



Л е к ц и я № 2

Типы данных и методы их обработки

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Цифровые технологии в биологии»

П л а н л е к ц и и

- Что такое «Тип данных»
- Скалярные и нескаларные данные
- Непрерывные и дискретные данные
- Бинарные и порядковые данные
- Текстовые данные
- Методы восстановления эмпирических закономерностей
- Регрессионные методы
- Распознавание образов и классификация
- Логические методы

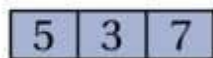
Что такое «Тип данных»

Множество допустимых значений и операций над этими значениями.

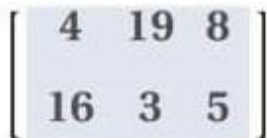
Скалярные и не скалярные данные

(11)

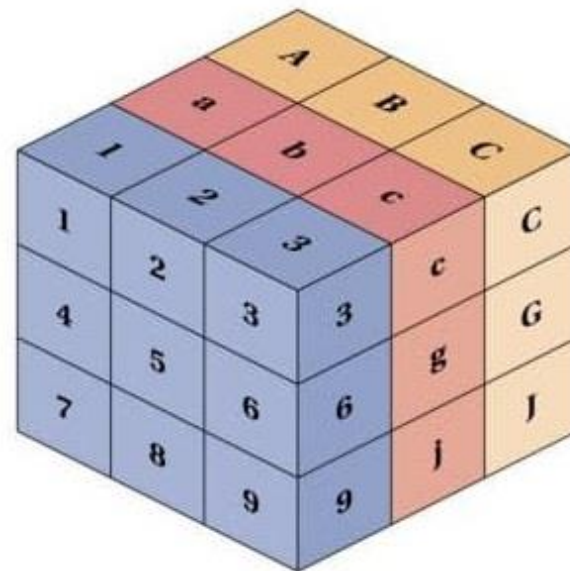
скаляр



вектор



матрица



тензор

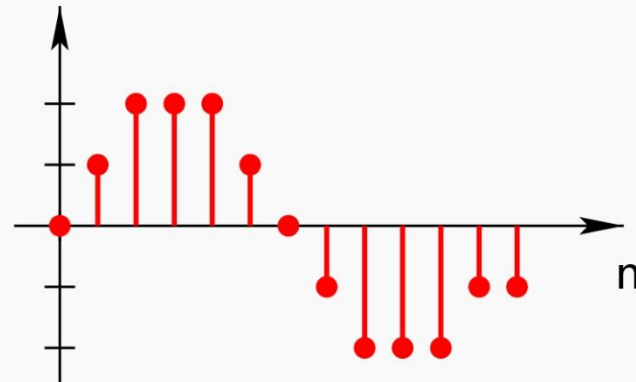
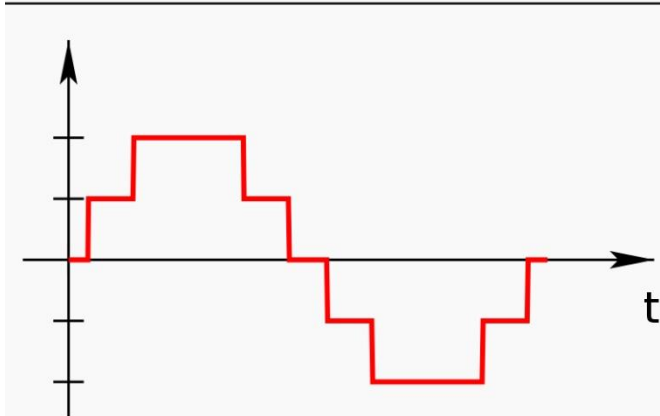
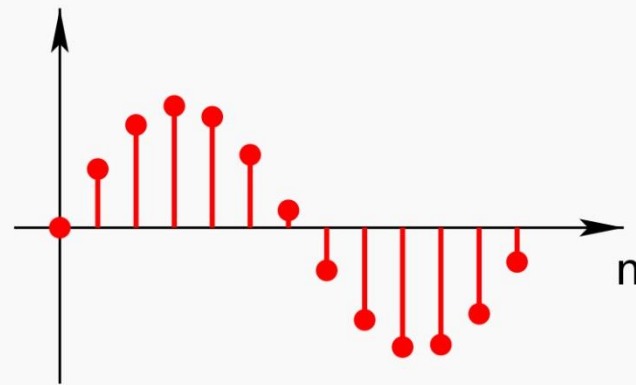
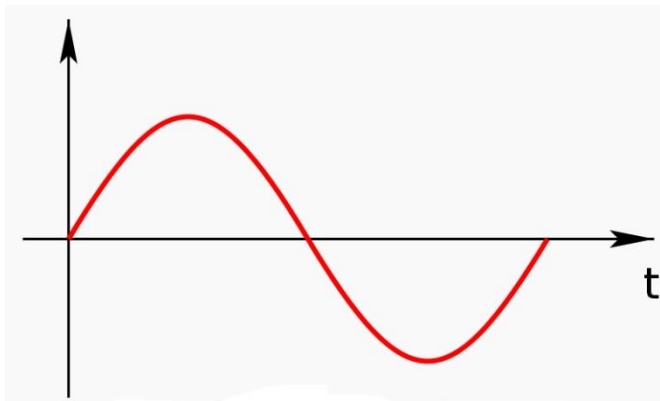
Скаляр — одно число

Вектор — два и более чисел

Матрица — таблица чисел

Тензор — кубик чисел

Непрерывные и дискретные данные



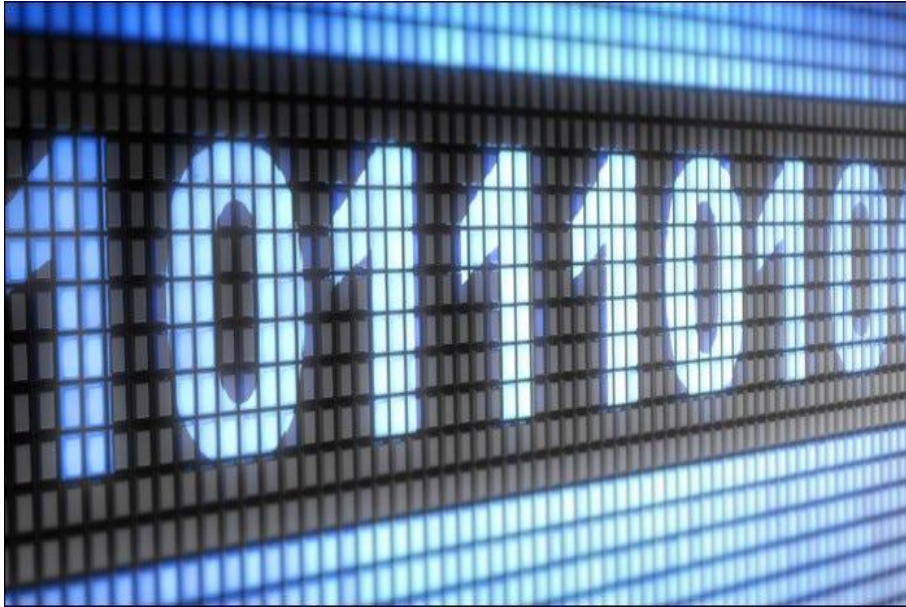
Непрерывные данные

Принимают любые значения
в некотором интервале

Дискретные данные

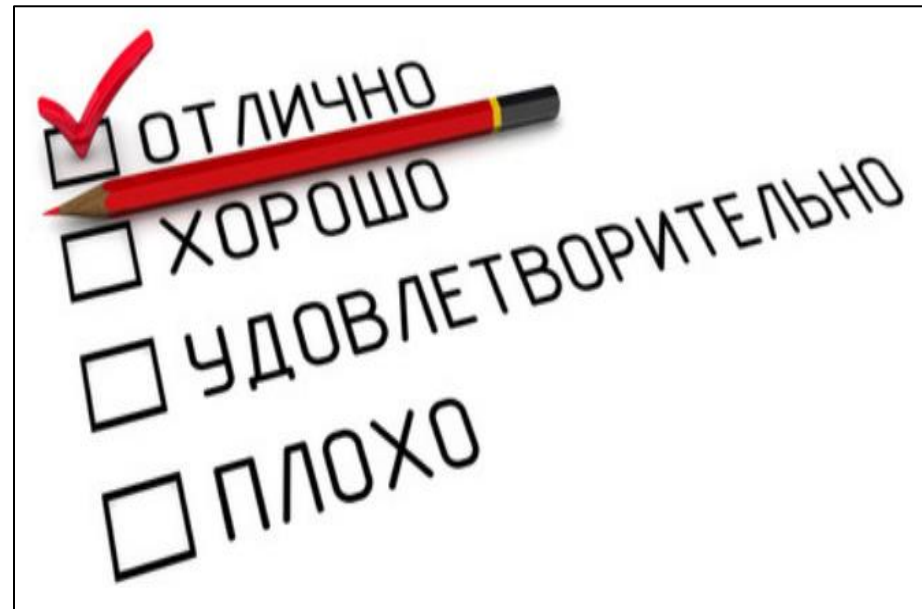
Принимают ограниченный
набор значений

Бинарные и порядковые данные

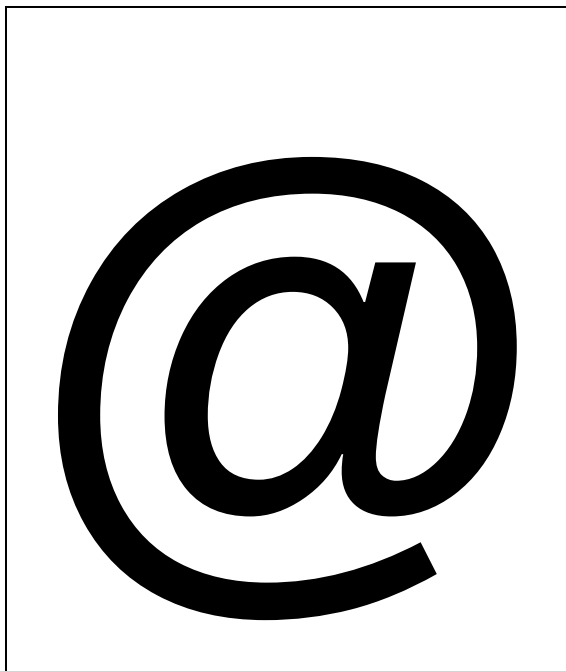


Бинарные данные
Принимают только
два значения

Порядковые данные
Ранжированные, но
интервал неизвестен



Текстовые данные



Maximum 75 characters

Cius mod mod est que molor moditios doluptiis destion et que que nonsen-
dam, nectoreperro il inus molorum qui dolorias aut molo milit, sundici lore,
officiae volupta spicto molorrovit, venecer enducip icidemo dem ea et fugitem
porit, evelita sintus, odit molore peria derum, cum quid eum et pa quatemolo-
rae acilicime porendipsant ut quod quiduciunt occae doluptatur, sus a dusan-
dam quatur sus re sin prerorum elloreperis dolupicias secto cone rem facidus
nonsequiati tem quam dolorep erfero experum neceat.Ullaut rent, vit ex est
volorecation pla soluptur sam il ilibus voluptatisin re consed quia quiandi ut
quis nones quos debis archillam doluptas quidiat.Rio. Ximet, esti dolupid es
que dest, qui sum sequias mi, odi cus, iundioriatum corepudae dolorem rem
aut prae nos amus, tem quature stiassinist ommodit mintio beres dempel inte-
mol uptur, alicae erfernam dolesci nisqui cus.

Wider columns are read more slowly.

Символы
Отдельные знаки

Строки
Последовательности
знаков фиксированной
длины

Методы восстановления эмпирических закономерностей

- **Регрессионные методы**
- **Распознавание образов**
- **Обучение без учителя**
- **Обучение с учителем**
- **Методы машинного обучения**
- **Логические методы**

Регрессионные методы

- **Метод наименьших квадратов**
- **Линейная регрессия**
- **Полиномиальная регрессия**
- **Логистическая регрессия**
- **Нейросетевая регрессия**

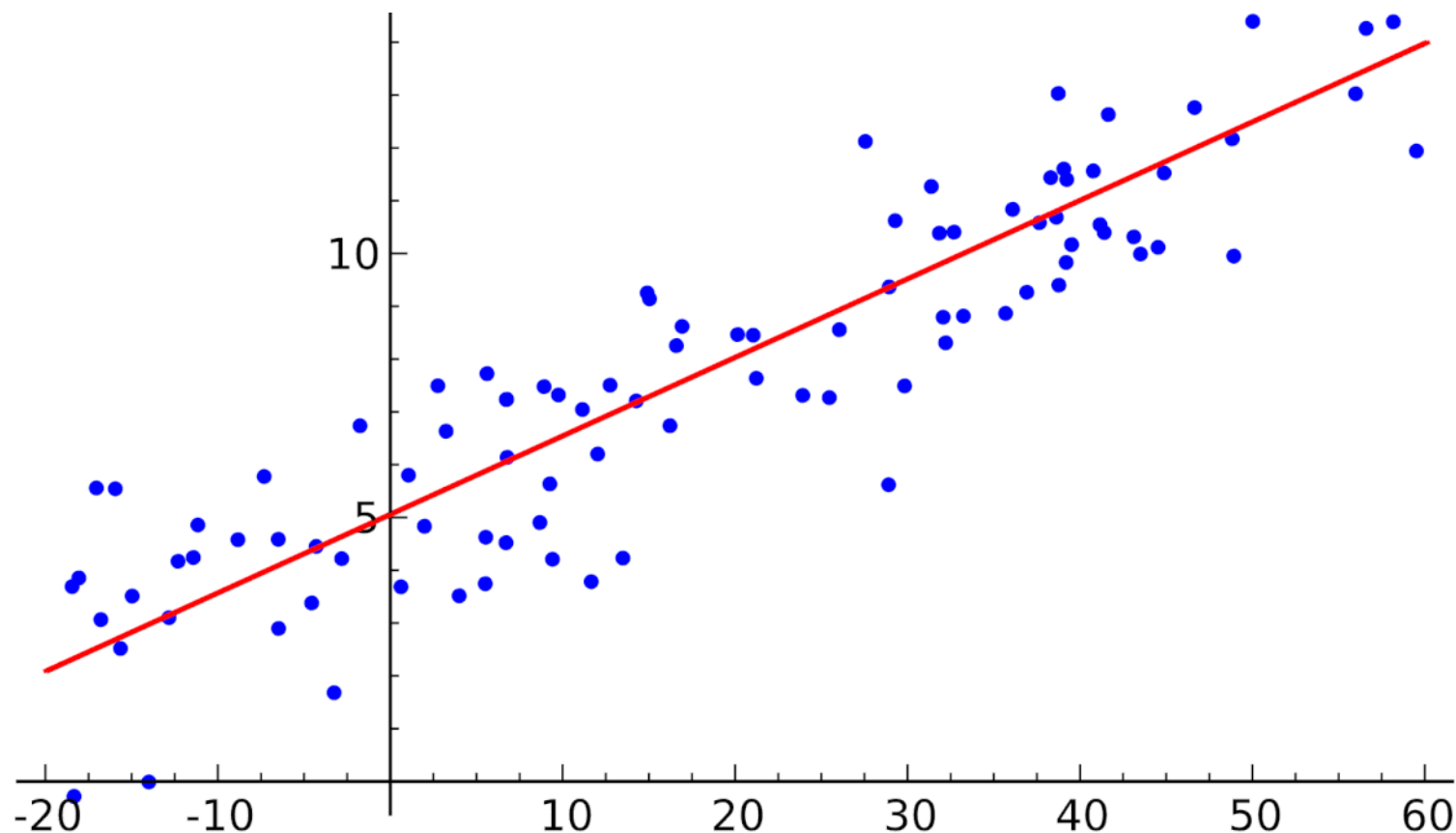
Метод наименьших квадратов

$$Q = \sum_{i=1}^N (Y_i - f_i(\theta))^2$$

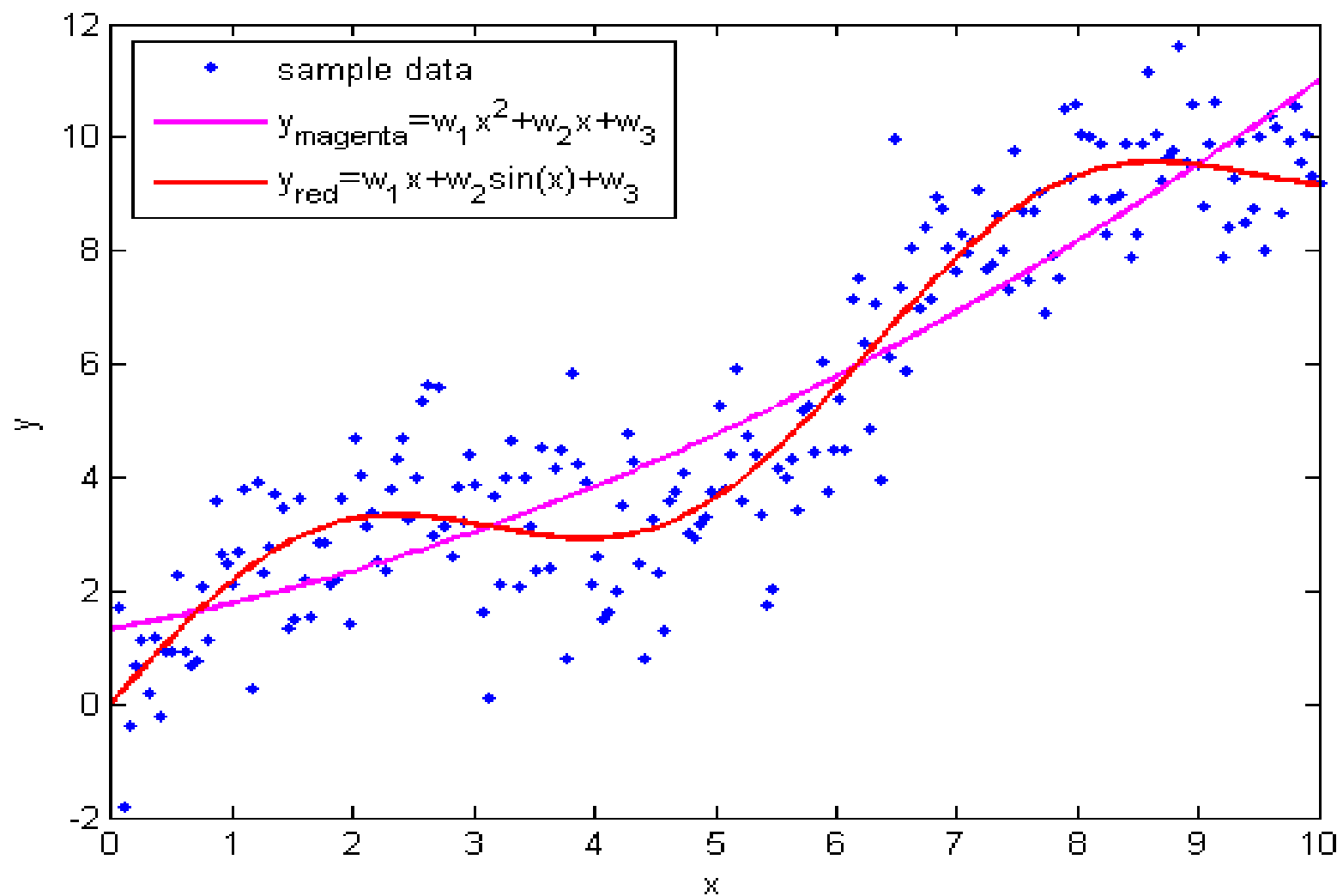
Минимизация суммы квадратов отклонений наблюдаемых значений зависимой переменной от значений, предсказанных моделью.

Простая линейная регрессия

$$Y = a + b * X$$

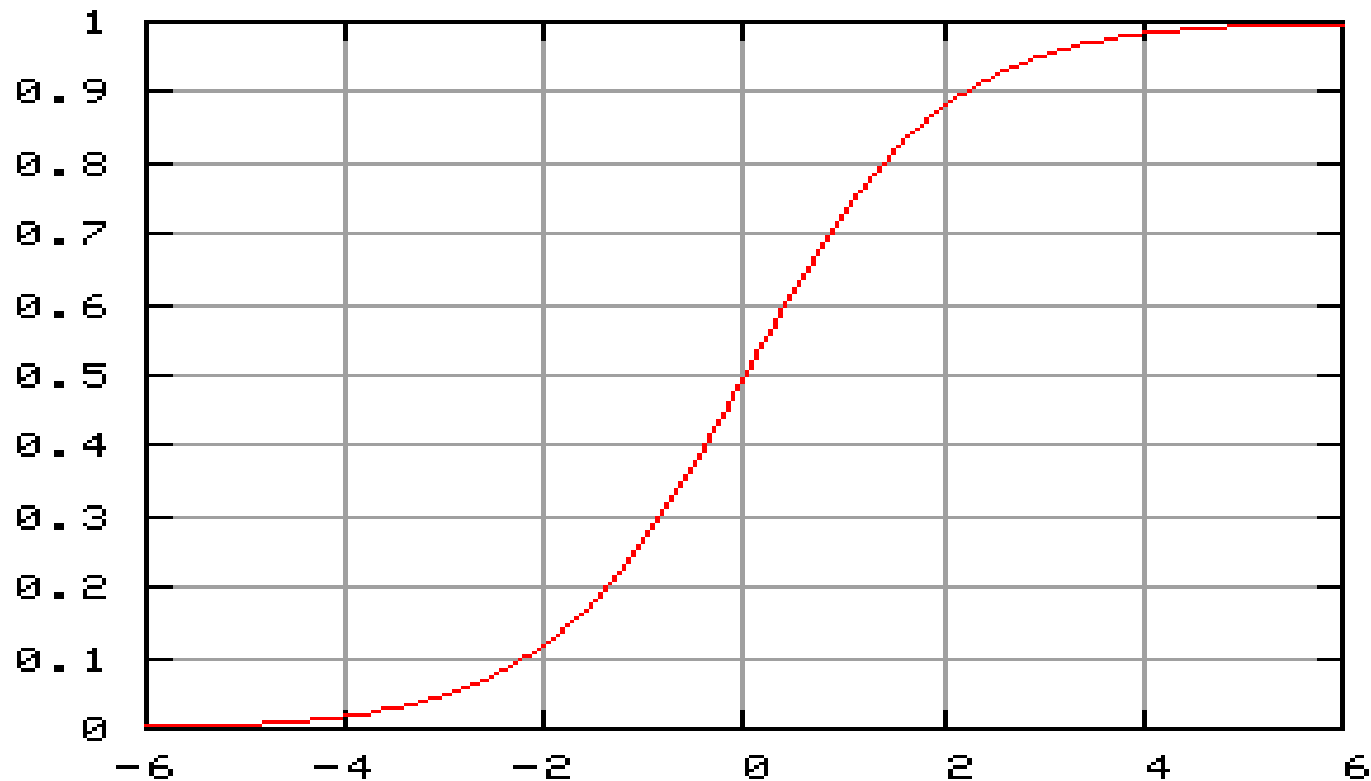


Полиномиальная регрессия

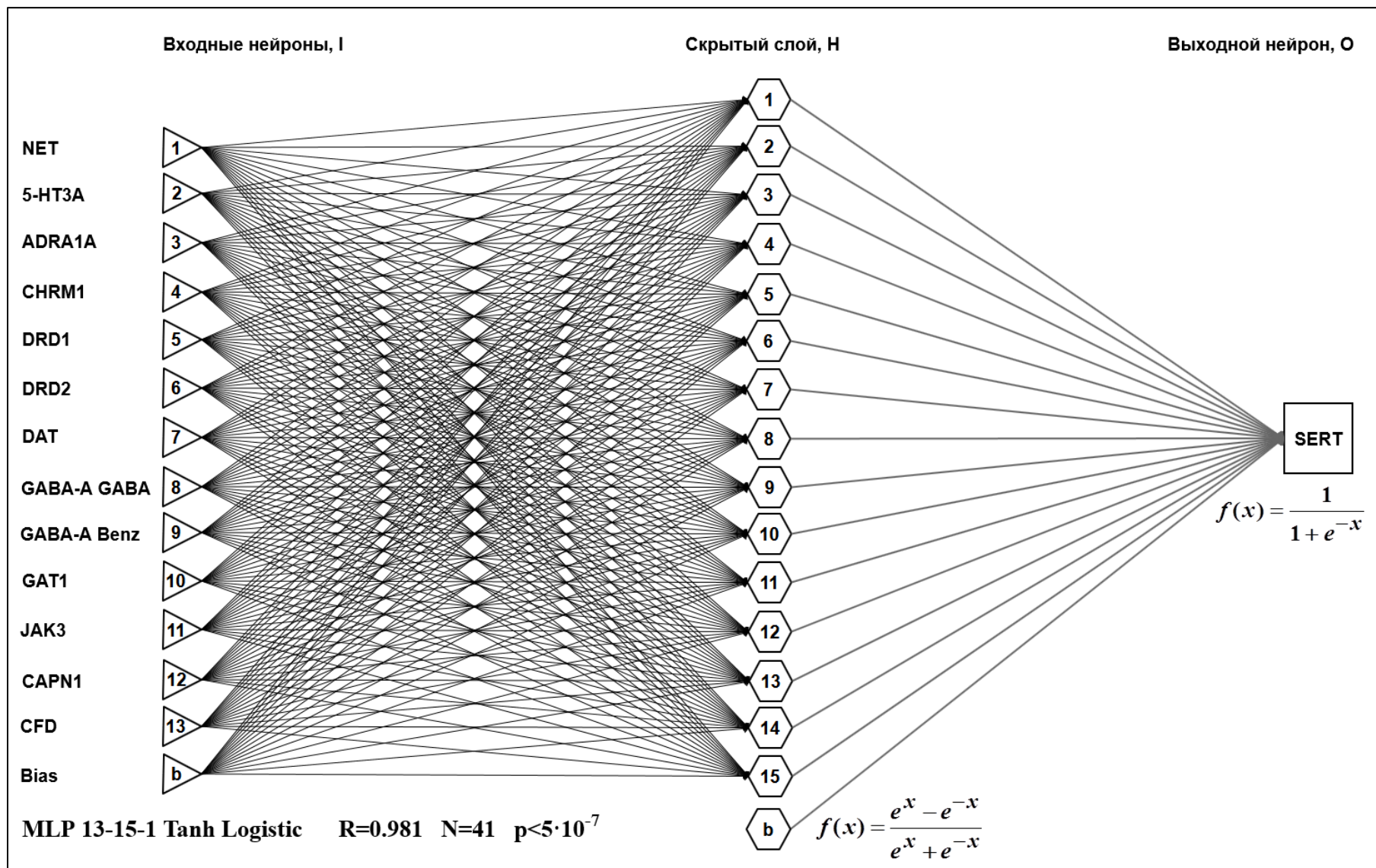


Логистическая регрессия

$$f(x) = \frac{1}{1 + e^{-x}}$$



Нейросетевая регрессия



Что такое «Образ»

Совокупность данных об объекте
или явлении, включающая
параметры и связи.

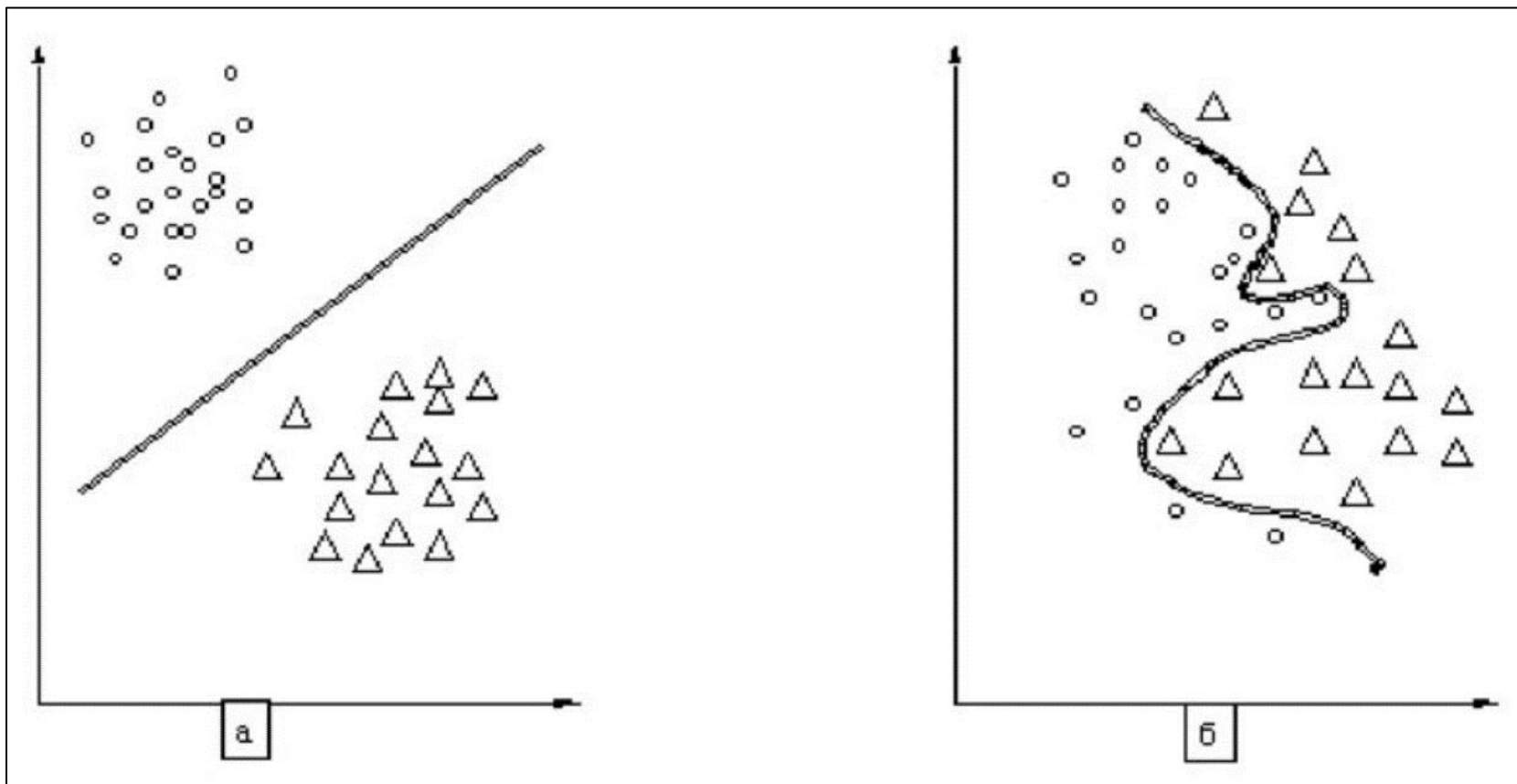
Распознавание образов

Сравнение и определение степени сходства образа данного конкретного объекта с образами других конкретных объектов или с обобщенными образами классов, в результате которой формируется рейтинг объектов или классов по убыванию сходства с распознаваемым объектом.

Гипотеза компактности (теорема о гадком утенке)

Расстояние между двумя объектами,
принадлежащими к одному и тому
же кластеру, всегда меньше
расстояния между двумя объектами,
принадлежащими к разным
кластерам.

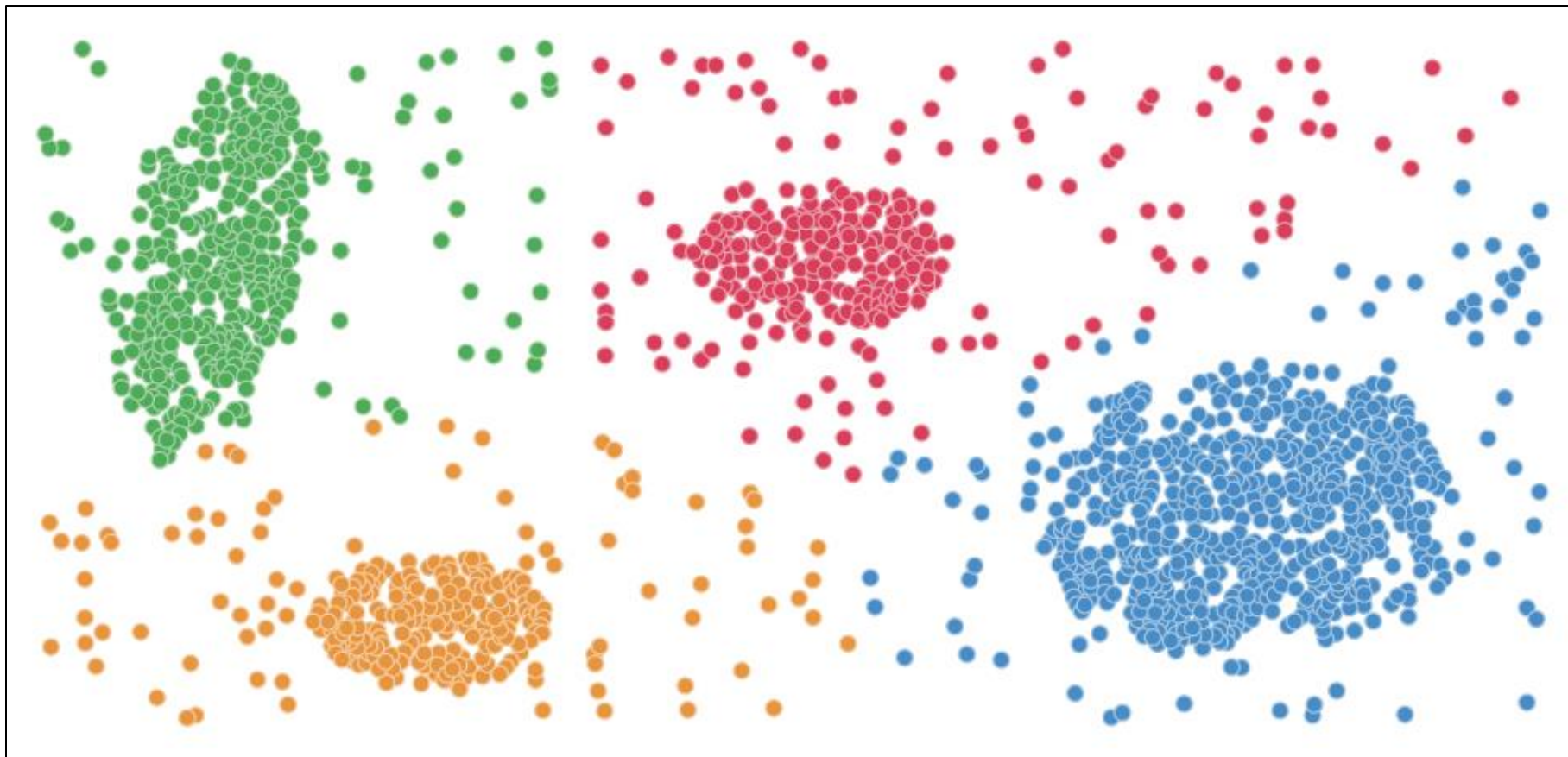
Компактные и некомпактные образы



Линейно разделимые

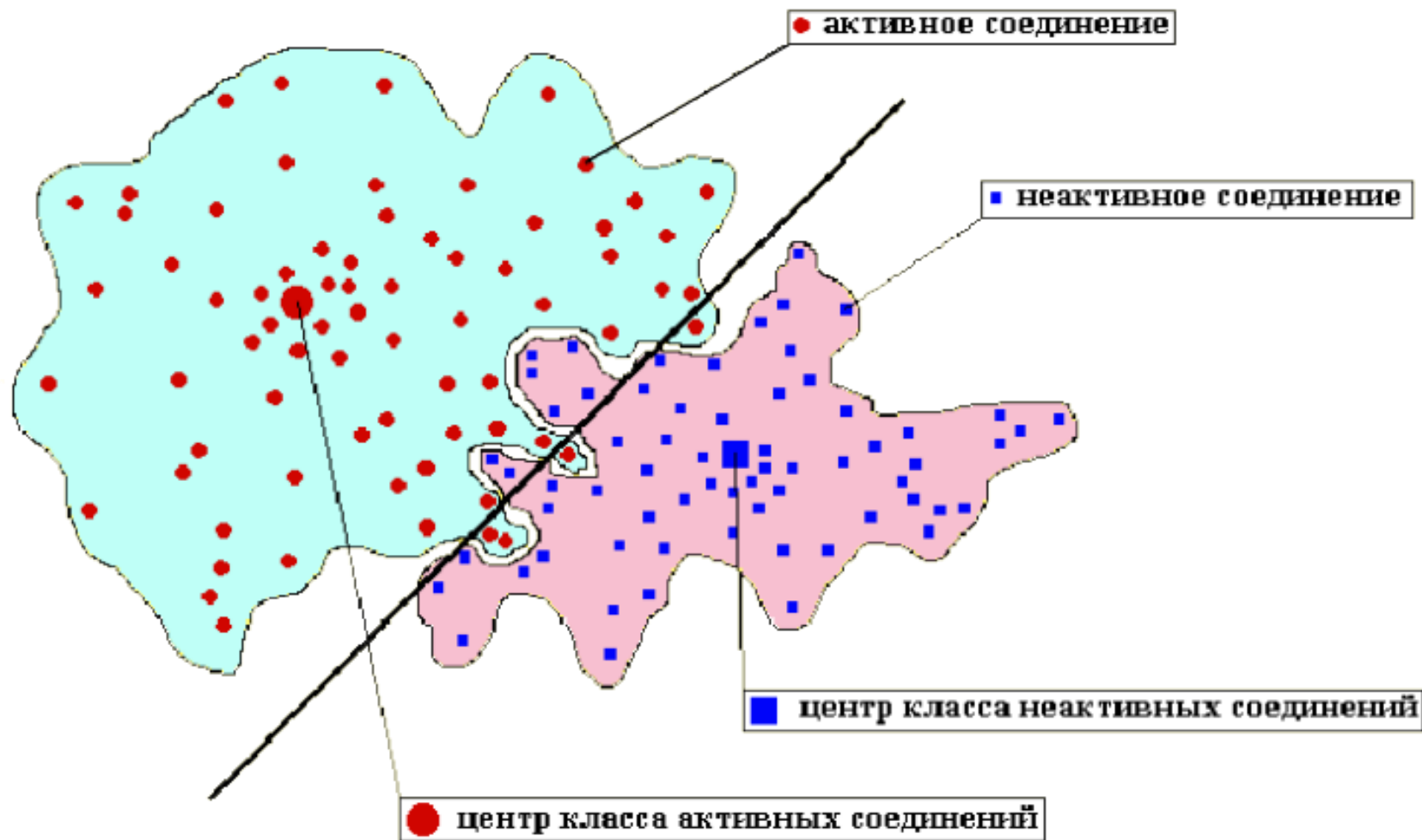
Линейно не разделимые

Обучение без учителя



Кластерный анализ
Выполняется гипотеза компактности

Обучение с учителем

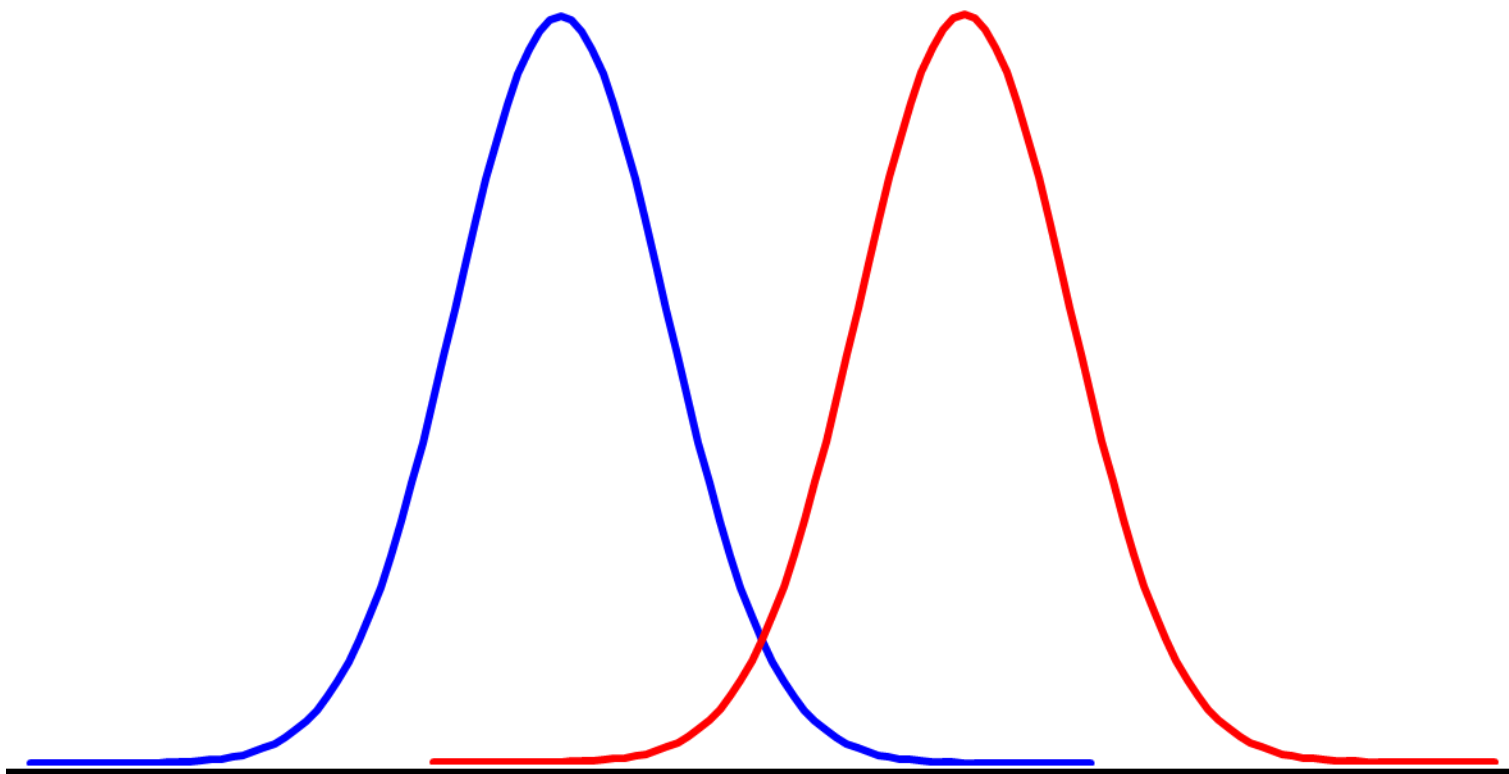


Методы классификации

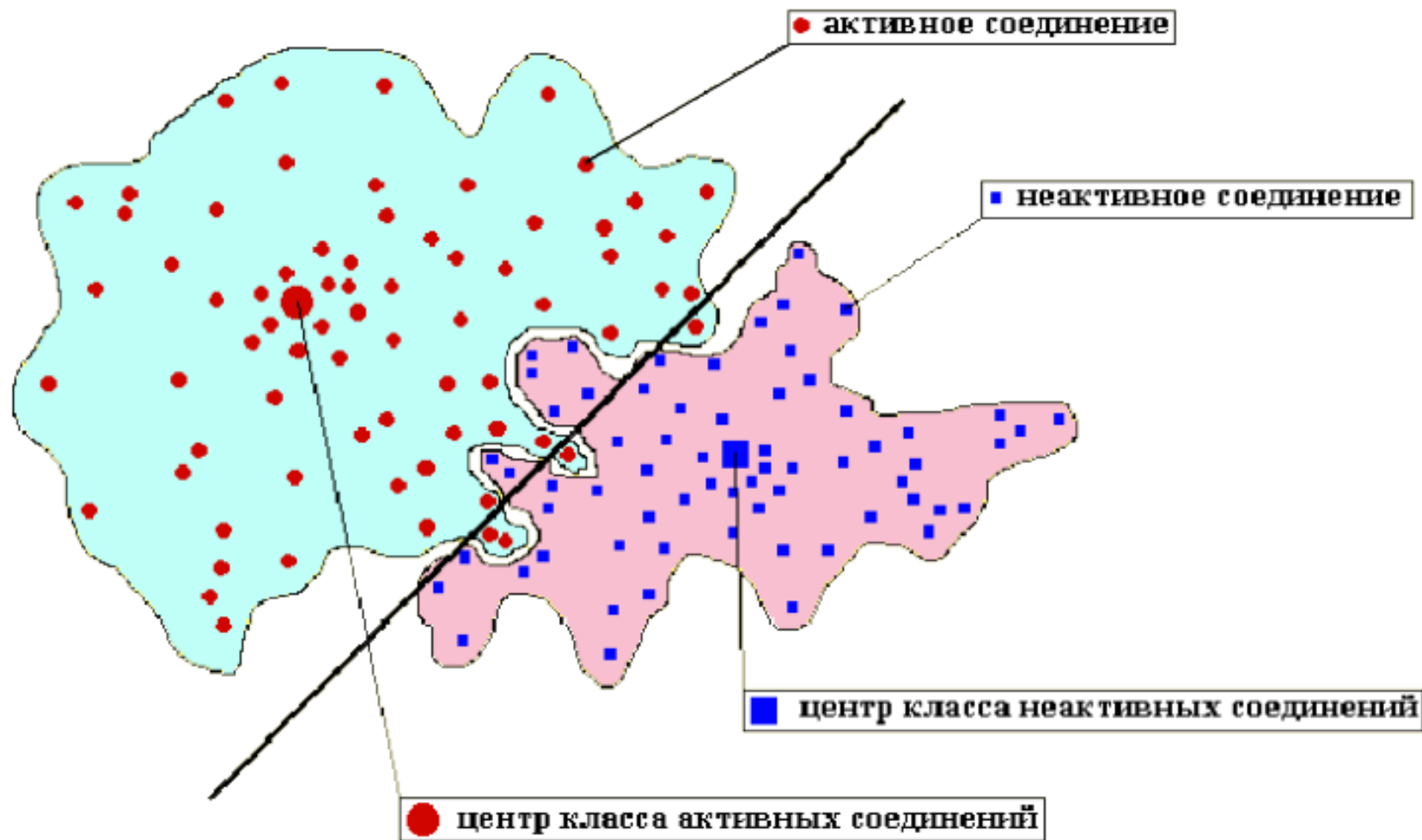
- **Вероятностные методы**
 - дискриминантный анализ
 - метод Байеса
 - методы локальных оценок
- **Геометрические методы**
 - расстояния до центров классов
 - метод k-ближайших соседей
- **Нейросетевые методы – ANN**
- **Метод опорных векторов – SVM**
- **«Ядерные» методы (kernels)**

Дискриминантный анализ

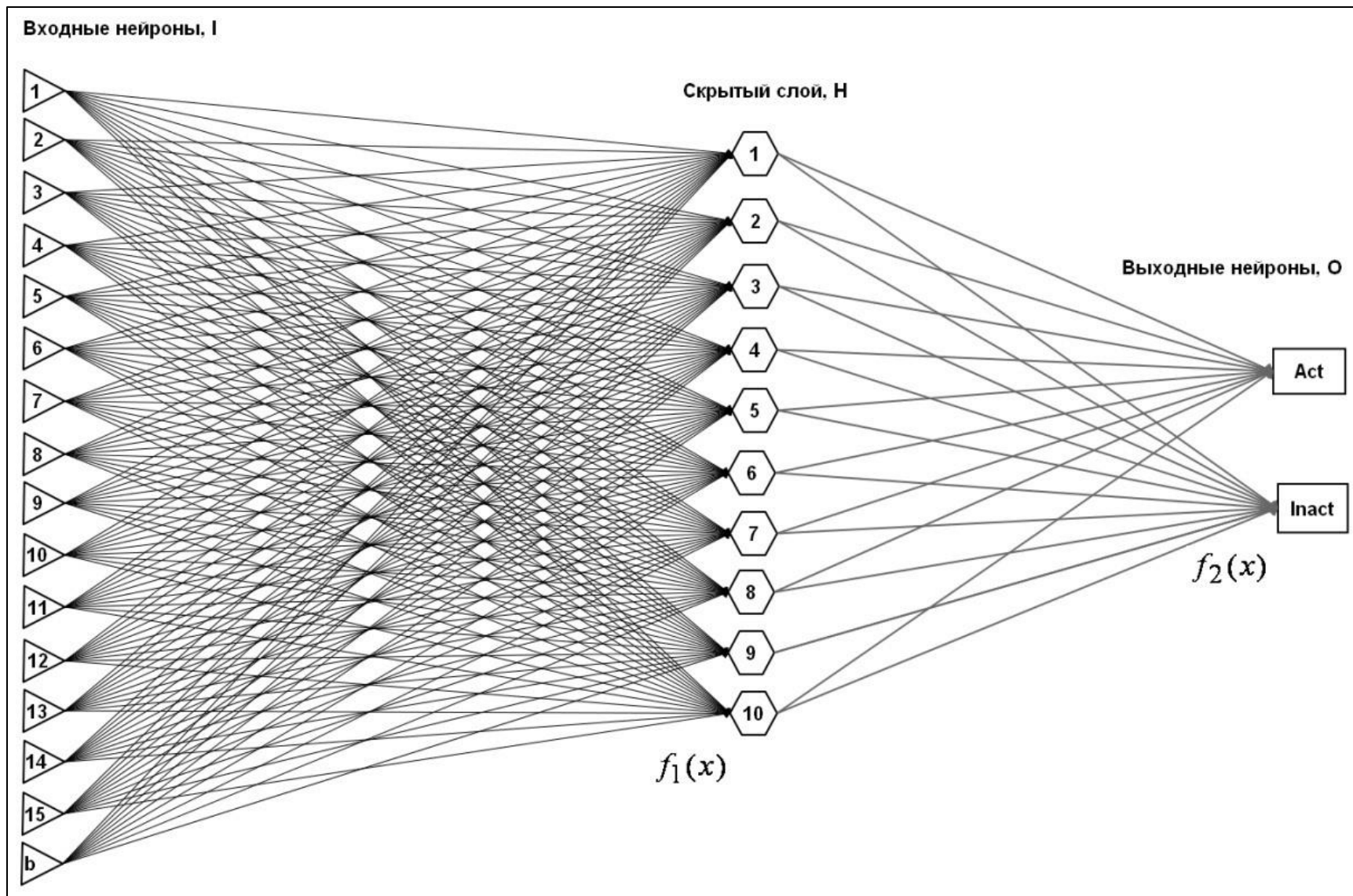
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Геометрические методы



Искусственные нейронные сети



Логические методы

Предикат

Выражение, использующее одну или более величину с результатом логического типа

Конструкция «если ... то»

Логические операции

$\wedge$, AND, & – логическое “И”, конъюнкция

$\vee$, OR – логическое “ИЛИ”, дизъюнкция

$\neg$, NOT – логическое “НЕ”, отрицание

To be continued ...

