

Обучение без учителя. Понижение размерности. Алгоритмы понижения размерности.

Урок 9

На этой лекции вы найдете ответы на такие вопросы как:

- Что такое обучение без учителя
- Понижение размерности как метод обучения без учителя
- Линейное понижение размерности
- Нелинейное понижение размерности
- Случайное понижение размерности



Булгакова Татьяна

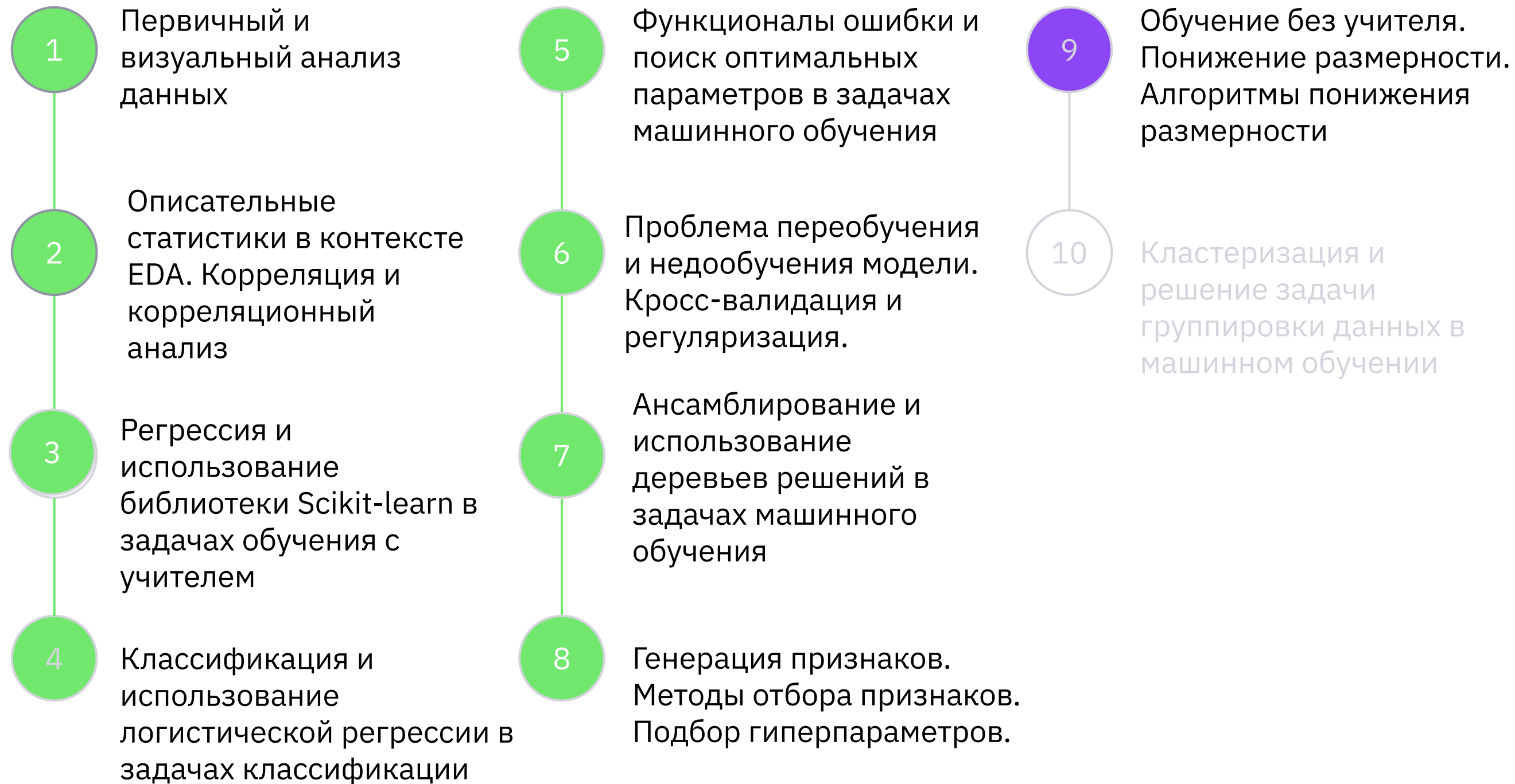
Преподаватель в GeekBrains, Нетология, Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям



План курса



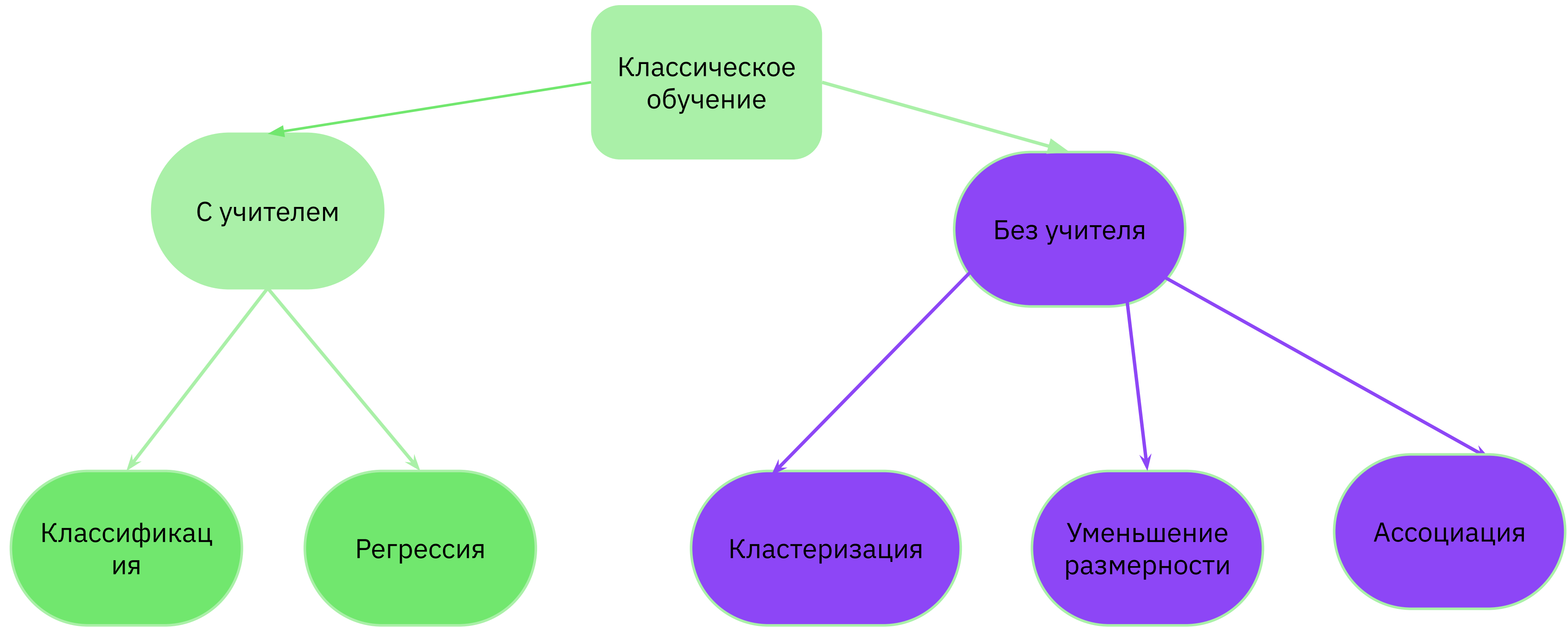


Что будет на уроке сегодня

- ? Что такое обучение без учителя
- ? Понижение размерности как метод обучения без учителя
- ? Линейное понижение размерности
- ? Нелинейное понижение размерности
- ? Случайное понижение размерности



Определение обучения без учителя.





Определение обучения без учителя.

Обучение без учителя – это метод машинного обучения, который позволяет модели самостоятельно находить закономерности и структуры в данных без явного присутствия учителя



позволяет работать с неразмеченными данными



выявления скрытых структур и зависимостей в данных



позволяет сократить количество признаков, несущих информацию в данных



можно использовать для генерации новых данных на основе имеющихся



отсутствие явной целевой переменной



невозможность оценить качество модели



неоднозначность интерпретации результатов

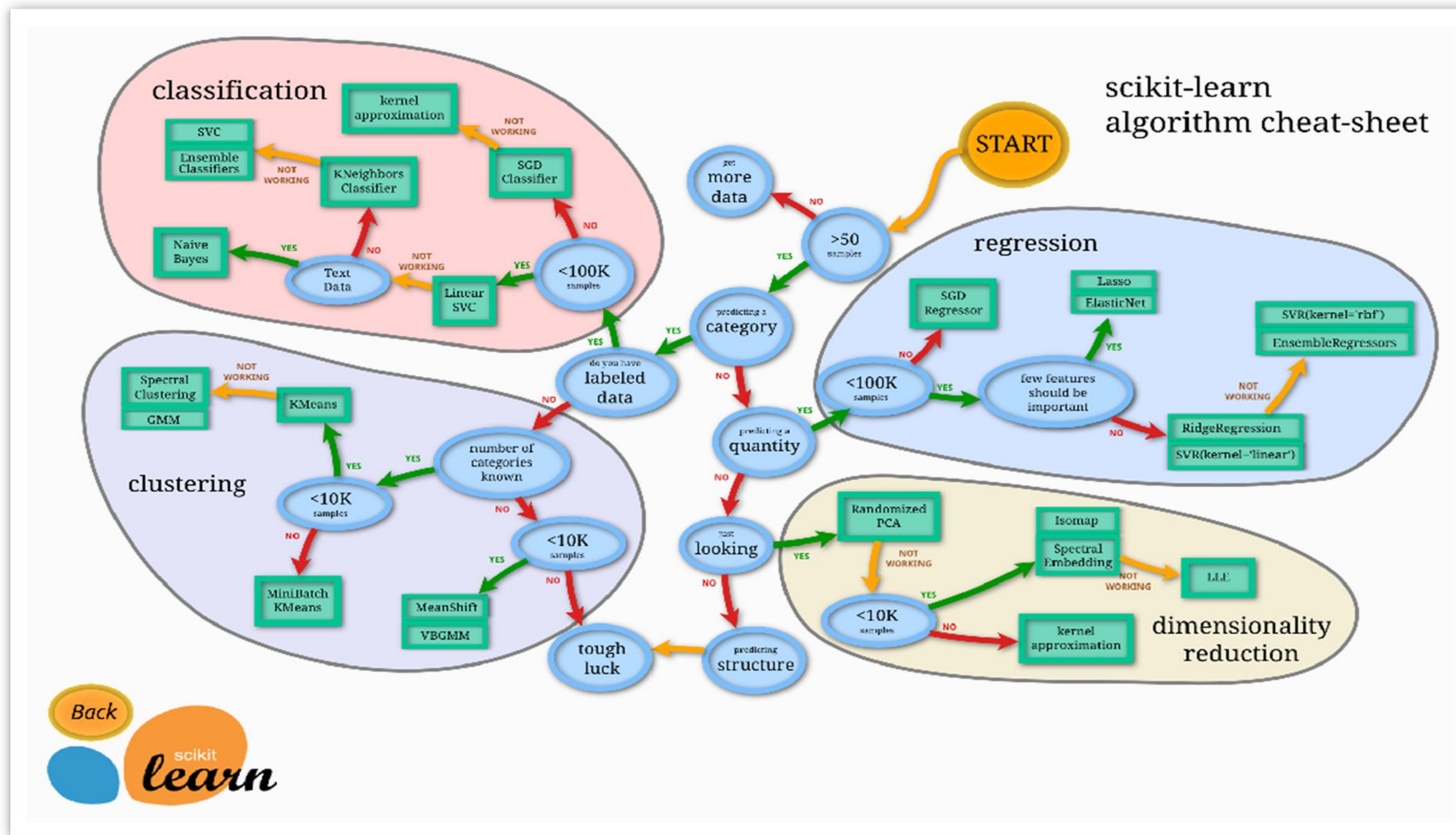


требуется больше вычислительных ресурсов



лимитированный контроль над процессом обучения

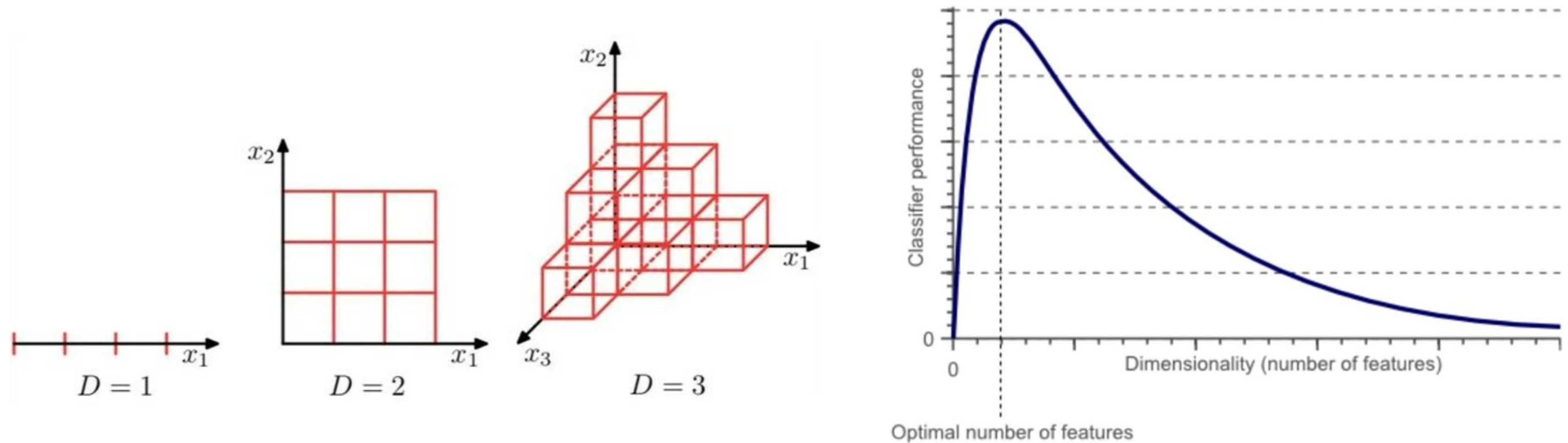
Определение обучения без учителя.





Понижение размерности как метод обучения без учителя

Проклятие размерности — это явление в машинном обучении, когда количество признаков (размерность) в данных намного больше, чем количество наблюдений (количество примеров), что может привести к переобучению модели.





Понижение размерности как метод обучения без учителя

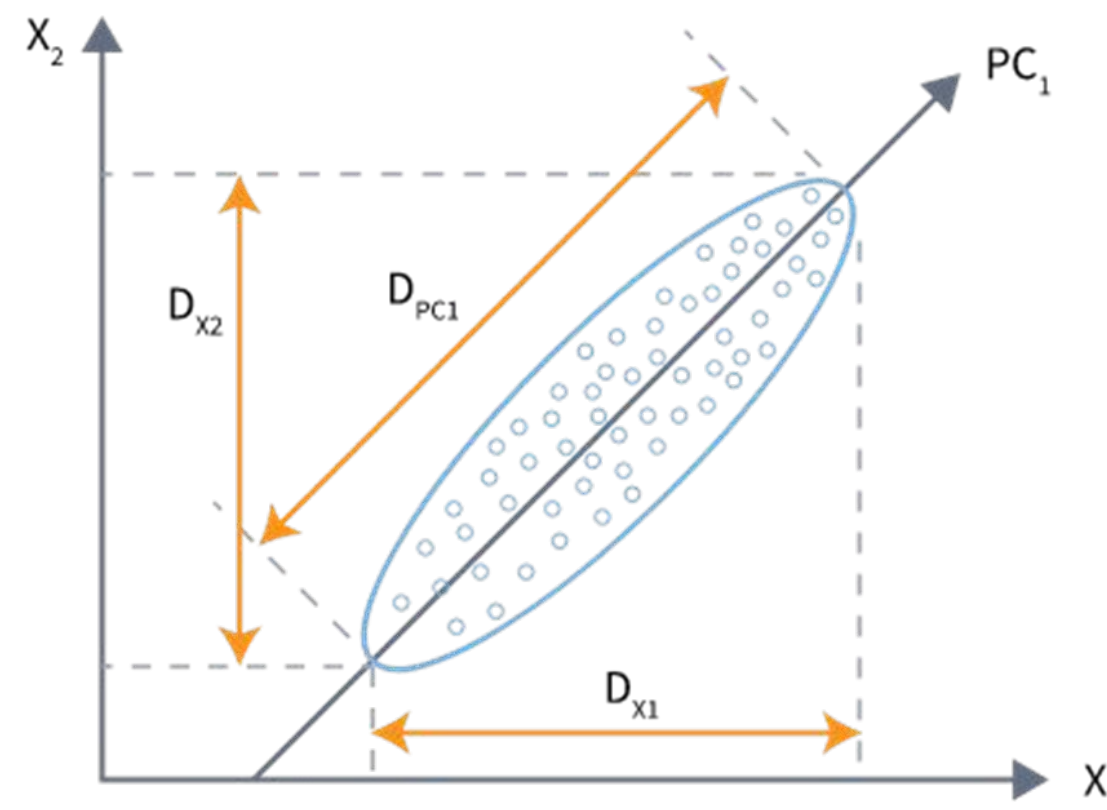
Выбор наилучшего метода понижения размерности зависит от множества факторов, включая специфику данных, вычислительные ресурсы и цели анализа.

1. Структура данных
2. Размерность и объём
3. Интерпретируемость
4. Необходимость визуализации
5. Устойчивость к шуму



Метод главных компонент (PCA)

Метод главных компонент (PCA) – это статистическая процедура, которая использует ортогональное преобразование для перевода набора возможно коррелированных переменных в набор значений линейно некоррелированных переменных, называемых главными компонентами





Метод главных компонент (РСА)

Алгоритмическая реализация РСА включает следующие шаги:

1. Центрирование данных путем вычитания среднего каждого признака
2. Вычисление ковариационной матрицы

$$C = \frac{1}{n-1} (X - \mu)^T (X - \mu)$$

где X — центрированная матрица данных,
 μ — вектор средних значений признаков.
 n — количество наблюдений.
 C — ковариационная матрица.

$$\begin{matrix} n \\ \boxed{A} \\ m \end{matrix} = \begin{matrix} m \\ \boxed{U} \\ m \end{matrix} \begin{matrix} n \\ \boxed{\Sigma} \\ m \end{matrix} \begin{matrix} n \\ \boxed{V^T} \\ n \end{matrix}$$



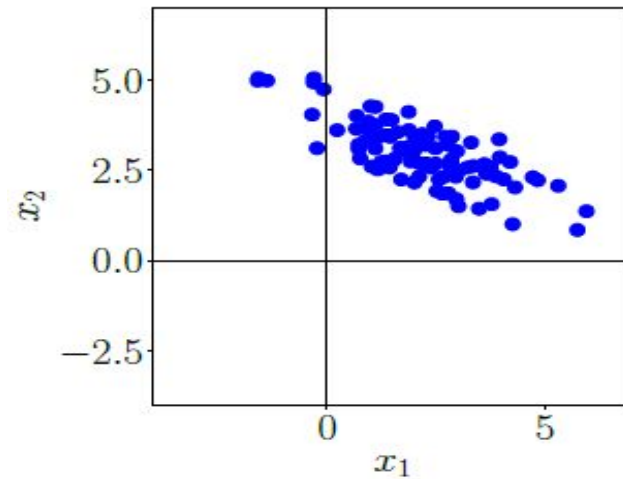
Метод главных компонент (РСА)

Алгоритмическая реализация РСА включает следующие шаги:

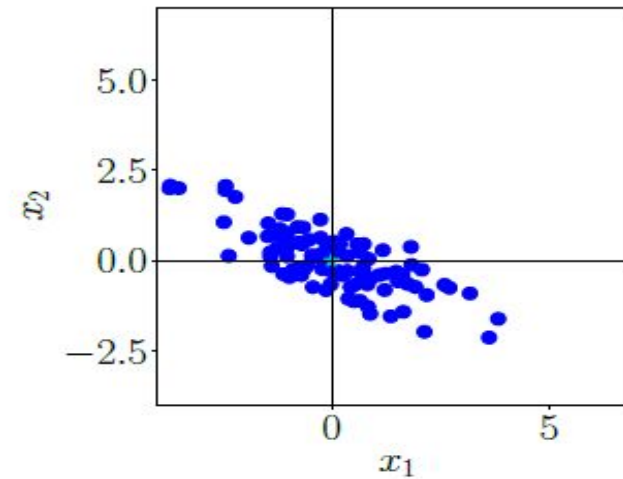
3. Нахождение собственных значений и собственных векторов ковариационной матрицы.
4. Сортировка собственных векторов по убыванию соответствующих собственных значений
5. Проекция данных на первые k главных компонент для понижения размерности



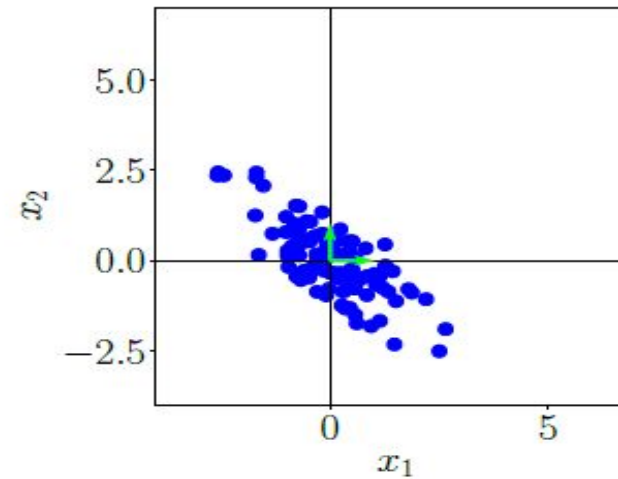
Метод главных компонент (PCA)



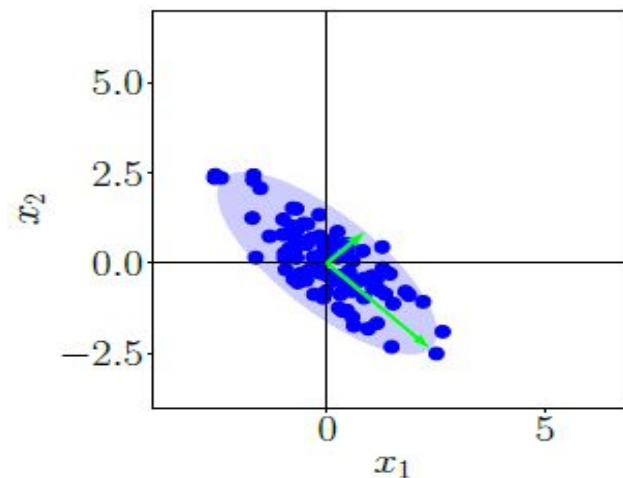
а) Исходный набор данных



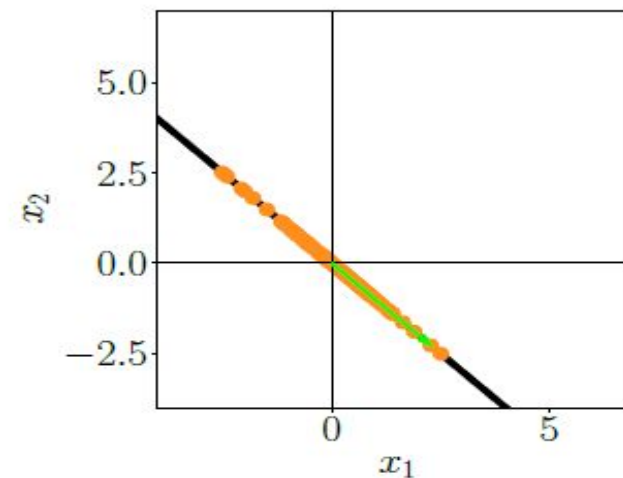
б) Шаг 1: Центрируем, вычитая из каждой точки среднее значение



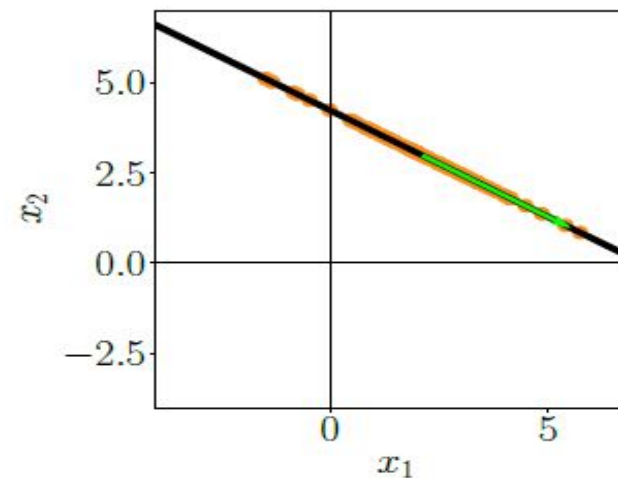
в) Шаг 2: Делим на стандартное отклонение, чтобы данные не зависели от единицы измерения. Теперь дисперсия данных по каждой оси равна 1.



г) Шаг 3: Вычисляем собственные значения и собственные векторы (стрелки) матрицы ковариации данных (эллипс)



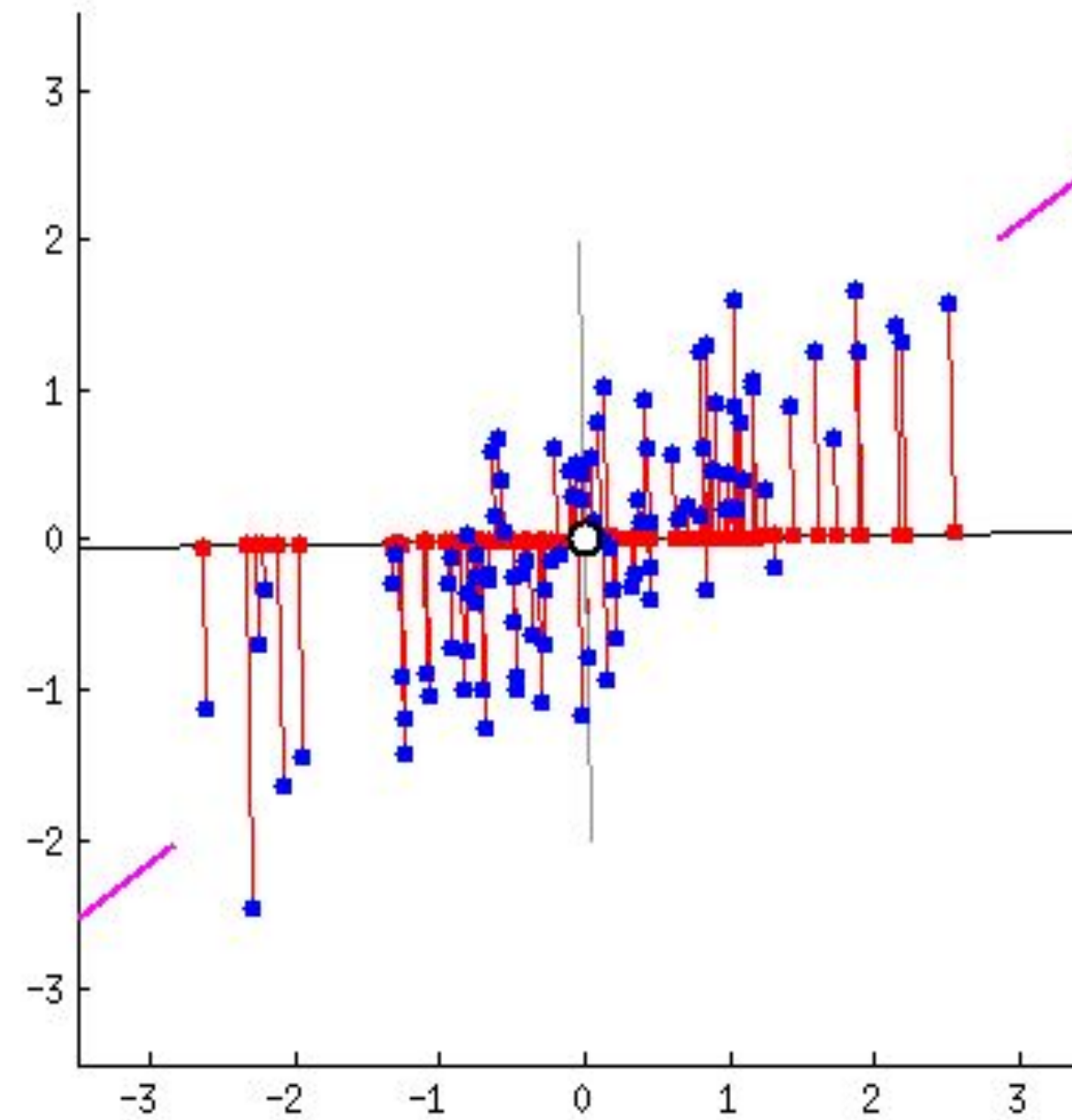
д) Шаг 4: Проецируем данные в подпространство главных компонент



е) Обращаем стандартизацию и перемещаем проецированные данные в исходное пространство

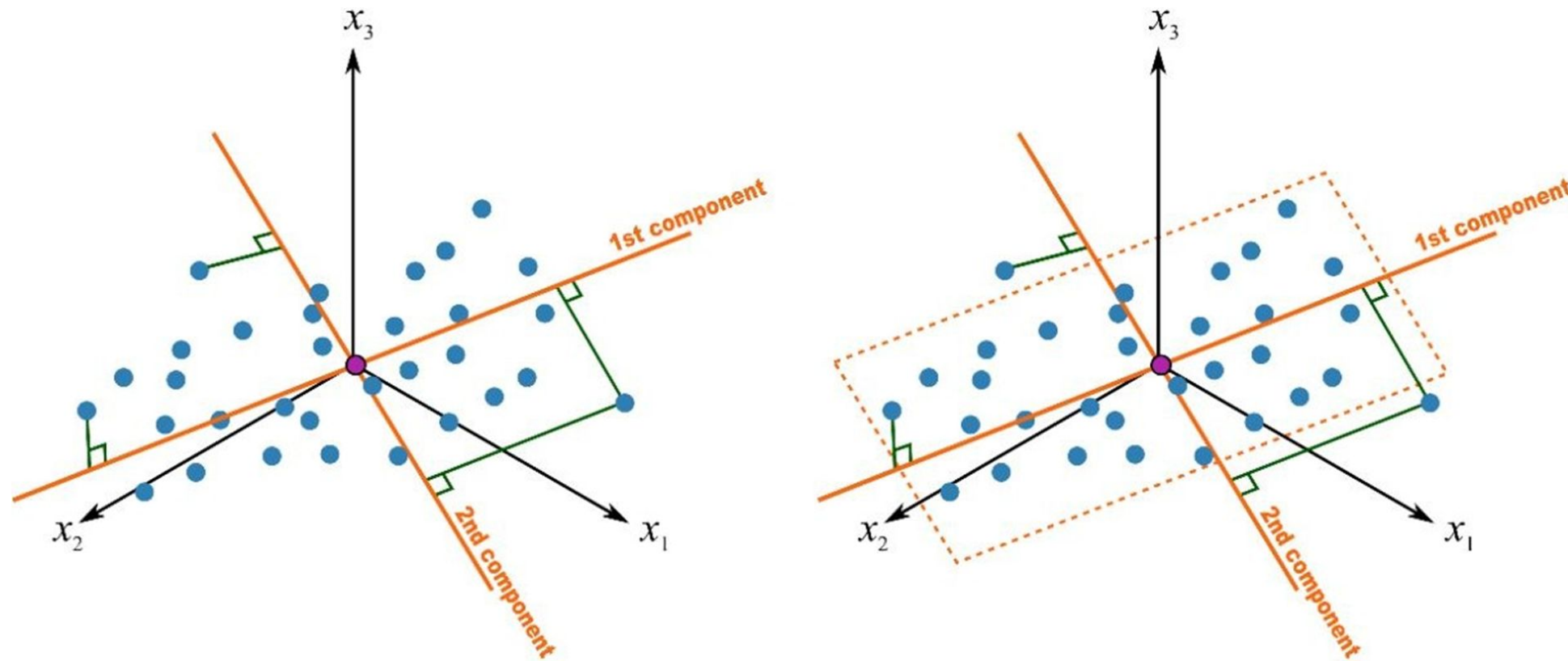


Метод главных компонент (PCA)





Метод главных компонент (PCA)









```
# Процент объясненной дисперсии для каждой компоненты
explained_variance_ratio = np.cumsum(explained_variance)

# Находим количество компонент, объясняющих не менее 95% дисперсии
num_components = np.argmax(explained_variance_ratio >= 0.95) + 1
```



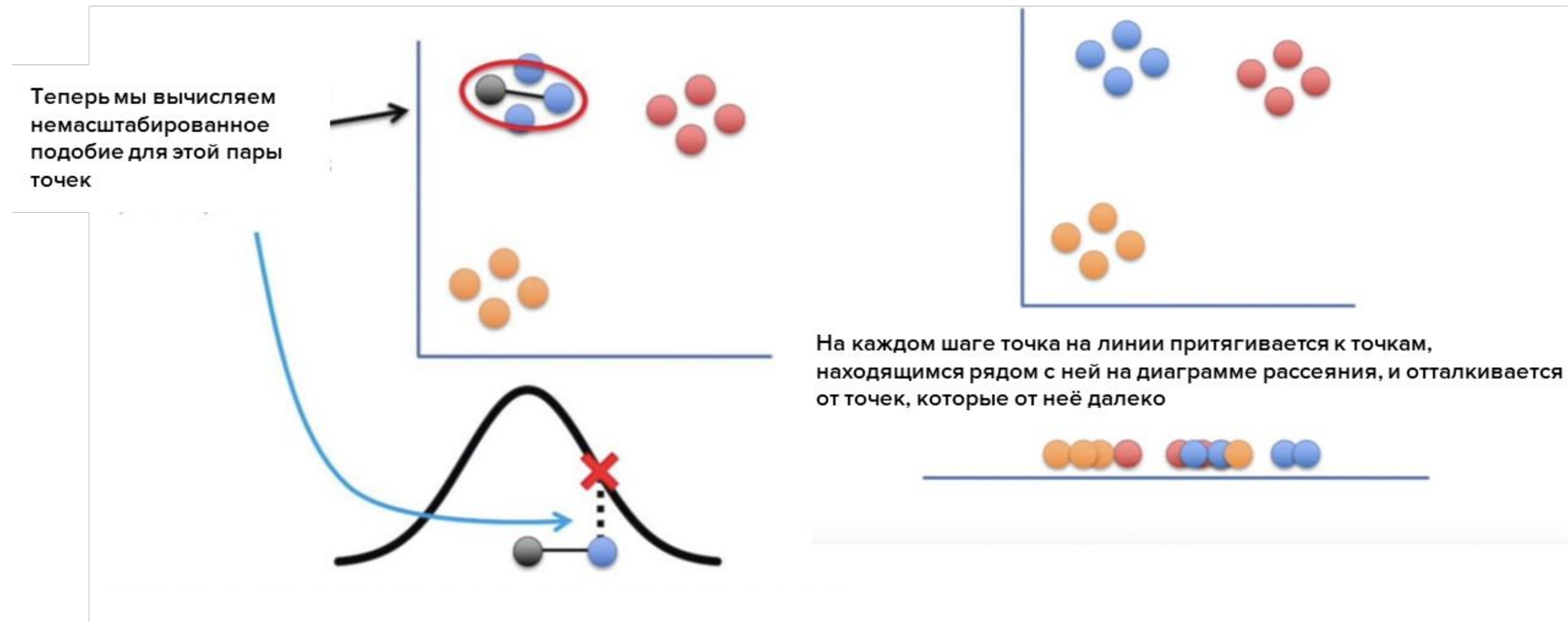
Метод главных компонент (РСА)

- | | | | |
|---|--------------------------------|---|------------------------------------|
|  | Уменьшение избыточности данных |  | Чувствительность к масштабированию |
|  | Улучшение визуализации |  | Потеря интерпретируемости |
|  | Оптимизация вычислений |  | Предположение линейности |



Метод t-SNE

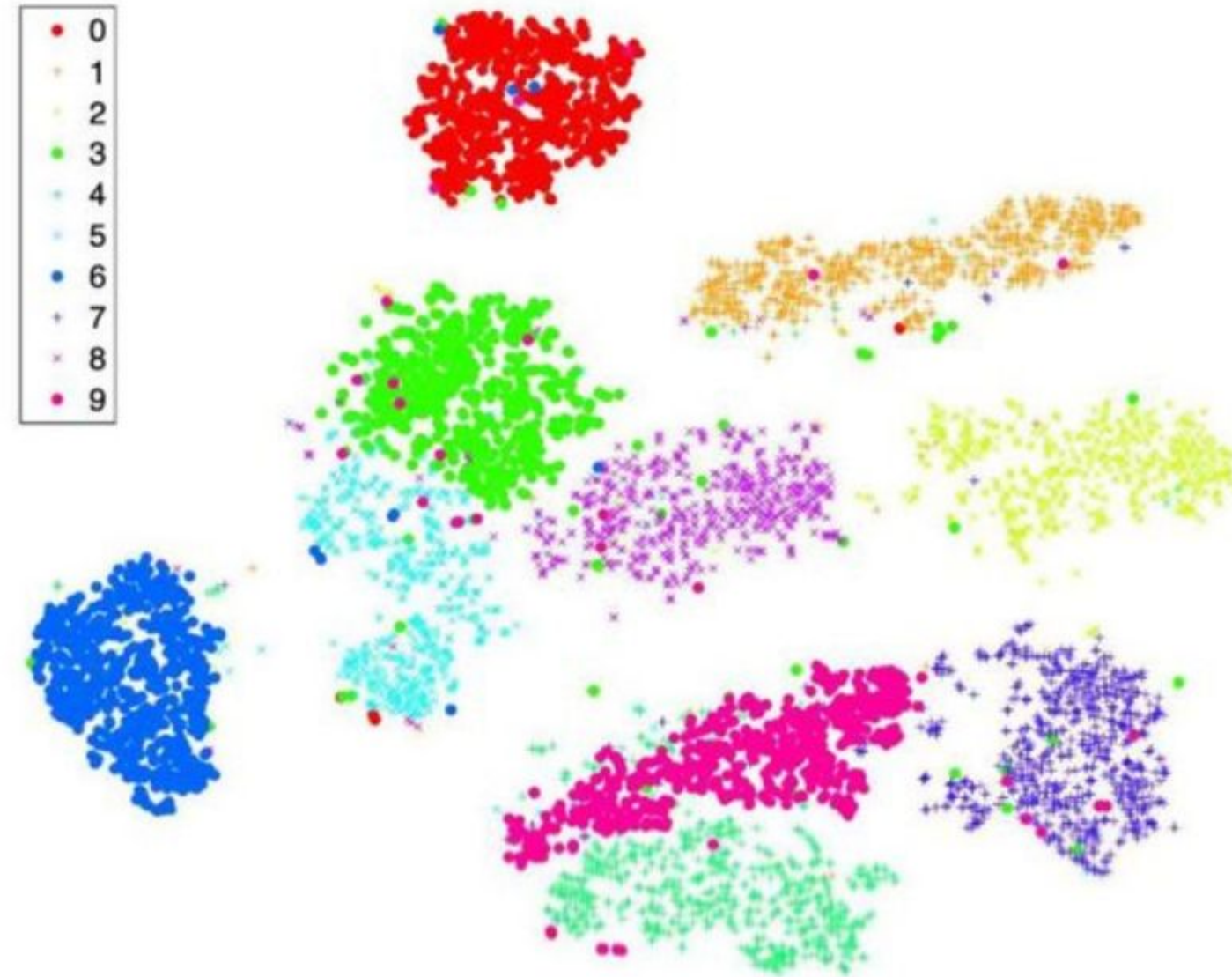
t-SNE (t-distributed Stochastic Neighbor Embedding) – это метод снижения размерности и визуализации данных, который позволяет сохранить локальные структуры данных и обнаруживать нелинейные зависимости.





Метод t-SNE

t-SNE on [MNIST dataset](#)





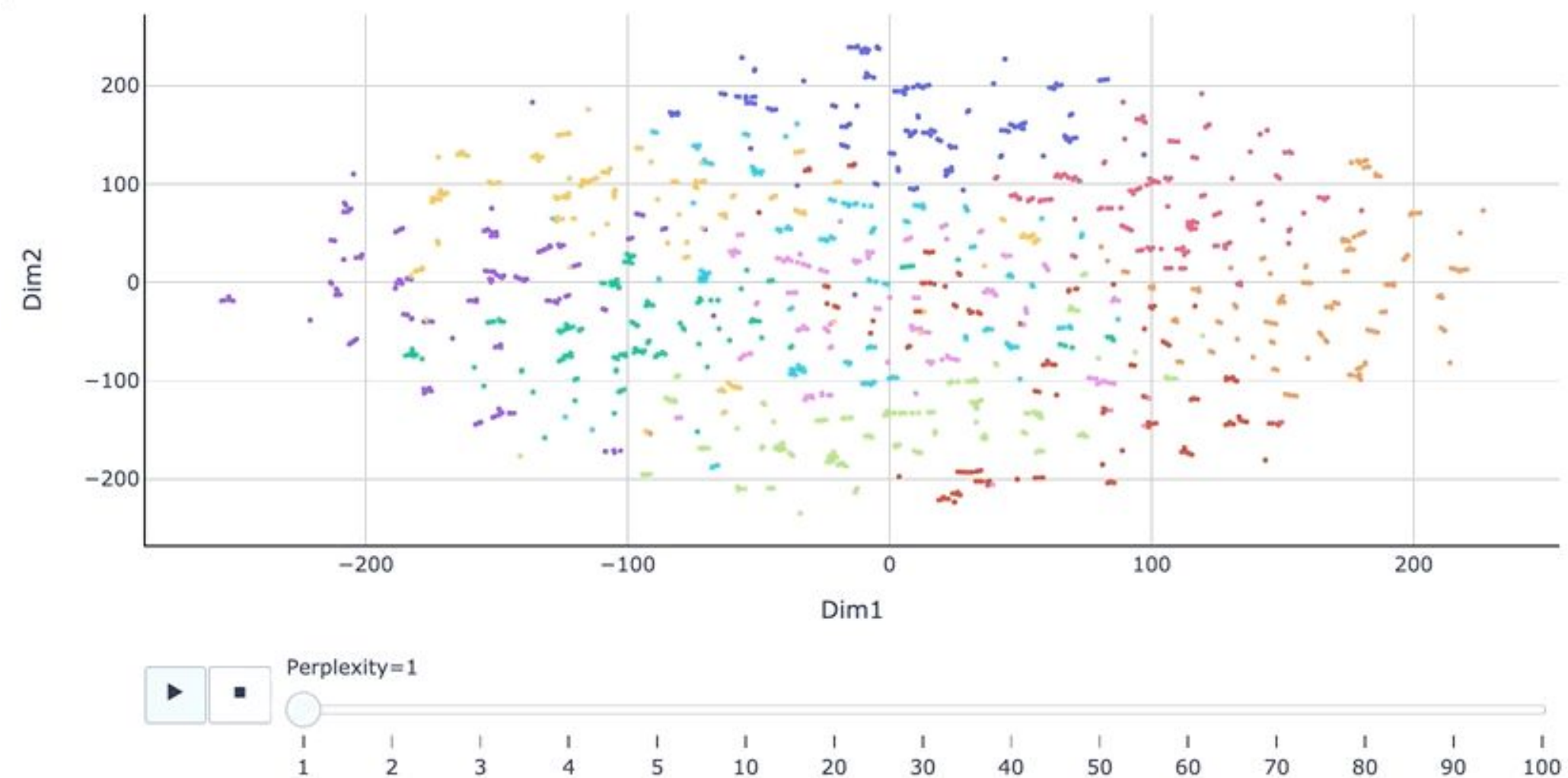
Метод t-SNE

9



Метод t-SNE

Один из ключевых параметров - это perplexity, который регулирует баланс между сохранением локальной и глобальной структуры данных.





Метод t-SNE



Сохранение глобальной структуры



Учет сложной нелинейной зависимости



Компактные и удобочитаемые визуализации



Гибкие параметры



Вычислительная сложность



Стохастические результаты



Не учитывает другие признаки



Неприменимость для временных данных



Случайное понижение размерности (Random Projection)

Случайное понижение размерности (Random Projection) – это метод снижения размерности данных, который основывается на идее проецирования исходных данных на случайно выбранные подпространства.



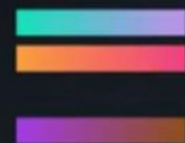


Случайное понижение размерности (Random Projection)

from d dimensions to k dimensions

$$X_{k \times N}^{RP} = R_{k \times d} X_{d \times N}$$

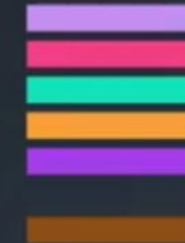
$$\begin{bmatrix} X_{11}^{RP} & X_{12}^{RP} & \dots & X_{1n}^{RP} \\ X_{21}^{RP} & X_{22}^{RP} & \dots & X_{2n}^{RP} \\ \dots & \dots & \dots & \dots \\ X_{k1}^{RP} & X_{k2}^{RP} & \dots & X_{kn}^{RP} \end{bmatrix}$$



=

$$\begin{bmatrix} r_{11} & r_{12} & \dots & r_{1d} \\ r_{21} & r_{22} & \dots & r_{2d} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kd} \end{bmatrix}$$

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ X_{31} & X_{32} & \dots & X_{3n} \\ X_{41} & X_{42} & \dots & X_{4n} \\ \dots & \dots & \dots & \dots \\ X_{d1} & X_{d2} & \dots & X_{dn} \end{bmatrix}$$





Случайное понижение размерности (Random Projection)



Простота и эффективность



Сохранение структуры данных



Универсальность



Потеря точности



Невозможность восстановления исходных данных



Зависимость от параметров

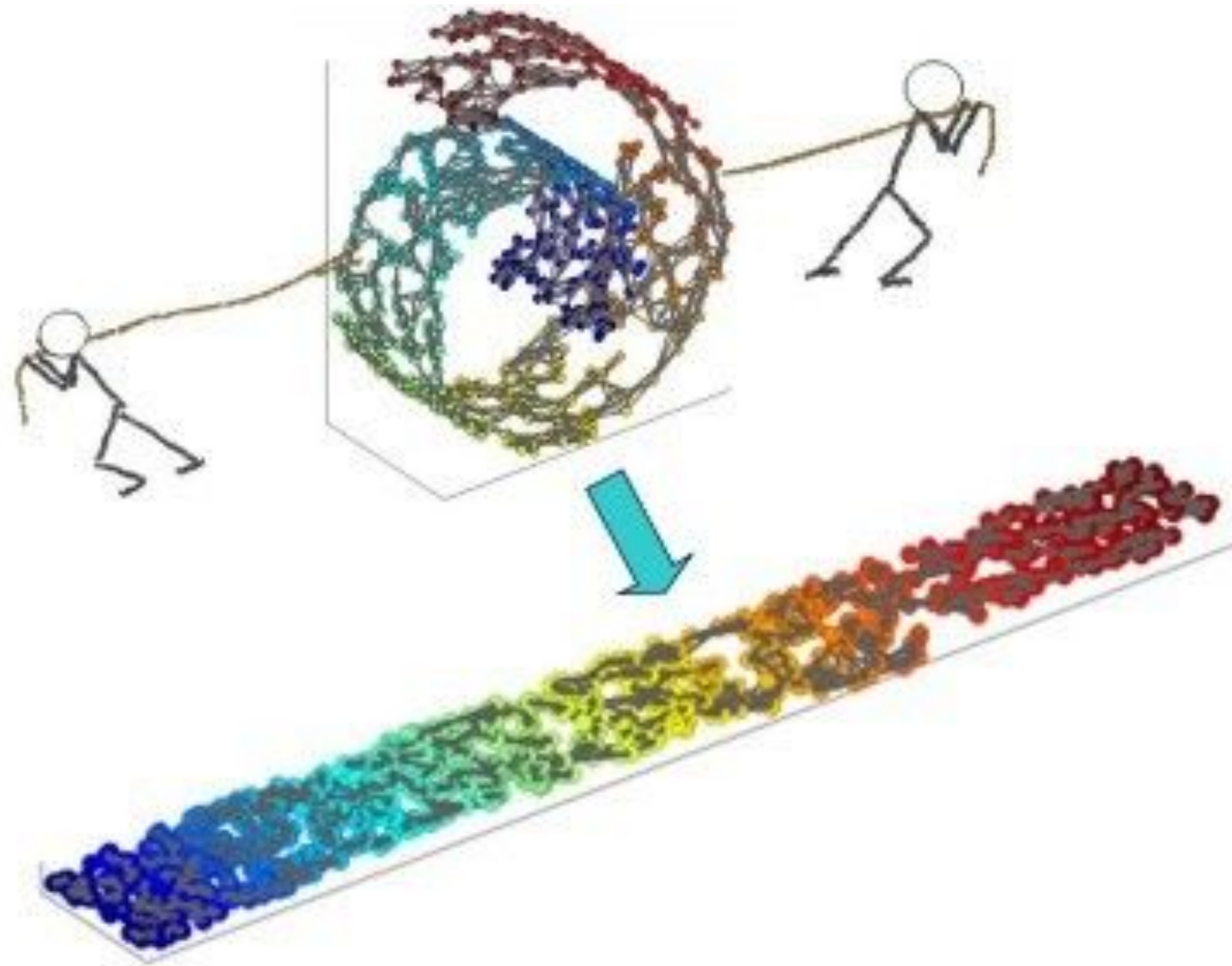


Влияние шума



LLE(Locally Linear Embedding)

LLE(Locally Linear Embedding) - это неконтролируемый подход, предназначенный для преобразования данных из исходного многомерного пространства в представление с меньшей размерностью, при этом стремясь сохранить основные геометрические характеристики лежащей в основе нелинейной структуры объектов





LLE(Locally Linear Embedding)

Вот поэтапный алгоритм LLE:

1. Шаг 1: Вычисление соседей

- задается параметр k - количество ближайших соседей для каждого объекта.
- Для каждого объекта находятся его k ближайших соседей с помощью метода поиска ближайших соседей

2. Шаг 2: Восстановление локальных весов

- Для каждого объекта находится наилучшее линейное приближение соседей.
- Для этого строится матрица весов W размером $k \times k$, где каждый столбец содержит координаты наилучшего линейного приближения для соответствующего соседа.

3. Шаг 3: Вычисление глобальных представлений

- Для каждого объекта находится его глобальное представление в низкоразмерном пространстве.

4. Шаг 4: Визуализация и анализ данных

- Полученные глобальные представления могут быть использованы для визуализации данных в низкоразмерном пространстве или для выполнения других анализов, таких как классификация или кластеризация.



LLE(Locally Linear Embedding)



Сохраняет локальную структуру данных



Сохраняет нелинейных отношений между данными



Работает хорошо на данных со сложной топологией



Требует достаточно большую выборку данных



Неустойчив к шуму в данных



Подбор параметра k - может быть нетривиальной задачей



Оценка качества методов понижения размерности

1. Объяснимая дисперсия (explained variance):

- Для метода PCA (Principal Component Analysis) можно использовать атрибут explained variance ratio _ после подгонки модели.
- Для метода t-SNE, может быть оценена сравнением дисперсии исходных данных и дисперсии данных после понижения размерности.

2. Сохранение информации восстановленных данных:

- Использование аппроксимационного обратного преобразования
- Доступ к исходным данным
- Контекстная информация



Итоги

Тема понижения размерности имеет практическую значимость в контексте машинного обучения по нескольким причинам:

- ✓ Сокращение вычислительной сложности
- ✓ Увеличение производительности алгоритмов машинного обучения
- ✓ Визуализация данных
- ✓ Избавление от шума и избыточности
- ✓ Улучшение интерпретируемости



Спасибо за внимание

