НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

# INTRODUCTION TO MACHINE LEARNING FOR PUBLIC POLICY AND ANALYTICS

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 07.09.2018

# ОБЗОР КУРСА

1. **Введение в область машинного обучения и обзор программных средств.**
2. **Обзор математического формализма, необходимого для машинного обучения.**
- Введение в пакет ‘Orange’, Препроцессинг данных, визуализация данных, общие принципы работы.
3. **Кластерный анализ.**
- K-means,
- Hierarchical clustering.
- Проблема выбора числа кластеров.
4. **Principal Component Analysis (PCA).**
5. **Классификация данных.**
- Обзор направлений.
- KNN, SVM, Оценка качества моделей
6. **Вероятностные модели**.
- Наивный Байесовский классификатор
- Сравнение классификаторов
7**. Тематическое моделирование.**
- Проблема выбора числа тем.
- Стабильность тематического моделирования

**8. Сентимент-анализ.**
- Словарный подход,
- применение классификаторов для сентимент анализа.

# BIG DATA — ВОЗМОЖНОСТИ ДЛЯ МОНИТОРИНГА И ПРИНЯТИЯ РЕШЕНИЙ ВЛАСТЬЮ

1. **Оценка общественного мнения об эффективности мер по реализации социально-экономической политики**: — присутствие информации и публикаций в информационном пространстве по всем направлениям деятельности, качественное и количественное сравнение публикаций в СМИ о результатах внедрения управленческих решений.
2. **Отслеживание динамики изменения общественного мнения в отношении деятельности органов государственной власти**: — характер публикаций в СМИ о деятельности органов государственной власти по конкретному набору актуальных тем: информации об изменении пенсионного возраста, применения материнского капитала, изменения минимального размера оплаты труда и т.д
3. **Эволюция образа органа власти или отдельного публичного лица**: — трассировка образа во времени и в региональном разрезе.
4. **Определение целевой аудитории по основным направлениям государственной политики**: — профиль потребителя услуг (здравоохранение, пенсионное обеспечение, социальная защита и т.д.).
5. **Измерение отклика отдельных социальных групп и реакции (интереса) по реализуемым мерам и направлениям государственной политики**: — количество публикаций в СМИ, (федеральные, региональные и т.п.), характер публикаций; — отзывы физических лиц (социальные сети, сайты и т.п.).

# ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

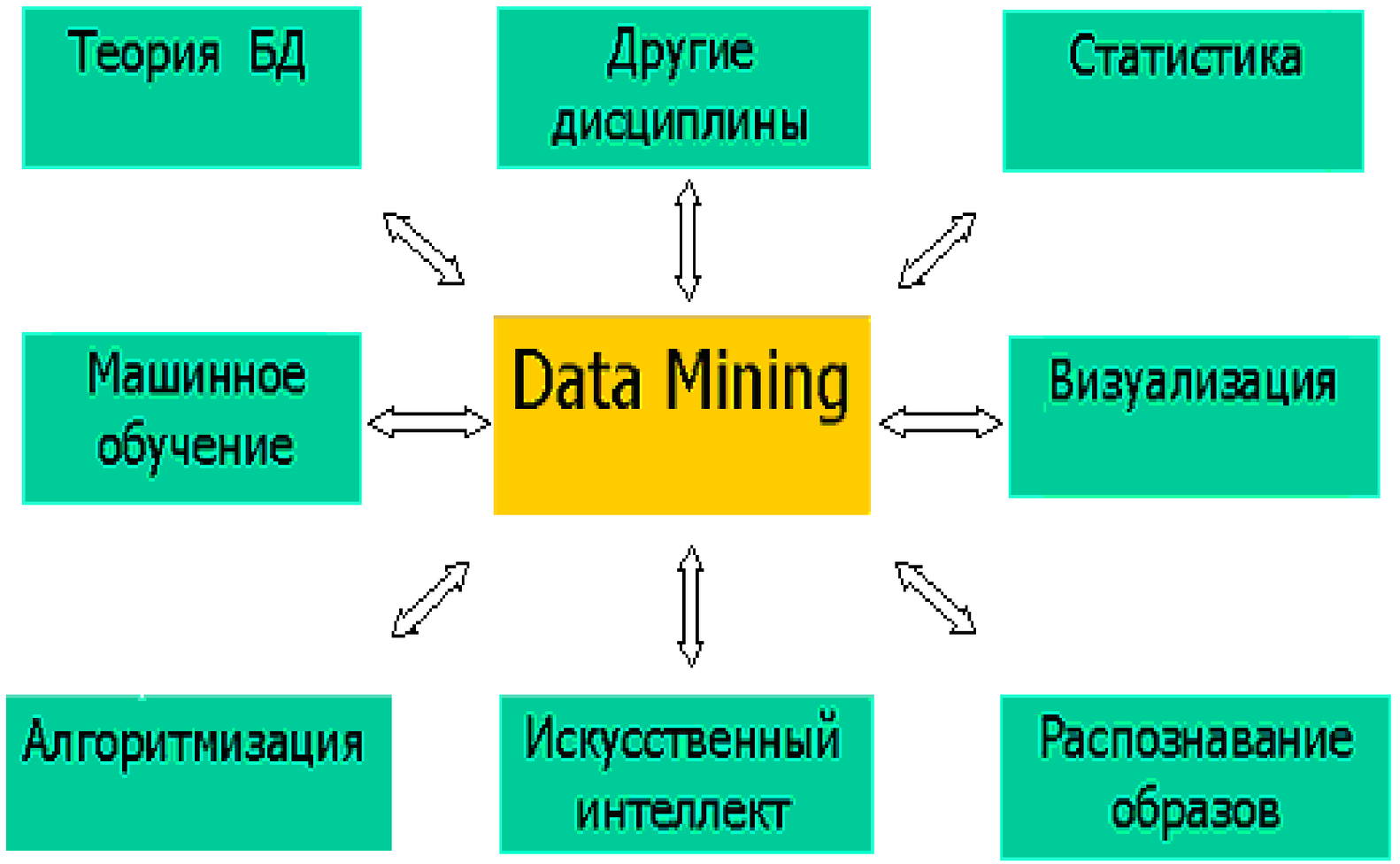
**Data Mining -** мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др..

Термин **Data Mining** часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний в базах данных, информационная проходка данных, "промывание" данных.

**Data Mining -** это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

**Data Mining** находится все еще на ранней стадии развития.
Многие IT-команды увлеклись мифом о том, что средства Data Mining просты в использовании. Предполагается, что достаточно запустить такой инструмент на терабайтной базе данных, и моментально появится полезная информация. На самом деле, успешный Data Mining - проект требует понимания сути деятельности, знания данных и инструментов, а также процесса анализа данных".

# ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

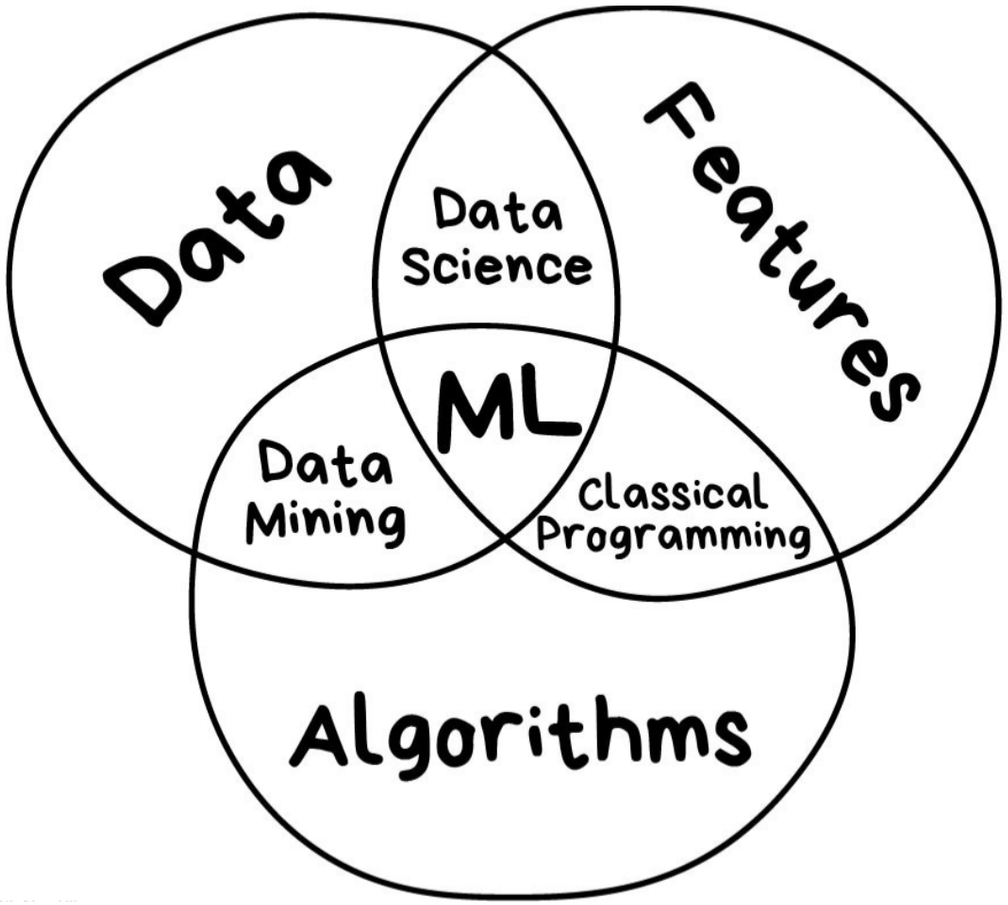


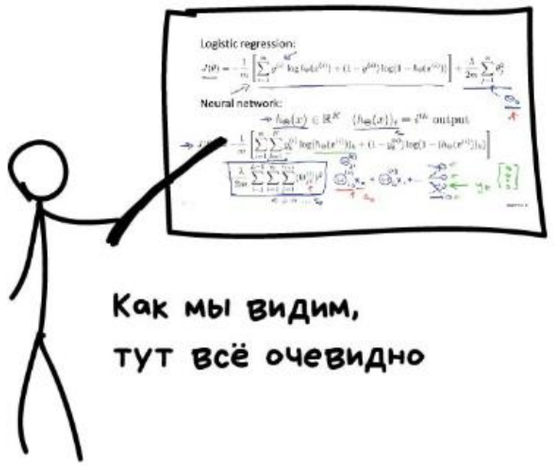

1. Data Mining технология не может заменить аналитика.
2. Технология не может дать ответы на те вопросы, которые не были заданы.
3. Извлечение полезных сведений невозможно без хорошего понимания сути данных

Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены в данных.

# ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

Машинное обучение заключается в извлечении знаний из данных. Это научная область, находящаяся на пересечении статистики, искусственного интеллекта и компьютерных наук и также известная как прогнозная аналитика или статистическое обучение.






Есть два типа статей про машинное обучение

# ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

**Три составляющие обучения**

Цель машинного обучения — предсказать результат по входным данным. Чем разнообразнее входные данные, тем проще машине найти закономерности и тем точнее результат.

**Данные**. Если хотим определять спам — нужны примеры спам-писем, предсказывать курс акций — нужна история цен, узнать интересы пользователя — нужны его лайки или посты. Данных нужно как можно больше. Данные собирают как могут. Кто-то вручную — получается дольше, меньше, зато без ошибок. Кто-то автоматически — просто сливает машине всё, что нашлось, и верит в лучшее. Самые хитрые, типа гугла, используют своих же пользователей для бесплатной разметки.

**Признаки**: Мы называем их фичами (features). Фичи, свойства, характеристики, признаки — ими могут быть пробег автомобиля, пол пользователя, цена акций, даже счетчик частоты появления слова в тексте может быть фичей.

**Алгоритм**. Одну задачу можно решить разными методами приблизительно всегда. От выбора метода зависит точность, скорость работы и размер готовой модели. Но есть один нюанс: если данные очень плохие, даже самый лучший алгоритм не поможет.

# КАРТА МАШИННОГО ОБУЧЕНИЯ

# КАРТА МАШИННОГО ОБУЧЕНИЯ

# ОБЗОР ПРОГРАММНЫХ СРЕДСТВ

## 1. Orange (https://orange.biolab.si/).


Open source фреймворка для анализа данных на основе визуального программирования. Пакет позволяет загружать данные, применять различные алгоритмы машинного обучения, а также визуализировать результаты работы алгоритмов.

## 2. Knime (https://www.knime.com/knime-analytics-platform/).

Open source фреймворка для анализа данных. Данный фреймворк позволяет реализовывать полный цикл анализа данных включающий чтение данных из различных источников, преобразование и фильтрацию, собственно анализ, визуализацию и экспорт.


## 3. R. (R studio: https://www.rstudio.com/ )

Язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом (Rstudio, Microsoft R)

## 4. Python (https://www.anaconda.com/).

Python стал общепринятым языком для многих сфер применения науки о данных (data science). В Python есть библиотеки для загрузки данных, визуализации, статистических вычислений, обработки естественного языка, обработки изображений и многого другого.

**Jupyter, Notebook, Pycharm, Spider.**

# ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ (SUPERVISED LEARNING)

Обучение с учителем используется, когда нужно, обучить алгоритм на основе пары объект-ответ. В этом случае, внутренние параметры алгоритма рассчитываются исходя из того, что есть соответствие между набором признаков, которые характеризуют объект, и ответом. После того как произошло обучение (настройка внутренних параметров алгоритма), можно использовать обученный алгоритм для получения предсказания на новых ранее не встречавшихся данных.

Обучение с учителем: классификация (classification) и регрессия (regression).
**Цель классификации** состоит в том, чтобы спрогнозировать метку класса (class label), которая представляет собой выбор из заранее определенного списка возможных вариантов.
**Цель регрессии** состоит в том, чтобы спрогнозировать непрерывное число в виде функции от заданных параметров.

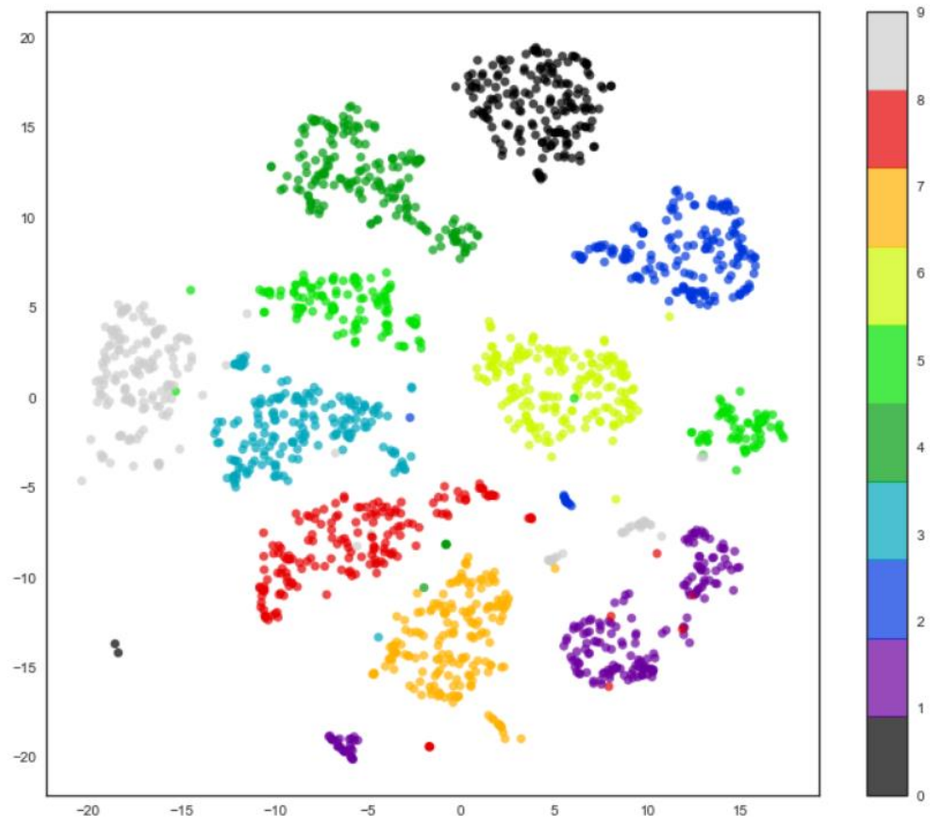
**Например**: Задача сентимент классификации текстов из социальных сетей на основе фиксированного набора оценок в зависимости от набора слов,

*Самый простой способ отличить классификацию от регрессии – спросить себя, заложена ли в полученном ответе определенная непрерывность (преемственность).*

# ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ (UNSUPERVISED LEARNING)

Обучение без учителя (самообучение, спонтанное обучение) — один из способов машинного обучения, при котором алгоритм спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора. Однако, самообучение происходит все же на основе заранее заданных метрик или правил обучения. Для каждого из алгоритмов существуют свои метрики.

Обучение без учителя:
Кластерный анализ (**K means**, **C means** и так далее)
Некоторые алгоритмы тематического моделирования (**Topic modeling**).
Задачи сокращения размерности (Метод главных компонент (Principal Components Analysis, **PCA**)), Метод независимых компонент (Independent component analysis (**ICA**)).

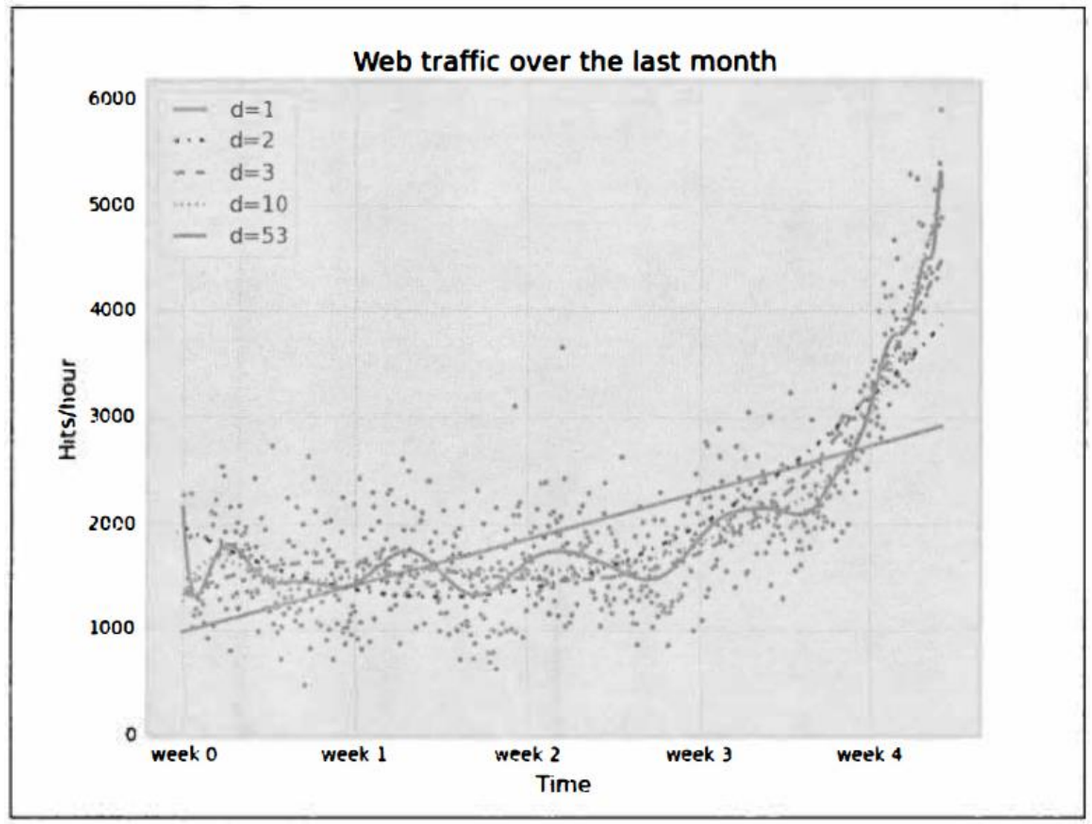


# ОБОБЩАЮЩАЯ СПОСОБНОСТЬ И ПЕРЕОБУЧЕНИЕ АЛГОРИТМОВ МО

**Обобщающая способность** (Generalization ability) это способность модели, построенной на основе обучения выдавать правильные результаты не только для примеров, участвовавших в процессе обучения, но и для любых новых, которые не участвовали в нем.

Если по какой-либо причине модель не приобрела способность к обобщению, ее практическое использование бессмысленно, поскольку на любой пример из обучающего множества она всегда будет выдавать правильный результат, а на любой новый пример – произвольное значение.

**Переобучение.** Построение модели, которая слишком сложна для имеющегося у нас объема информации называется переобучением (overfitting). Переобучение происходит, когда модель слишком точно подстраивается под особенности обучающего набора и вы получаете модель, которая хорошо работает на обучающем наборе, но не умеет обобщать результат на новые данные.



# МАШИННОЕ ОБУЧЕНИЕ В ИССЛЕДОВАТЕЛЬСКОМ ПРОЕКТЕ

Создавая модель машинного обучения, нужно ответить, или, по крайней мере, задуматься над следующими вопросами:

**1.** На какой вопрос(ы) я пытаюсь ответить? Собранные данные могут ответить на этот вопрос?

**2.** Как лучше всего сформулировать свой вопрос(ы) с точки зрения задач машинного обучения?

**3.** У меня собрано достаточно данных, чтобы составить представление о задаче, которую я хочу решить?

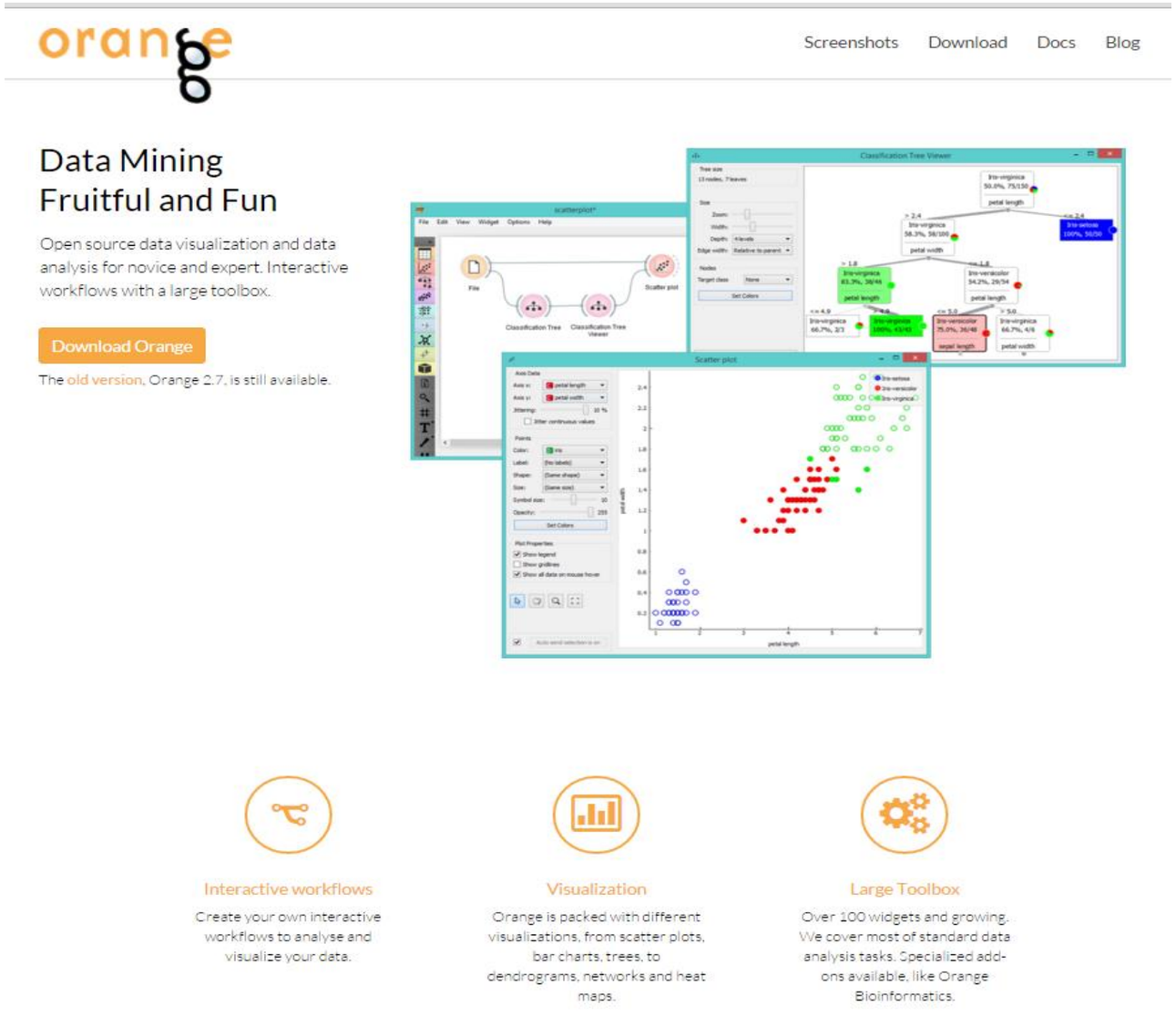
**4.** Какие признаки я извлек и помогут ли они мне получить правильные прогнозы?

**5.** Как я буду измерять эффективность решения задачи?

**6.** Как решение, полученное с помощью машинного обучения, будет взаимодействовать с другими компонентами моего исследования или бизнес-продукта?

В более широком контексте, алгоритмы и методы машинного обучения могут являются лишь этапом более крупного процесса, призванного решить конкретную задачу, и поэтому необходимо всегда держать схему этого процесса в голове.

# ORANGE – OPEN SOURCE SOFTWARE FOR DATA VISUALIZATION AND DATA ANALYSIS




Можно скачать по адресу: http://orange.biolab.si/

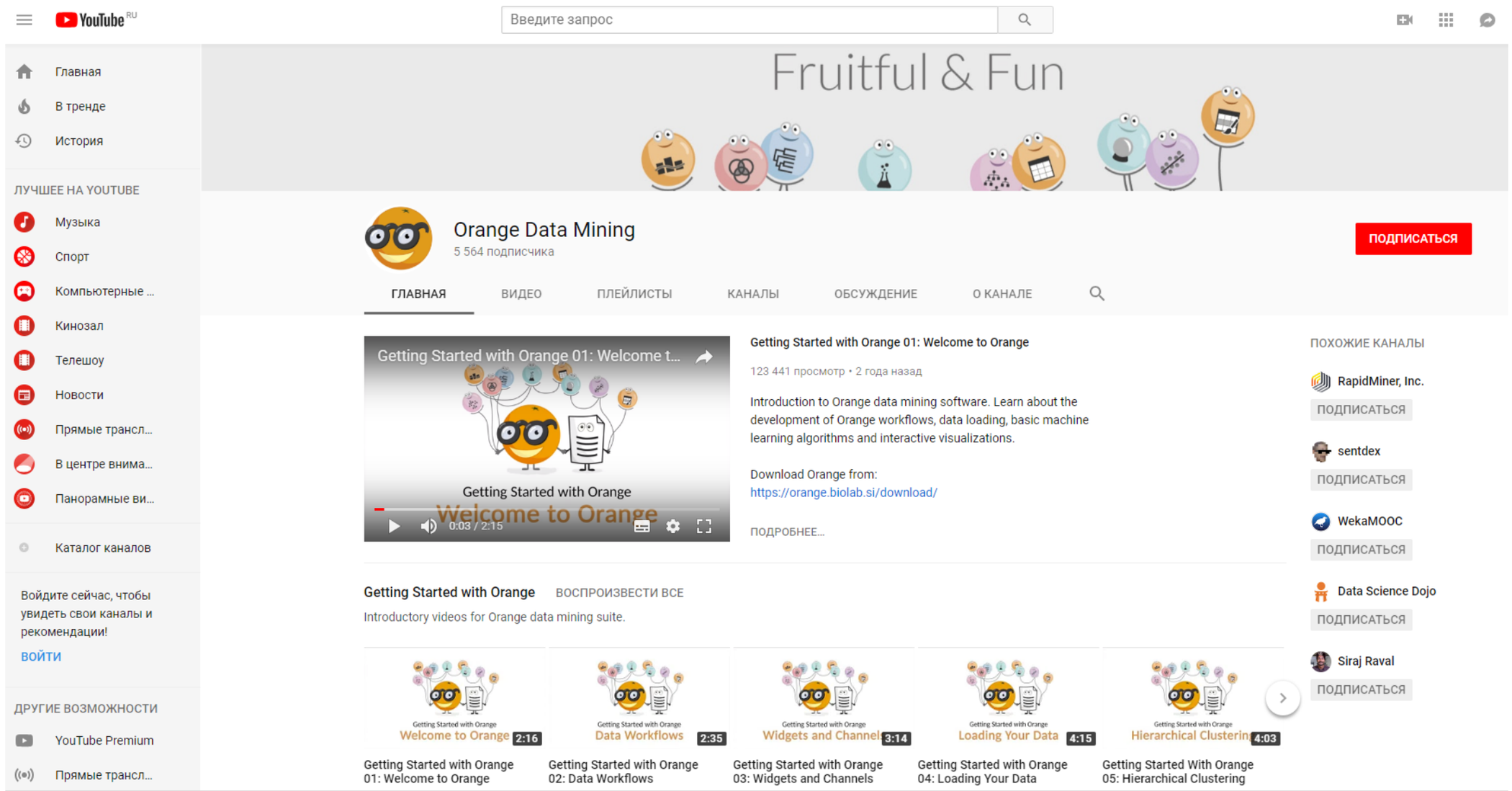
Faculty of Computer and Information Science University of Ljubljana

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python.

Journal of Machine Learning Research

# ORANGE – OPEN SOURCE SOFTWARE FOR DATA VISUALIZATION AND DATA ANALYSIS

## YOUTUBE ВИДЕО КУРС




НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru