

Основы работы с данными: сбор, анализ, визуализация

Лекция 1. Введение в data-журналистику.
Логистика курса, актуальность и мотивация.

Максим Карпов
mekarpov@hse.ru



HSE
University

Команда курса

Лекции

Карпов Максим Евгеньевич

Аспирант, Младший научный сотрудник: [Факультет компьютерных наук / Научно-учебная лаборатория методов анализа больших данных](#)

Преподаватель: [Факультет компьютерных наук / Департамент больших данных и информационного поиска](#)

Начал работать в НИУ ВШЭ в 2017 году.

Научно-педагогический стаж: 2 года.



Семинары 191, 192

Максимовская Анастасия Максимовна

Приглашенный преподаватель: [Факультет компьютерных наук / Департамент больших данных и информационного поиска](#)

Начала работать в НИУ ВШЭ в 2021 году.



Команда курса

Семинары 193

Быков Кирилл Валерьевич

Приглашенный преподаватель: [Факультет коммуникаций, медиа и дизайна /](#)
[Департамент медиа](#)

Начал работать в НИУ ВШЭ в 2021 году.



Семинары 194, 195

Перевышина Татьяна Олеговна

Инженер: [Факультет городского и регионального развития /](#) [Институт экономики транспорта и транспортной политики /](#)
[Центр исследований транспортных проблем мегаполисов](#)

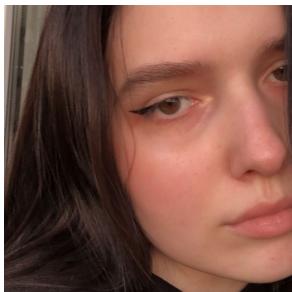
Начала работать в НИУ ВШЭ в 2020 году.



Команда курса

Семинары

Группа	Преподаватель	Учебный ассистент
191	Максимовская Анастасия Максимовна	Стрельцов Тёма
192	Максимовская Анастасия Максимовна	Никулина Женя
193	Быков Кирилл Валерьевич	Егорова Настя
194	Перевышина Татьяна Олеговна	Кордзахия Натела
195	Перевышина Татьяна Олеговна	Васильев Коля



Общевышкинский проект Data Culture

- Наш курс реализуется в рамках общеуниверситетского проекта **Data Culture**
- Этот проект стартовал в 2017/2018 уч. г.
- Охватывает **все бакалаврские программы** Вышки, часть магистерских и аспирантских программ
- Обучение **культуре работы с данными**, основам программирования, ключевым цифровым компетенциям современности



Национальный исследовательский университет «Высшая школа экономики» → Data Culture

RU EN 🔍 🌐

Data Culture

Что такое Data Culture?

Data Culture — это общий термин для обозначения навыков и культуры работы с данными. Современный специалист в любой области постоянно сталкивается с задачами по обработке данных: юрист изучает сотни дел, историк — тысячи документов на различных языках, экономист — разрабатывает прогнозную модель на основе большого объема данных.

Новости 3

2 июля 2020

От основ информационной безопасности до методов машинного обучения

Какие цифровые компетенции будут развивать студенты Вышки

29 июня 2020

Вышкинский курс по цифровой грамотности открывается для внешних слушателей

Цели курса

- Закрепление базовых навыков программирования на языке Python
- Формирование навыков работы со специализированными библиотеками для сбора, обработки, визуализации и анализа данных
- Развитие навыков чтения, интерпретации и оценки качества анализа и визуализации количественных данных.

Про что не будем говорить на курсе: про иллюстративную инфографику

АНАТОМИЯ ПЕТЕРБУРГА как устроен город

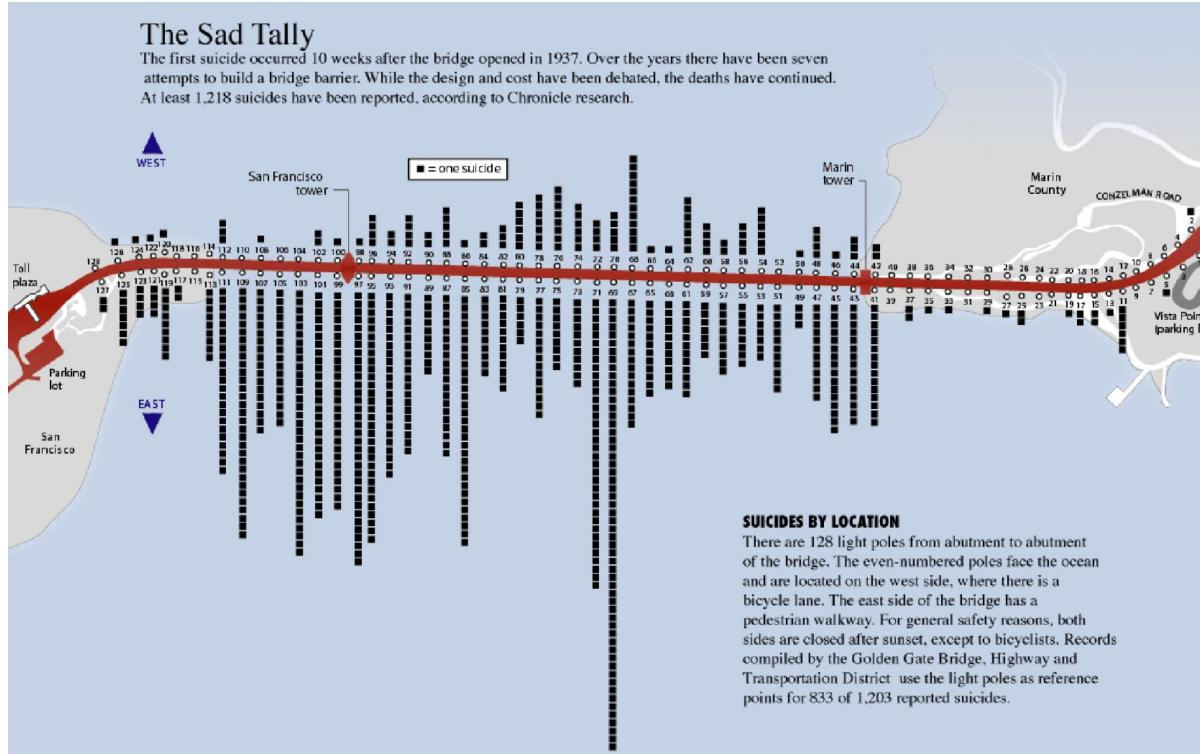
Исх. №: 275
от "17" 2014 г.

Небесная линия

Еще в начале строительства Петербурга дома ставили «по линиям и в один горизонт», то есть по периметру кварталов и одинаковой высоты. Здесь же на больших расстояниях друг от друга вырастали намногочисленные сооружения, сильно возвышающиеся над средним уровнем застройки.



Про что не будем говорить на курсе: Про то как графически вписывать статистические графики в контекст



Но при этом научимся
делать статистически
корректные графики, с
которыми потом смогут
работать дизайнеры.

Как этот курс будет полезен?

- Можно прокачать навыки, которые впоследствии пригодятся в научной, экономической или расследовательской журналистике (такой, где обычно работают эксперты области)
- Можно научиться технически визуализировать статистические данные. Мы не будем делать шедевры инфографики, но шедевры инфографики часто могут быть бессмысленны, если данные принесены в жертву красоте.

Место дисциплины в учебном плане образовательной программы “Журналистика”

Базовые инструменты журналистики



Python для извлечения и обработки данных



Основы работы с данными: сбор, анализ, визуализация



Проектный семинар "Журналистика данных"

Проектный семинар "Визуализация данных"

Когда читается:

1-й курс, 1-4 модуль

Когда читается:

2-й курс, 3 модуль

Когда читается:

3-й курс, 1, 2 модуль

Когда читается:

3-й курс, 3 модуль

Когда читается:

3-й курс, 4 модуль



Выпускные квалификационные работы студентов НИУ ВШЭ

Всего по вашему запросу работ **108**

Эмоциональный аспект в визуализации данных в онлайн-медиа в период пандемии COVID-19

Факультет коммуникаций, медиа и дизайна

Программа: [Журналистика данных](#) (Магистратура)

ФИО студента: Голубева Анастасия Леонидовна

Руководитель: Бережная Валентина Сергеевна

Год защиты: 2021

Использование журналистики данных в правительенных СМИ Иордании

Факультет коммуникаций, медиа и дизайна

Программа: [Журналистика данных](#) (Магистратура)

ФИО студента: Алхарасис Сулейман Хасан Сулейман -

Руководитель: [Давыдов Сергей Геннадьевич](#)

Год защиты: 2021

Анализ тематического разнообразия статей как элемента поддержания интереса к спорту во время пандемии при помощи инструментов журналистики данных

Факультет коммуникаций, медиа и дизайна

Программа: [Журналистика данных](#) (Магистратура)

ФИО студента: Газыева Элина Рустамовна

Руководитель: [Дмитриев Олег Аркадьевич](#)

Год защиты: 2021

Кампус/факультет

Факультет коммуникаций, медиа и дизайна

ФИО студента

Голубева Анастасия Леонидовна

Руководитель

Бережная Валентина Сергеевна

Год защиты

2021

Название работы

Использование журналистики данных в правительенных СМИ Иордании

Уровень образования

бакалавриат

магистратура

Не важно

Оценка

Не имеет значения

Программа

Журналистика данных

Доступен полный текст

Не имеет значения

Язык

Все

Русский

Английский

Поиск

сбросить фильтры

Опрос по Питону

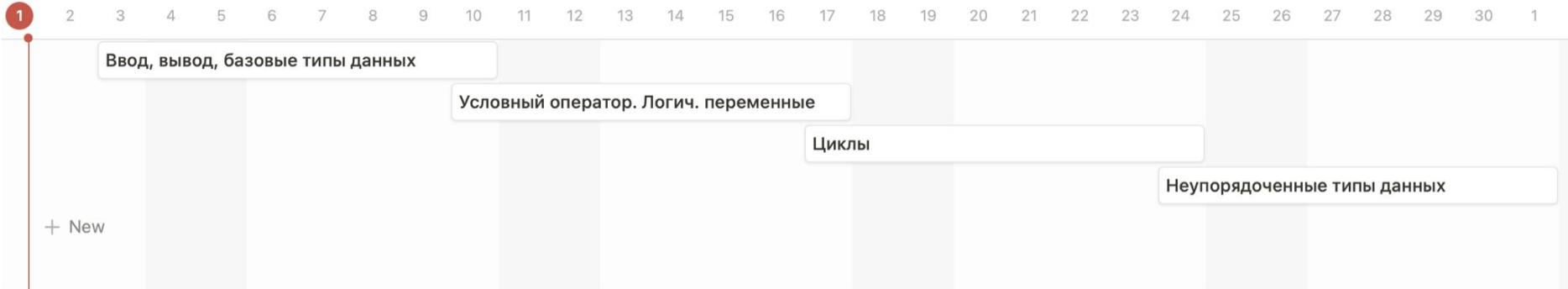
- От 1 до 5 насколько хорошо был усвоен материал онлайн курса по Питону в 3 модуле прошлого уч.г.?
- От 1 до 5 насколько сложным был материал онлайн курса по Питону в 3 модуле прошлого уч.г.?
- Установлена ли программа Jupyter Notebook для работы с Питоном на компьютере?

Темы курса

1) Основы программирования на Python

» September 2021

Oct Mc



Темы курса

- 2) Сбор новостей в автоматическом режиме в один текстовый файл (октябрь)
- 3) Предобработка текстов, сбор количественных данных в табличном формате, разведывательный анализ (ноябрь)
- 4) Визуализация данных и интерпретация полученных результатов (ноябрь-декабрь)

Формула оценки

0.1 * Онлайн курс (с 11 сентября)

- + 0.1 * Работа на семинарах (решение небольшой задачи, написание не-ко строчек кода в большом алгоритме, знание методов/функций)
- + 0.1 * Тесты на лекциях (10 вопросов по материалу предыдущей лекции)
- + 0.5 * ДЗ (всего 5 штук)
- + 0.15 * Проект (строго в командах по два человека, три этапа исследования с дедлайнами)
 - 0.05 Парсинг данных (новости, статьи, посты),
 - 0.05 предобработка данных и разведывательный анализ,
 - 0.05 визуализация и интерпретация результатов исследовательской задачи
- + 0.15 * Экзамен (дистанционно на онлайн платформе с прокторингом)

В сумме теоретически
можно набрать 11 баллов
=). В ведомость пойдет
10.

Сдача домашних заданий

- У каждого дз и этапа проекта будет свой жесткий **дедлайн**.
После жесткого дедлайна работы не принимаются.
- В течение семестра каждый студент может не более **2 раз** сдать задание после жесткого дедлайна – в этом случае за каждый день просрочки вычитается по два балла.
- В таком случае, студент должен предупредить своего ассистента о выбранной опции.

Академическая этика



В НИУ ВШЭ **не допускаются**: списывание, двойная сдача письменных работ, плагиат, подлоги, покупка выполнения работы на стороне (“наймиты”).

Проверка через закрытую систему антиплагиат **Яндекс** нетривиальных алгоритмических задач. Очевидные случаи – **зануление**, подозрительные – устный опрос по заданию в согласованное с ассистентом время в зуме.

Важно! В случае подлога: оба студента получают зануление работы (кто предоставил работу и кто принял её)

Пара похожих решений MVP для приложения "Пенькофф Инвестиции" (20 баллов)

Пример подлога

Факторы remaining-sequence
сравнения:

id посылки: 49119552

49199725

статус new
пары:

вердикт: ok

runtime-error

автор: dap_econ_2021_140

dap_econ_2021_160

отправлено 2021.03.03 22:46:21
в:

2021.03.06 17:36:50

исходный
код:

COMP = {}

**ИДЕНТИЧНО С ТОЧНОСТЬЮ ДО
ЗАМЕНЫ ПЕРЕМЕННЫХ**

```
class Client_one:
    def __init__(self, name):
        self.name = name
        self.invest = {}

    def sell(self, value, company):
        if self.invest.get(company) is not None:
            if self.invest[company] > value:
                self.invest[company] -= value
            else:
                self.invest[company] = 0
        if COMP.get(company) is None:
            COMP[company] = 1

    def get_balance(self):
        balance = 0
        for active in self.invest.items():
            balance += COMP.get(active[0]) * active[1]
        print(round(balance))

    def buy(self, value, company):
        if self.invest.get(company) is None:
            self.invest[company] = value
        else:
            self.invest[company] += value
        if COMP.get(company) is None:
            COMP[company] = 1
```

```
COMP = {}

class First:
    def __init__(he, name):
        he.name = name
        he.inv = {}

    def s(he, v, company):
        if he.inv.get(company) is not None:
            if he.inv[company] > v:
                he.inv[company] -= v
            else:
                he.inv[company] = 0
        if COMP.get(company) is None:
            COMP[company] = 1

    def gb(he):
        balance = 0
        for a in he.inv.items():
            balance += COMP.get(a[0]) * a[1]
        print(round(balance))

    def buy(he, v, company):
        if he.inv.get(company) is None:
            he.inv[company] = v
        else:
            he.inv[company] += v
        if COMP.get(company) is None:
            COMP[company] = 1
```

Почему Python?



Язык программирования Python



- Высокоуровневый язык программирования, созданный в 1980-х годах.
- Наиболее актуальные версии - 3.xx (более старые версии могут отличаться в синтаксисе).
- Популярность стремительно растет.
- Гид по стилю: <https://www.python.org/dev/peps/pep-0008/>

Почему Python?



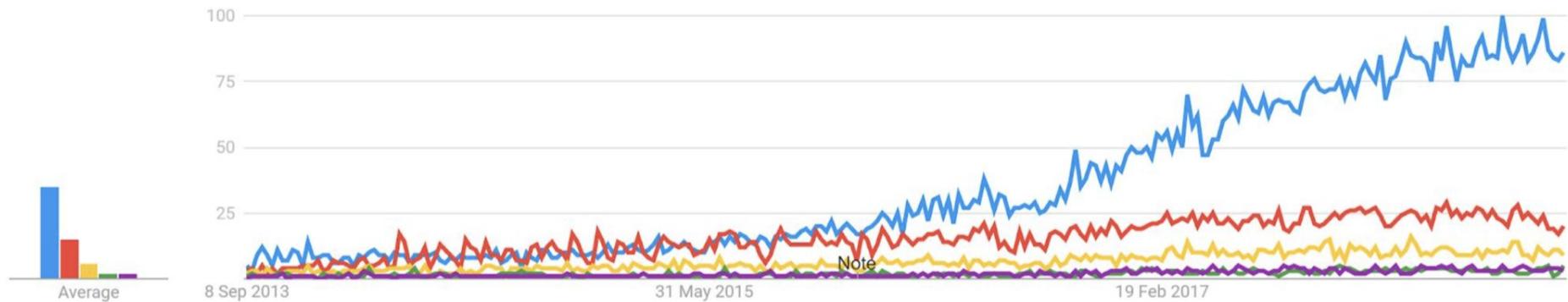


Worldwide ▾

Past 5 years ▾

All categories ▾

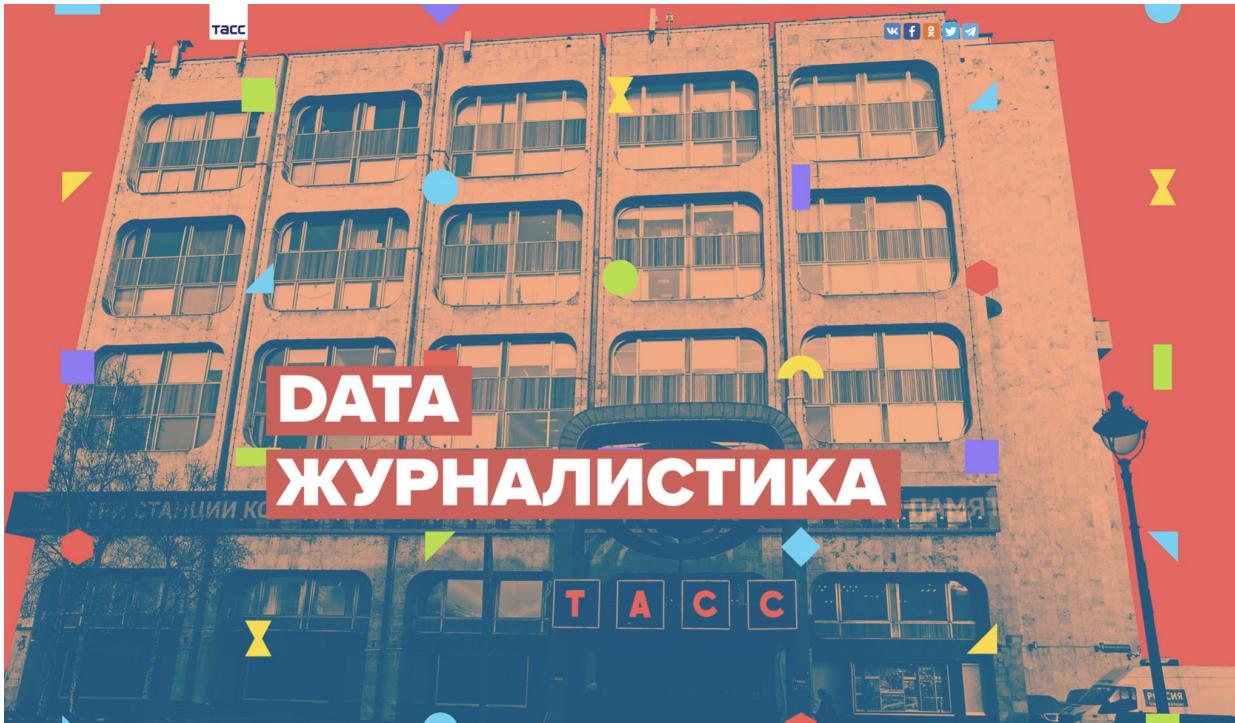
Web Search ▾

Interest over time ?

Как Python может пригодиться вам?

- Сбор данных (scraping и API)
- Работа с табличным представлением данных (pandas), возможность выполнения SQL запросов
- Создание визуализаций
- Анализ текста
- Статистический анализ
- Использование нейронных сетей и машинного обучения для анализа больших данных или автоматизации анализа (замена кодеров)
- Оформление готовых отчетов из Jupyter notebook (с публикацией кода или нет)

Актуальность



<https://data-journalism.profi.tass.ru>

Навык работы с данными — это мощное знание, которое нужно и в публицистике, и в аналитике журналистской активности.

Новая старая журналистика

- СМИ всегда стремились подкреплять статьи данными и фактами.
- Журналистика данных — это принципиально новый механизм организации общения с аудиторией



Сложные выводы из простых данных

- Для объяснения действительности data-журналисты используют базы данных
- Для широкой аудитории они оказываются сложны и не понятны.
- Аудитории нужен журналист, чтобы увидеть хоть что-то значимое посреди нарастающего шума информации



Вернуть доверие

- Согласно исследованию «Edelman Trust Barometer 2020», 57% респондентов считают, что медиа транслируют **непроверенную и недостоверную информацию**
- Журналистика данных, основанная на **фактах**, доступных любому пользователю, способна вернуть доверие к СМИ

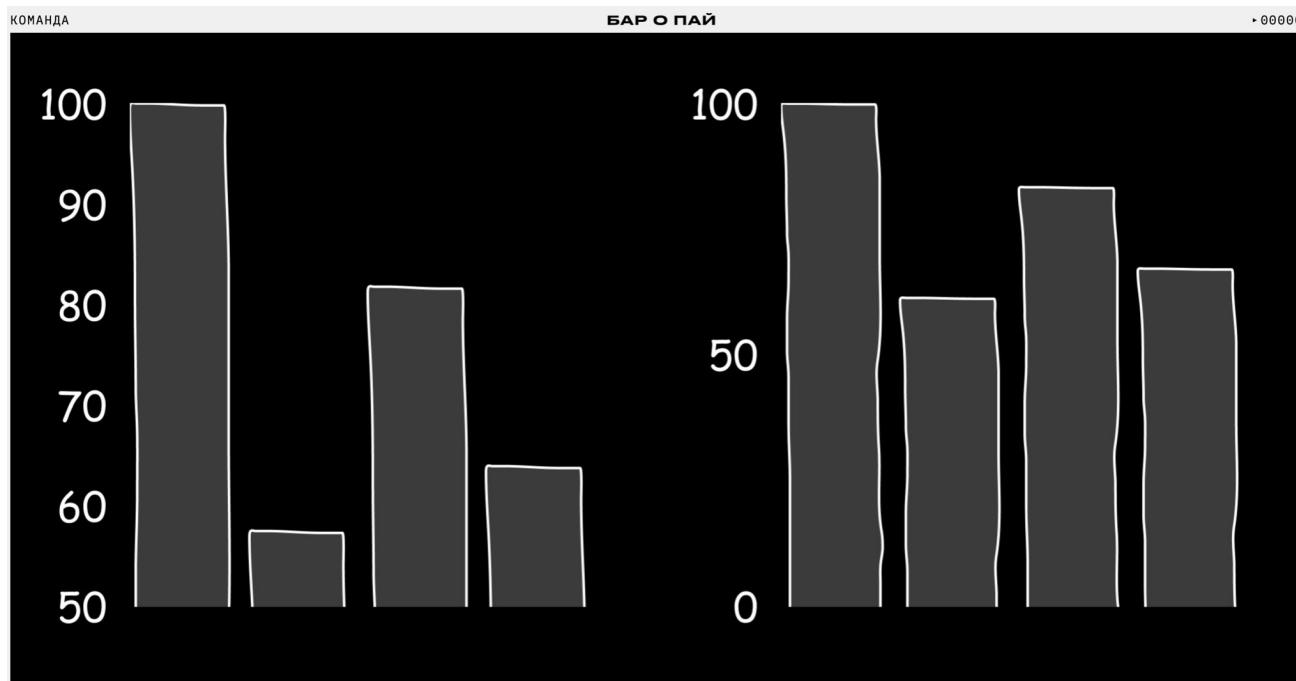


Данные как история

- Сухие данные наглядны, они дают возможность с помощью поиска, фильтрации, анализа и визуализации превращать абстрактное в конкретное, но лишены эмоциональности
- Задача журналистов данных преподнести информацию так, чтобы *создать историю*, которая звучит увлекательно и убедительно, очеловечить ее



Базовая грамотность по визуализации



<https://bar-or-pie.dianov.org>



Грамотность работы с данными

Понимайте, анализируйте и применяйте

Руководителям

Аналитикам

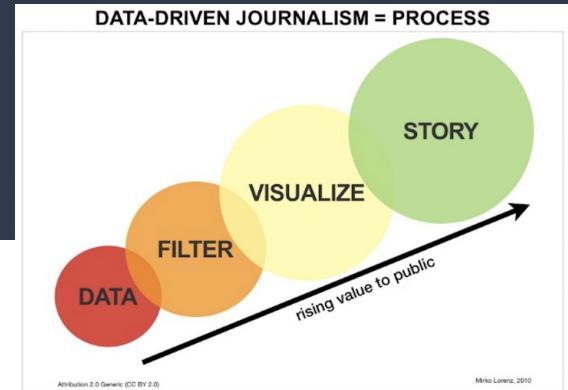
Начинающим

<https://dataliteracy.ru>

Логистика курса

Главный информационный ресурс – [wiki-страница](#) курса. На ней находятся все полезные ссылки, а именно:

- [Канал](#) в *Telegram* для объявлений
- [Чат](#) в *Telegram* для обсуждений



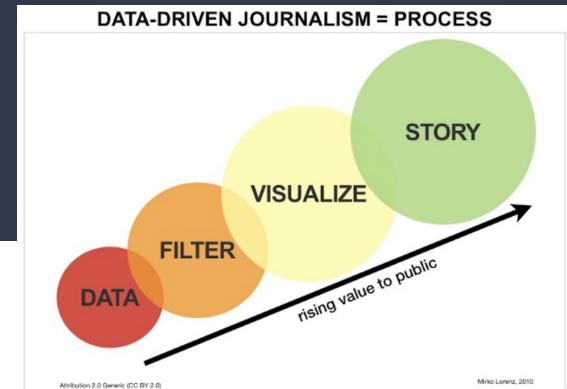
Логистика курса

Главный информационный ресурс – [wiki-страница](#) курса. На ней находятся все полезные ссылки, а именно:

- Инструкция по установке дистрибутива **Anaconda** на [Windows](#) и [MacOS](#).



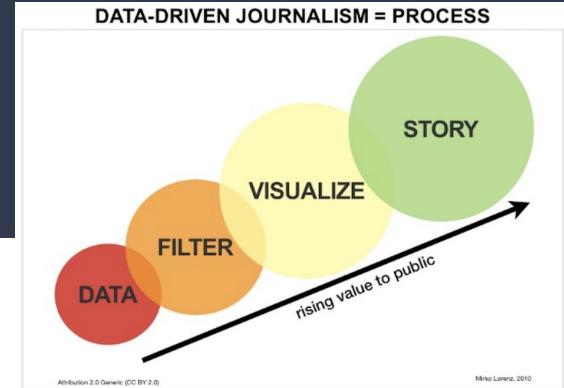
Это программа, которая позволяет работать с Python и удобным редактором-запускаторм кода **Jupyter Notebook**.



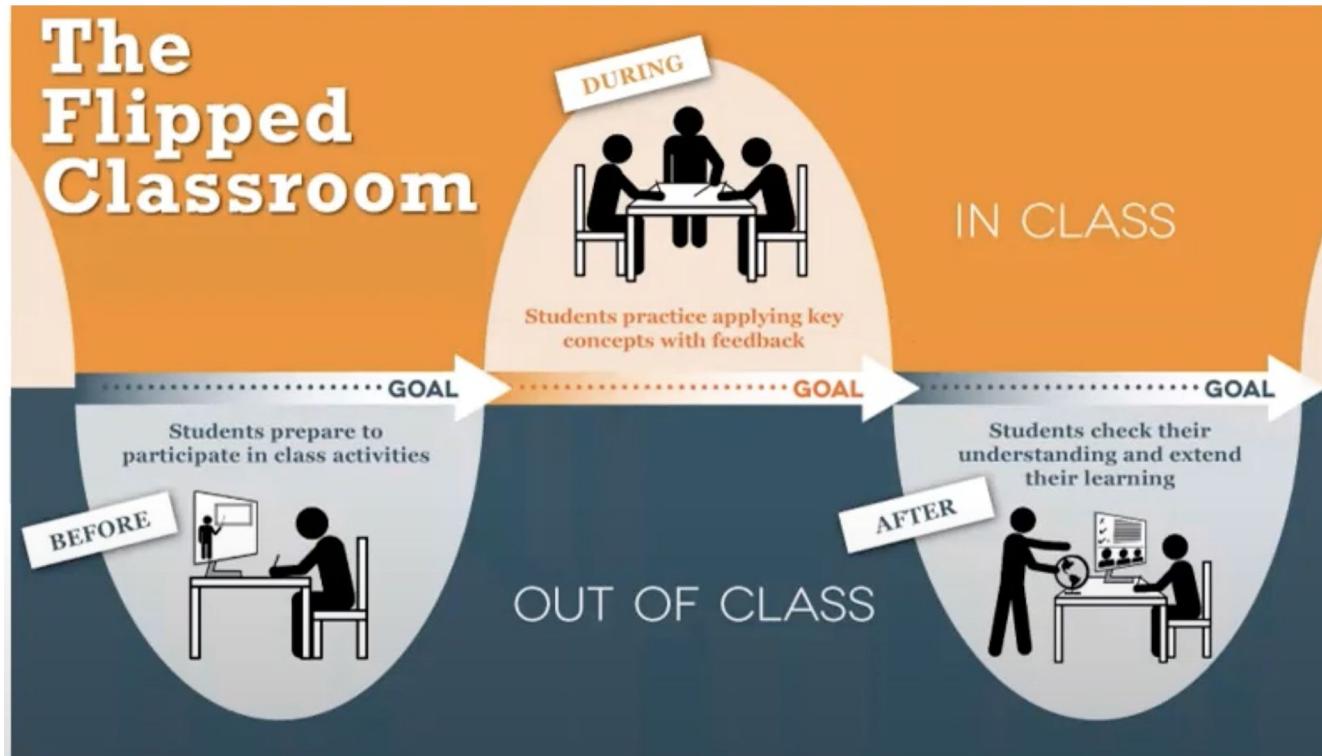
Логистика курса

Главный информационный ресурс – [wiki-страница](#) курса. На ней находятся все полезные ссылки, а именно:

- Карточка курса и ПУД
- Репозиторий с материалами на GitHub
- и многое другое.

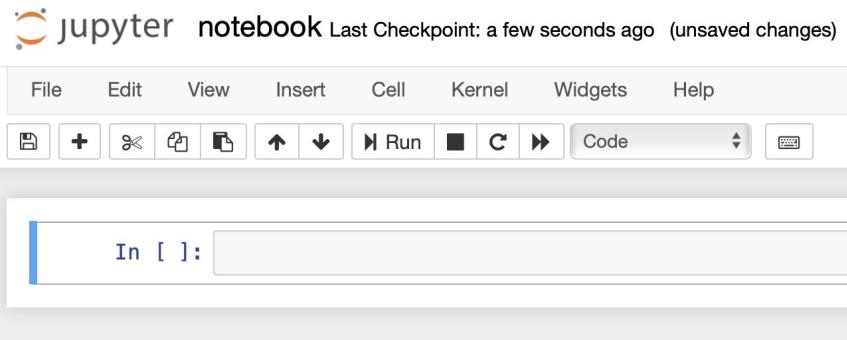


Концепция обучения “Перевернутый класс”



Полезные горячие клавиши в джupyterе

Находясь во внешней области ячейки
(загорается **голубая** вертикальная линия)



shift + enter #запуск ячейки

x #удаление ячейки

a (above) #добавление ячейки сверху

b (below) #добавление ячейки снизу

ДЗ к первому семинару

- 1) Поставить Anaconda с установленным Jupyter Notebook, согласно инструкции на вики
- 2) Прочитать ноутбуки на [гитхабе](#) про:
 - a) Введение в Python
 - b) Основы работы в Jupyter Notebook
 - c) Code style Питона