Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №6

за 7 семестр

По дисциплине: «КМиАД»

Выполнил:

Студент 4 курса

Группы ПО-4(1)

Калиновский В. Е.

Проверил:

Чичурин А. В.

Брест 2022

**Визуальный анализ данных**

**Основной идеей визуального анализа** данных является представление данных в некоторой визуальной форме, позволяющей человеку погрузиться в [данные](https://intellect.icu/informatsiya-znaniya-i-dannye-otnosheniya-mezhdu-nimi-otlichiya-i-skhodstva-preobrazovanie-priznaki-znanij-6173#term-dannye), работать с их визуальным представлением, понять их суть, сделать выводы и напрямую взаимодействовать с данными.

В настоящее время существует достаточно большое количество различных видов графических образов, позволяющих представлять результаты анализа в виде, удобном для понимания человеком.

С помощью новых технологий пользователи способны оценивать: большие [объекты](https://intellect.icu/obekty-i-elementy-zashhity-v-kompyuternykh-sistemakh-obrabotki-dannykh-765#term-obekty)или маленькие, далеко они находятся или близко. Пользователь в реальном времени может двигаться вокруг объектов или кластеров объектов и рассматривать их cо всех сторон. Это позволяет использовать для анализа естественные человеческие перцепционные навыки в обнаружении неопределенных образцов в визуальном трехмерном представлении данных.

**Визуальный анализ данных особенно полезен**, когда о самих данных мало известно и цели исследования до конца непонятны. За счет того, что пользователь напрямую работает с данными, представленными в виде визуальных образов, которые он может рассматривать с разных сторон и под любыми углами зрения, в прямом смысле этого слова, он может получить дополнительную информацию, которая поможет ему более четко сформулировать цели исследования.

Таким образом, визуальный анализ данных можно представить как **процесс генерации гипотез.** При этом сгенерированные гипотезы можно проверить или автоматическими средствами (методами статистического анализа или методами Data Mining), или средствами визуального анализа. Кроме того, прямое **вовлечение пользователя** в визуальный анализ имеет **два основных преимущества** перед автоматическими методами:

- визуальный анализ данных позволяет легко работать с неоднородными и зашумленными данными, в то время как не все автоматические методы могут работать с такими данными и давать удовлетворительные результаты;

- визуальный анализ данных интуитивно понятен и не требует сложных математических или статистических алгоритмов.

**Три основные характеристики визуализации:**

- характер отображаемых данных, которые нужно визуализировать с помощью данного средства;

- методы визуализации и образы, в виде которых могут быть представлены данные;

- возможности взаимодействия с визуальными образами и методами для лучшего анализа данных.

Наборы визуализируемых данных, как и в Data Mining, представляют собой матрицы, в которых [ряды](https://intellect.icu/category/ryady#term-ryady)являются данными (например, записями об экспериментах, покупки в магазине и т. п.), а колонки — атрибутами данных. При этом данные могут характеризоваться одним или несколькими атрибутами. Кроме того, сами данные могут иметь более сложную структуру: иерархическую, текстовую, графическую и т. п.

Для визуализации перечисленных типов данных используются различные визуальные образы и методы их создания. Очевидно, что количество визуальных образов, которыми могут представляться данные, ограничиваются только человеческой фантазией. Основное требование к ним — это наглядность и удобство анализа данных, которые они представляют. Методы визуализации могут быть как самые простые (линейные графики, диаграммы, гистограммы и т. п.), так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов.

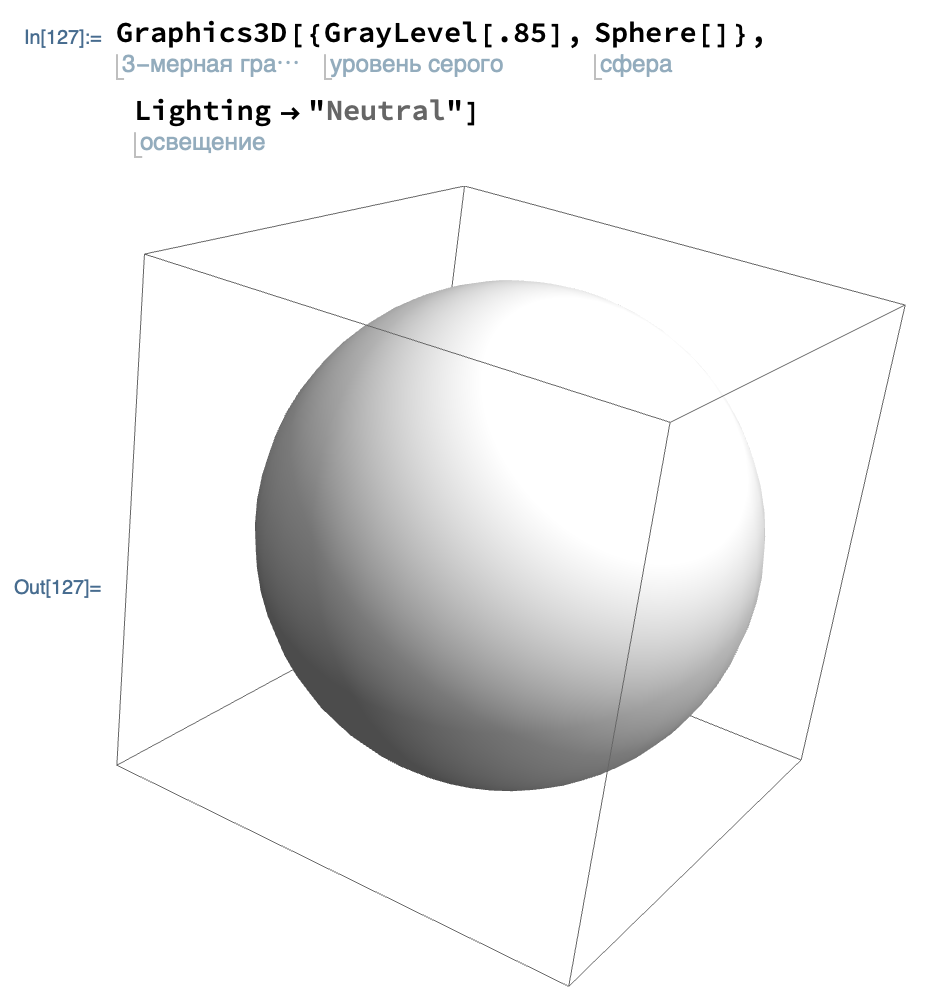
Простейшие методы визуализации, к которым относятся 2D/3D-образы, широко используются в существующих системах (например, в Microsoft Excel). К этим методам относятся: графики, диаграммы, гистограммы и т. п. Основным их недостатком является невозможность приемлемой визуализации сложных данных и большого количества данных.

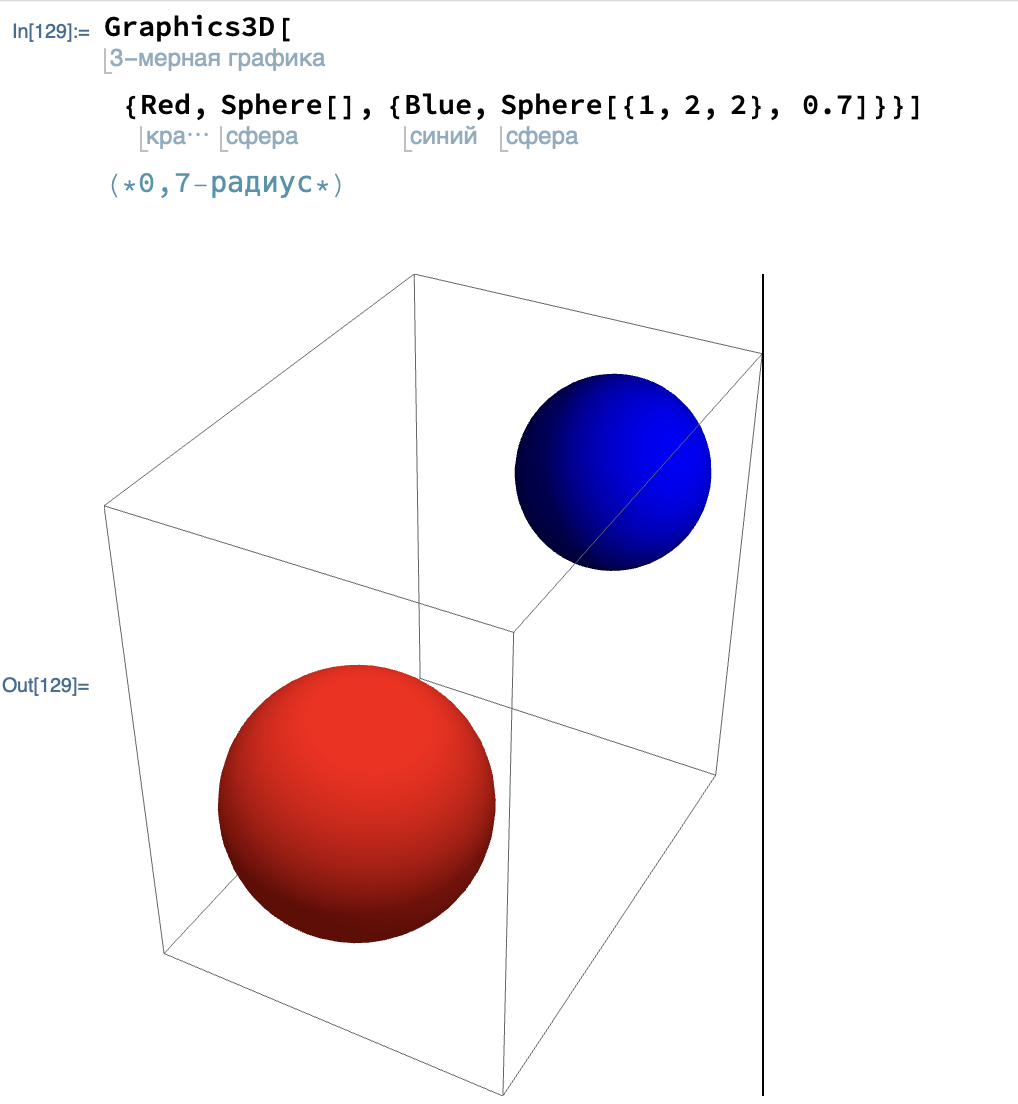
При исследовании большого количества данных важно иметь возможность разделять наборы данных и выделять интересующие поднаборы — фильтровать образы. При этом важно, чтобы данная возможность предоставлялась в режиме реального времени работы с визуальными образами (т. е. интерактивно). Выбор поднабора может осуществляться или напрямую из списка, или с помощью определения свойств интересующего поднабора. Выбор из списка неудобен при большом количестве поднаборов, в то же время запросы не всегда позволяют получить желаемый результат.

**Graphics3D**

Mathematica включает в себя графические примитивы, при помощи которых Вы можете создавать двумерную и трехмерную графику любой сложности.

Функция Graphics3D преобразует трехмерные примитивы в трехмерные графические объекты, добавляя в структуру примитива стили (цвет, толщину и т.д.).



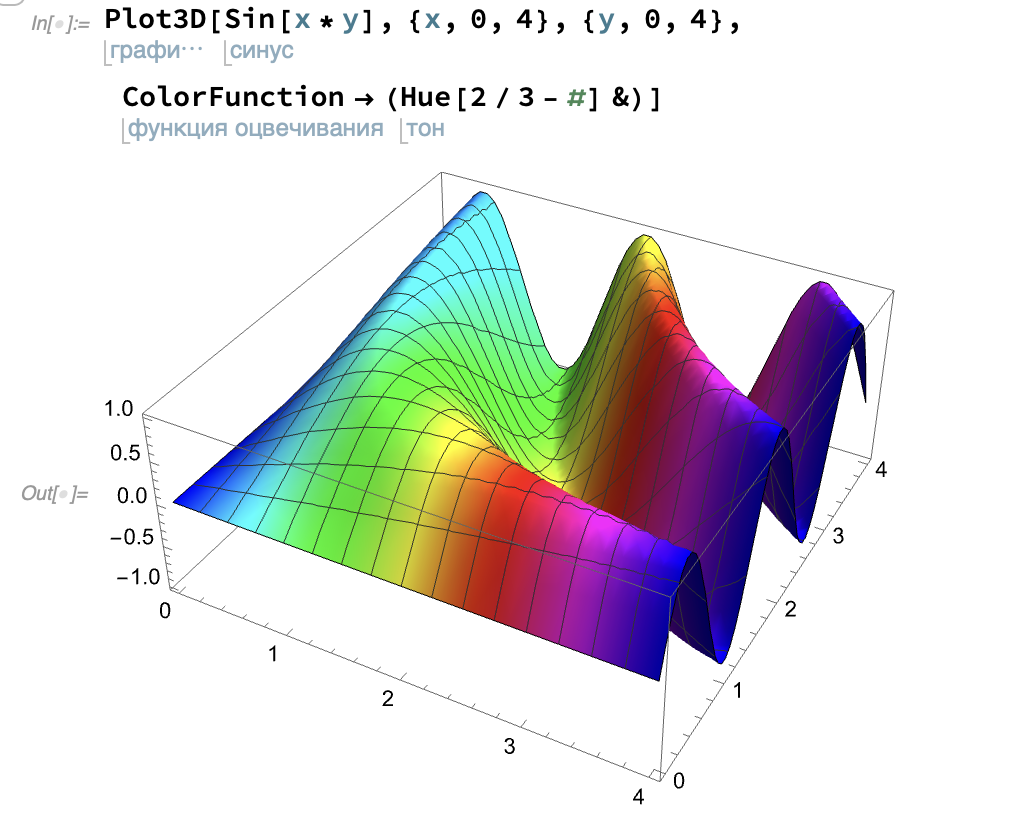


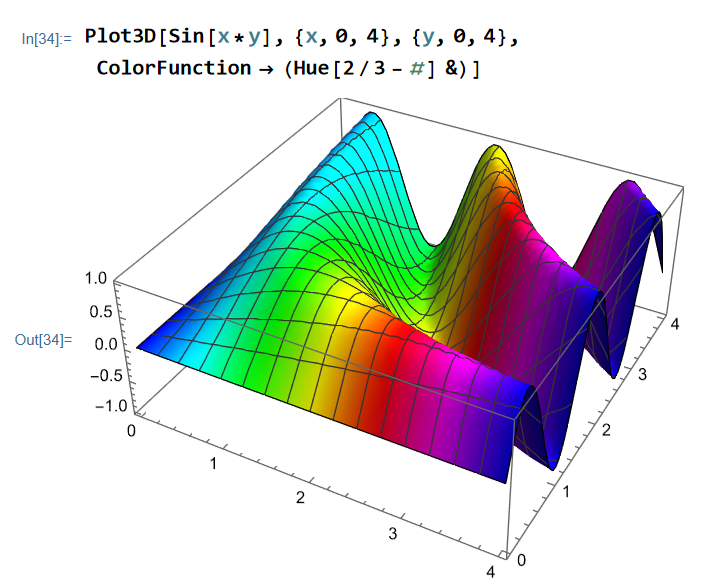
**Plot3D**

Mathematica позволяет создавать не только статичную, но и анимированную графику.

Первый аргумент – выражение f, график которого должен быть построен.

Второй и третий аргументы определяют границы изменения переменных x и y задаваемые в виде списков. Обычно функция вычисляется на сетке 15 на 15 и каждый кусок закрашивается согласно светоотражающей модели.

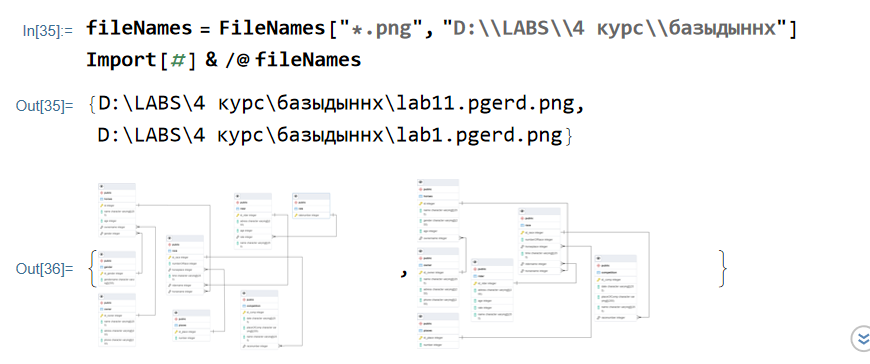




Импорт и экспорт может обрабатывать не только табличные данные, но и данные, соответствующие графике, звукам, выражениям и даже целым документам. Импорт и экспорт часто могут вывести соответствующий формат для данных, просто посмотрев на расширение имени файла для файла, в котором хранятся данные.

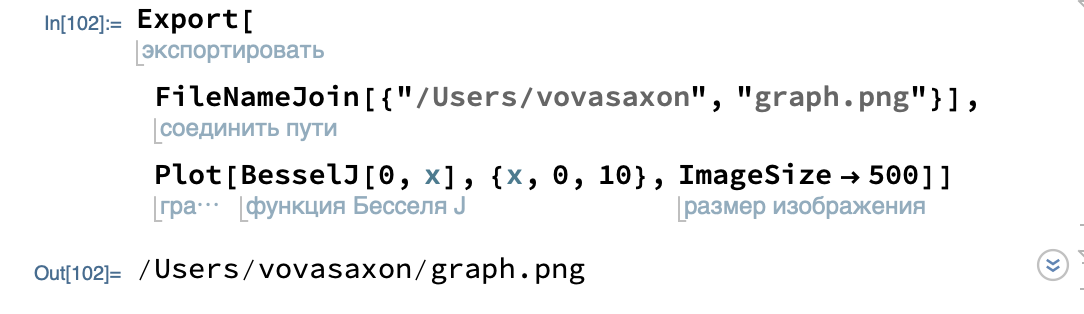
**Import**

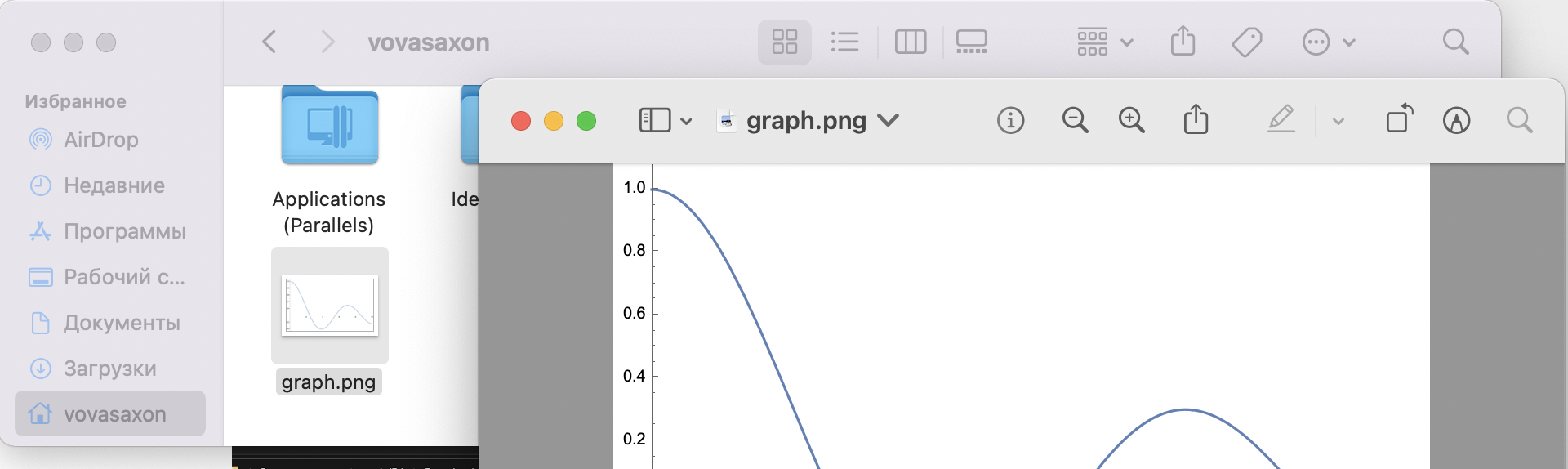
Импортирует данные из источника, возвращая их представление.



**Export**

Экспортирует данные в файл, преобразуя их в формат, соответствующий расширению файла.

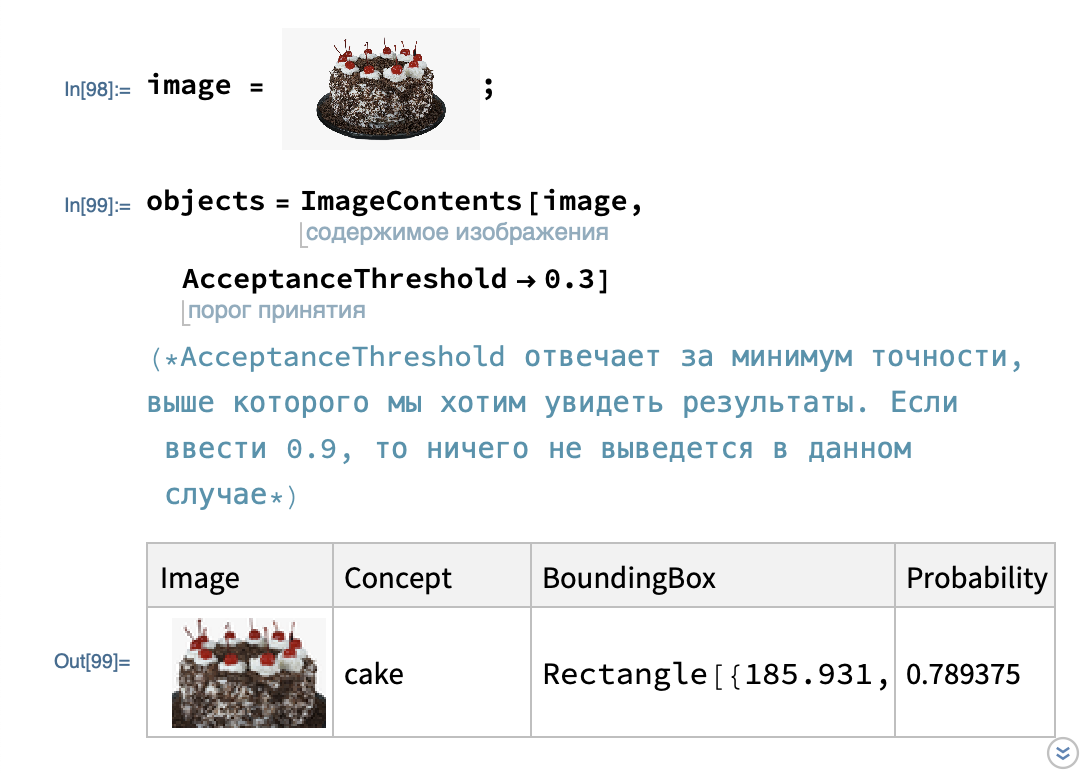




Используя различные современные методы, Wolfram Language обеспечивает немедленные функции для идентификации изображений и обнаружения и распознавания объектов, а также извлечения признаков. Wolfram Language поддерживает определенные геометрические объекты, такие как края и углы, а также общие ключевые моменты, которые можно использовать для регистрации и сравнения изображений.

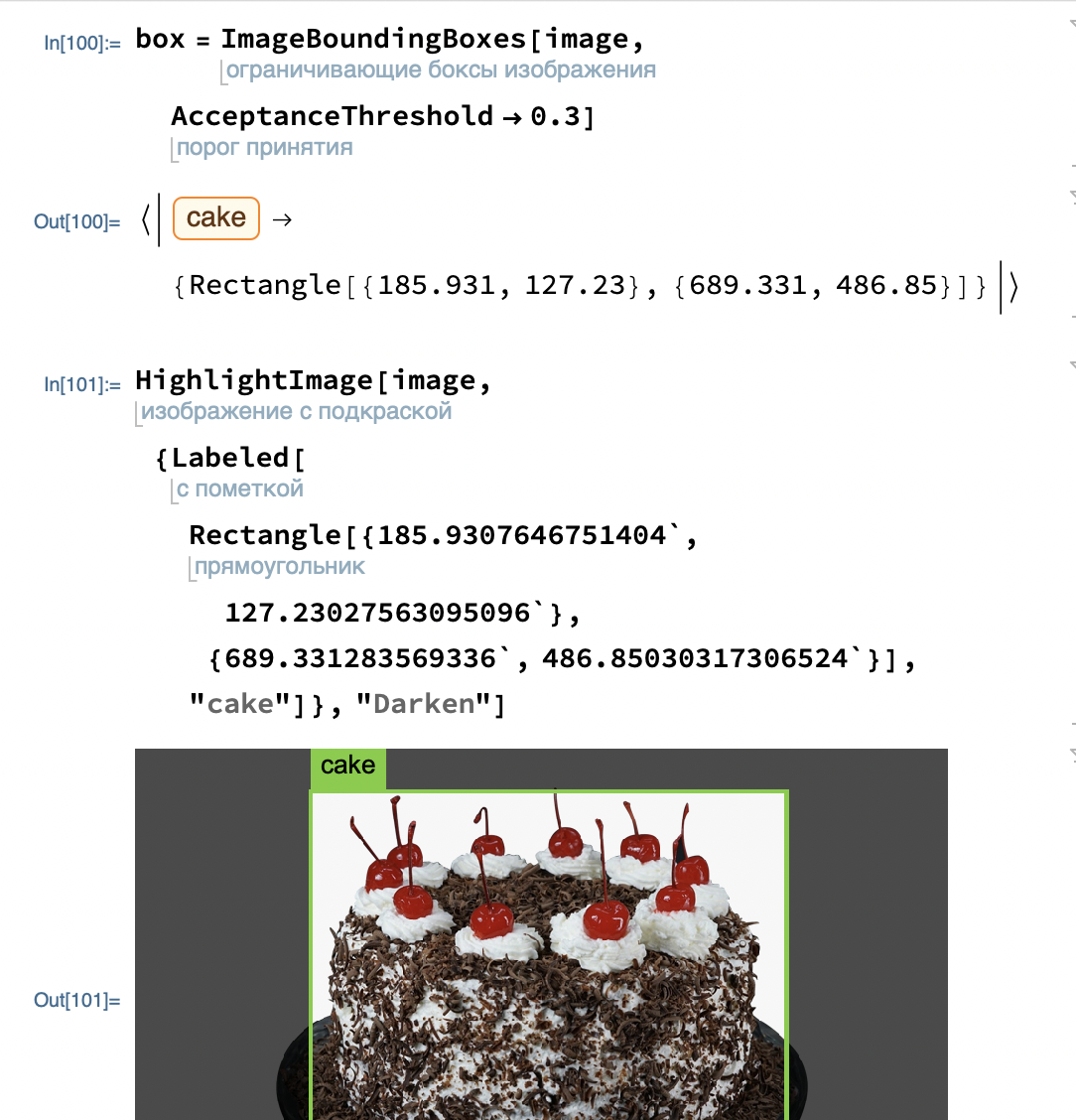
**ImageContents, ImageBoundingBoxes**

Дает набор данных идентифицированных сущностей на изображении.



**HighlightImage**

Дает сопоставление списков ограничительных рамок для каждой идентифицированной категории объектов на изображении.



**Извлечение знаний из Web – Web Mining**

1. Проблемы анализа информации из Web.

Всемирная сеть сейчас содержит огромное количество информации, знаний. Пользователи на различных условиях могут просматривать всевозможные документы, аудио- и видеофайлы. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в Интернет.

1. Этапы Web Mining.

Web Mining – технология, использующая методы Data Mining для исследования и извлечения информации из Web-документов и сервисов. Выделяют следующие этапы применения Web Mining:

1. Поиск ресурсов – локализация неизвестных документов и сервисов в Web.  
2. Извлечение информации – автоматическое извлечение определенной информации из найденных Web-ресурсов.   
3. Обобщение – обнаружение общих шаблонов в отдельных и пересекающихся множествах сайтов.   
4. Анализ – интерпретация найденных шаблонов.

1. Web Mining и другие интернет-технологии.

Web Mining, являясь инструментом для обработки и анализа Web-ресурсов, рассматривается в одном ряду с такими интернет-технологиями, как получение информации (Information Retrieval – IR) и извлечение информации (Information Extraction – IE).

1. Извлечение Web-контента в процессе информационного поиска.

Методы извлечения Web-контента в процессе информационного поиска во многом зависят от типа анализируемых документов. Различают два основных типа: неструктурированные и почти структурированные. К неструктурированному типу относятся все текстовые документы, не имеющие определенной структуры. К почти структурированным относятся документы, имеющие структуру в целом, но позволяющую вхождение в структурный элемент неструктурированного текста. К таким документам относятся HTML, XML и др.

1. Слабоструктурированные документы.

Извлечение Web-контента из слабоструктурированных документов использует более развитые средства представления текста. Это в первую очередь связано с тем, что в документах уже выделены некоторые структурные элементы. Практически все методы в этой области для представления документа используют HTML-структуры внутри документов. Некоторые методы используют также для представления гиперссылки между документами.

1. Извлечение Web-контента для формирования баз данных.

Задача извлечения Web-контента для его размещения в базе данных относится к проблеме управления информацией и обработки запросов к ней. Существуют три класса задач, относящихся к этой проблеме:

* моделирование и формирование запросов к Web;
* извлечение информации и интеграция;
* создание и реструктуризация Web-сайта.

7) Исследование использования Web-ресурсов.  
Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:  
1. Препроцессиг  
2. Извлечение шаблонов  
3. Анализ шаблонов.

8) Этап препроцессинга.

Для решения задачи исследования использования Web на этапе препроцессинга в массиве анализируемых данных должны быть выделены перечисленные сущности. Это значительно усложняет неполнота данных, получаемых с одного источника.

9) Этап извлечения шаблонов.   
Для извлечения шаблонов из информации об использовании Web-ресурсов применяются различные методы как классической статистики, так и относящиеся к области Data Mining.

10) Этап анализа шаблонов и их применение

Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. Целью анализа является отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие шаблоны. Методология анализа во многом зависит от области применения, в которой он выполняется.

Используйте базу данных рукописных цифр MNIST для обучения свёрточной сети определять цифры в изображении.

Начните с получения данных режима обучения и контрольных данных.

In[1]:=

Click for copyable input

resource = ResourceObject["MNIST"];

trainingData = ResourceData[resource, "TrainingData"];

testData = ResourceData[resource, "TestData"];

In[2]:=

Click for copyable input

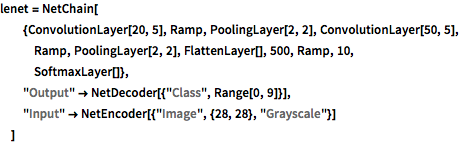
RandomSample[trainingData, 5]

Out[2]=



Задайте свёрточную нейронную сеть, которая принимает входные данные в виде полутоновых изображений размером 28×28.

In[3]:=



lenet = NetChain[

{ConvolutionLayer[20, 5], Ramp, PoolingLayer[2, 2],

ConvolutionLayer[50, 5], Ramp, PoolingLayer[2, 2], FlattenLayer[],

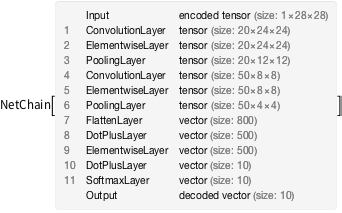
500, Ramp, 10, SoftmaxLayer[]},

"Output" -> NetDecoder[{"Class", Range[0, 9]}],

"Input" -> NetEncoder[{"Image", {28, 28}, "Grayscale"}]

]

Out[3]=



Проведите три раунда тренировки нейронной сети и восемь для сравнения.

In[4]:=

Click for copyable input

lenet = NetTrain[lenet, trainingData, ValidationSet -> testData,

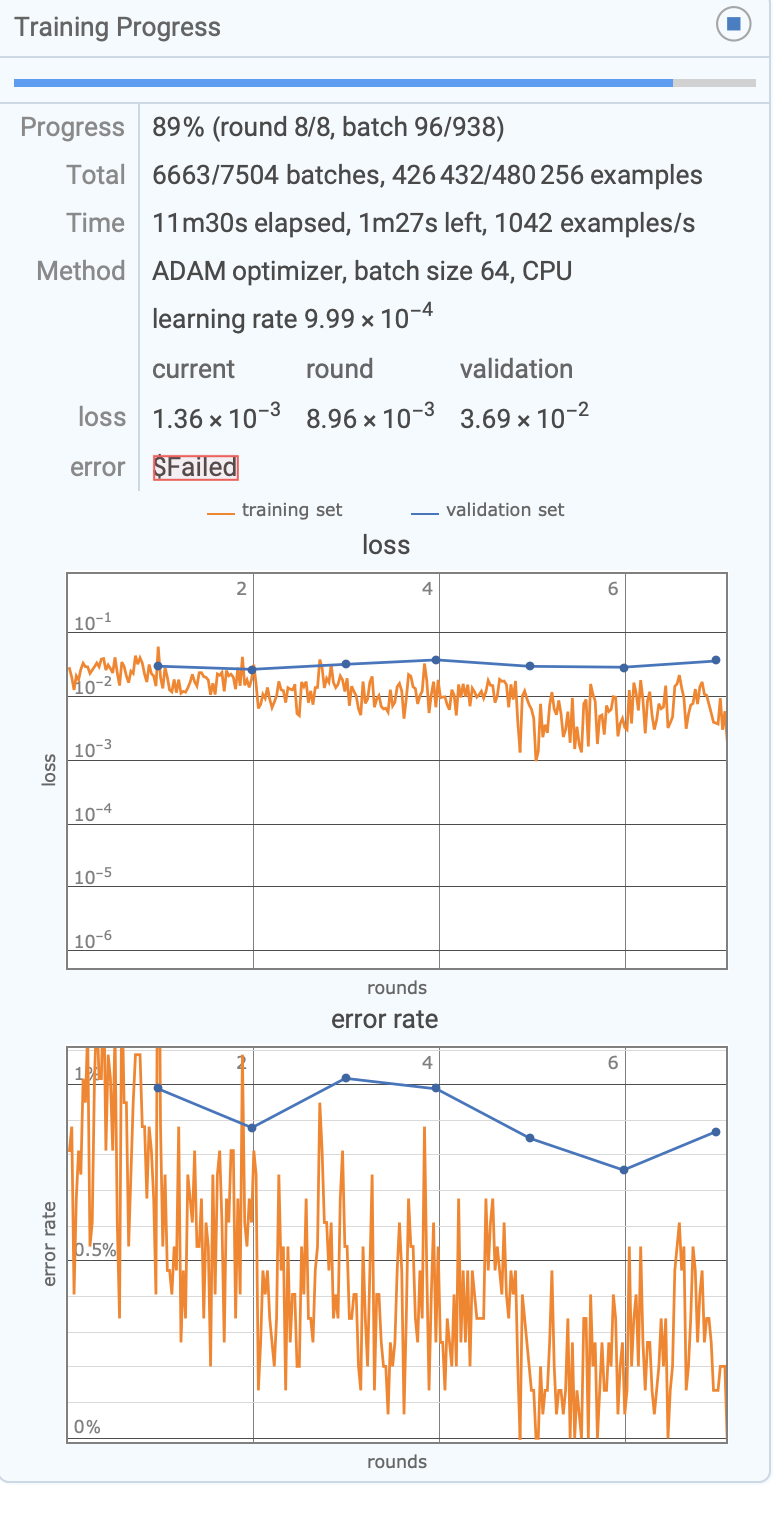
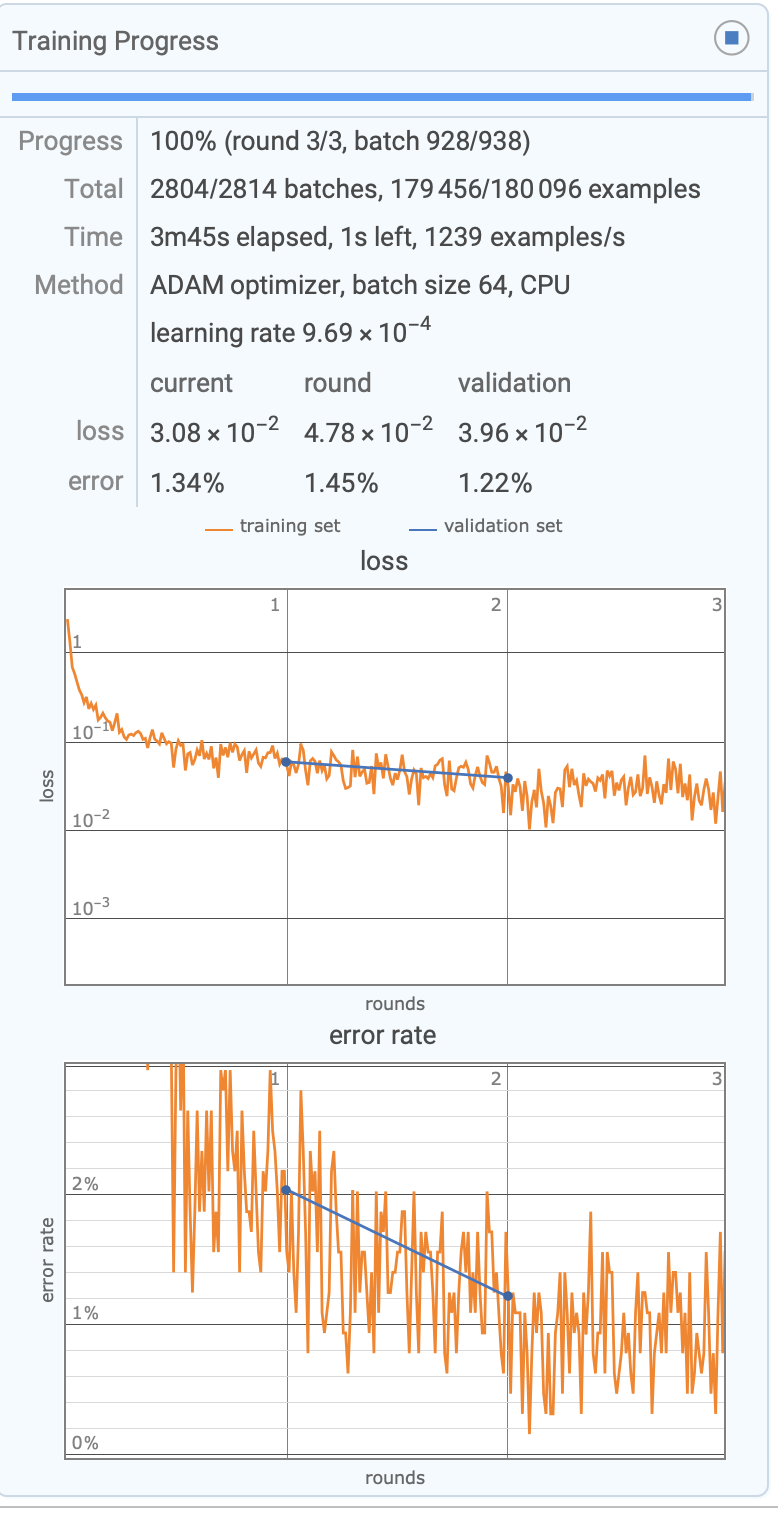
MaxTrainingRounds -> 3];

Click for copyable input

lenet = NetTrain[lenet, trainingData, ValidationSet -> testData,

MaxTrainingRounds -> 8];

Out[5]=



Оцените обученную сеть непосредственно на изображениях, случайно отобранных из контрольного набора.

In[6]:=

Click for copyable input

imgs = Keys @ RandomSample[testData, 5];

Thread[imgs -> lenet[imgs]]

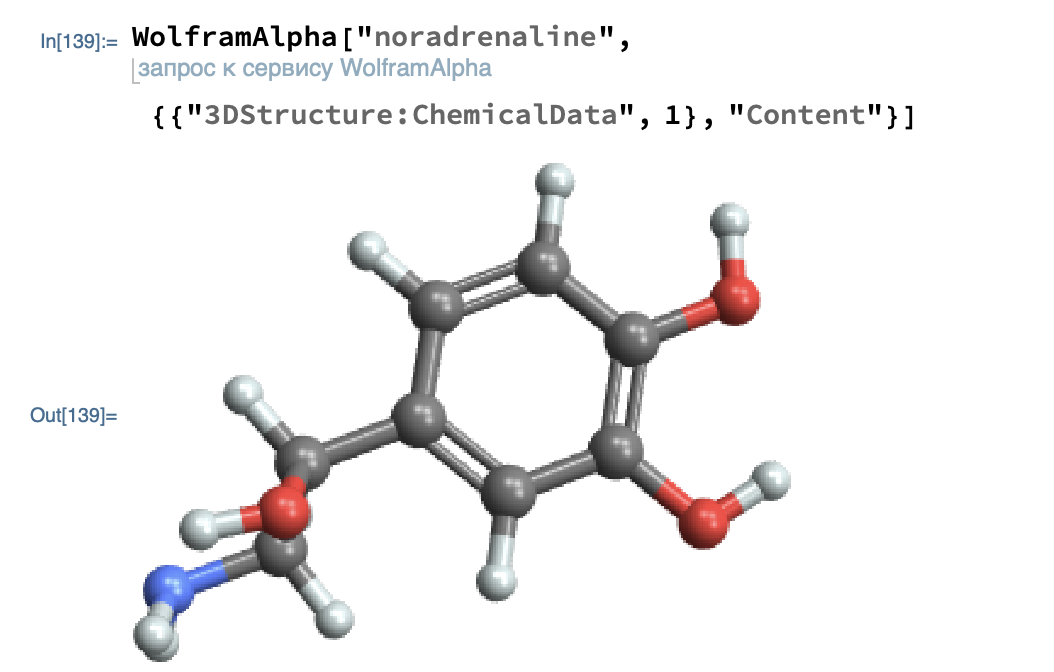
Out[6]=



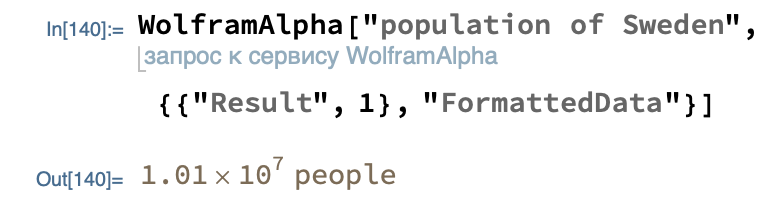
**WolframAlpha** – это смесь алгоритмов и использование искусственного интеллекта для поиска нужной информации. То есть он сам найдёт на основе входных данных нужную информацию. Всё, что нужно будет сделать – это обработать её.

**WolframAlpha**

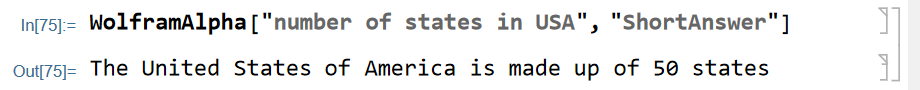
Oтправляет запрос в адрес Wolfram| Альфа и импортирует выходные данные.



Вывод информации населении Норвегии:



Получение информации о количестве штатов в США:



**Вывод:**

- Web Mining включает в себя следующие этапы: поиск ресурсов, извлечение информации, обобщение и анализ.

- Различают следующие категории задач Web Mining: извлечение Web-контента, извлечение Web-структур и исследование использования Web-ресурсов.

- В решении задачи извлечения структуры Web используются подходы из области социальных сетей, библиометрики, ранжирования документов и т. п.

- В задаче исследования использования Web анализу подвергаются вторичные данные о взаимодействии пользователя с Web: протоколы работы, куки, авторизация и т. п.

- Существуют два основных подхода анализа использования Web-ресурсов: преобразование данных использования Web-сервера в реляционные таблицы до выполнения адаптированных методов Data Mining и использование информации из файла протокола непосредственно, применяя специальные методы предварительной обработки.