

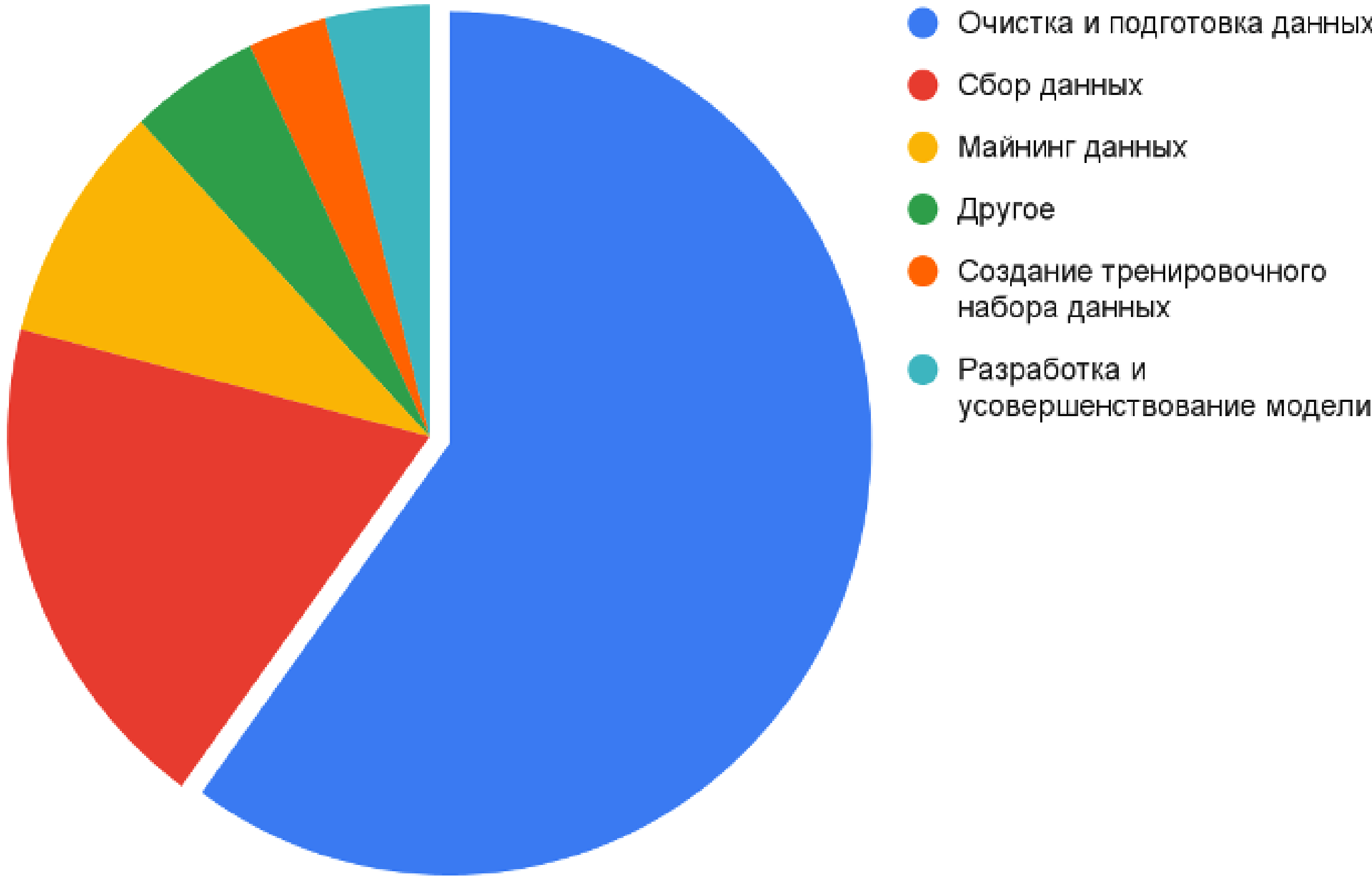
Занятие 1

Разведочный анализ данных (Exploratory Data Analysis)

План занятия

- Про EDA и почему он так важен
- Feature Engineering, основные шаги
- Частые ошибки и проблемы при анализе данных

Как проводит время среднестатистический DS



Как проводит время среднестатистический DS

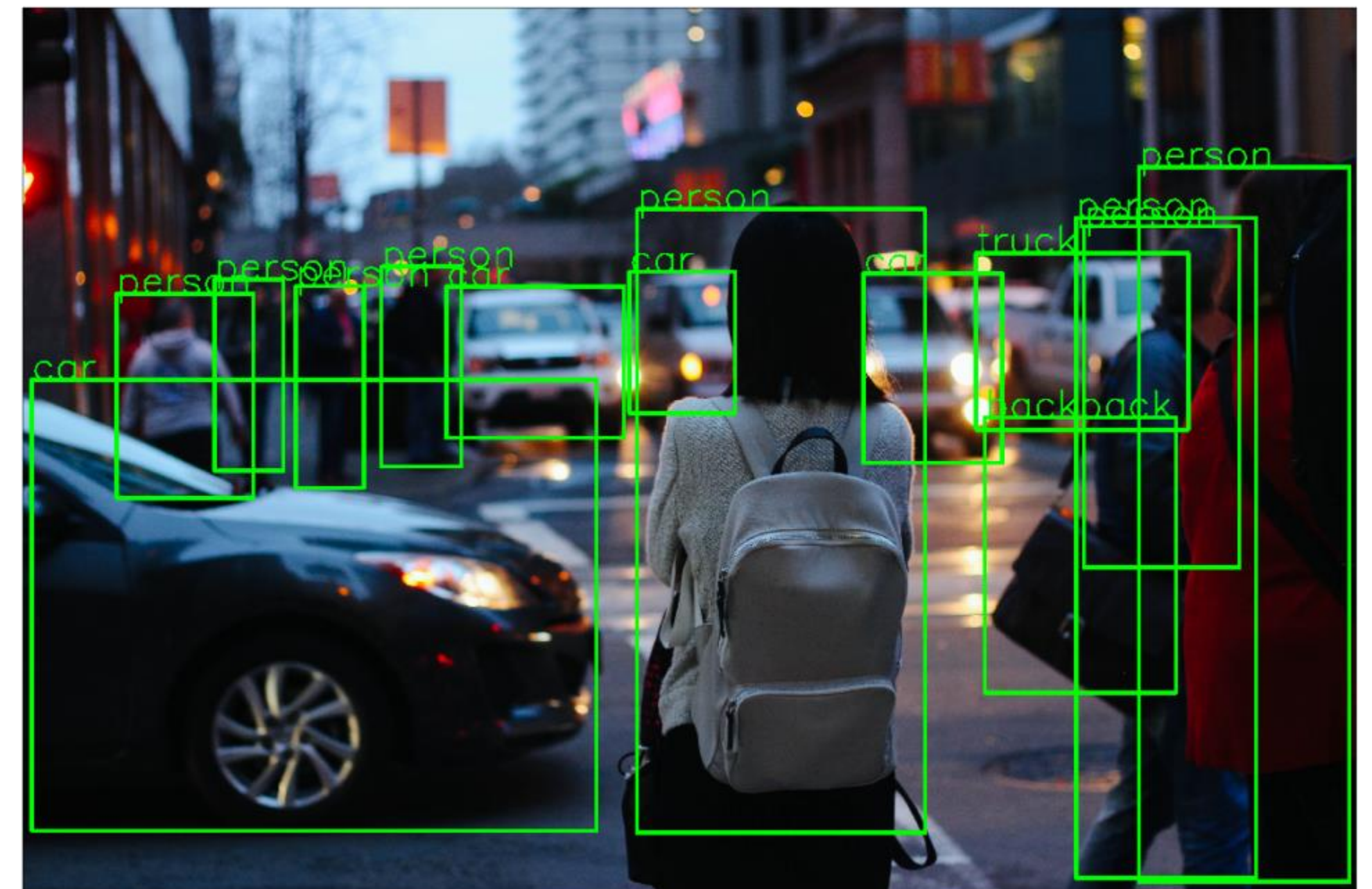
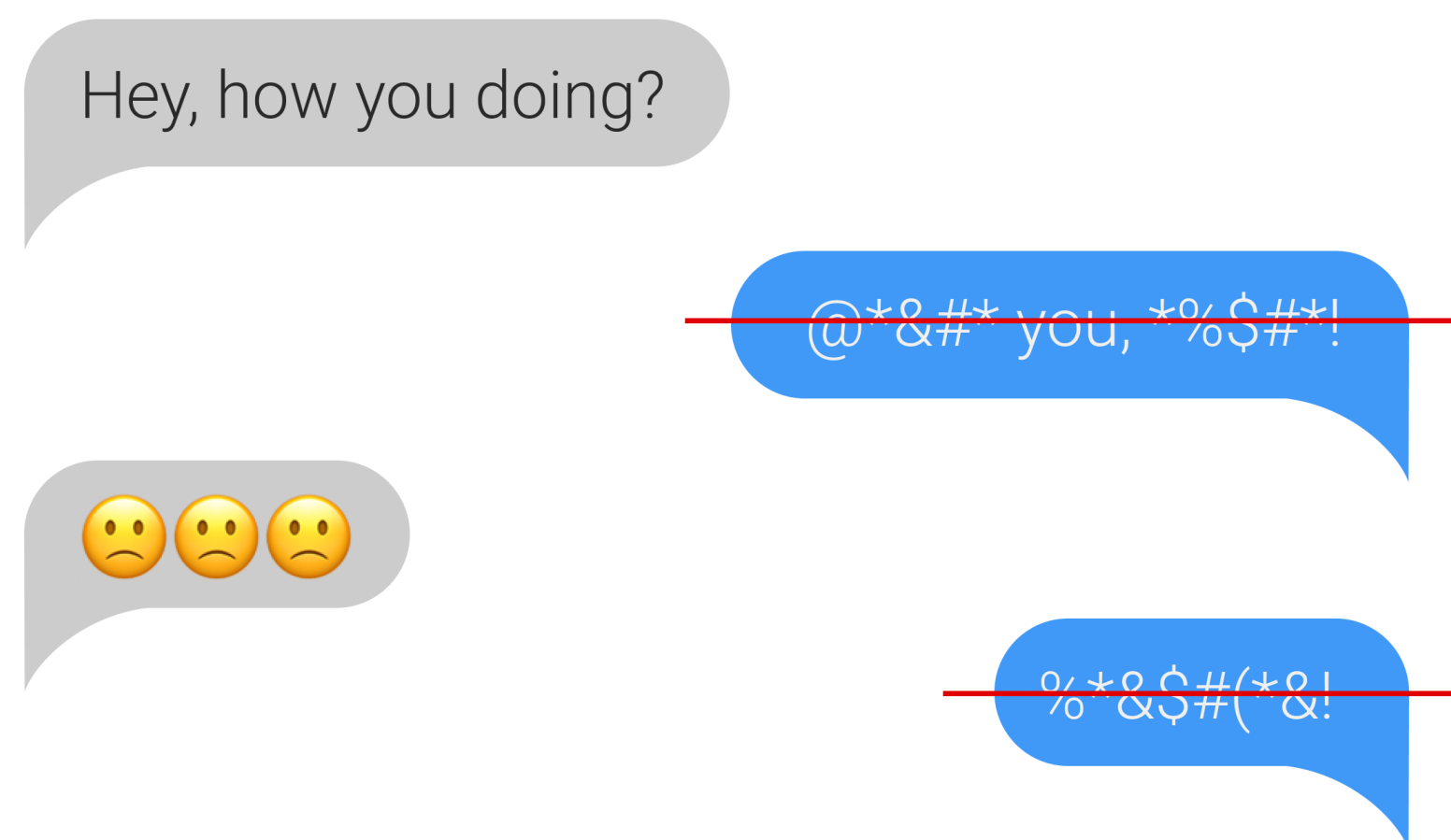


Зачем машинное обучение?

Обычно ML используют для:

- Замены человека при решении задач (автоматизация).
Пример: модерация постов в интернете (токсичный/нет),
распознавание изображений.
- Поиска закономерностей в данных, которых человек не находит

Пример: выделение наиболее важных симптомов при болезни



Зачем машинное обучение?

Обычно ML используют для:

- Замены человека при решении задач (автоматизация).
Пример: модерация постов в интернете (токсичный/нет),
распознавание изображений.
- Поиска закономерностей в данных, которых человек не находит
Пример: выделение наиболее важных симптомов при болезни

Задача ML требует:

- Данные (признаки + целевая переменная)
- Алгоритм
- Критерий качества алгоритма

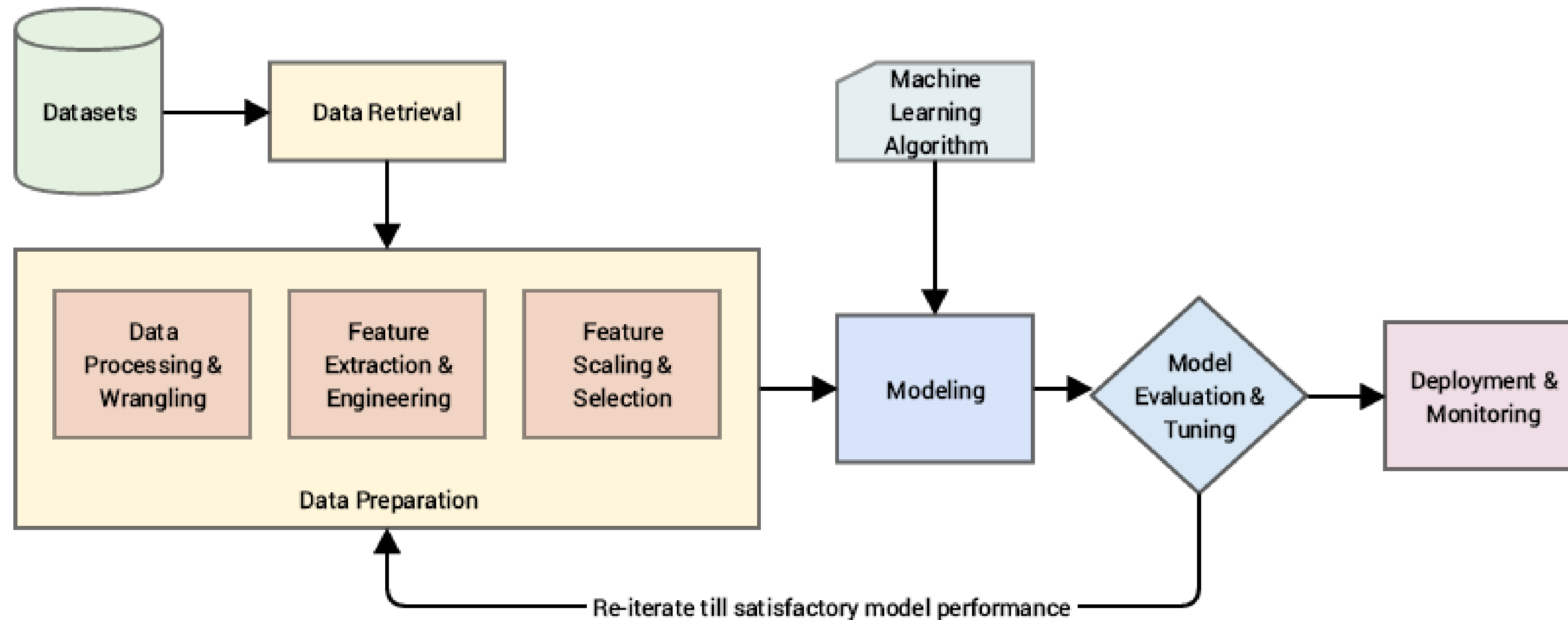
Пример данных

Case	Attributes				Decision
	Length	Height	Width	Weight	Quality
1	4.7	1.8	1.7	1.7	high
2	4.5	1.4	1.8	0.9	high
3	4.7	1.8	1.9	1.3	high
4	4.5	1.8	1.7	1.3	medium
5	4.3	1.6	1.9	1.7	medium
6	4.3	1.4	1.7	0.9	low
7	4.5	1.6	1.9	0.9	very-low
8	4.5	1.4	1.8	1.3	very-low

Признаки (атрибуты, фичи, features)

Целевая переменная (Target)

Работа с признаками как часть ML



Правильно обработанные данные – это основа успешного запуска модели машинного обучения!!!

Основные определения

- **Exploratory Data Analysis (EDA, разведочный анализ данных)** - анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей.
- **Feature Engineering (конструирование признаков)** - трансформация существующих признаков или создание на их основе новых с целью улучшения работы алгоритмов машинного обучения.

Состоит из 3 основных вещей:

1. **Feature Extraction (выделение признаков)** - снижение размерности, при котором набор исходных переменных сокращается до более управляемых групп (признаков) для дальнейшей обработки, оставаясь при этом достаточным набором для точного и полного описания исходного набора данных.
2. **Feature Transformation (трансформация признаков)** - процедуры предобработки признаков, например: нормализация, стандартизация, дискретизация, и тд.
3. **Feature Selection (отбор признаков)** - отбор переменных (фичей) для моделирования

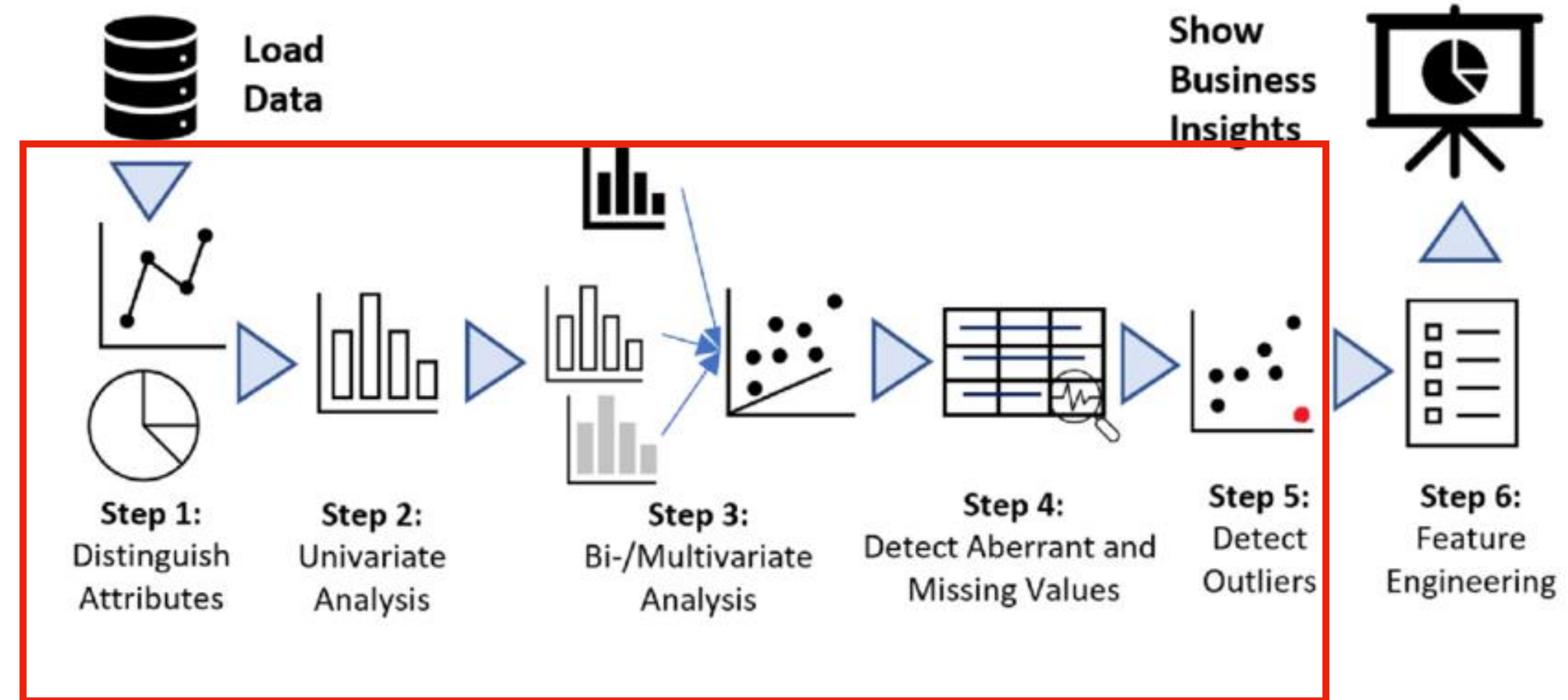
Exploratory Data Analysis (он же EDA)



Для EDA, в основном, используется визуализация.

Основные **цели** анализа:

- Выявление основных структур.
- Обнаружение отклонений и аномалий.
- Проверка основных гипотез (предположений).
- Разработка начальных моделей.



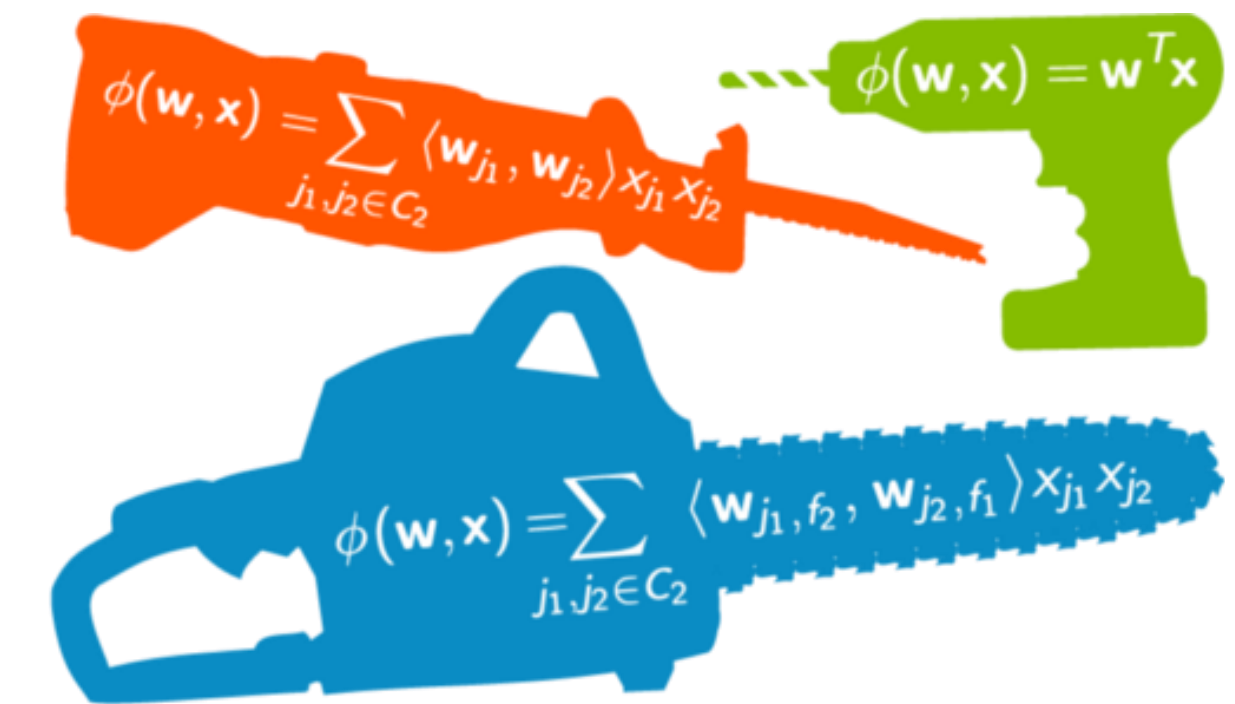
**Основные шаги в
EDA**

Feature Engineering

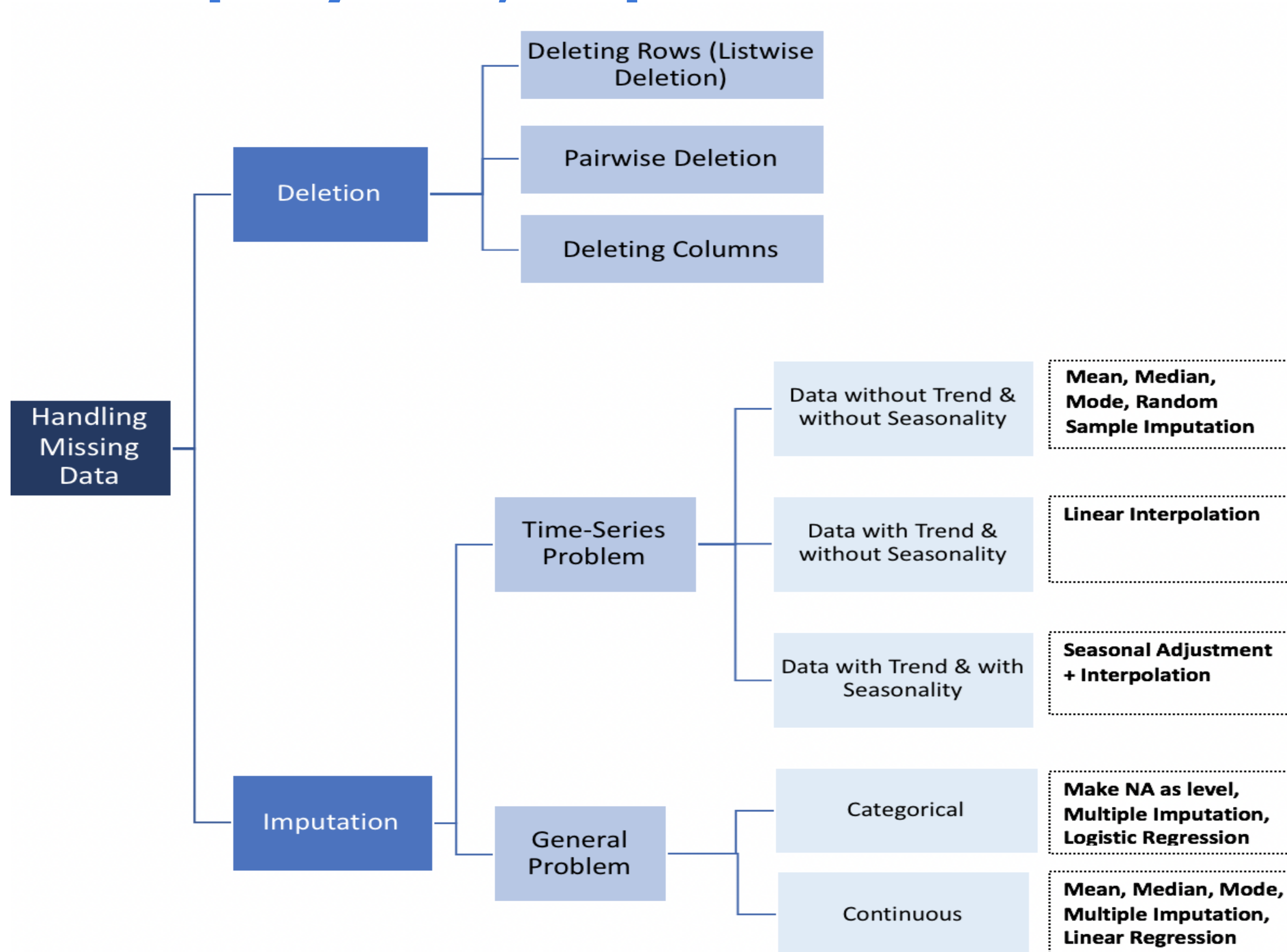
Процесс конструирования признаков:

1. Проверка пригодности признаков на уровне понимания бизнес-процесса (Метод мозгового штурма)
2. Решение, какие признаки генерировать;
3. Генерация признаков;
4. Проверка, какие признаки работают с вашей моделью;
5. Улучшение признаков, если требуется;
6. Возврат к методу мозгового штурма, пока работа не будет завершена.

**Самый долгий
процесс, может
повторяться
множество раз**



Заполнение пропусков, Imputation



Заполнение пропусков, Imputation

Заполнение средним значением

- **Заполнение пропуска средним значением** (Mean Substitution) (другие варианты: заполнение нулем, медианой и тому подобные) — название метода говорит само за себя.

Всем вариантам данного метода свойственны одни и те же недостатки

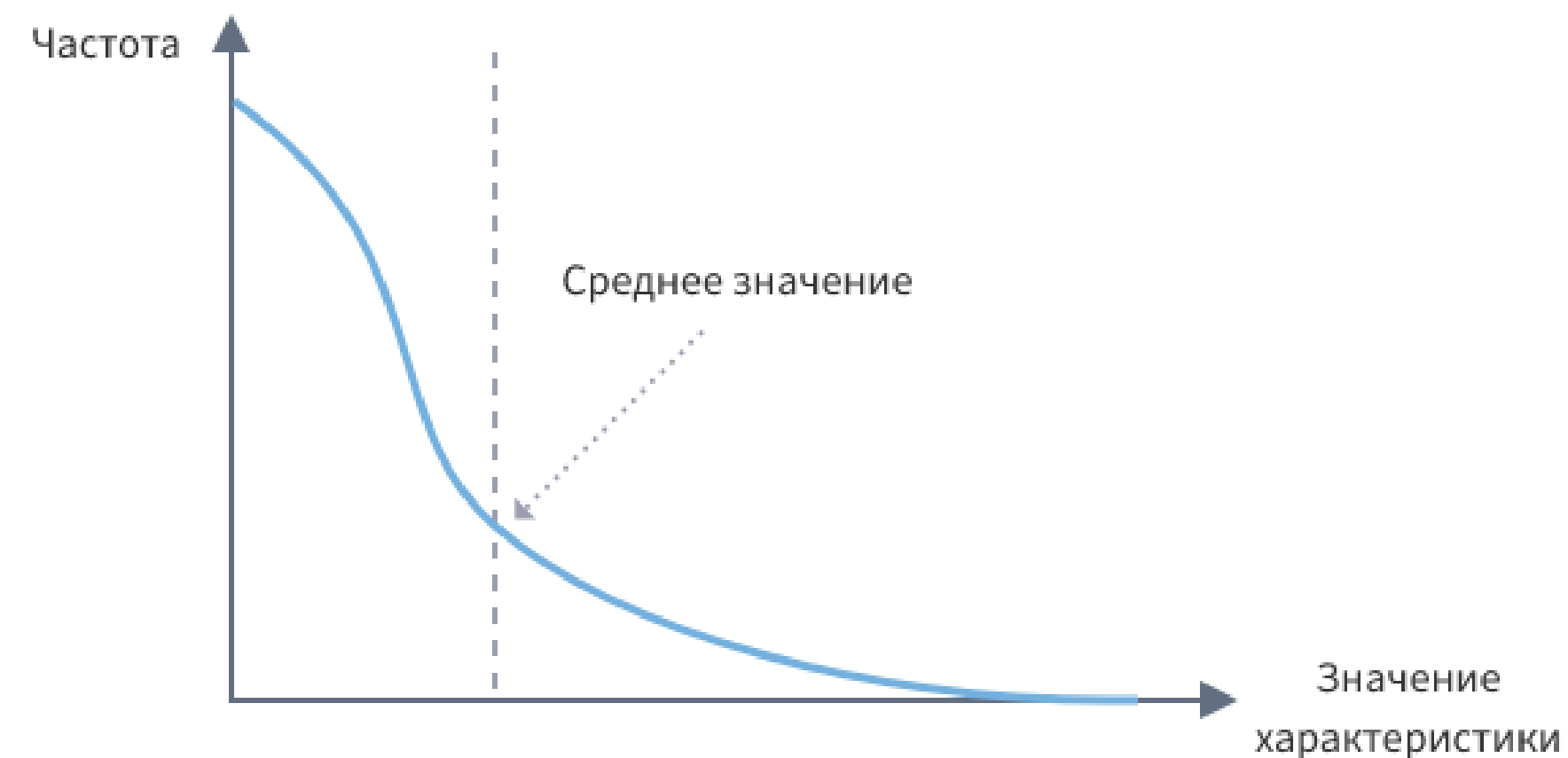


Рисунок 1а — Распределение значений непрерывной характеристики **до** заполнения пропусков

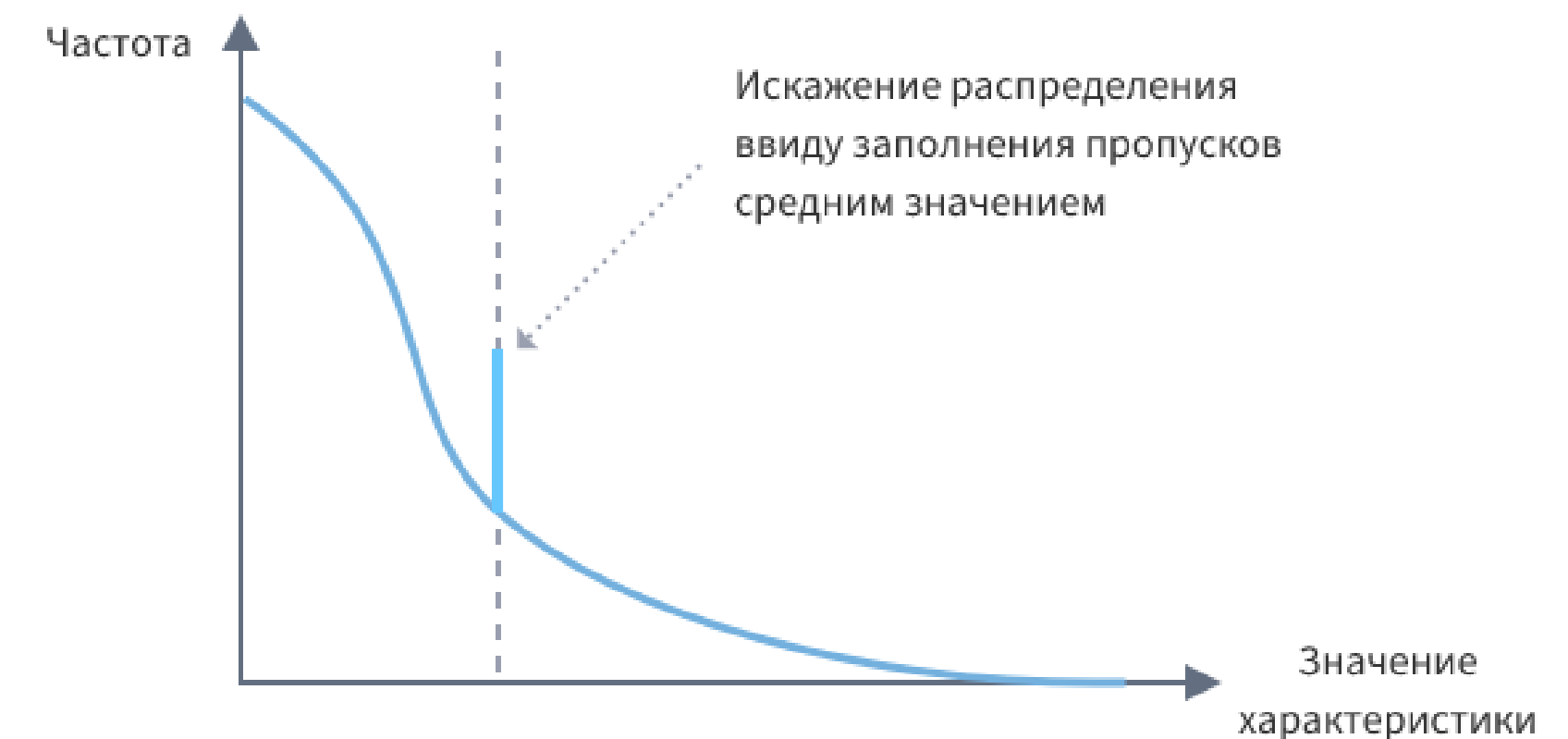


Рисунок 1б — Распределение значений непрерывной характеристики **после** заполнения пропусков

Заполнение пропусков, Imputation

Восстановление на основе регрессионных моделей

- Данный метод заключается в том, что пропущенные значения заполняются с помощью модели линейной регрессии, построенной на известных значениях набора данных.

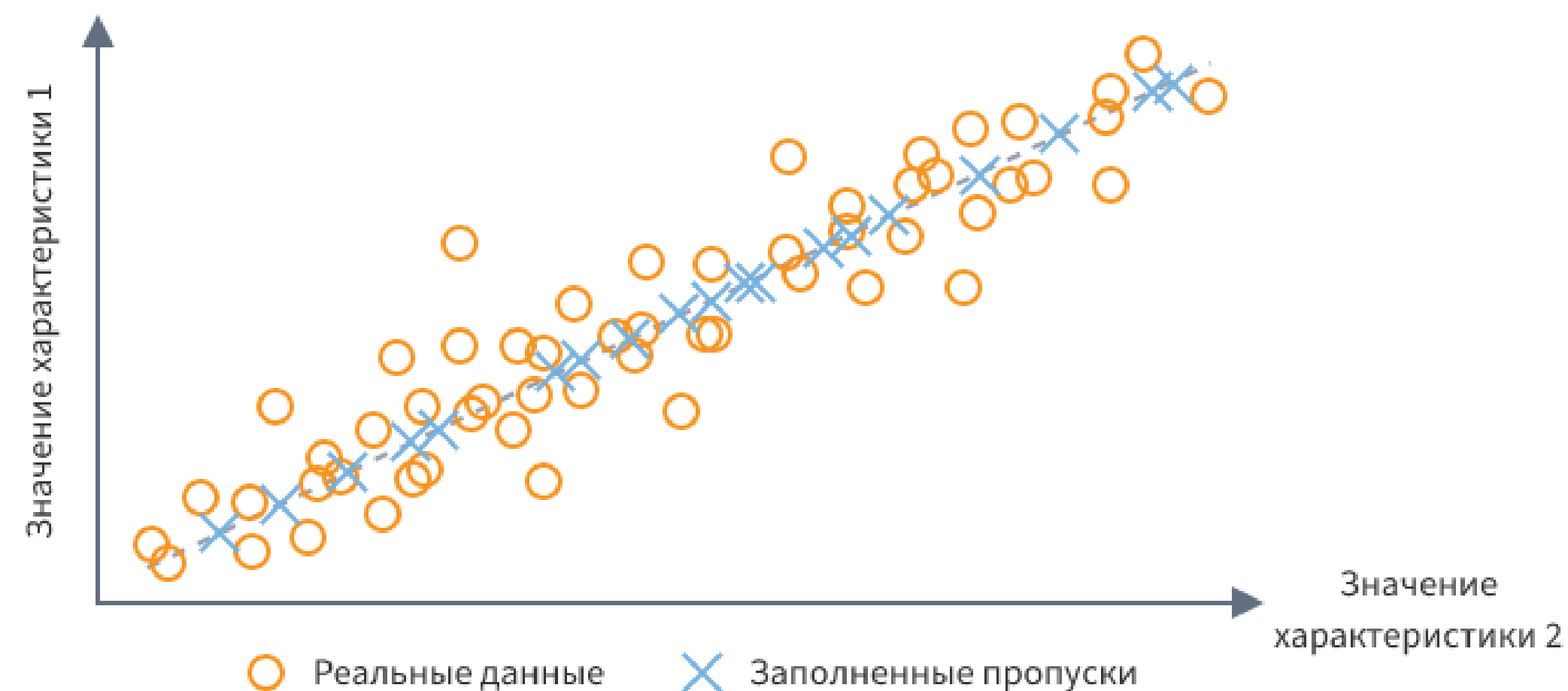


Рисунок 2а — Заполнение пропусков на основе **линейной регрессии**

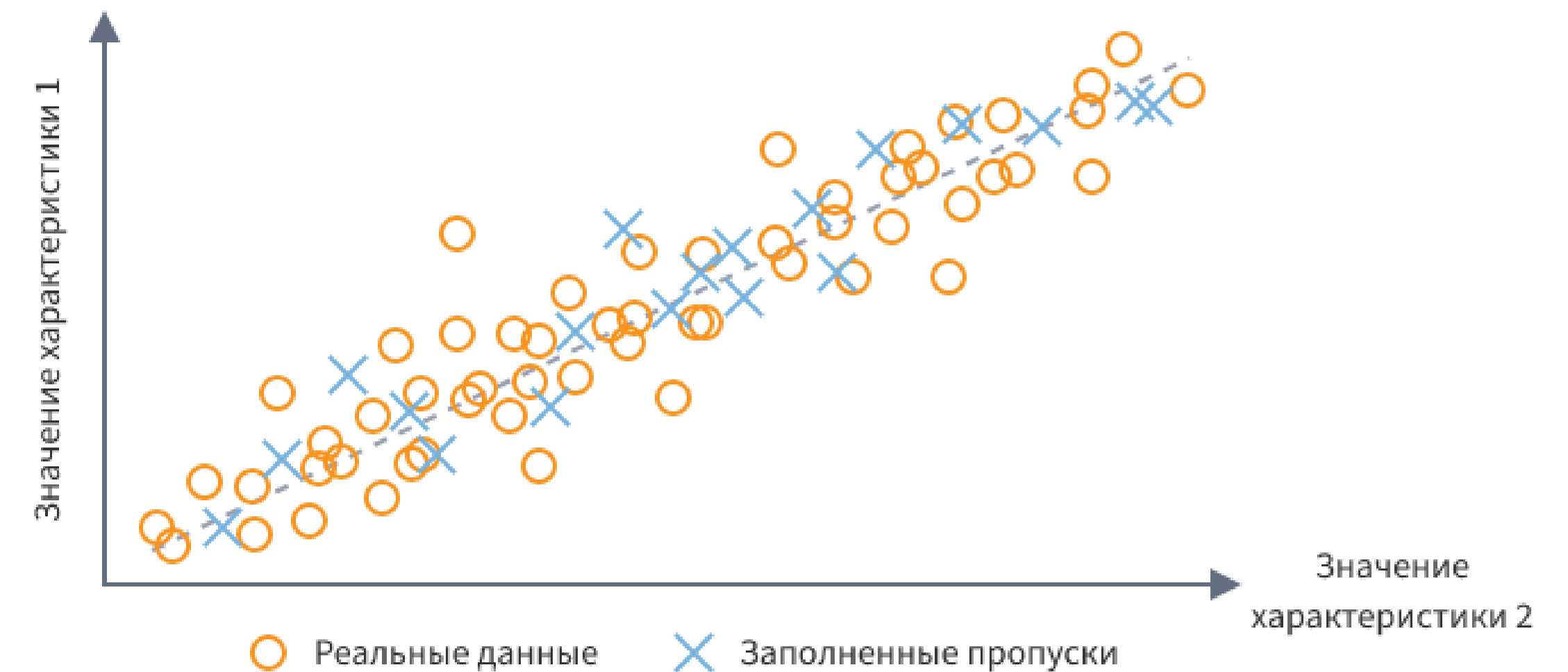


Рисунок 2б — Заполнение пропусков на основе **стохастической линейной регрессии**

Pandas Profiling

Пример отчета для одного из признаков:

Parent Property Id

Categorical

HIGH CARDINALITY

Distinct	102
Distinct (%)	0.9%
Missing	0
Missing (%)	0.0%
Memory size	1.0 MiB

Pandas Profiling

Пример отчета для одного из признаков:

Parent Property Id

Categorical

HIGH CARDINALITY

Distinct	102
Distinct (%)	0.9%
Missing	0
Missing (%)	0.0%
Memory size	1.0 MiB

Cardinality характеризует уникальность данных.
Высокая кардинальность - уникальные данные, низкая кардинальность - повторяющиеся данные.

Pandas Profiling

Пример отчета для одного из признаков:

Parent Property Id

Categorical

HIGH CARDINALITY

Distinct	102
Distinct (%)	0.9%
Missing	0
Missing (%)	0.0%
Memory size	1.0 MiB

Различные значения (отличные от других в выборке)

Пропуски в данных

Cardinality характеризует уникальность данных.
Высокая кардинальность - уникальные данные, низкая кардинальность - повторяющиеся данные.

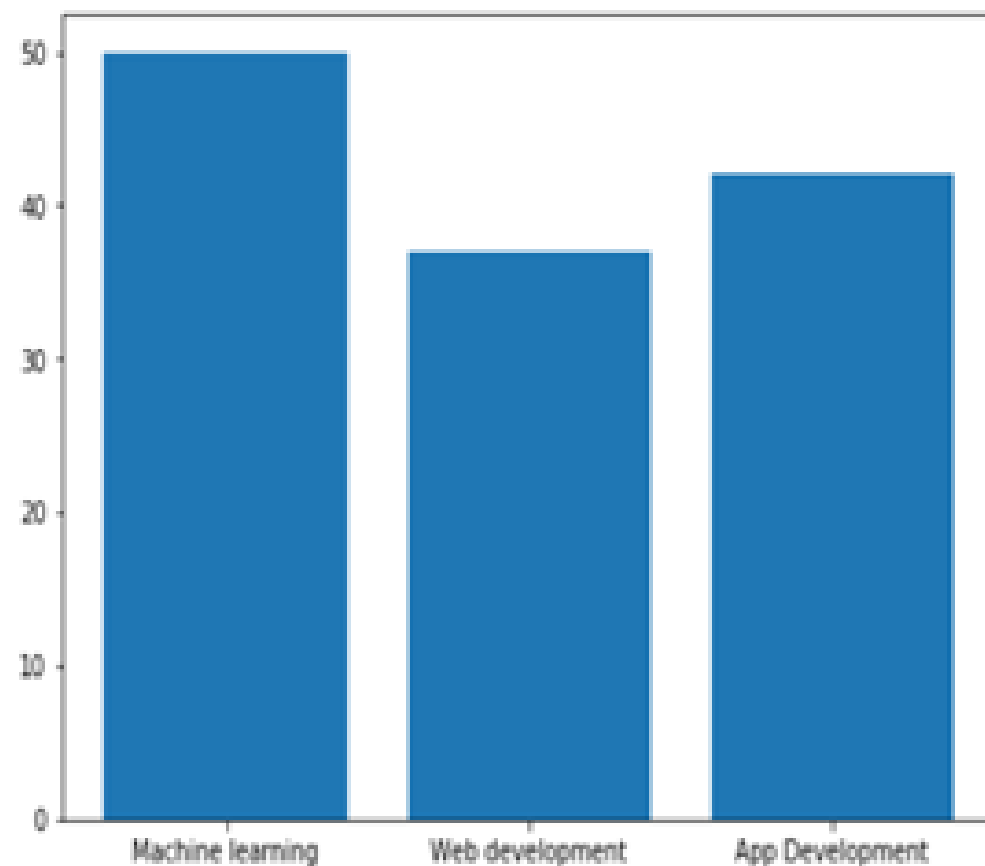
Univariate analysis

Анализ одной переменной

Чаще всего используются: PDF (Probability density function), CDF (Cumulative distribution function), Boxplot, Violin

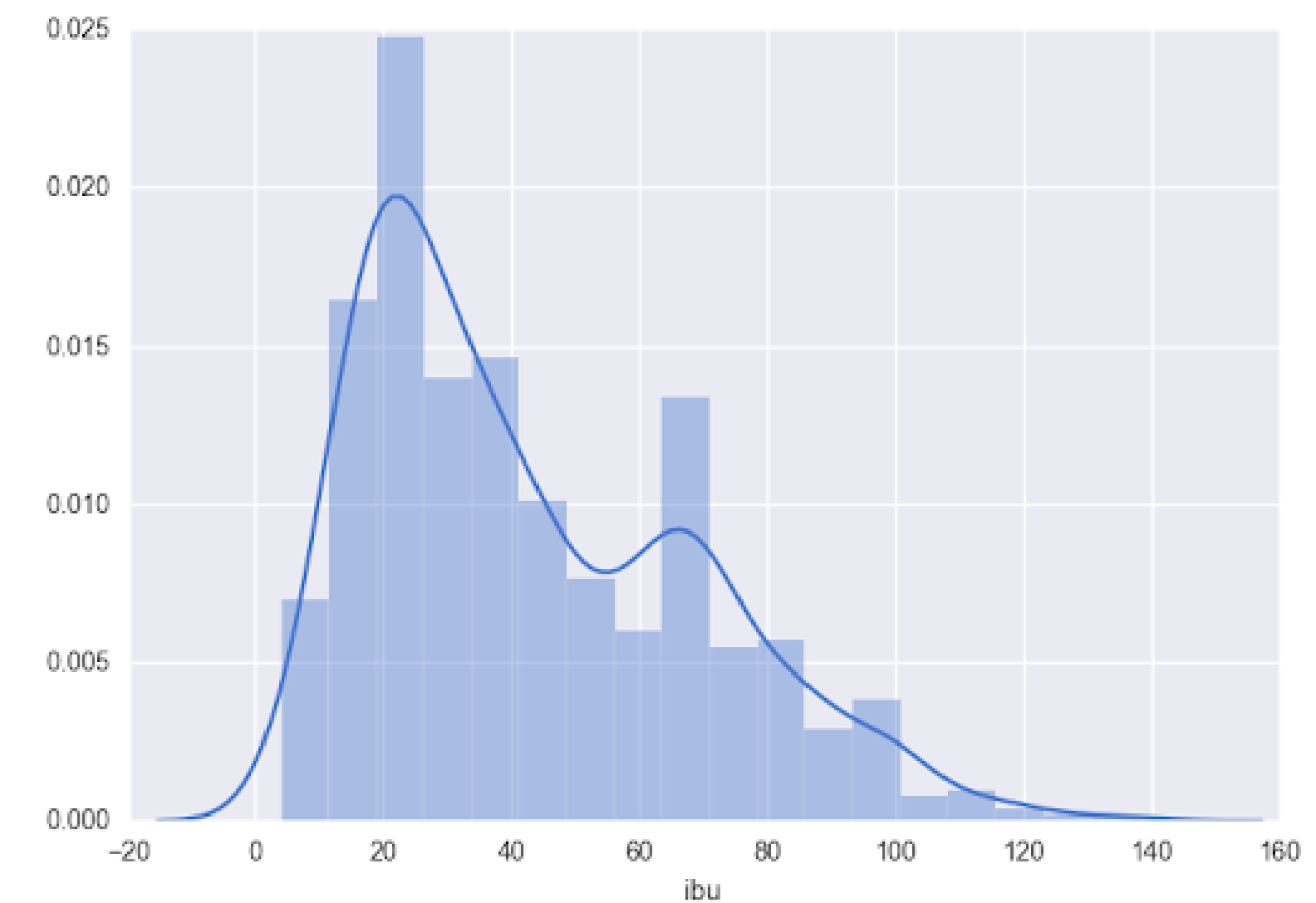
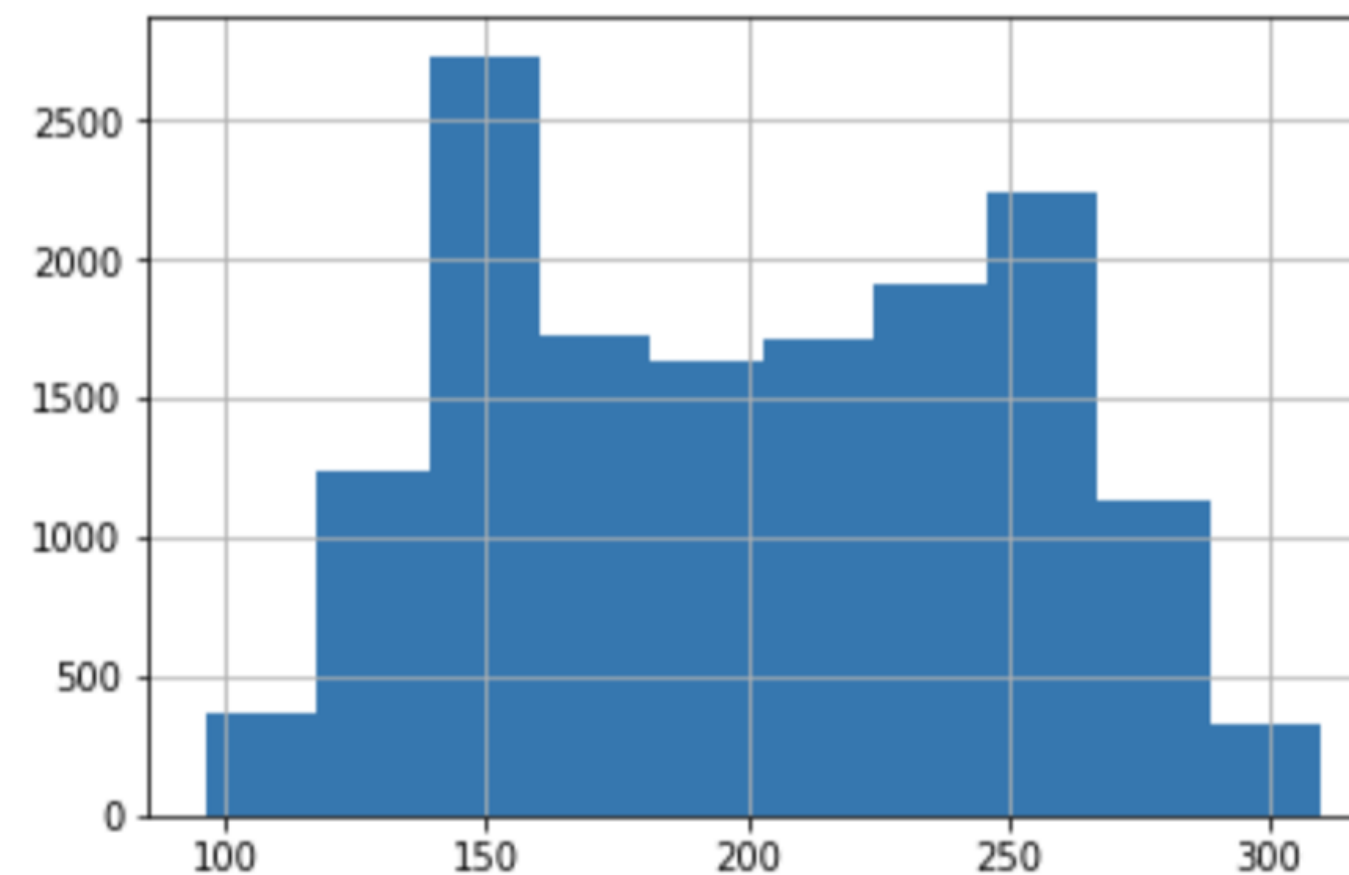
plots, Distribution plots

```
In [5]: import matplotlib.pyplot as plt
fig=plt.figure()
ax=fig.add_axes([0,0,1,1])
courses=['Machine learning','Web development','App Development']
students_enrolled=[50,37,42]
ax.bar(courses,students_enrolled)
plt.show()
```



```
df.average_monthly_hours.hist()
plt.plot()
```

```
[ ]
```



Распределение, основные характеристики

Определение. Законом распределения (или просто распределением) случайной величины $\xi(w)$ называют правило, позволяющее находить вероятность попадания значений $\xi(w)$ в любой заданный числовой промежуток.

Определение. Случайная величина — в теории вероятностей, величина, принимающая в зависимости от случая те или иные значения с определёнными вероятностями. Примером случайной величины может служить число, выпавшее на игральной кости или расстояние от точки падения снаряда до цели.

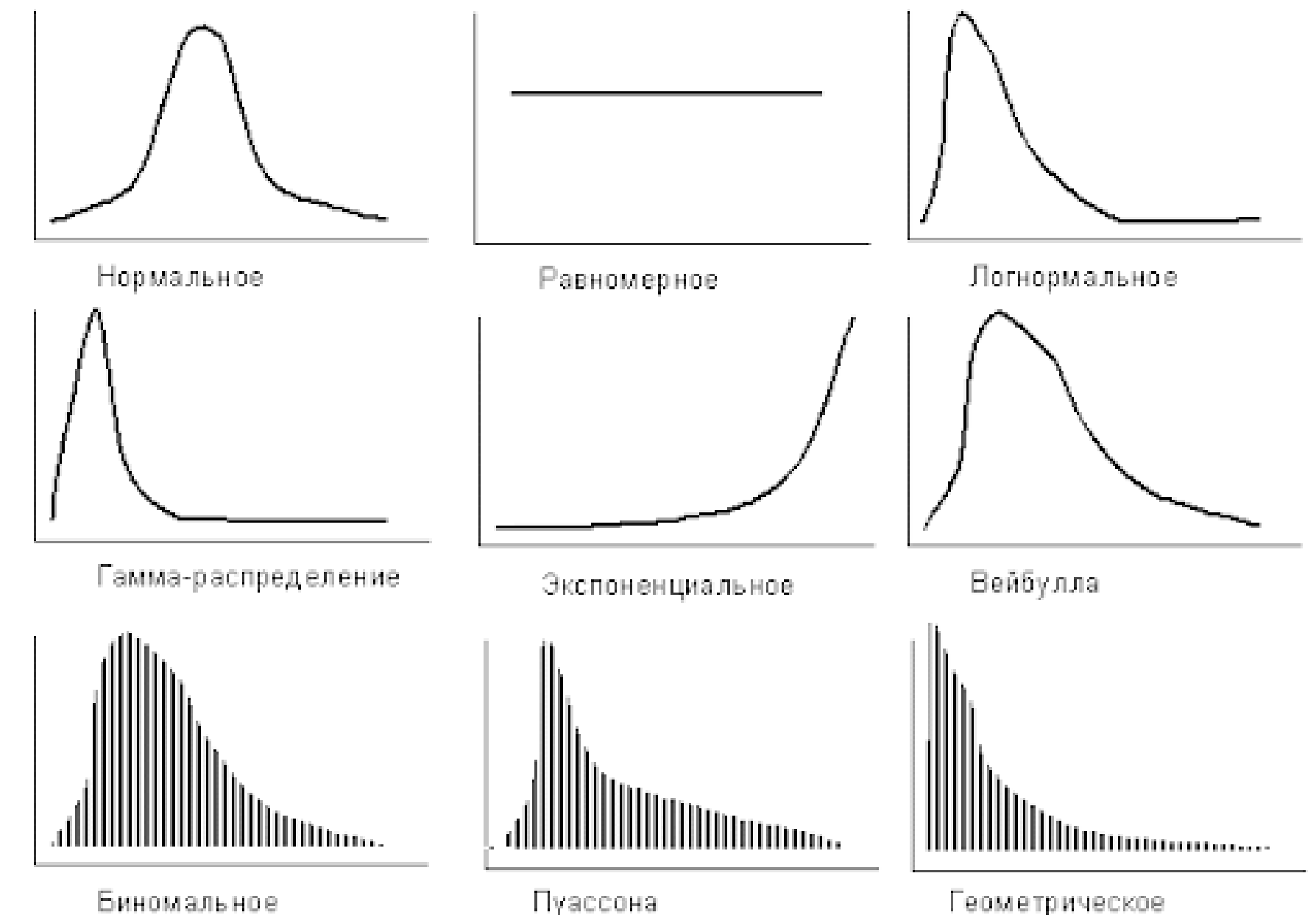
Распределения

непрерывная
случайная величина

- Нормальный закон распределения (закон Гаусса)
- Логарифмически нормальное распределение
- Гамма-распределение
- Экспоненциальный закон распределения

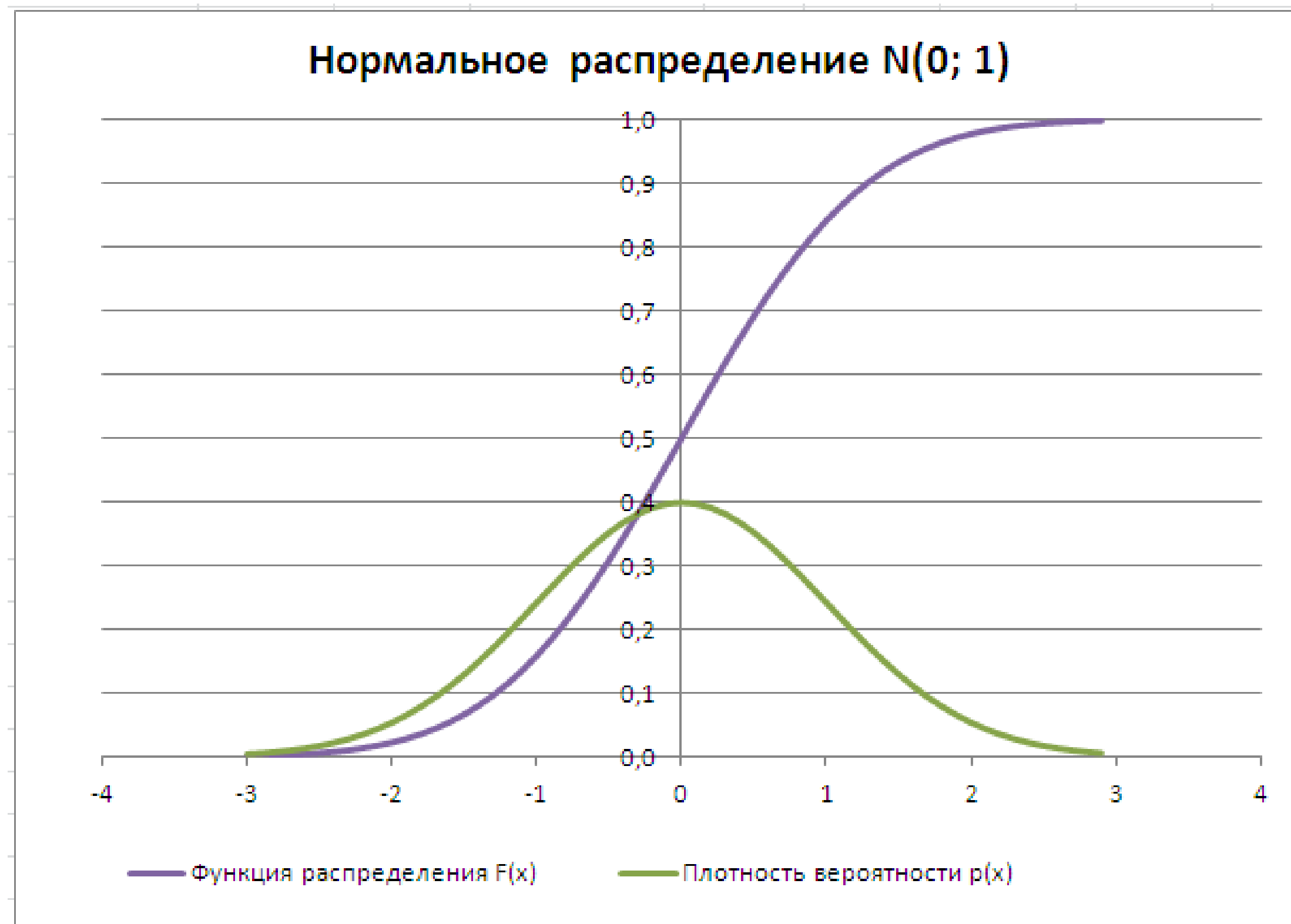
дискретная случайная
величина

- Распределение Пуассона
- Геометрическое распределение
- Биномиальное распределение
- Бернулли



Функция распределения

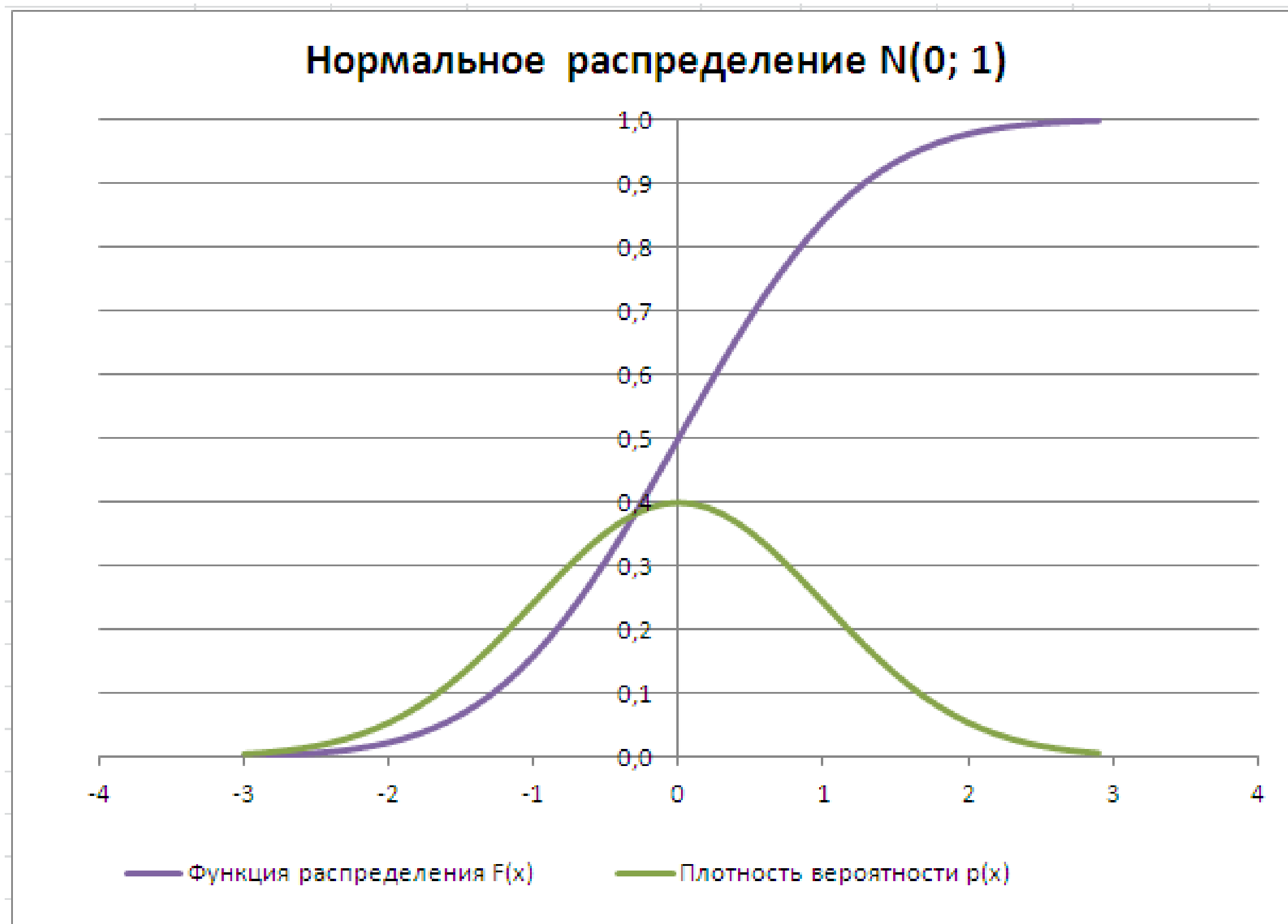
Функцией распределения случайной величины X называется функция $F(x)$, выражающая для каждого x вероятность того, что случайная величина X примет значение, меньшее x



Плотность распределения

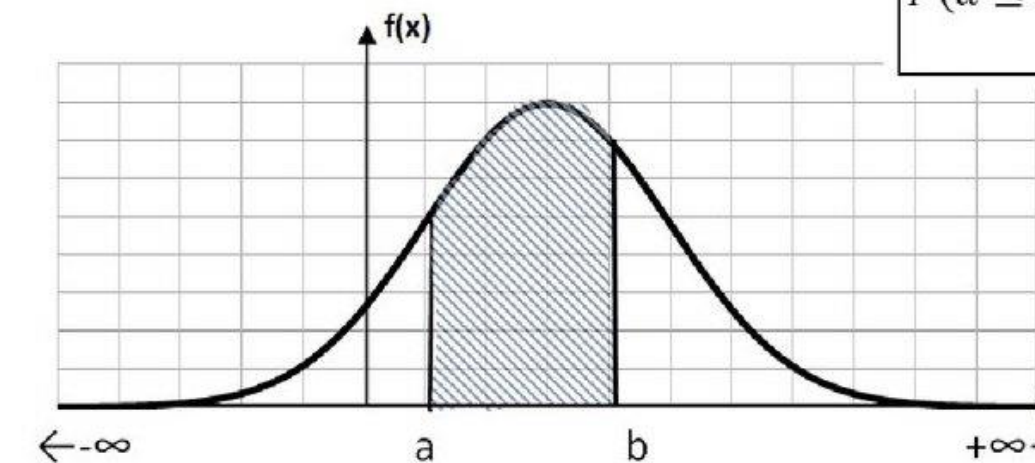
Плотность вероятности (плотностью распределения или просто плотностью) $p(x)$ непрерывной случайной величины X называется производная ее функции распределения. Один из способов задания распределения случайной величины.

График плотности вероятности $p(x)$ называется *кривой распределения*



Вероятность попадания случайной величины в интервал $[a, b]$:

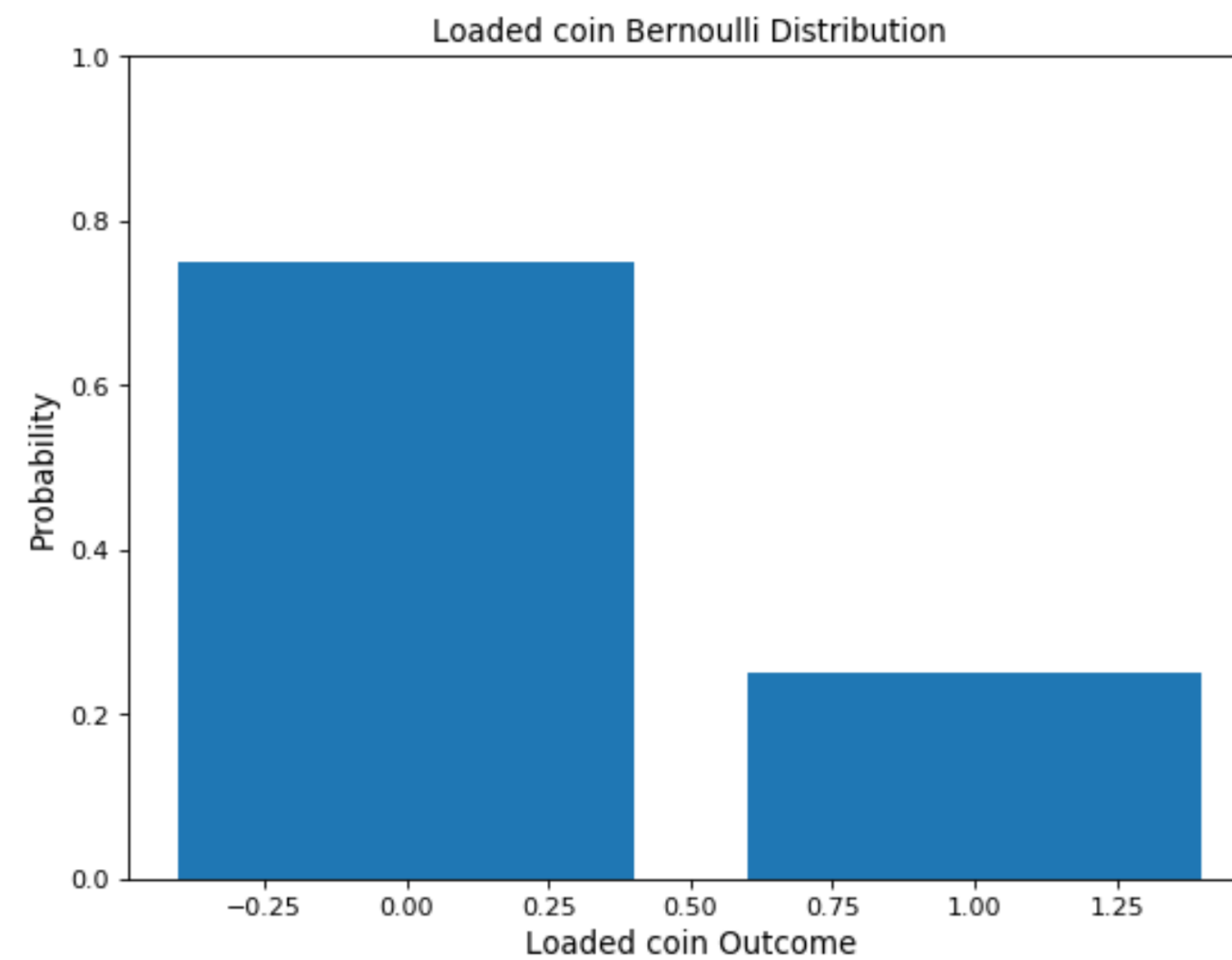
$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$$



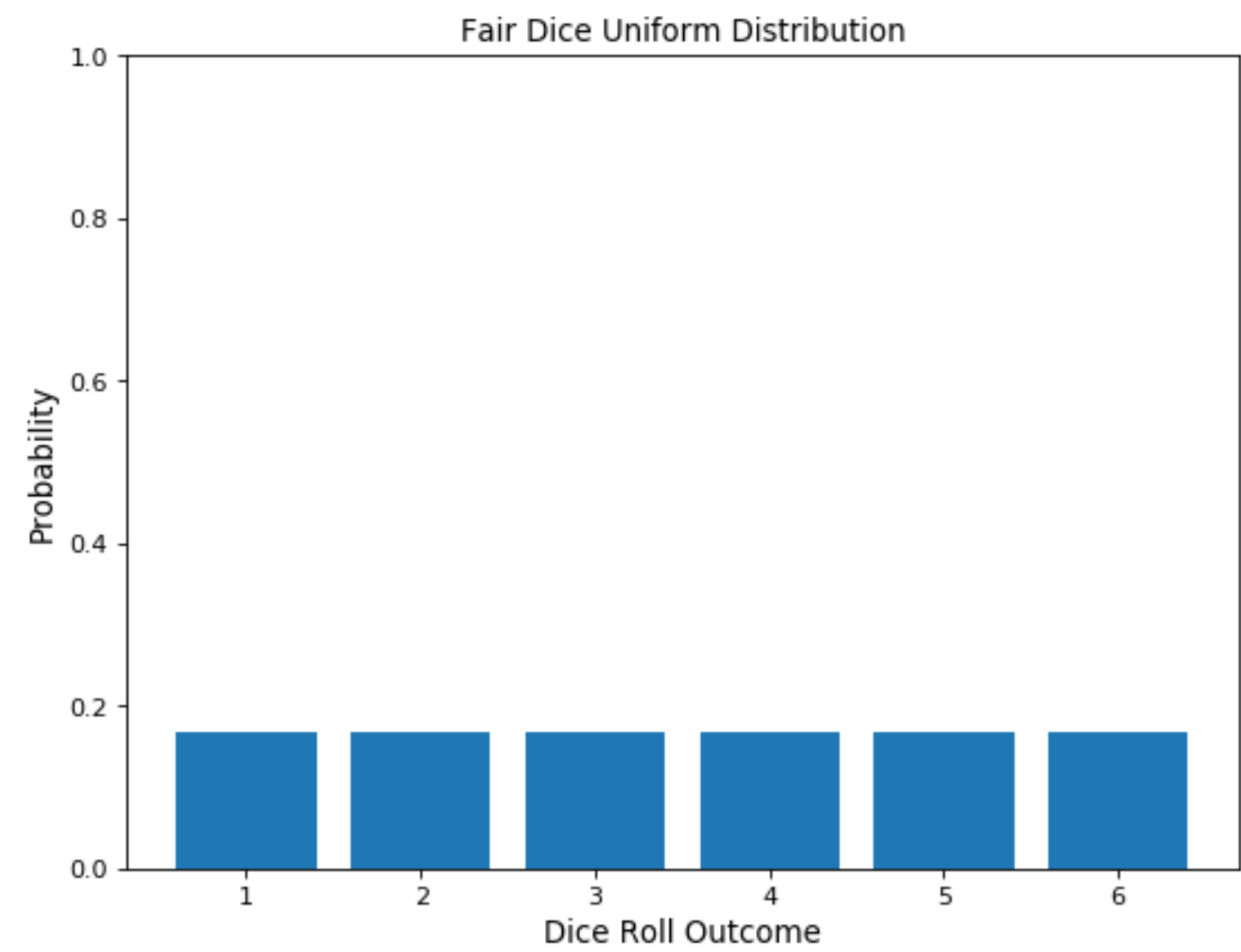
$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

Бернулли распределение

```
probs = np.array([0.75, 0.25])  
face = [0, 1]  
plt.bar(face, probs)  
plt.title('Loaded coin Bernoulli Distribution', fontsize=12)  
plt.ylabel('Probability', fontsize=12)  
plt.xlabel('Loaded coin Outcome', fontsize=12)  
axes = plt.gca()  
axes.set_ylim([0,1])
```

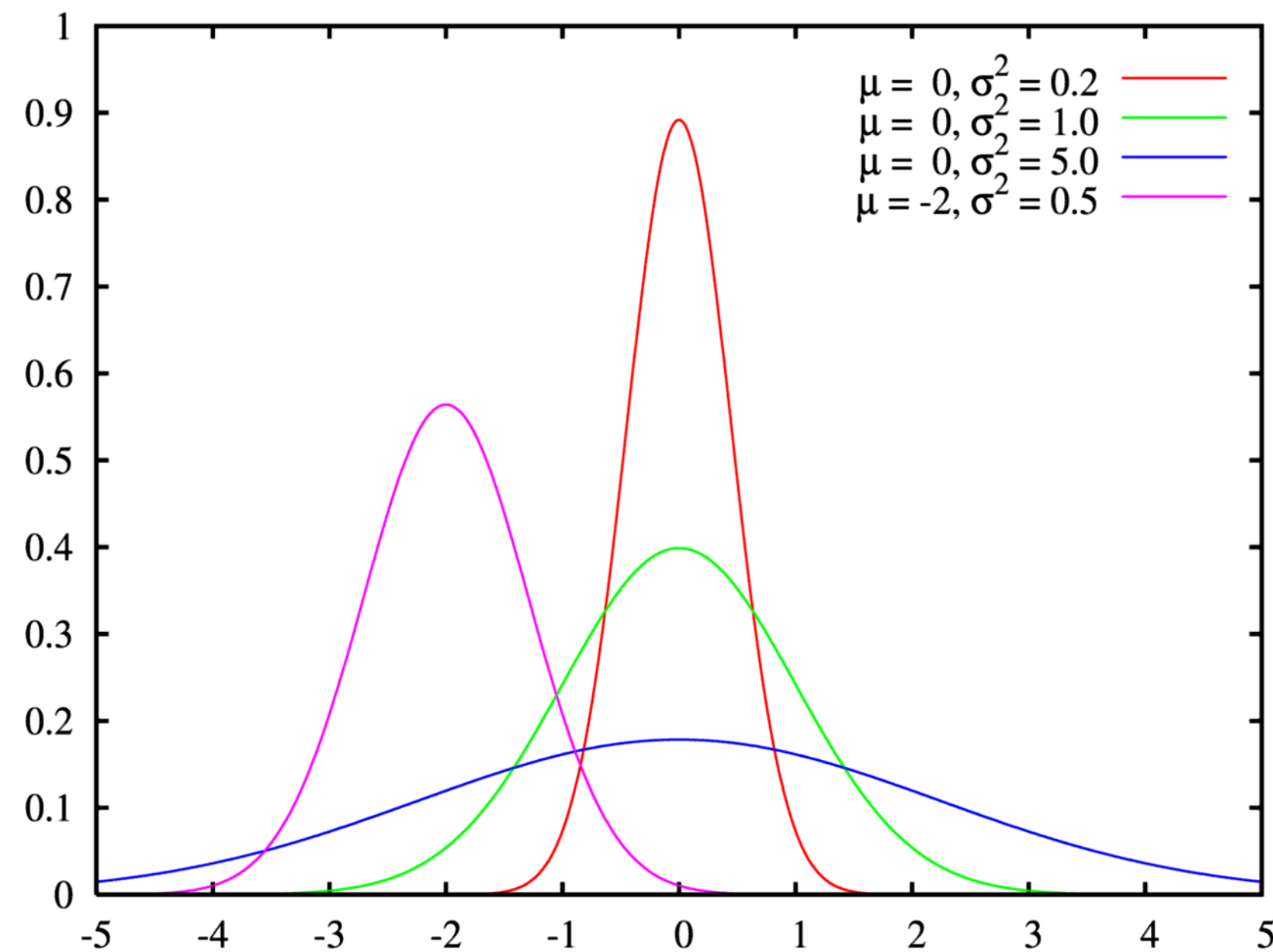


Равномерное распределение



Нормальное (Гауссово) распределение

Физическая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех.



Зеленой линией обозначено стандартное нормальное распределение

Нормальное распределение может быть получено с использованием следующей формулы:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = Mean

σ = Standard Deviation

$\pi \approx 3.14159 \dots$

$e \approx 2.71828 \dots$

Алгоритмы ML, которым необходимо нормальное распределение признаков:

- Наивный Байесовский классификатор
- Модели на основе МНК
- Квадратичный дискриминантный анализ
- Линейный дискриминантный анализ

Основные определения для случайной величины

Математическим ожиданием случайной величины называют сумму произведений её значений на те вероятности, с которыми она эти значения принимает. На практике при анализе выборок математическое ожидание, как правило, не известно и вместо него используют среднее арифметическое.

Дисперсия случайной величины – это один из основных показателей в статистике. Он отражает меру разброса данных вокруг средней арифметической.

Если из дисперсии извлечь квадратный корень, получится **среднеквадратичное (стандартное) отклонение** (сокращенно **СКО**). Встречается название **среднее квадратичное отклонение** и **сигма** (от названия греческой буквы).

Нормальное распределение может быть получено с использованием следующей формулы:

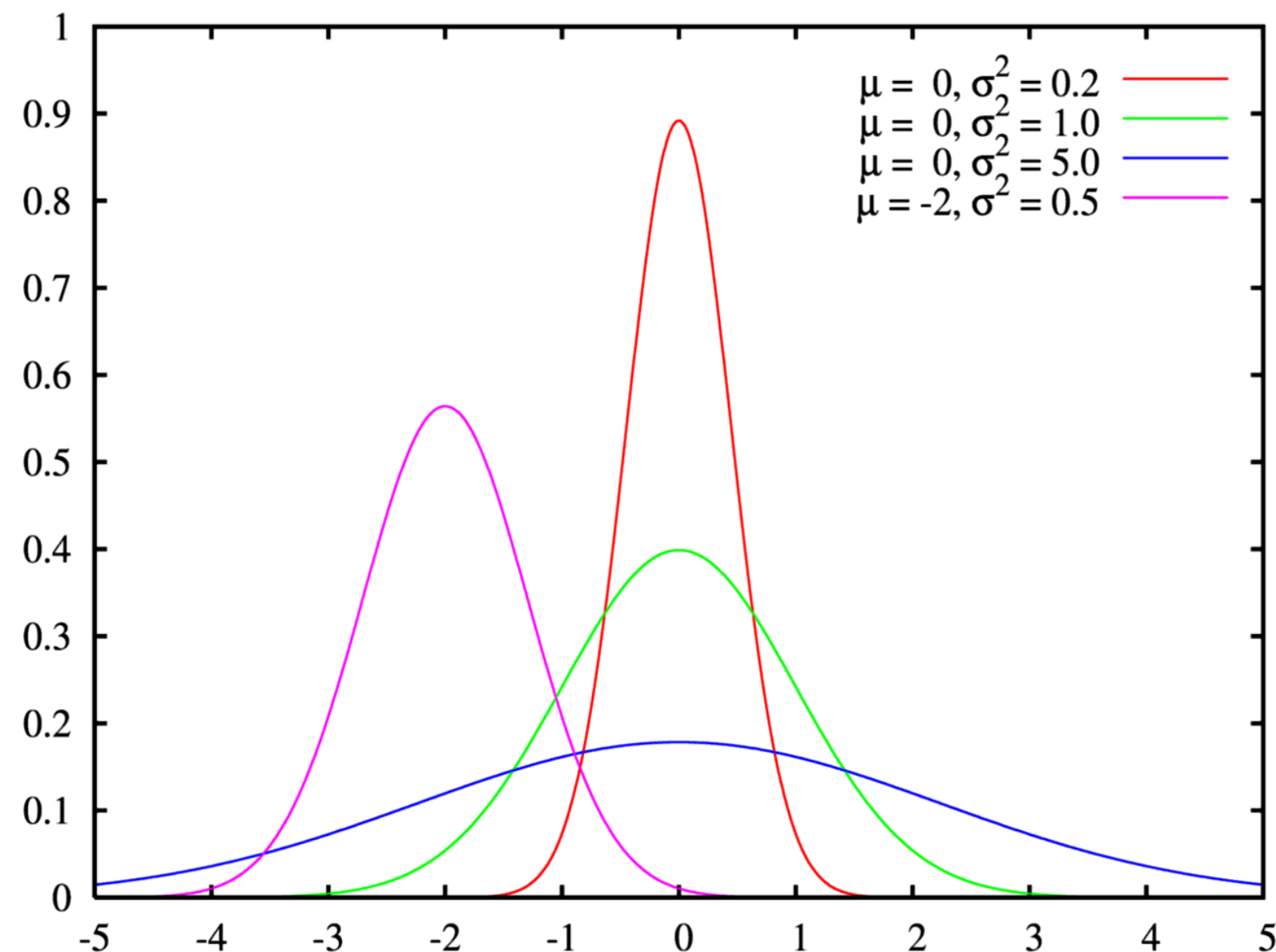
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

μ = Mean

σ = Standard Deviation

$\pi \approx 3.14159 \dots$

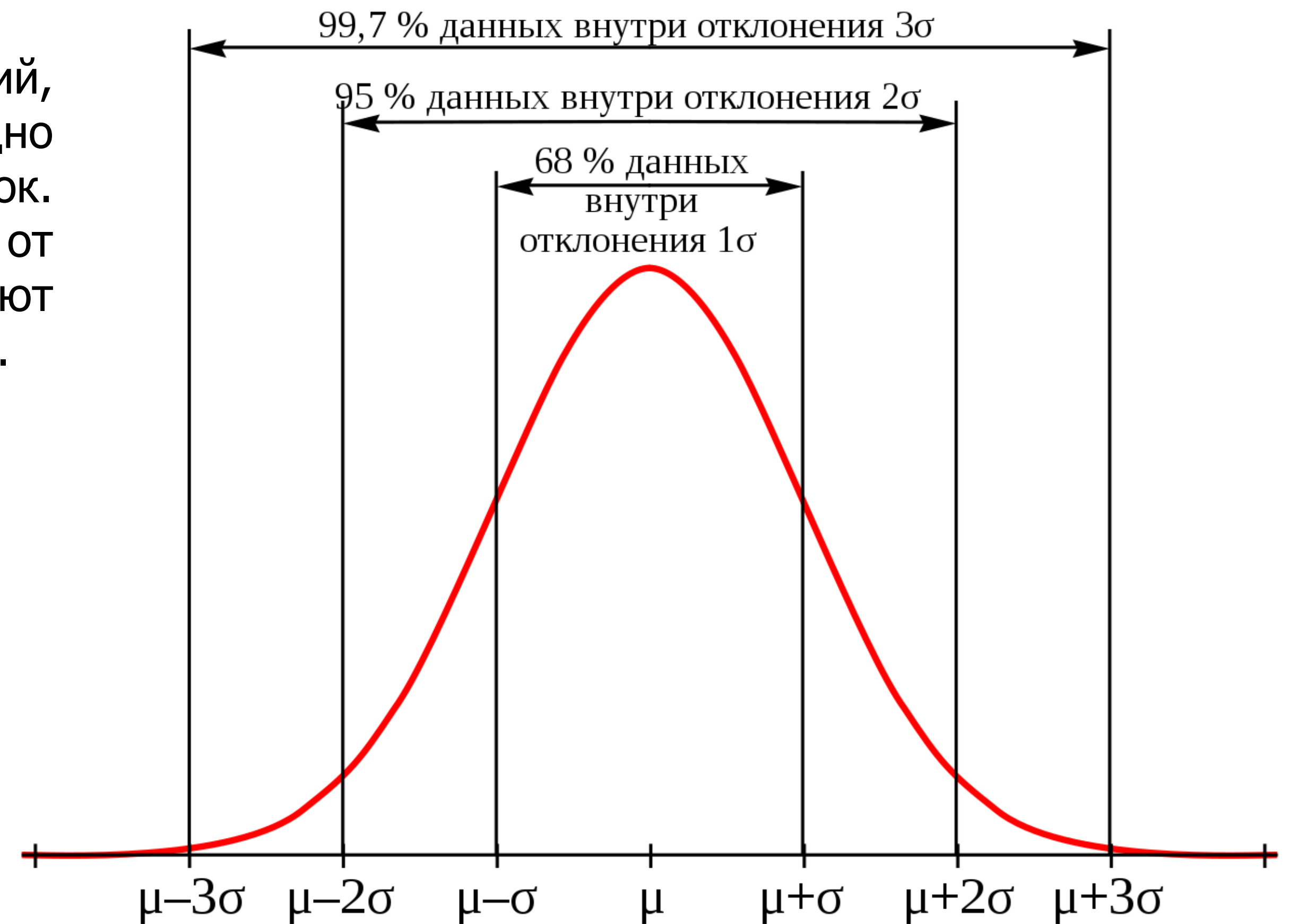
$e \approx 2.71828 \dots$



Правило 68-95-99,7 для нормального (Гауссового) распределения

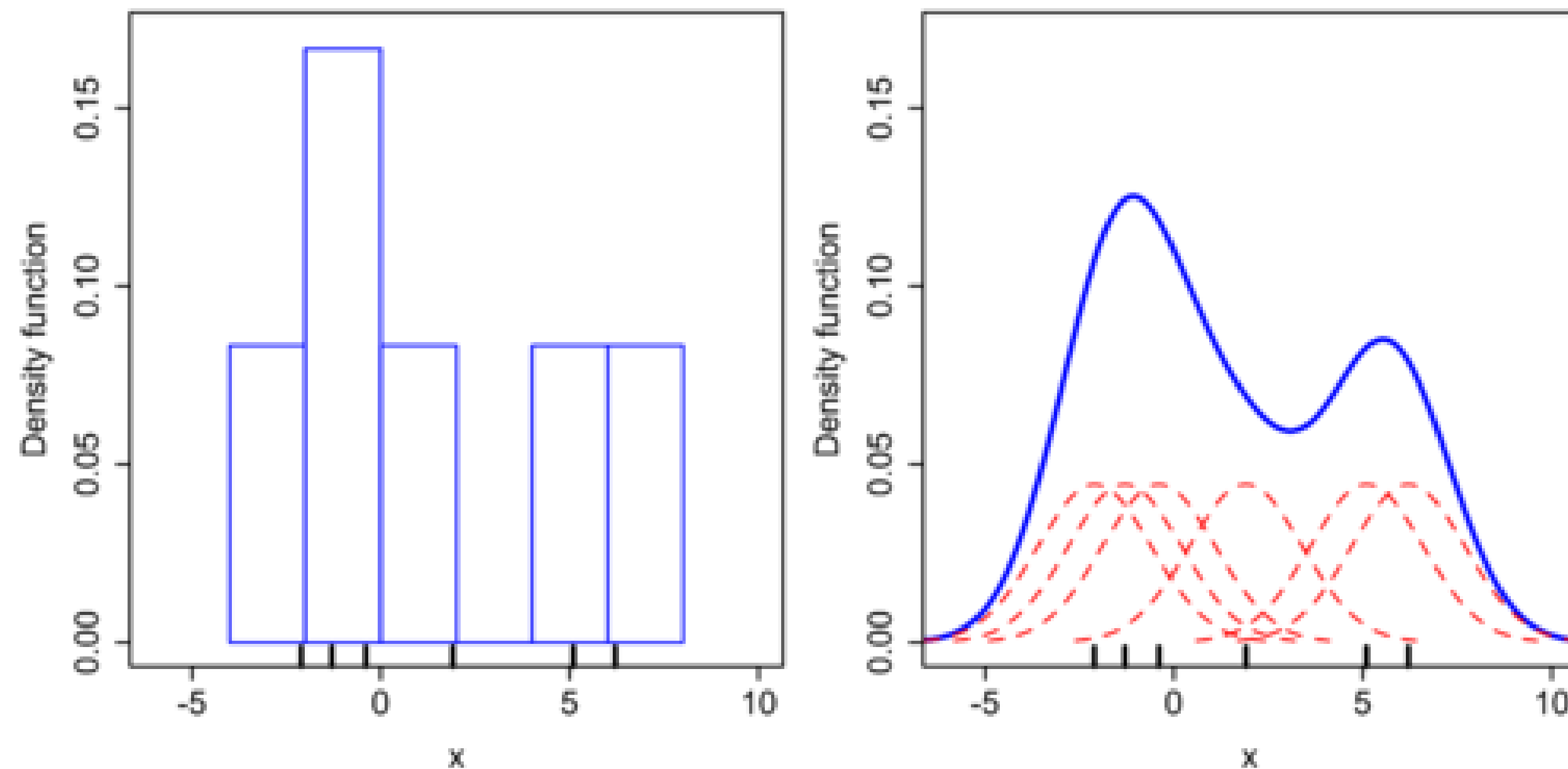
Правило:

Для нормального распределения количество значений, отличающиеся от среднего на число, меньшее чем одно стандартное отклонение, составляют 68,27 % выборок. В то же время количество значений, отличающиеся от среднего на два стандартных отклонения, составляют 95,45 %, а на три стандартных отклонения — 99,73 %.

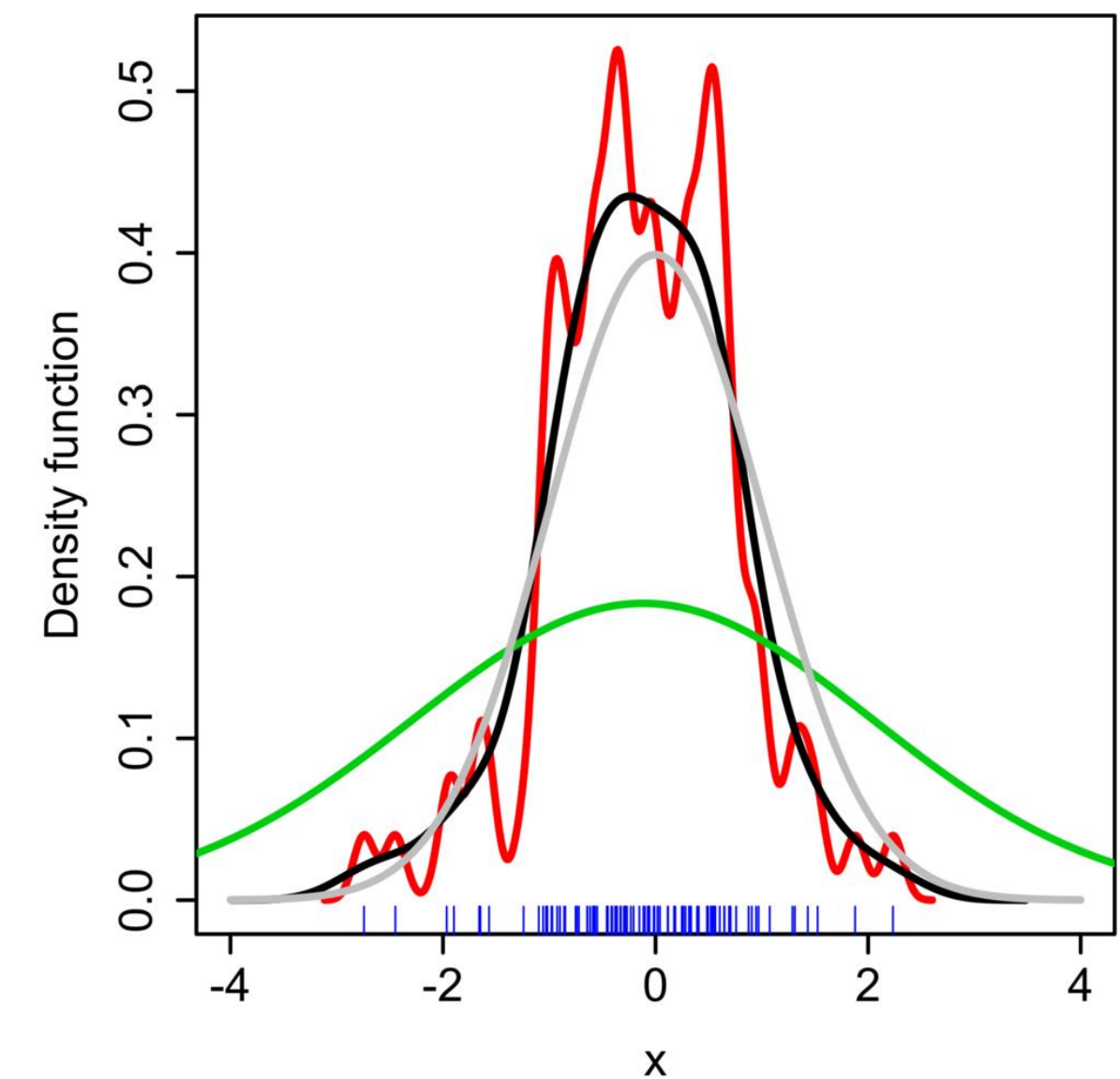


Kernel Density Estimation

Ядерная оценка плотности



Сравнение гистограммы (слева) и ядерной оценки плотности (справа), построенных из тех же самых данных. 6 индивидуальных ядер показаны красными пунктирными линиями, ядерная оценка плотности показана синей кривой. Точки данных показаны чёрточками на ленточной диаграмме по горизонтальной оси.



Ядерная оценка плотности с различными полосами пропускания случайной выборки 100 точек из стандартного нормального распределения.

Серая кривая: истинная плотность (нормальное распределение)

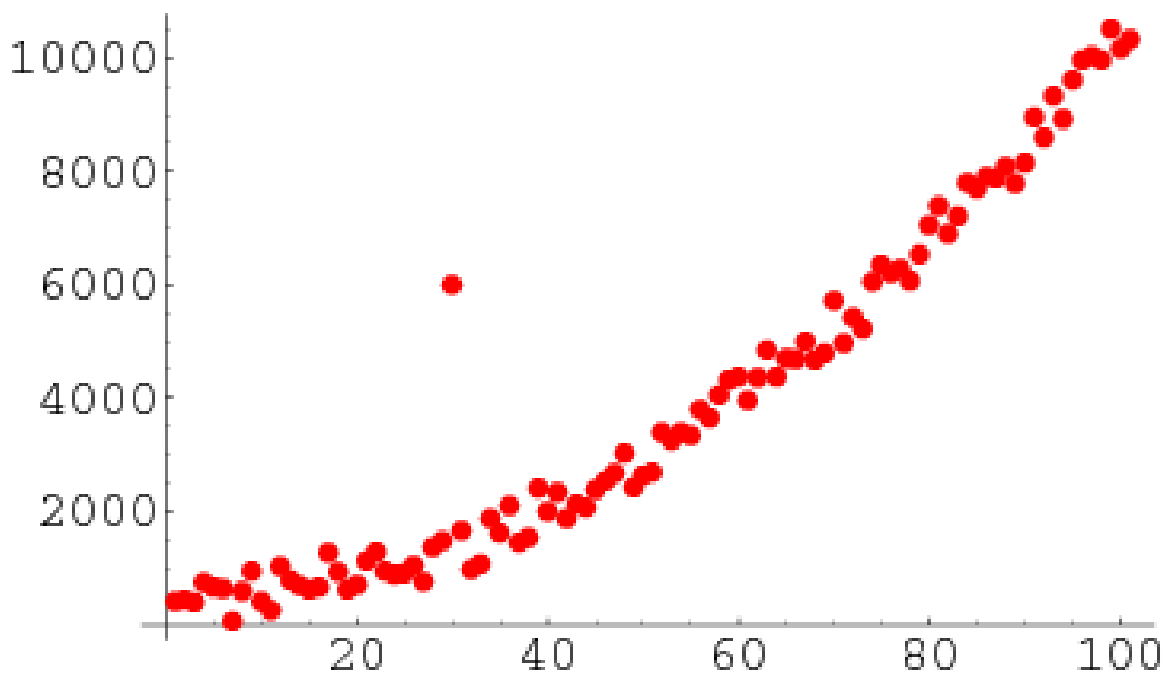
Красная кривая: KDE с $h=0,05$.

Чёрная кривая: KDE с $h=0,337$.

Зелёная кривая: KDE с $h=2$.

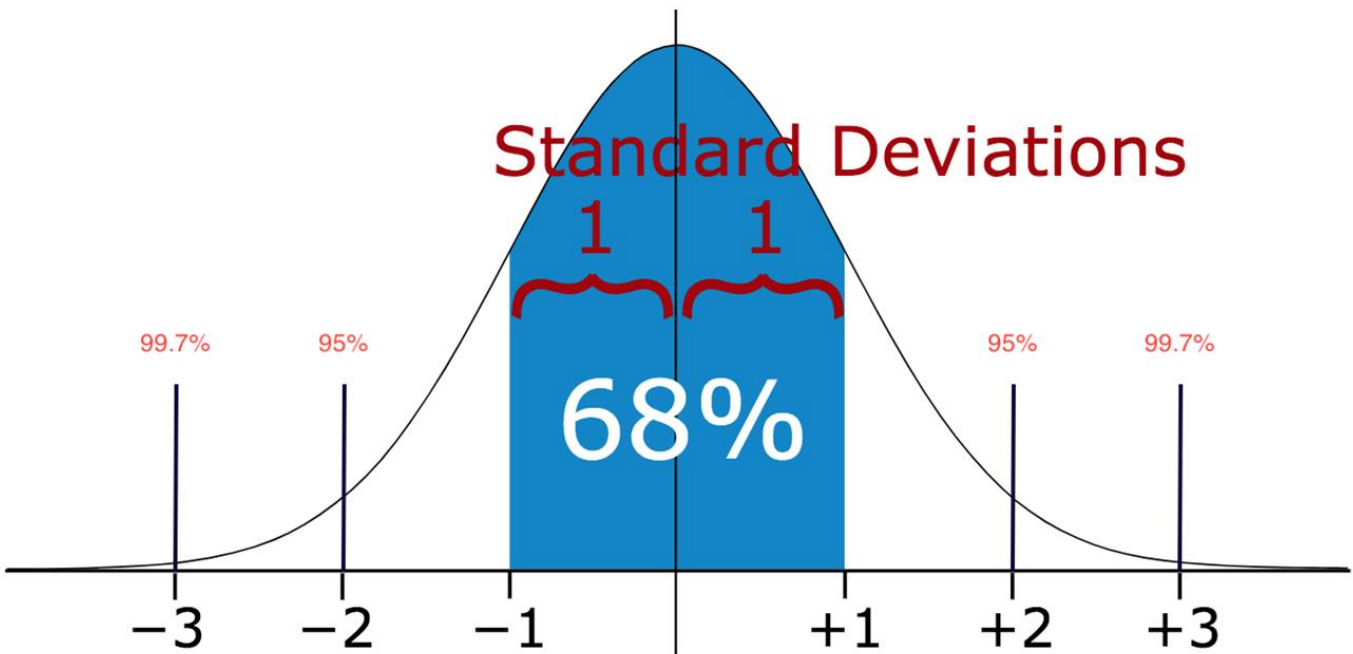
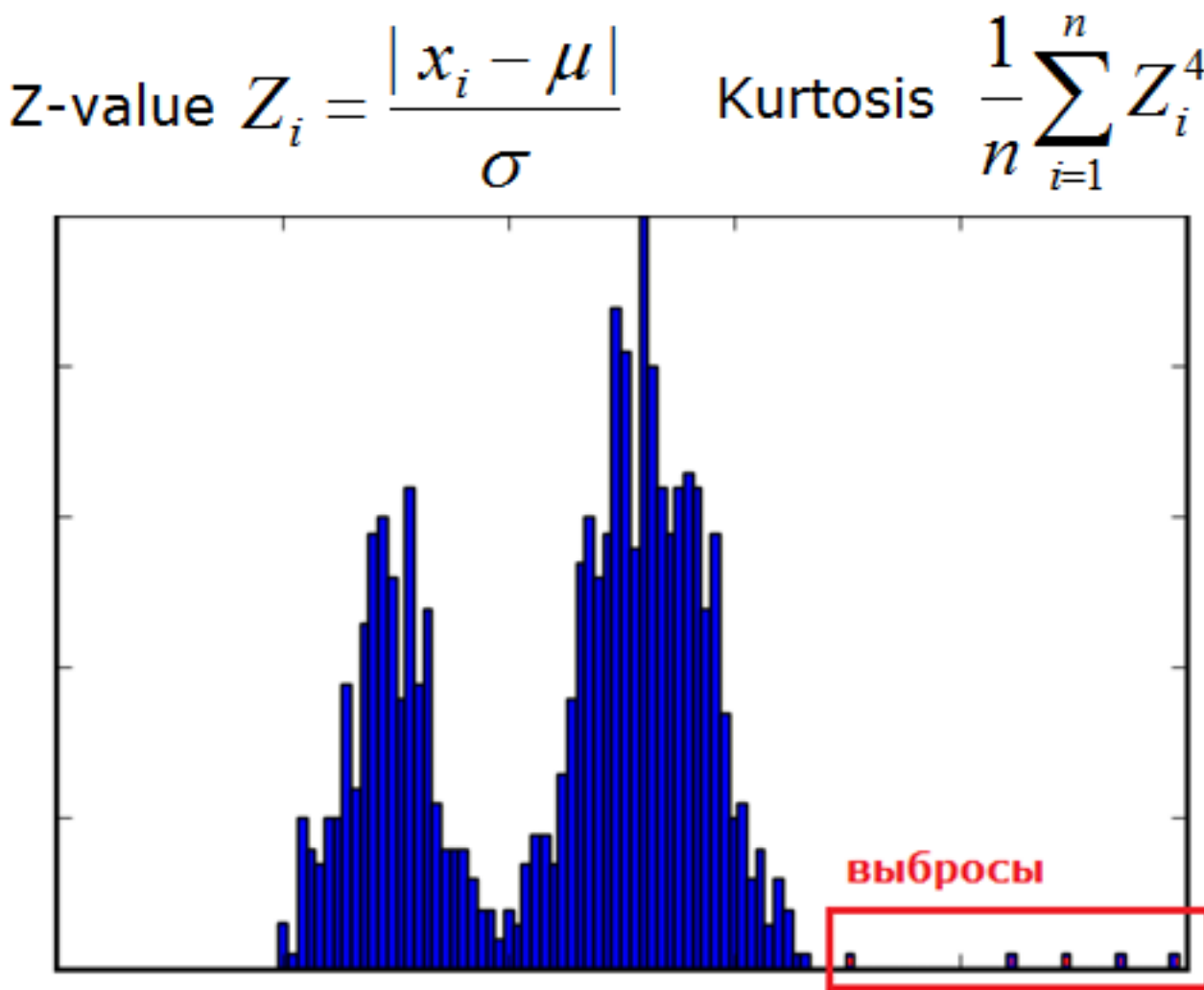
Выбросы в данных

В статистике, **выбросы** - это экстремальные значения во входных данных, которые находятся далеко за пределами других наблюдений. Это ненормальное наблюдение, которое находится далеко от других.

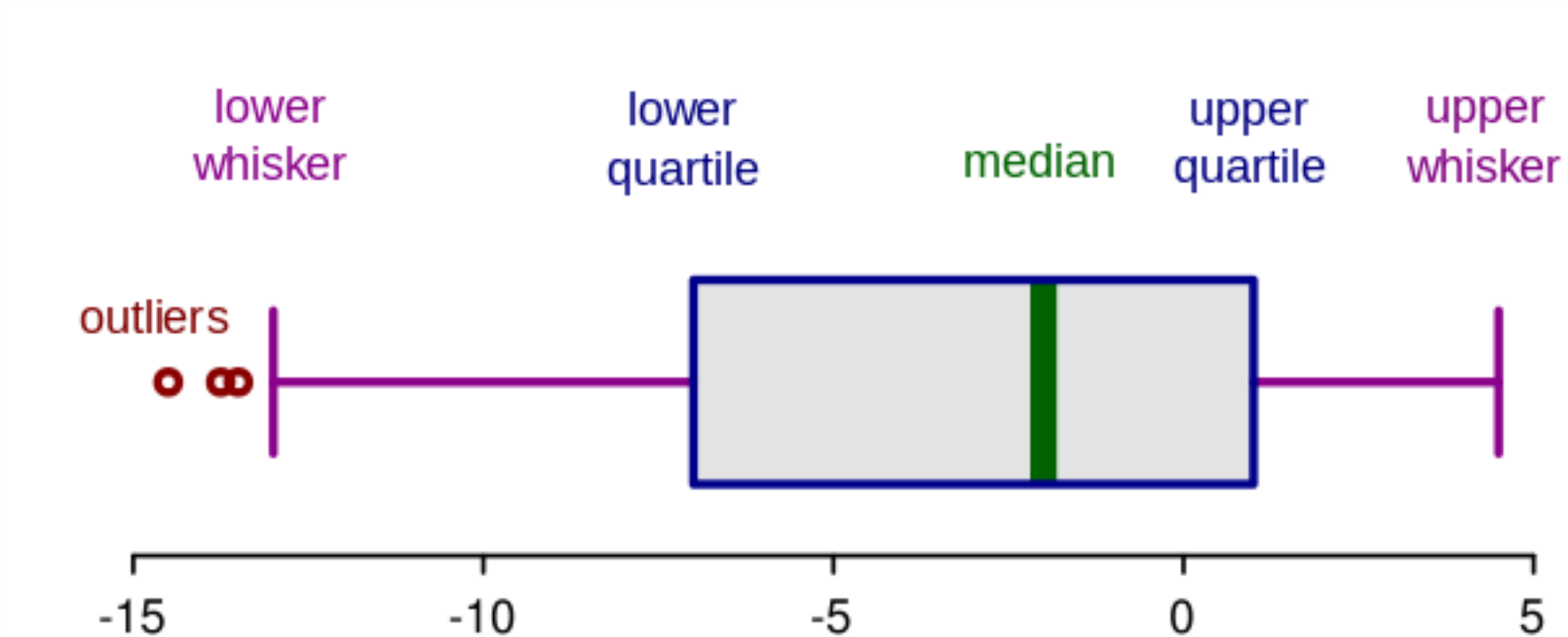


Методы поиска

1. Статистические тесты



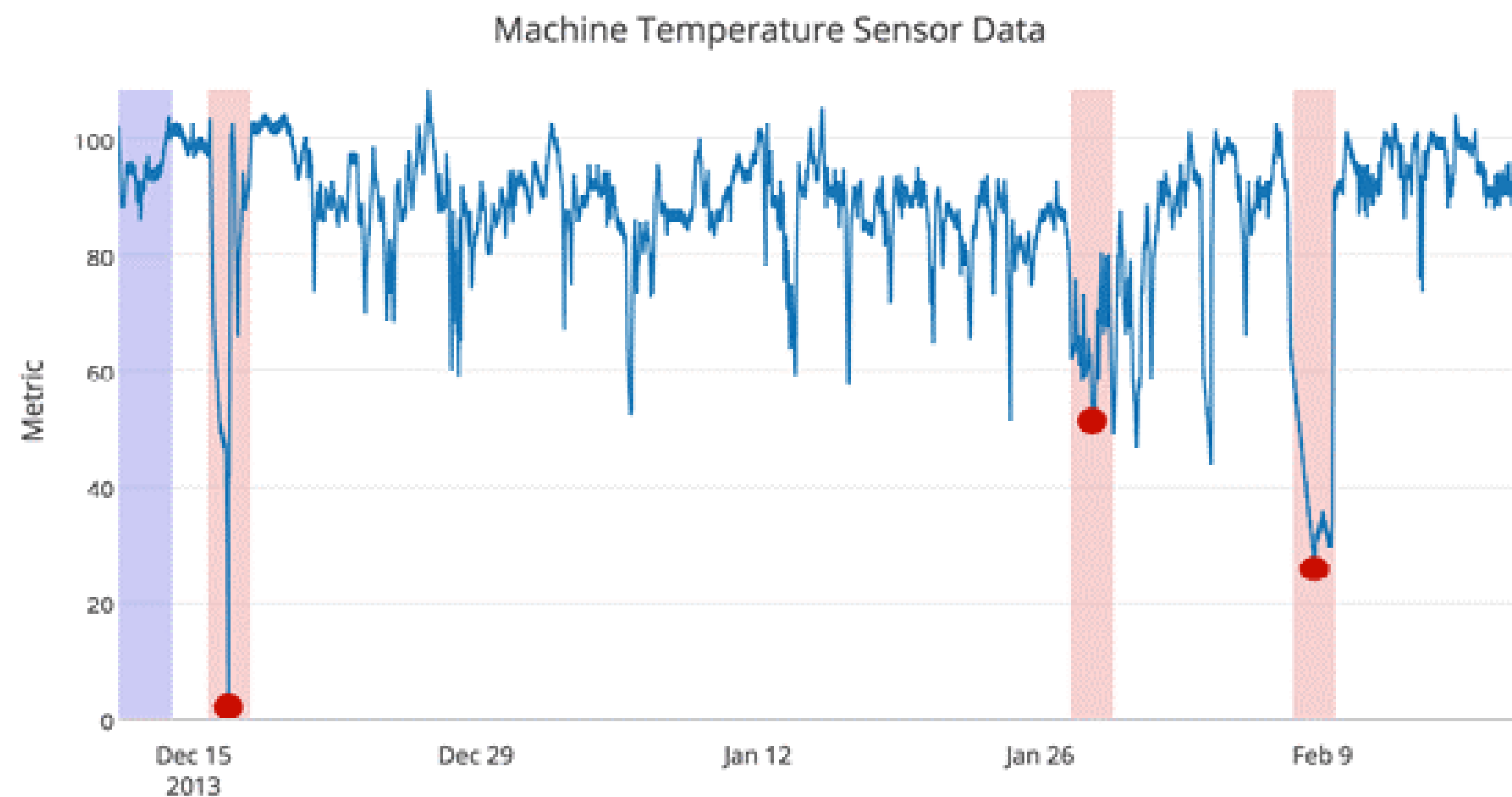
2. BoxPlot



Выбросы в данных

Выбросам посвящен целый раздел в ML - Anomaly detection

Выявление аномалий (также обнаружение выбросов) — это опознавание во время интеллектуального анализа данных редких данных, событий или наблюдений, которые вызывают подозрения ввиду существенного отличия от большей части данных. Обычно аномальные данные характеризуют некоторый вид проблемы, такой как мошенничество в банке, структурный дефект, медицинские проблемы или ошибки в тексте.

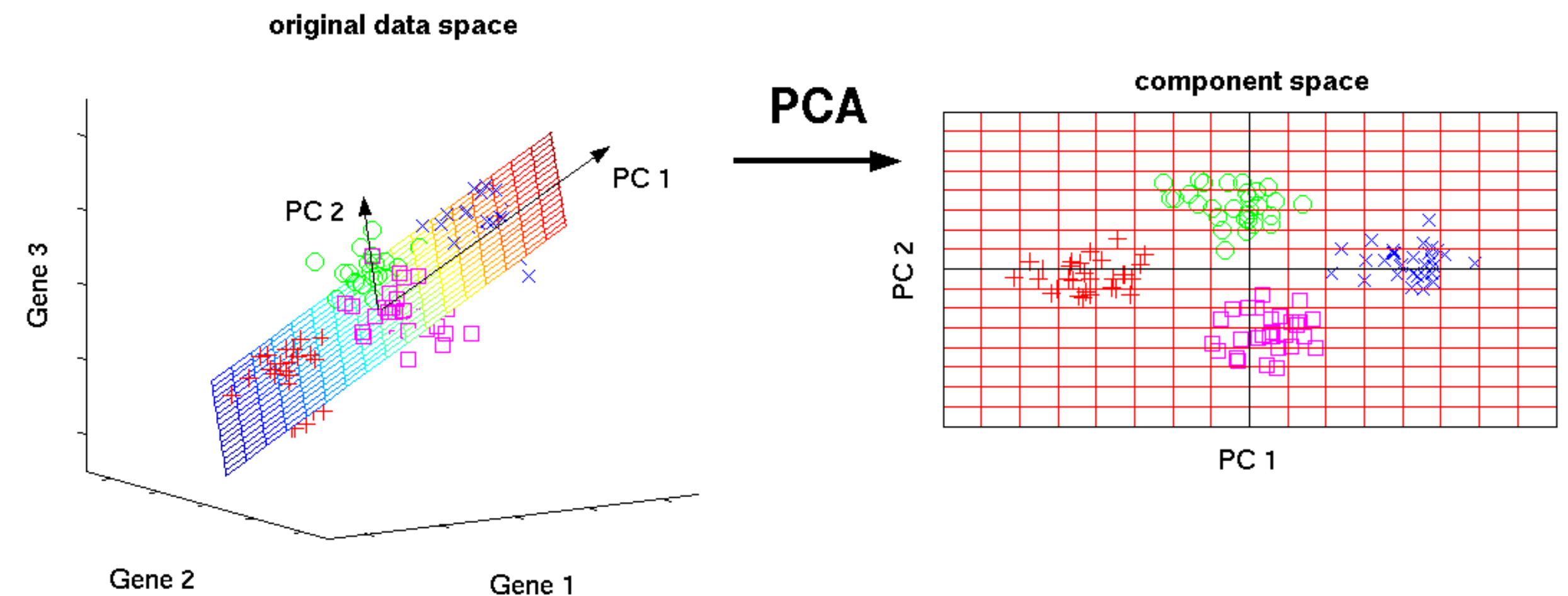
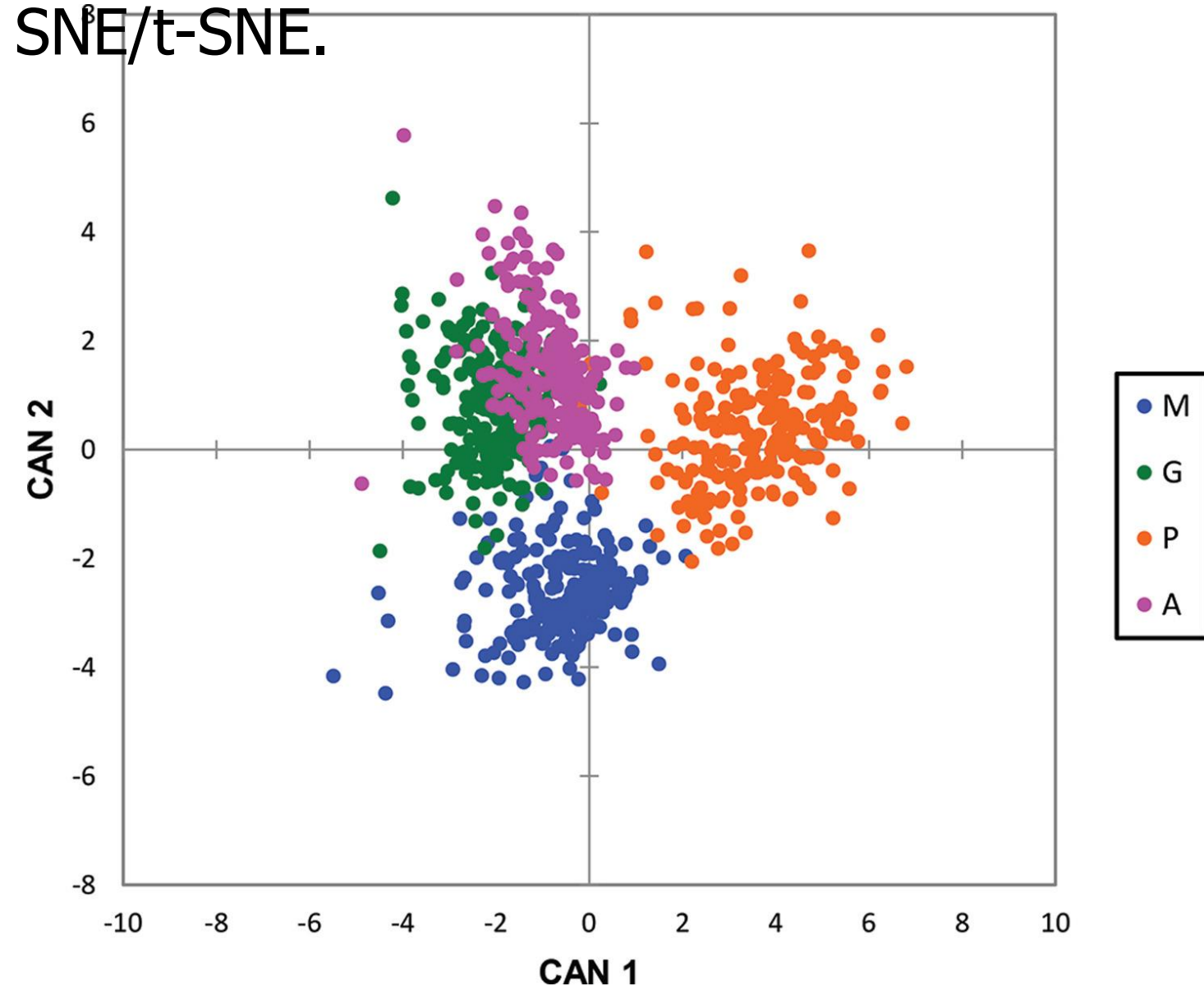


Multivariate analysis

Анализ многих переменных

Чаще всего используются:

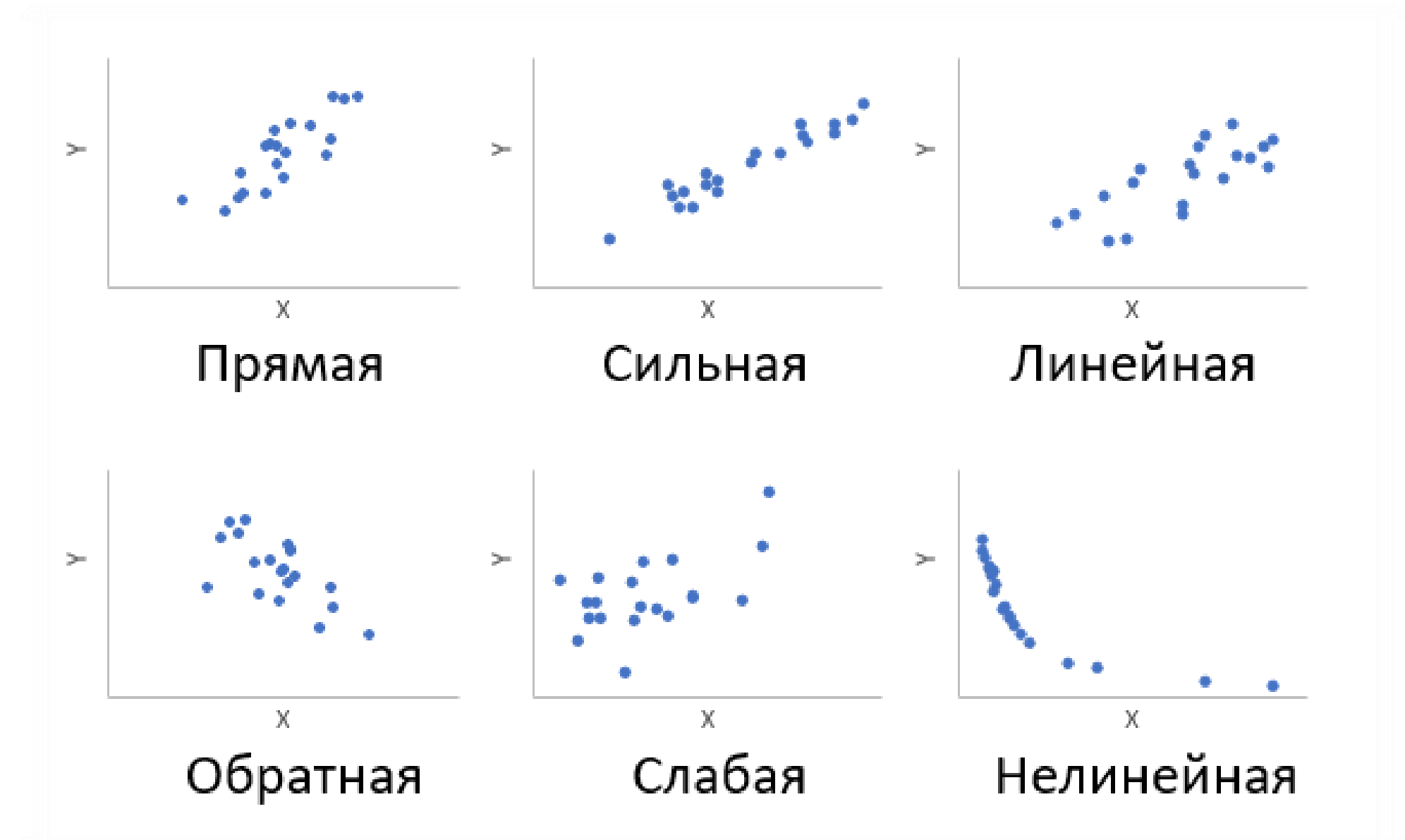
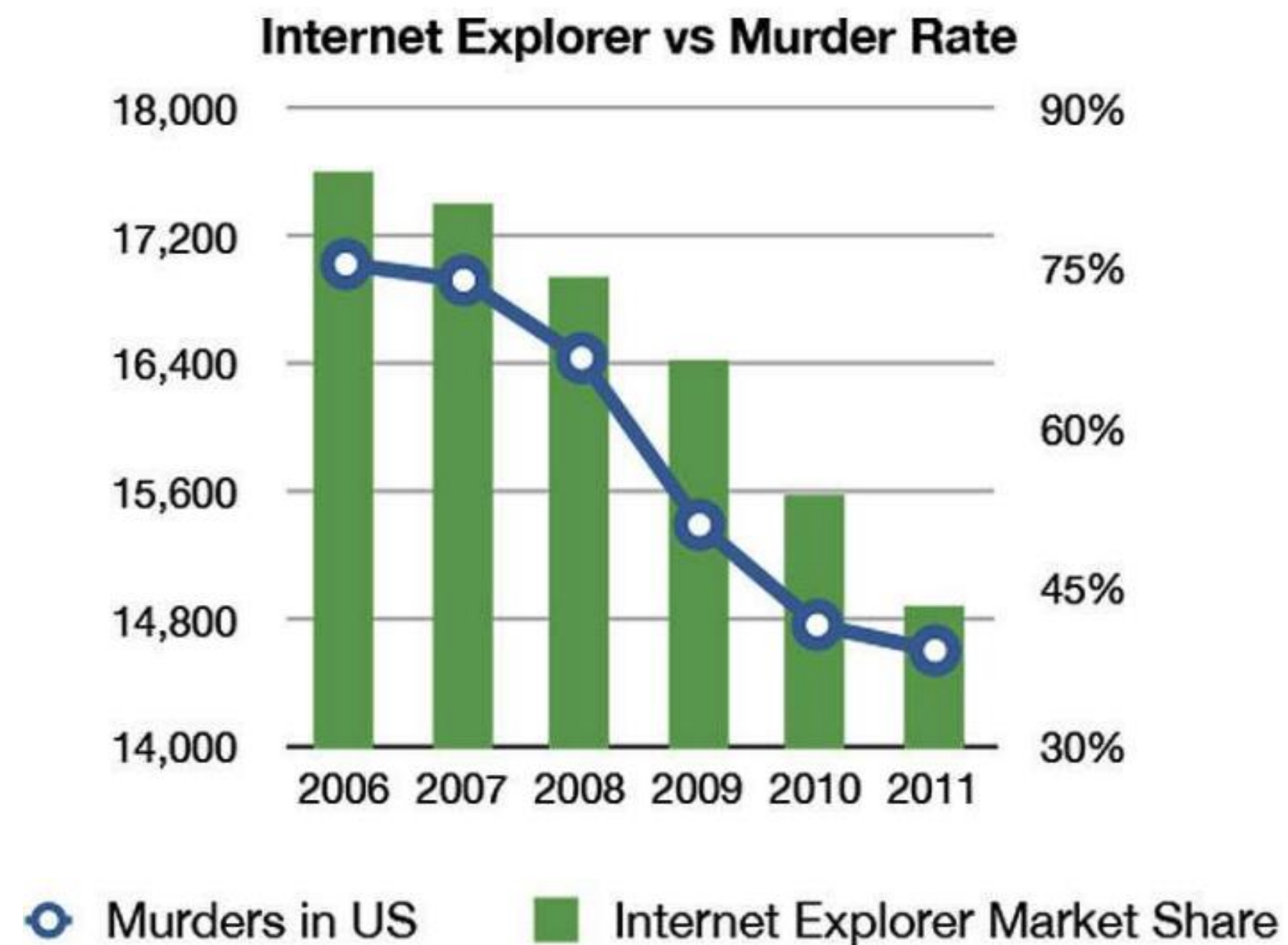
- Scatter plot (визуализация),
- PCA - метод главных компонент, многомерный дисперсионный анализ (о нем, совсем скоро, мы будем подробно говорить),
- SNE/t-SNE.



Коллинеарность признаков

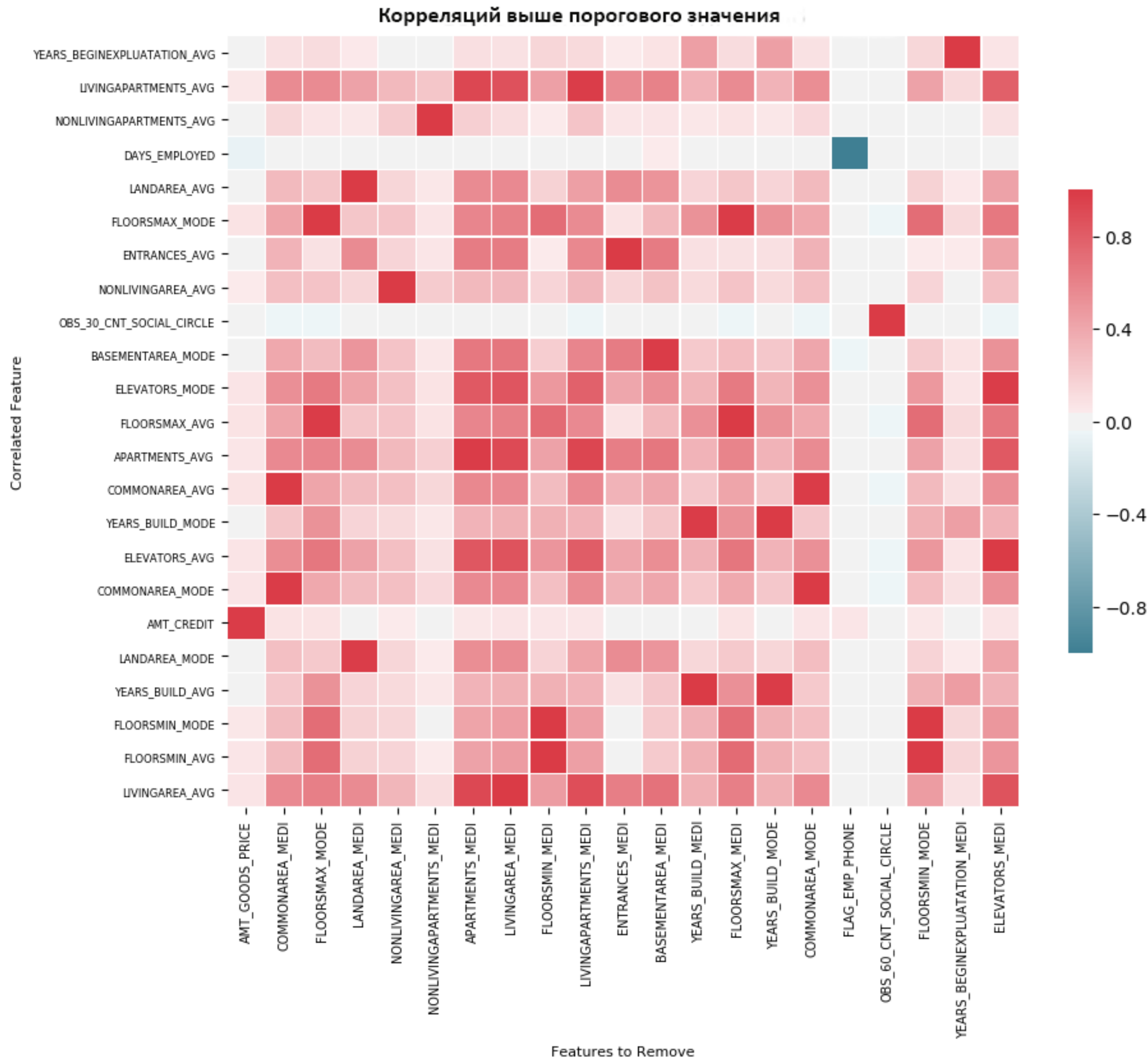
Коллинеарными называются **признаки**, которые сильно коррелируют друг с другом. В машинном обучении это приводит к снижению производительности обобщения данных из-за высокой дисперсии и меньшей интерпретируемости модели.

Корреляция — статистическая взаимосвязь между случайными величинами; не всегда является достаточным условием причинно-следственной связи.



Коллинеарность признаков

Коэффициент корреляции принимает значения от -1 до +1. Чем выше значение коэффициента корреляции, тем больше зависимость между величинами. **Корреляция** бывает положительной и отрицательной.



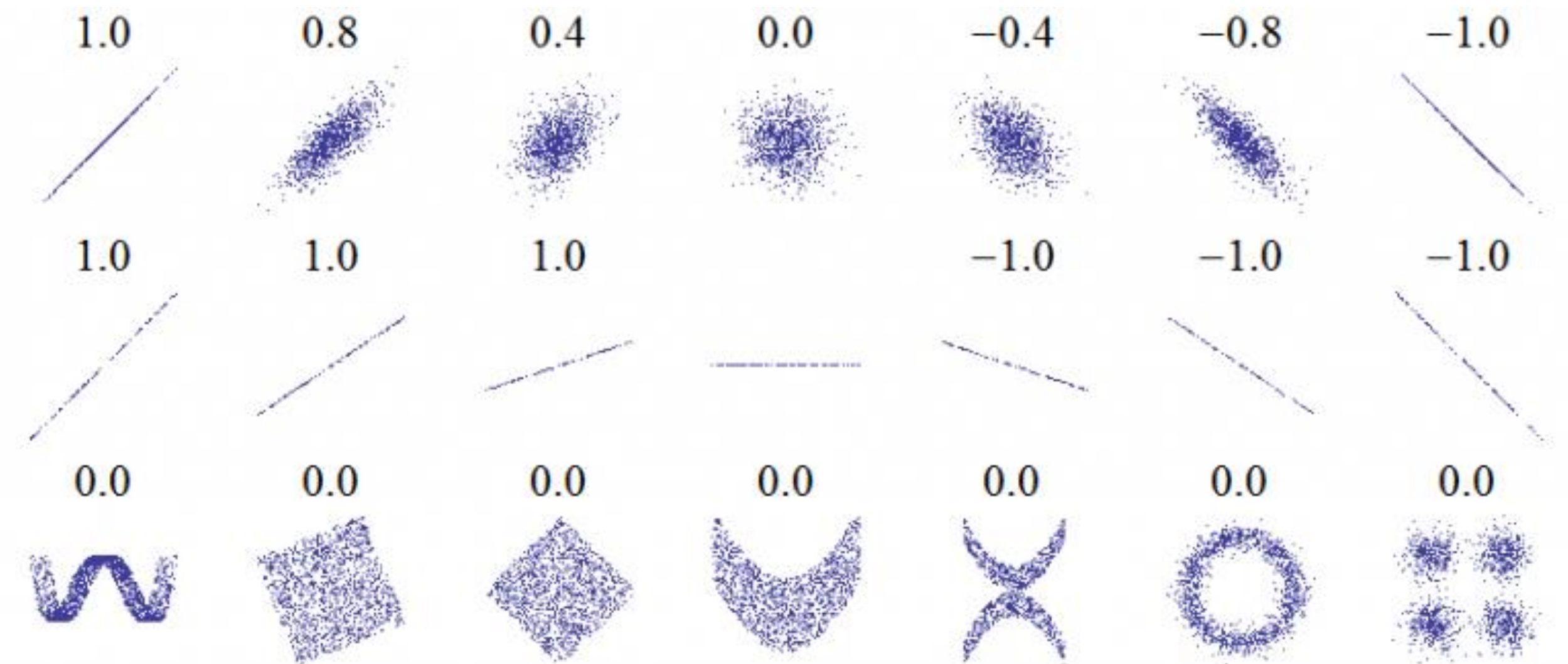
Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона (коэффициент корреляции)- представляет собой статистику, которая из- меряет величину линейной связи (корреляцию) между двумя переменными. Он принимает значения от -1 до +1. Значение коэффициента +1 означает наличие полной положительной линейной связи, а значение -1 – наличие полной отрицательной линейной связи.

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$

\bar{x} – математическое ожидание ряда x ;

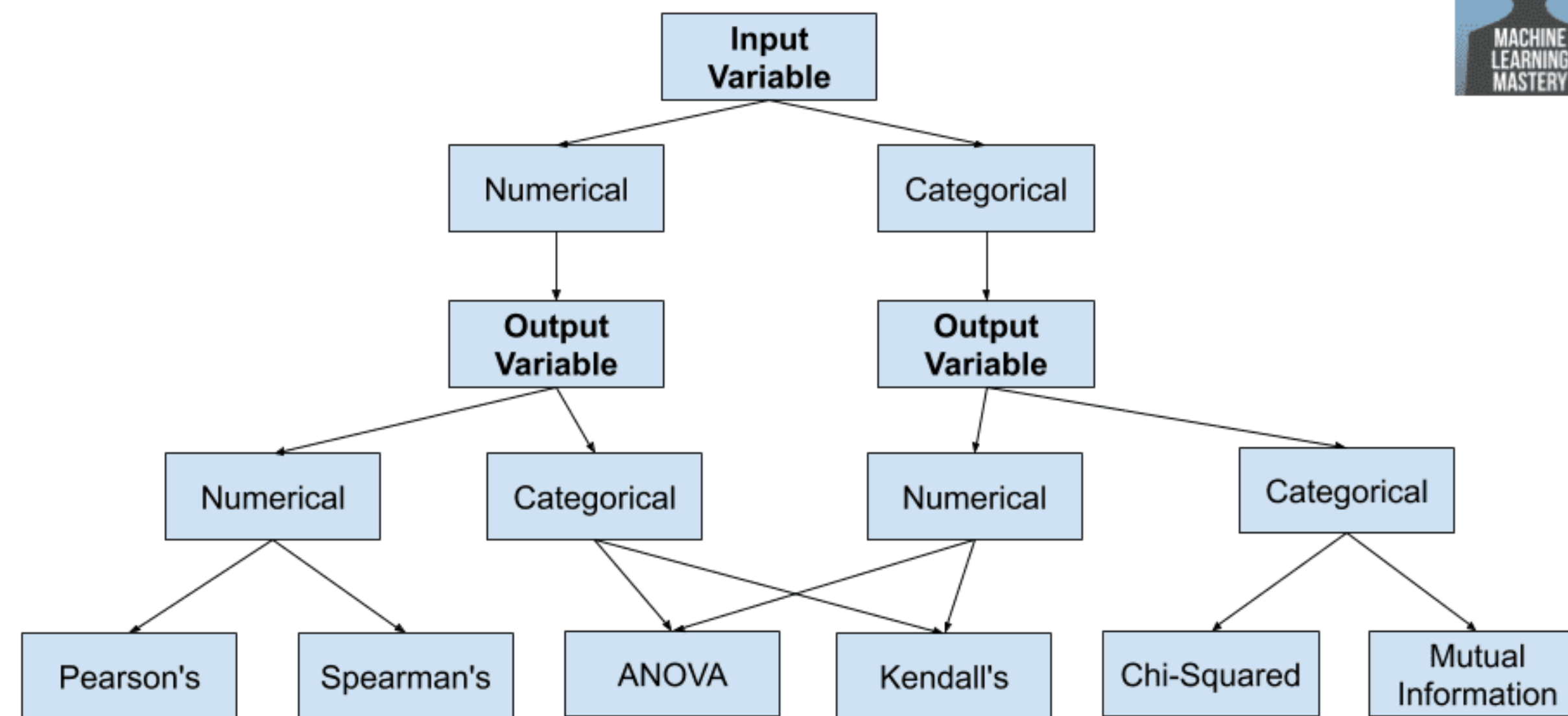
\bar{y} – математическое ожидание ряда y ;



Другие коэффициенты корреляции

- Коэффициент ранговой корреляции Кендалла (для нелинейных зависимостей)
- Коэффициент ранговой корреляции Спирмена (для нелинейных зависимостей)

How to Choose a Feature Selection Method



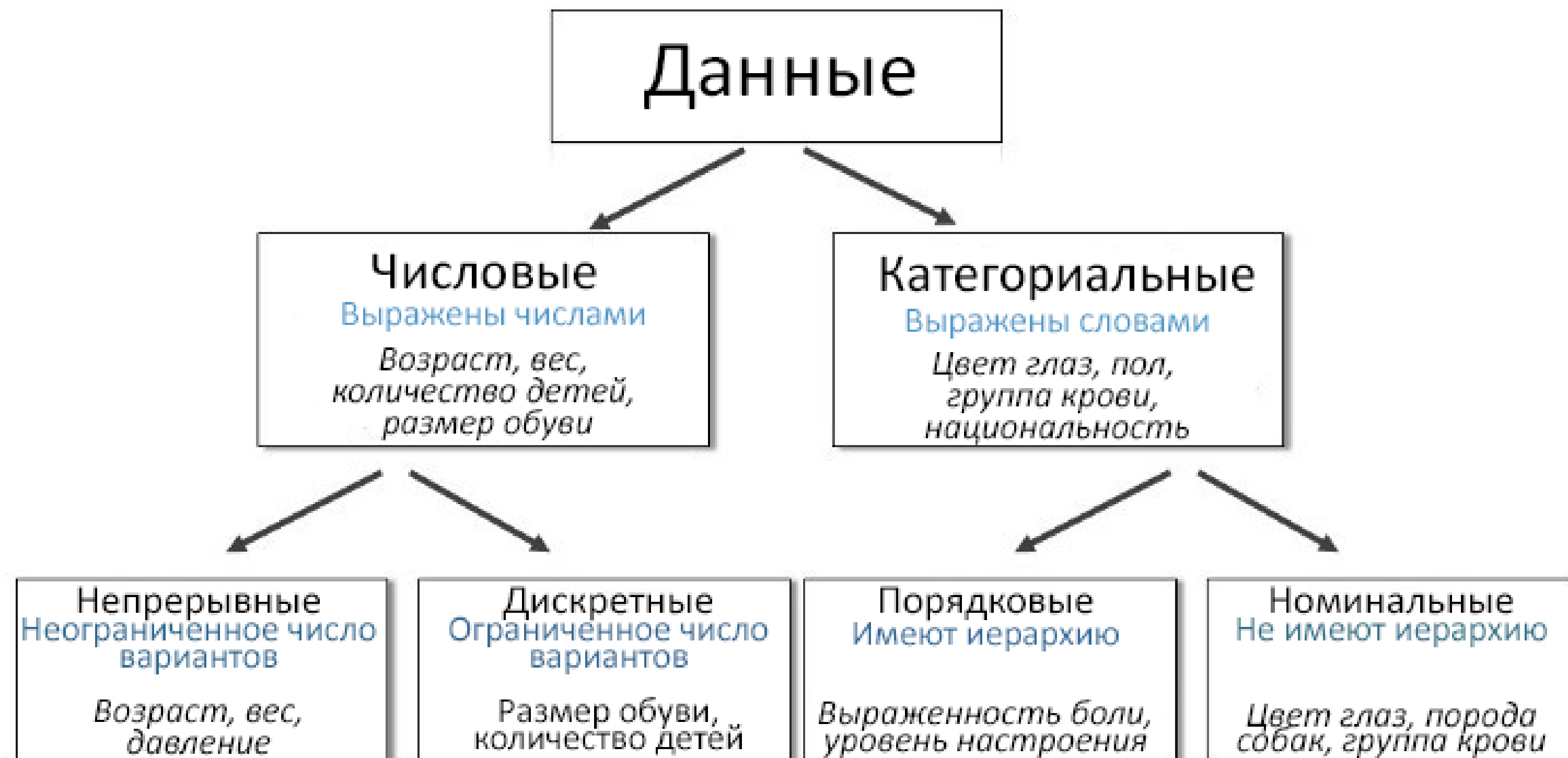
Copyright © MachineLearningMastery.com

<https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D1%8F>

<https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>

Кодирование признаков

Основные типы данных в ML



Кодирование признаков

Как подать модели на вход такие данные?

	Name	Platform	Year	Genre	Publisher
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo
5	Tetris	GB	1989.0	Puzzle	Nintendo
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo

Кодирование признаков

Как подать модели на вход такие данные?

	Name	Platform	Year	Genre	Publisher
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo
5	Tetris	GB	1989.0	Puzzle	Nintendo
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo

Категориальные данные

Кодирование категориальных признаков

Ordinal encoding

Метод реализован в классе `sklearn.preprocessing.LabelEncoder`.


Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Кодирование категориальных признаков

One-hot-encoding

Следующий способ – dummy-кодирование, также называемое — one-hot. Суть заключается в создании дополнительных N признаков (столбцов), где N – количество уникальных категорий. Новые признаки принимают значения 0 или 1 в зависимости от принадлежности к категории. One-hot encoder значительно увеличивает объем данных, что делает его неэффективным с точки зрения памяти, частично эту проблему решает применение разреженных матриц. Метод реализован в классе `sklearn.preprocessing.OneHotEncoder`.

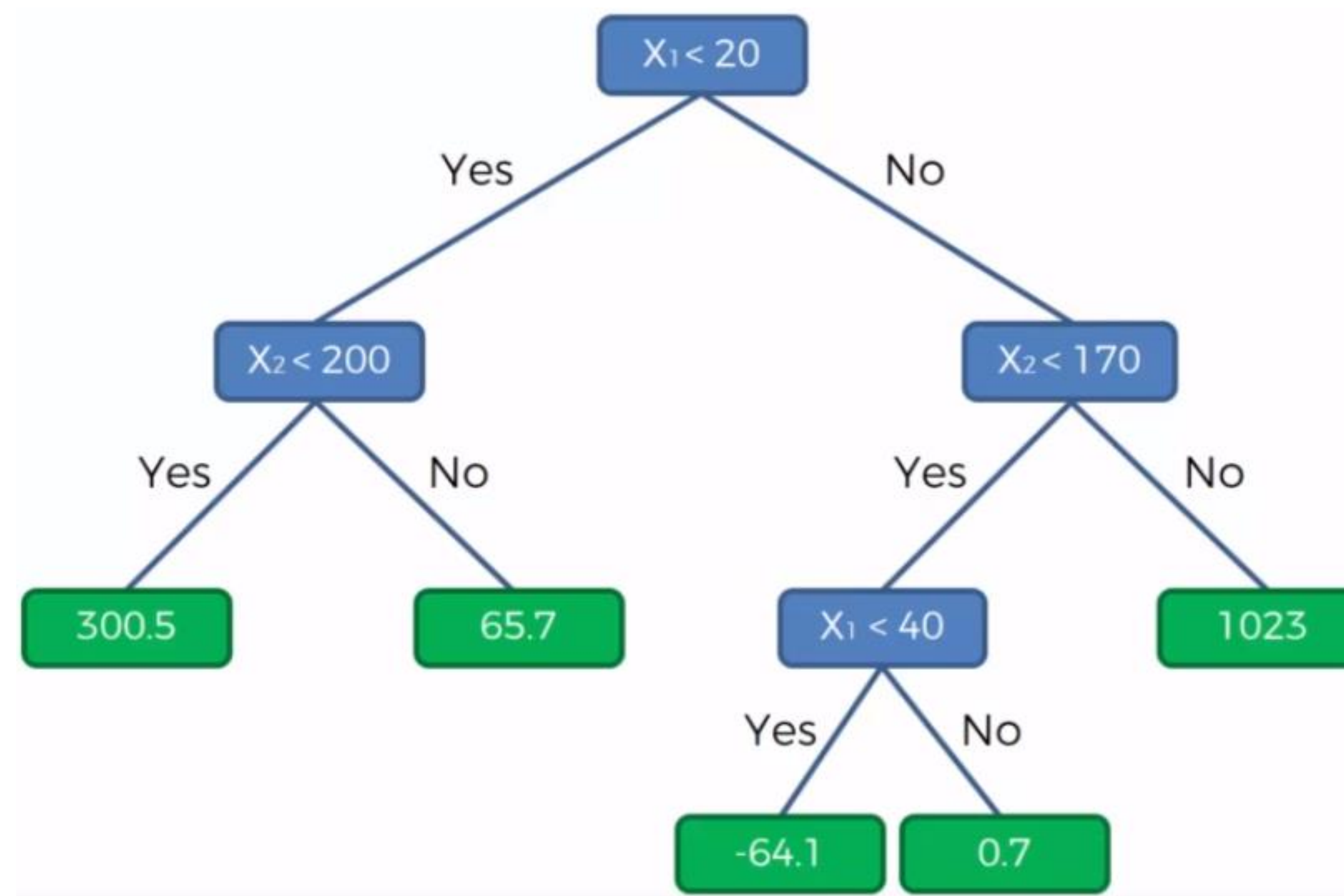
id	color
1	red
2	blue
3	green
4	blue



id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

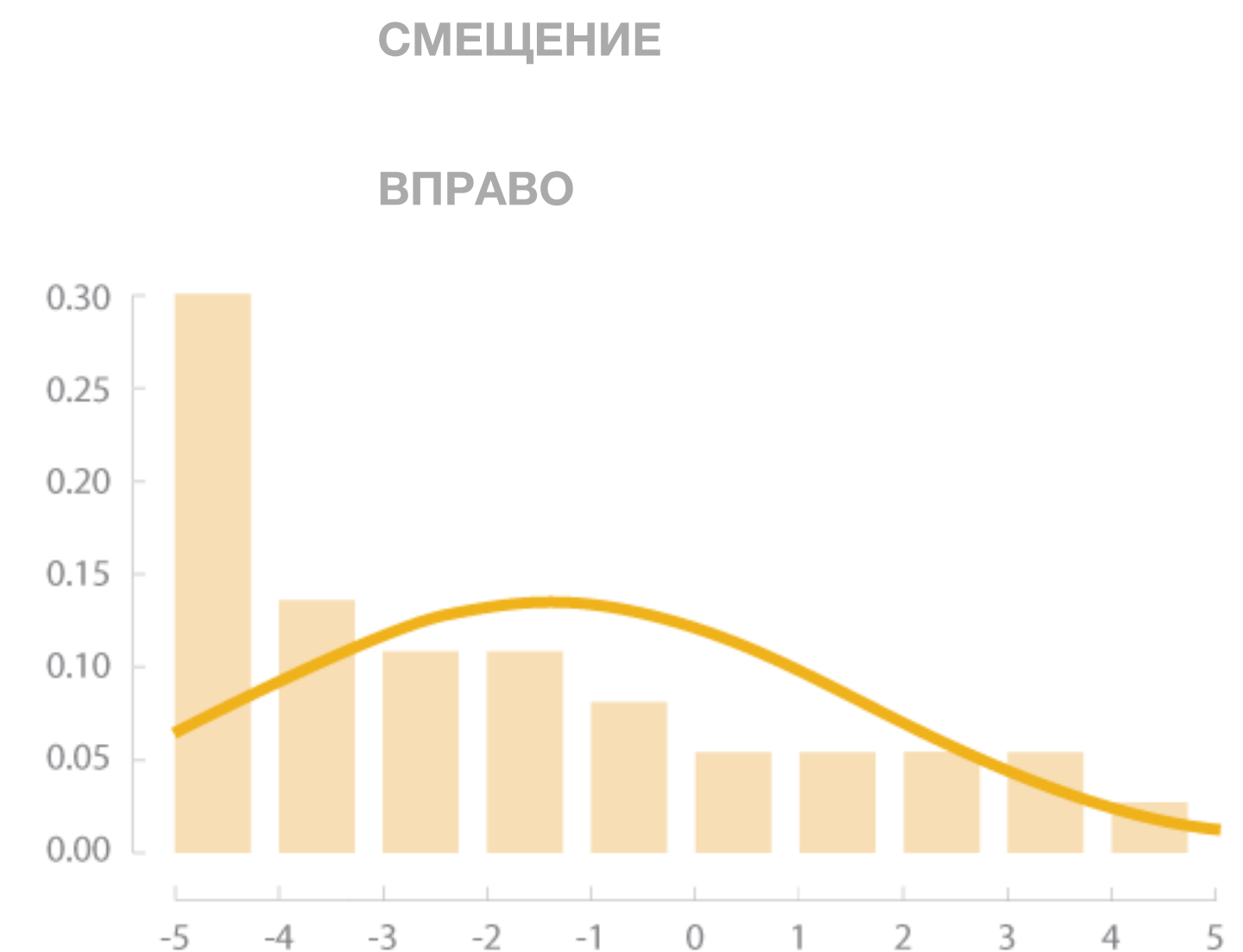
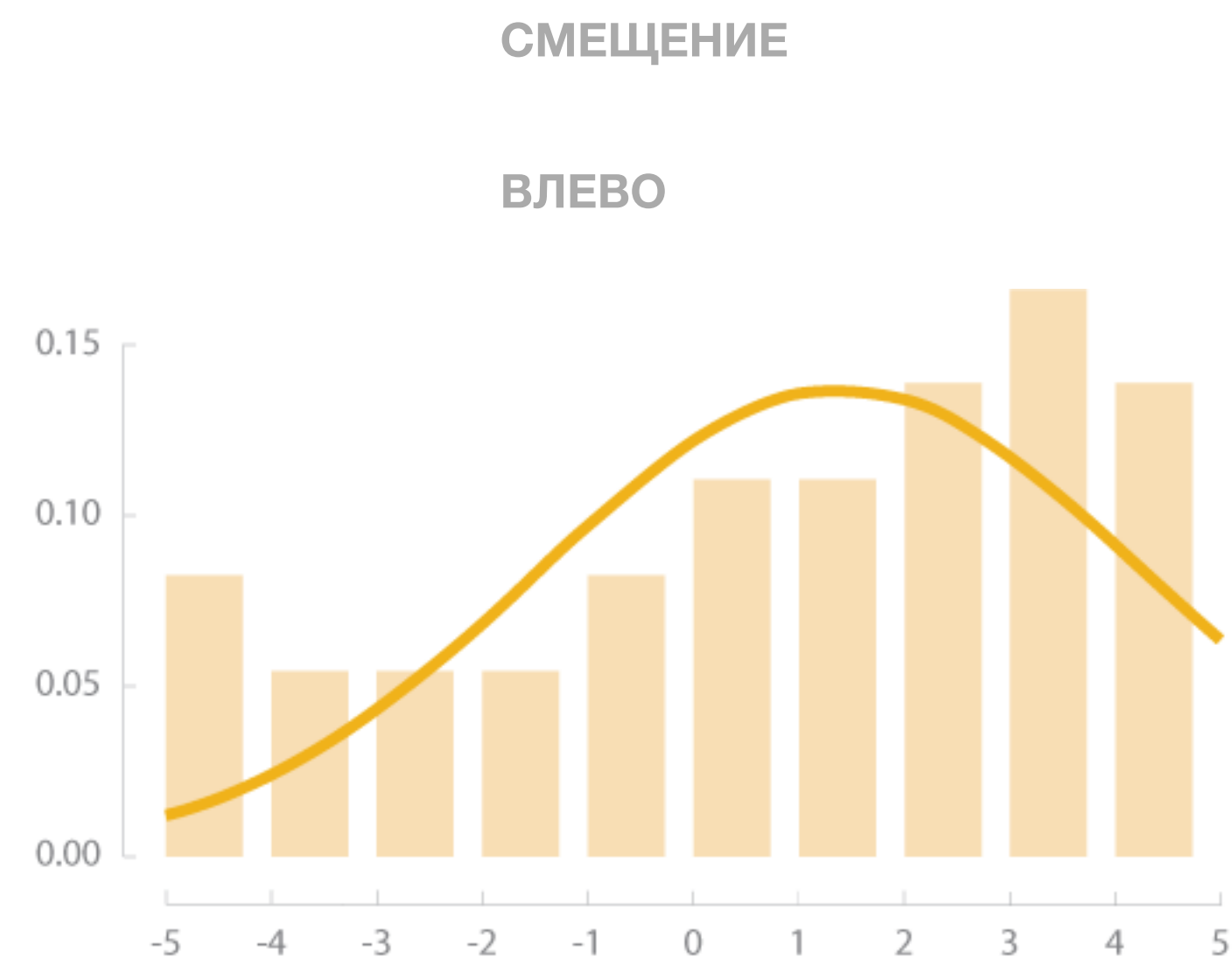
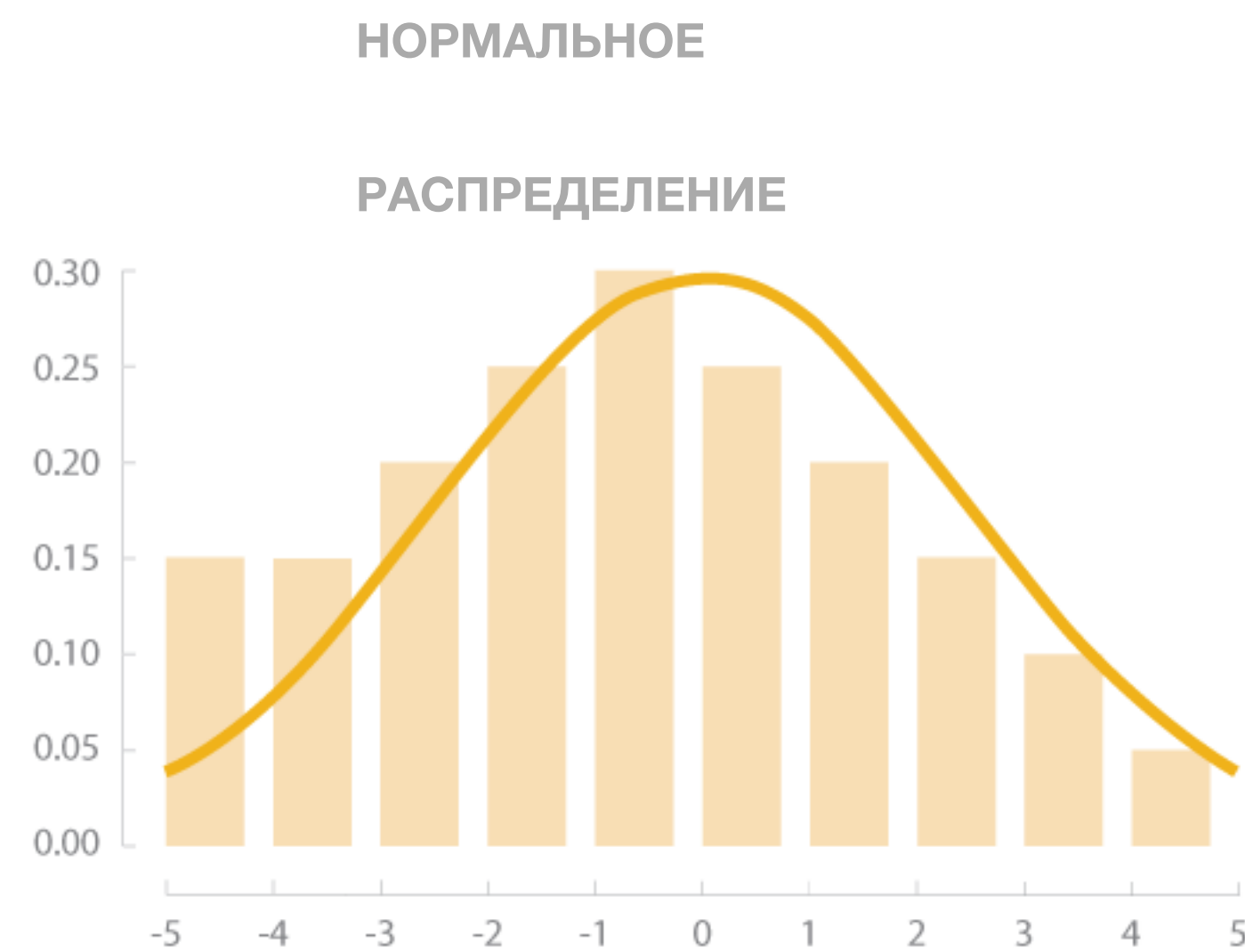
Кодирование категориальных признаков

Вопрос: One-hot-encoding подходит для использования с деревьями решений?



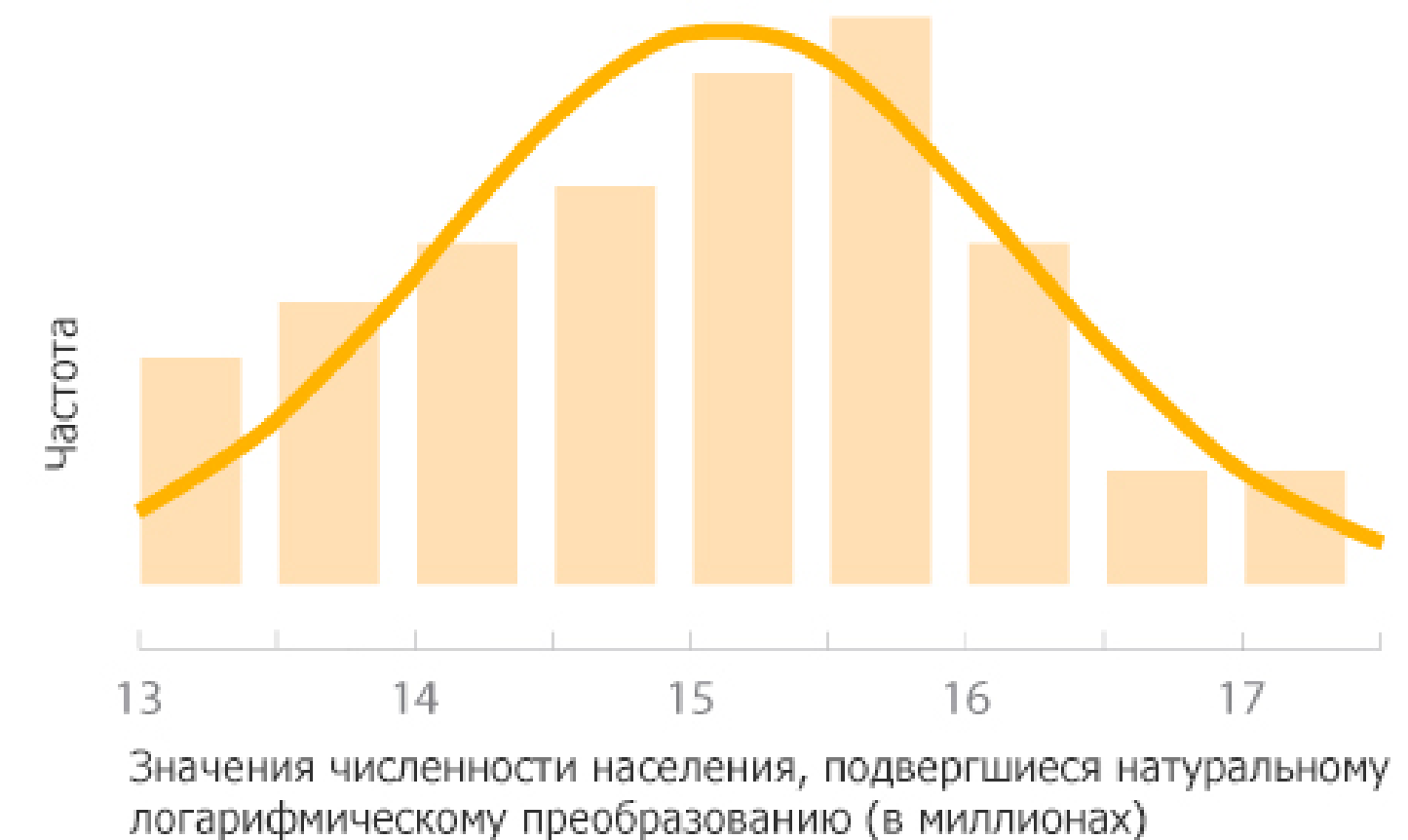
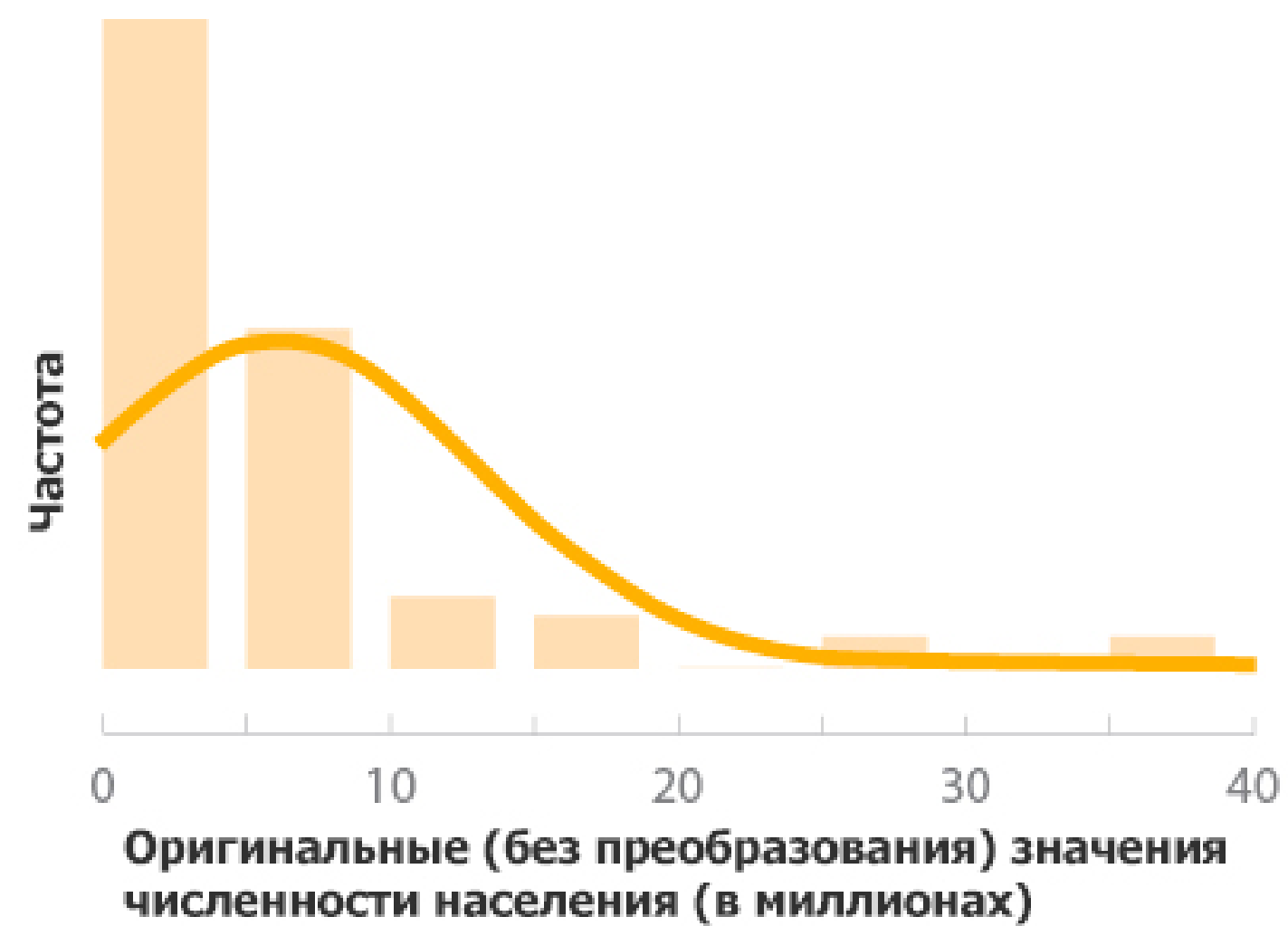
Распределения данных

Одна из наиболее часто встречающихся предпосылок статистических тестов заключается в том, что данные должны быть нормально распределены.



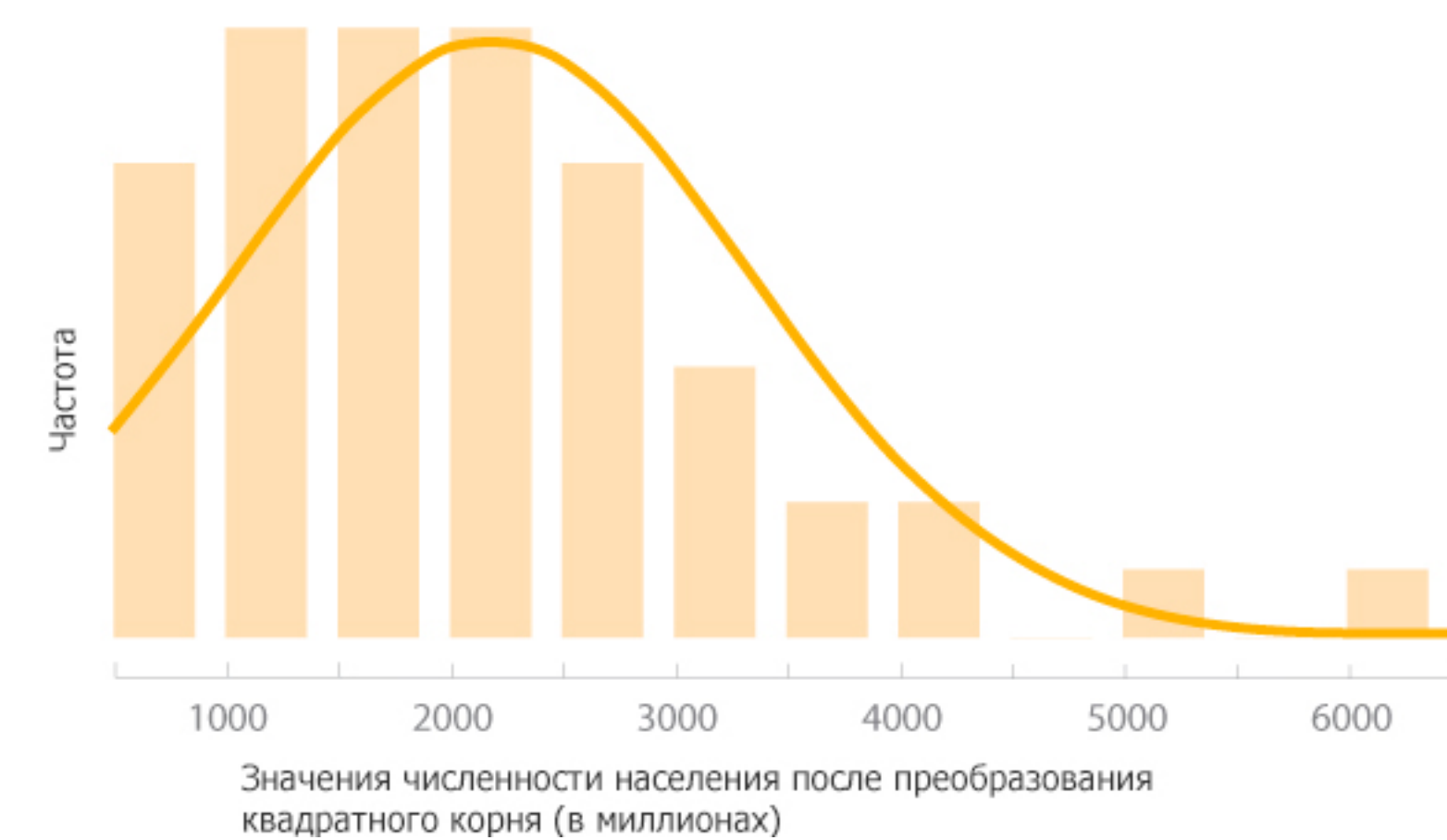
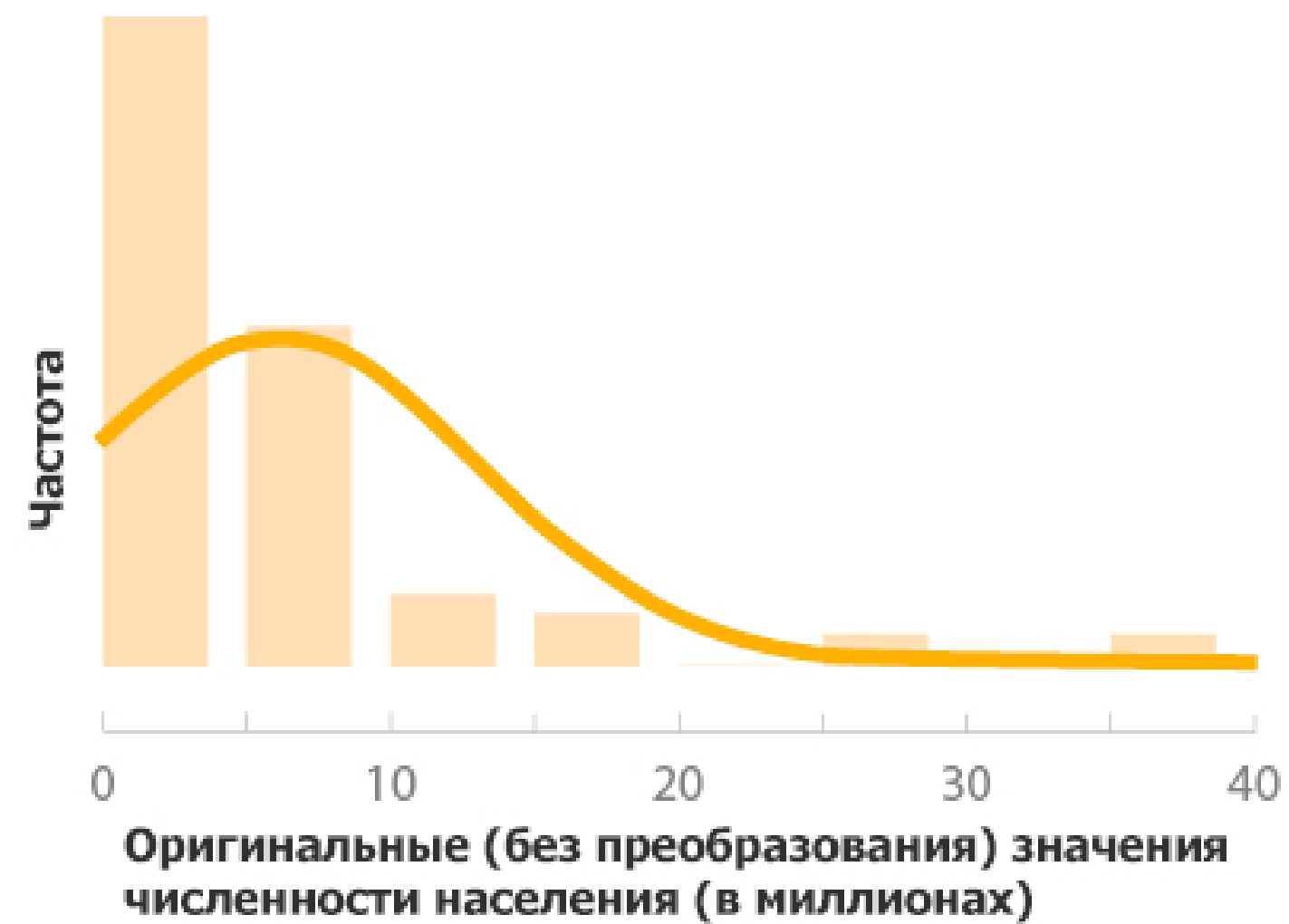
Кодирование числовых признаков

Log-transform



Кодирование числовых признаков

Sqrt transform



РАСПРОСТРАНЕННЫЕ ПРЕОБРАЗОВАНИЯ

Метод	Математическая операция	Подходит для:	Не подходит для:
Логарифм	$\ln(x)$ $\log_{10}(x)$	Данных, смещенных вправо	Нулевых значений
Квадратный корень	\sqrt{x}	особенно хорошо работает со степенями 10 более высокого порядка (например, 1000, 100000)	Отрицательных значений
Квадрат	x^2	Данных, смещенных вправо	Отрицательных значений
Корень кубический	$x^{1/3}$	Данных, смещенных влево	Отрицательных значений
Обратная дробь	$1/x$	Данных, смещенных вправо	Не так эффективен при нормализации, как логарифмическое преобразование

Литература

- **Топ 9 подходов в Feature Engineering:**

- <https://rubikscube.net/2021/06/29/top-9-feature-engineering-techniques/>

- **Детекция аномалий:**

- <https://dyakonov.org/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/>

- **Гайд по EDA и FE:**

- <https://towardsdatascience.com/exploratory-data-analysis-feature-engineering-and-modelling-using-supermarket-sales-data-part-1-228140f89298>
- <https://habr.com/ru/company/ods/blog/325422/>
- <https://nagornyy.me/it/rabota-s-priznakami-kak-chast-mashinnogo-obucheniia/>
- <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114#3abe>
- <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>

- **Гайд по Pandas profiling на русском:**

- <https://www.helenkapatsa.ru/pandas-profiling/>

- **Случайные величины и их распределения:**

- [https://mipt.ru/education/chair/mathematics/study/methods/%D0%A1%D0%92%D0%B8%D0%A0_%D0%A1%D0%B0%D0%BC%D0%B0%D1%80%D0%BE%D0%B2%D0%B0\(2\).pdf](https://mipt.ru/education/chair/mathematics/study/methods/%D0%A1%D0%92%D0%B8%D0%A0_%D0%A1%D0%B0%D0%BC%D0%B0%D1%80%D0%BE%D0%B2%D0%B0(2).pdf)