К **машинному обучению** относится класс методов **искусственного интеллекта**, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

В последние годы вычислительные возможности компьютерной техники значительно улучшились, что позволяет эффективно обрабатывать и анализировать огромные объемы данных. Это позволяет моделям машинного обучения быстро обучаться и применяться в реальном времени.

Чтобы применять алгоритмы машинного обучения, крайне важно понимать **различные типы данных**, что помогает выбрать подходящие алгоритмы, правильно предварительно обработать данные, правильно интерпретировать результаты и обеспечить качество данных, что в итоге приводит к более эффективным и точным моделям машинного обучения.

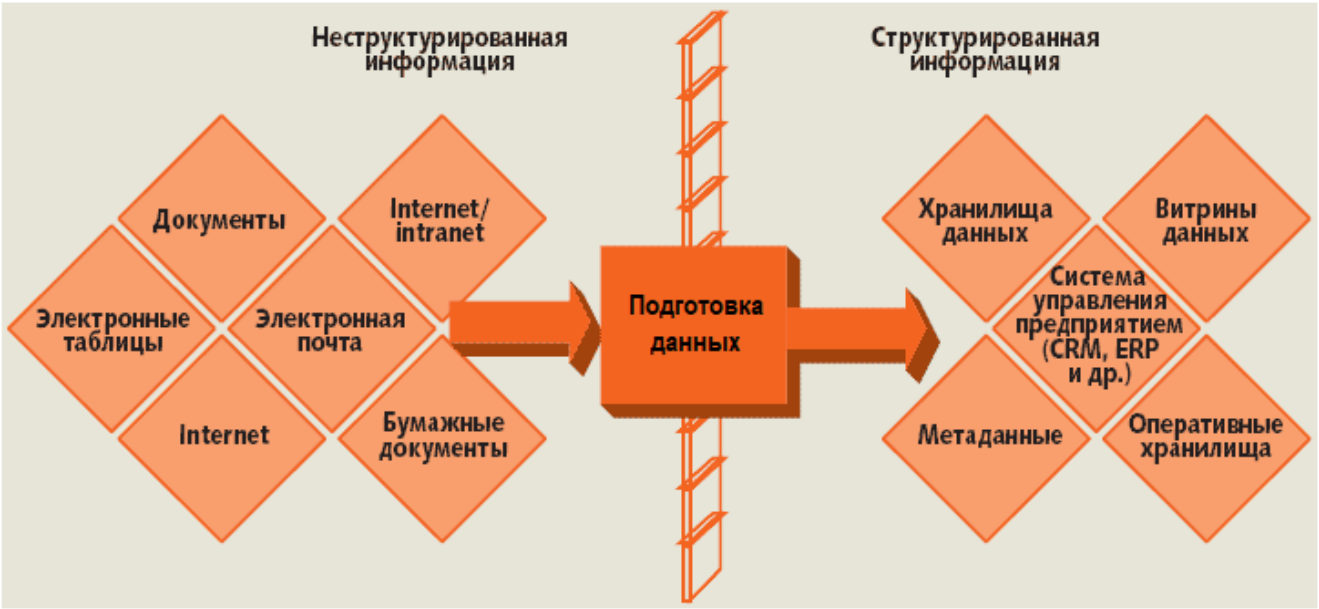
Данные должны быть преобразованы в числовое представление, чтобы машины могли изучать закономерности в данных.

Понимание данных

Первым шагом к решению поставленной задачи машинного обучения является понимание того, с какими данными приходится иметь дело.

Исходные данные обычно представлены файлами разного типа, содержащими информацию разного характера.

Обычно данные делятся на **структурированные** и **неструктурированные**.



Структурированные данные.

В классическом случае имеют вид таблицы

	marital	income	agecat
1	Женат	77680.0	<31
2	Женат	37111.5	<31
3	Женат	NA	<NA>
4	Одинокий	NA	<31
5	Одинокий	16829.6	<31
6	Одинокий	57272.7	<31
7	Женат	NA	<31
8	<NA>	NA	<31
9	Одинокий	92167.3	<31
10	Женат	37135.5	<31

Расшифровка признаков:

marital – *отношение к замужеству*

income – *заработок*

agecat – *возрастная категория*

1. Объект - Признак:

   каждая строчка - объект,

   каждый столбец - некоторый признак.

2. Сенсорные данные (временные ряды): каждый столбец - некоторый сенсор (параметр), каждая строчка - показатели сенсоров на некоторой временной отметке.

Алгоритмы ML работают только со структурированными данными

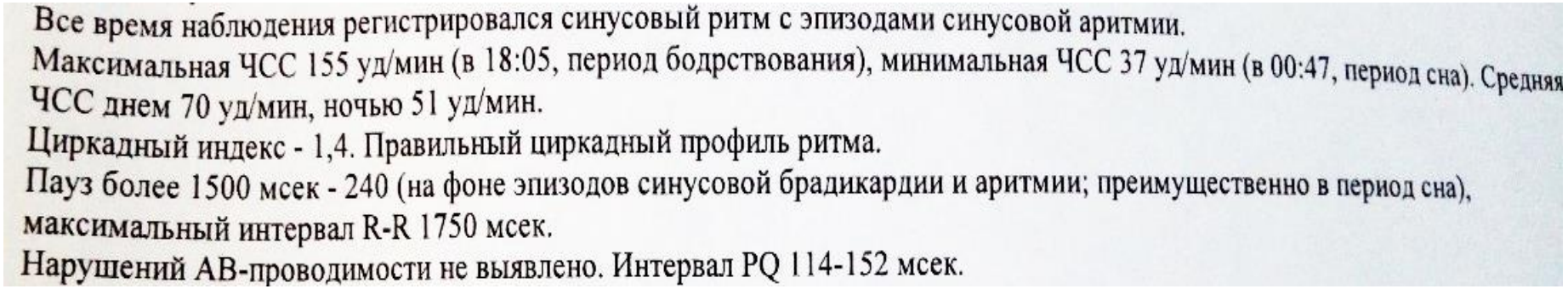
Неструктурированные данные.

Обычно это некоторые документы:

- Простые тексты
- Файлы формата Word
- Презентации
- PDF - документы

Неструктурированные данные – самые «неприятные» для машинного обучения. При работе с ними требуется каким-то фантастическим образом привести их к структурированному (табличному) формату.

Пример – записи врачей об осмотре пациентов.



Все время наблюдения регистрировался синусовый ритм с эпизодами синусовой аритмии.
Максимальная ЧСС 155 уд/мин (в 18:05, период бодрствования), минимальная ЧСС 37 уд/мин (в 00:47, период сна). Средняя ЧСС днем 70 уд/мин, ночью 51 уд/мин.
Циркадный индекс - 1,4. Правильный циркадный профиль ритма.
Пауз более 1500 мсек - 240 (на фоне эпизодов синусовой брадикардии и аритмии; преимущественно в период сна), максимальный интервал R-R 1750 мсек.
Нарушений АВ-проводимости не выявлено. Интервал PQ 114-152 мсек.

Полуструктурированные данные.

Так называют те данные, которые имеют некоторую структуру, но либо она далека от табличного представления либо не является строгой и стабильной во всём документе.

Обычно это различные описания в специализированных форматах.

1. Логи (журналы событий): каждая строчка - это событие, представленное в формализованном виде.

2. Описания объектов

```
{person: &o1{name: “Mary”, age: 45,
            child: &o2, child: &o3},
person: &o2{name: “John”, age: 17,
            relatives: {mother: &o1, sister: &o3}},
person: &o3{name: “Jane”, country: “Canada”, mother: &o1}
}
```

Какая структура здесь описана?

3. Файлы форматов HTML, XML, JSON

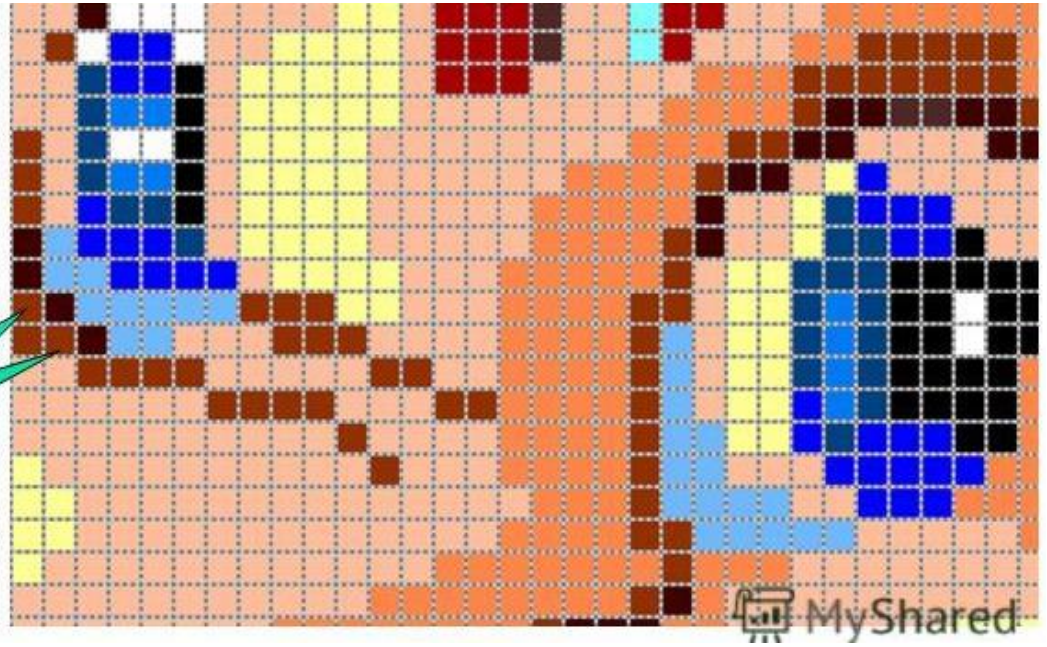
```
<HTML>
<HEAD>
<TITLE> Exercise 1 </TITLE>
</HEAD>
<BODY bgcolor=n#cccc99">
<FONT size=14 color="blue"> Tag «Font»
  </FONT>
</BODY>
</HTML>
```

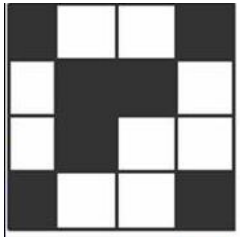
```
<?xml version="1.0" encoding="UTF-8"?>

<bookstore>
  <book>
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

  . . .
</bookstore>
```

```
{
        "firstName": "John",
        "lastName": "Smith",
        "age": 25,
        "address": {
                "streetAddress": "21 2nd Street",
                "city": "New York",
                "state": "NY",
                "postalCode": 10021
        },
        "phoneNumbers": [
                {
                        "type": "home",
                        "number": "212 555-1234"
                },
                {
                        "type": "fax",
                        "number": "646 555-4567"
                }
        ]
}
```

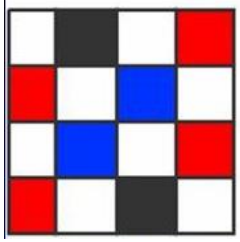
4. Программный код

5. Файлы изображений - каждый пиксель закодирован некоторым образом (RGB, YCbCr)

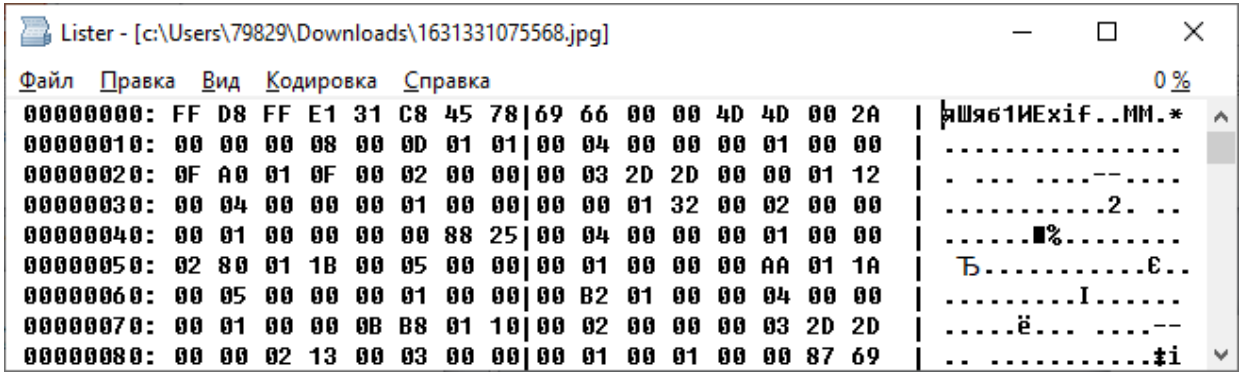




1 0 0 1
0 1 1 0
0 1 0 0
1 0 0 1



00 11 00 01
01 00 10 00
00 10 00 01
01 00 11 00



Типы данных

В машинном обучении принято разделять данные на два принципиально различных типа – количественные (числовые) и категориальные.

Количественные данные состоят из чисел, т.е. из математических объектов, к которым **разумно применять арифметические операции** (допустим, находить среднее значение). Количественными данными являются, например, *цена*, *расстояние*, *площадь*, *артериальное давление*, *рост*, *вес*, *длина*, *размер стипендии*, результат любого измерения, представленный числом.

К **категориальным** относятся те данные, к которым **невозможно** (или бессмысленно) **применять арифметические операции** (допустим, находить среднее значение). Например, невозможно найти среднее *имя студентов* одной группы и совершенно бессмысленно (хотя математически можно) вычислить среднее *номеров их студенческих билетов*.

Количественные данные

Количественные данные могут быть далее разбиты на **дискретные** и **непрерывные** данные.

Дискретные данные подсчитываются, а непрерывные - измеряются.

При броске кубика мы получаем дискретное значение, например 1, 2, 3, 4, 5 или 6. А после подсчета количества студентов в группе получаем 12, 20, 22 и т. п.

Дискретные значения обычно задаются целыми числами. Для дискретных данных бессмысленно брать более мелкую градацию. Так, никогда при броске кости не выпадет число «полтора» (1,5), и нельзя четверть студента направить на олимпиаду.

Примерами могут служить *количество домов* в городе, *количество покупателей* в продуктовом магазине за последний месяц, *количество ваших друзей* во ВКонтакте и так далее.

Непрерывные данные измеряются по шкале, масштаб которой может быть **конечным** или **бесконечным**.

Непрерывные данные можно разбить на более мелкие измерения в зависимости от точности инструмента, которым выполняются измерения. Например, единицами измерения времени могут быть года, месяцы, сутки, часы, минуты, секунды, миллисекунды, микросекунды, наносекунды.

Другими словами, непрерывные данные – данные, величины которых могут принимать какое угодно значение в некотором интервале (рост, вес, длина, время, температура, давление, скорость, плотность).

Т.о., если для того, чтобы найти значения, используется счет, то это дискретные данные. Однако, если используется шкала для измерения значения, тогда это будут непрерывные данные.

Категориальные данные

Традиционно, к категориальному типу относят данные, которые могут быть представлены словами. Название указывает, что этот тип данных определяет группы или категории.

Некоторыми примерами являются названия всех товаров в супермаркете, рейтинги фильмов (хорошие, средние, плохие), страна рождения людей, национальность (русский, татарин, манси), пол (мужской, женский), цвет (красный, синий, зеленый), присутствие на паре (присутствовал, отсутствовал), статус болезни и т.п.

Категориальные данные - это описание или измерение, не имеющее числового значения.

Категориальными являются некоторые данные, записанные цифрами, например, номер паспорта, почтовый индекс (560034), номер автобуса. Их нельзя считать числовыми, т.к. для них бессмысленно выполнение арифметических операций.

Т.к. в моделях машинного обучения данные для обработки должны быть представлены только в числовом виде, возникает необходимость преобразования категориальных данных в числовой тип. На практике применяются разные подходы к такому преобразованию, при этом выбор подхода определяется видом категориальных данных (порядковые, номинальные).

Порядковые данные

Этот вид данных подразумевает порядок, присутствующий в категориях. Например, если рассматриваются рейтинги фильмов с хорошим, средним и плохим как разные категории, то «хороший» имеет более высокий рейтинг, чем «средний», который выше, чем «плохой». При этом существует фиксированное конечное число категорий/групп. Примерами могут быть уровень образования, оценки учащихся («хорошо», «удовлетворительно»), этапы болезни, производительность сотрудников и так далее.

Обычно порядковому значению числовой аналог присваивается по правилу: самому «маленькому» – 0 (ноль), следующему – 1 (один), потом 2 и т.д.

Номинальные данные

Этот вид данных не имеет числового значения или порядка. Это просто **название категории**. Тут невозможно однозначно определить, какая категория больше (нет упорядоченности категорий по какому-нибудь правилу). Например, тип жилья (дом, квартира, вилла), материал стен (кирпич, дерево, бетон, шлакоблок) религиозные типы (мусульмане, индуисты, христиане, буддисты, иудеи), цвет (фиолетовый, розовый, салатный).

Общее число категорий и в этом типе данных обычно конечно.

Для номинальных данных кодирование (перевод в числовой тип) выполняется с помощью довольно сложных процедур, например, «One-Hot Encoder».

Бинарные данные

Категориальная переменная, имеющая только два возможных значения, называется бинарной (дихотомической) переменной.

Например, возраст: («до 60» - «60 и старше»), пол («мужской» - «женский»).

Обычно бинарным значениям присваивается тривиальный числовой аналог: одному из них 0 (ноль), а другому 1 (один). Часто это наличие или отсутствие некоторого признака, например, присутствие на паре (0 или 1), наличие зачёта, облачность (0 или 1) и т.п.

Пример данных

Рассмотрим известный набор данных о пассажирах теплохода «Титаник». Данные собраны в таблицу (показаны данные последних 6-ти пассажиров):

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.125	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

Рассмотрим тип данных представленных 12-ти признаков:

Количественные дискретные:

- **SibSp** — количество братьев, сестер, сводных братьев, сводных сестер, супругов на борту;
- **Parch** — количество родителей, детей (в том числе приемных) на борту.

Количественные непрерывные:

- **Age** — возраст;
- **Fare** — плата за проезд.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.125	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

Категориальные номинальные (уникальные, без повторения)

- **Name** — имя;
- **Ticket** — номер билета;
- **Cabin** — номер каюты.

Категориальные номинальные

- **Embarked** — порт посадки (C — Шербур; Q — Квинстаун; S — Саутгемптон).

Категориальные порядковые

- **Pclass** — класс пассажира (1 — высший, 2 — средний, 3 — низший).

Категориальные бинарные

- **Survived** — выжил ли данный пассажир (0 для умерших, 1 для выживших);
- **Sex** — пол.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.125	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

Особо стоит отметить признак **PassengerId** — номер пассажира в списке. Такой признак скорее не характеристика пассажира, а характеристика строчки в списке пассажиров. В терминах таблиц данных такой атрибут называется индекс (порядковый номер). Обратите внимание, что кроме **PassengerId** каждой строке присвоен автоматический индекс. Он записан в самом левом безымянном столбце. Автоматический индекс часто обозначают **Id**. Он начинается с нуля, поэтому последний 891 пассажир списка имеет **Id** = 890.

Важная особенность индекса – он уникален (разный у любых двух объектов списка). В данном наборе также уникальны признаки **Name** и **Ticket**. Они также могут играть роль индексов.

Некоторые данные в таблице не определены (неизвестны, отсутствуют). Они представлены специальным словом NaN (Not a Number).