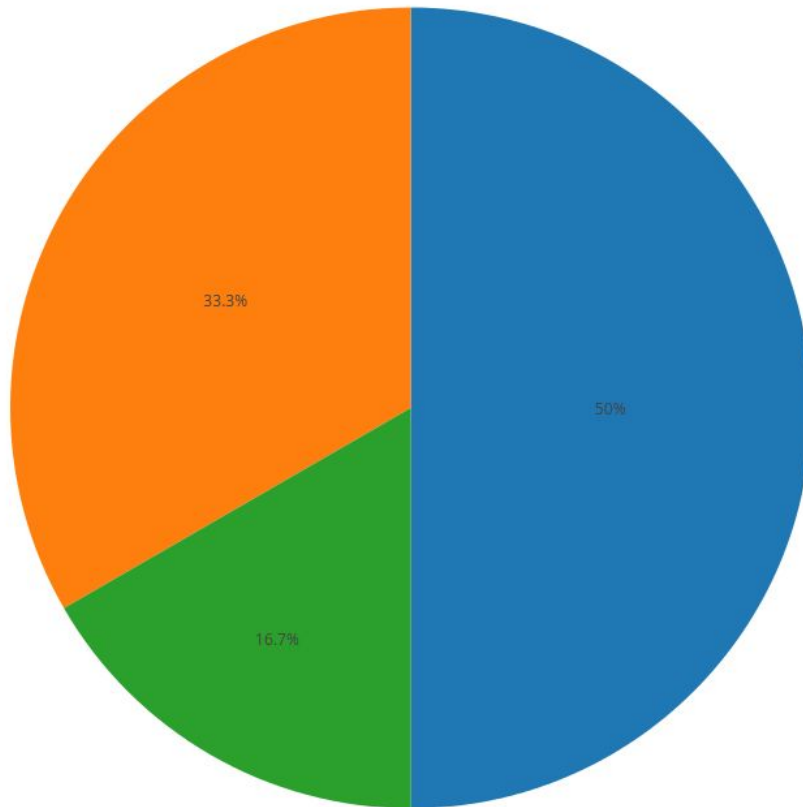
Ожидается ~2 миллиардов геномов к 2025 году

Чем займемся  
мы?

1. Вводная лекция (вы на ней :) )
2. Блок работы с данными (2 - 7 занятия)
3. Алгоритмы машинного обучения (8 - 13 занятия)
4. Внедрение и контроль результатов (14 занятие)
5. Финальный проект (15 - 16 занятия)



## Структура уроков



- Лекция
- Практика
- Домашнее задание

1. Индивидуальные или небольшие команды (3 человека максимум)
2. Две презентации (10-15 минут)
3. Каждый этап - неделя
4. Первый этап - формулировка задачи и сбор данных.  
Презентация рассказывает о цели проекта, источниках, результатах сбора и анализа данных
5. Второй этап - финальный продукт. Постройка модели, презентация результатов

Главный канал - <https://datascienceua.slack.com/>

Линк приглашение будет разослан всем участникам после занятия.

Социальные сети - группы курсов, meetup, facebook.

Домашние задания сдавать в личном сообщении мне в слаке.

Просьба не писать в личные сообщения без исключительной надобности.

Все материалы будут выкладываться на гитхаб

[https://github.com/Templarr/datascienceua\\_2017](https://github.com/Templarr/datascienceua_2017)

# Домашнее задание

Настройка окружения в основной системе.

Плюсы :

- Опыт
- Эффективное использование ресурсов

Минусы:

- Сложнее

Настройка окружения на виртуальной машине.

Плюсы :

- Не нужно возиться с установкой и настройкой, всё готов

Минусы:

- Медленнее
- Отсутствие опыта установки

Демонстрация

Вопросы?