спецкурс
“СУЧАСНІ КОМП’ЮТЕРНІ МЕТОДИ В ХІМІЇ”

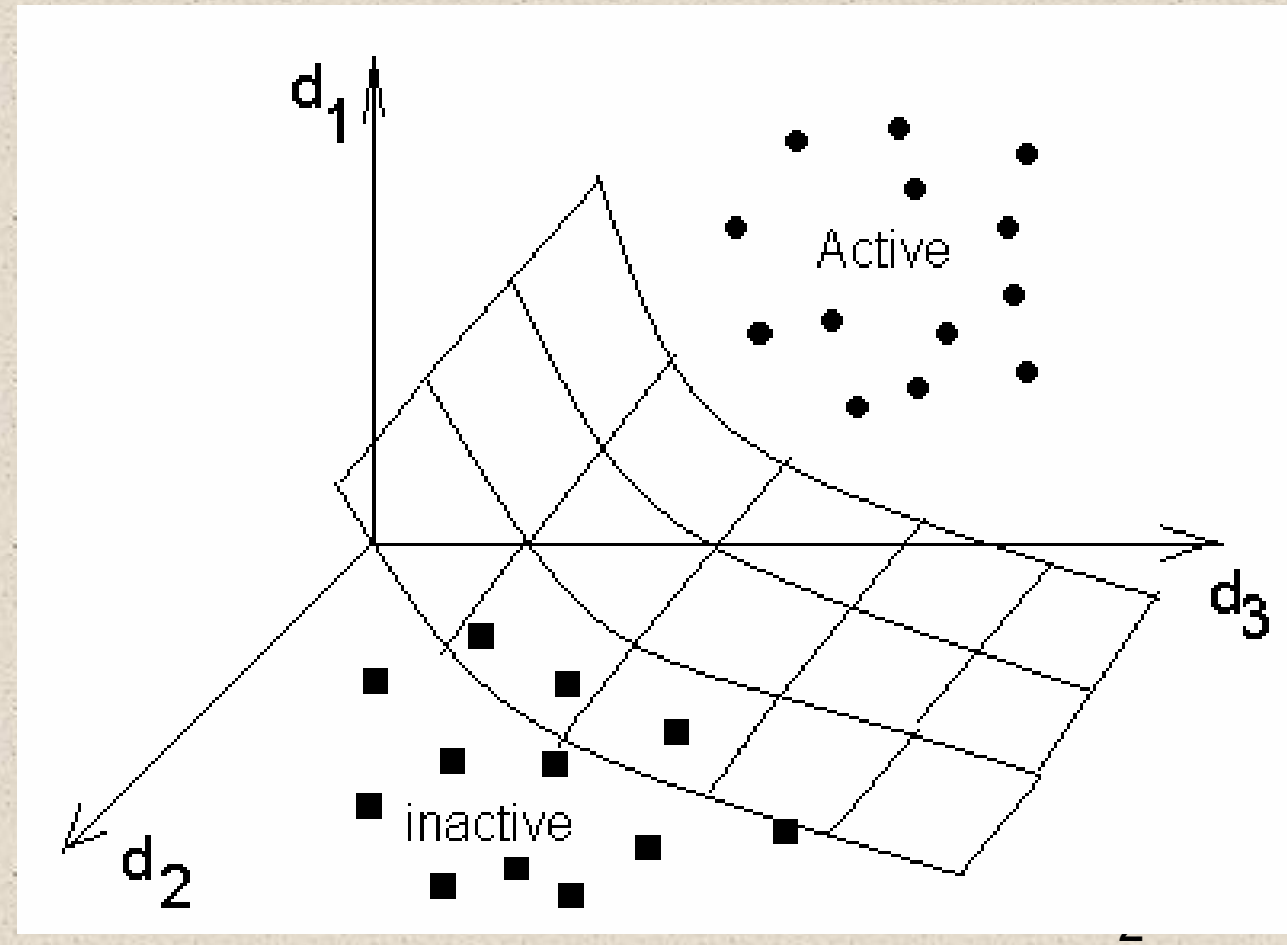
Методы классификации

В.В.Иванов

Materials Chemistry Department
V. N. Karazin National University,
61077, Kharkiv, Ukraine
ivv34@yahoo.com

ДИСКРИМИНАНТНЫЙ (дискриминационный) АНАЛИЗ

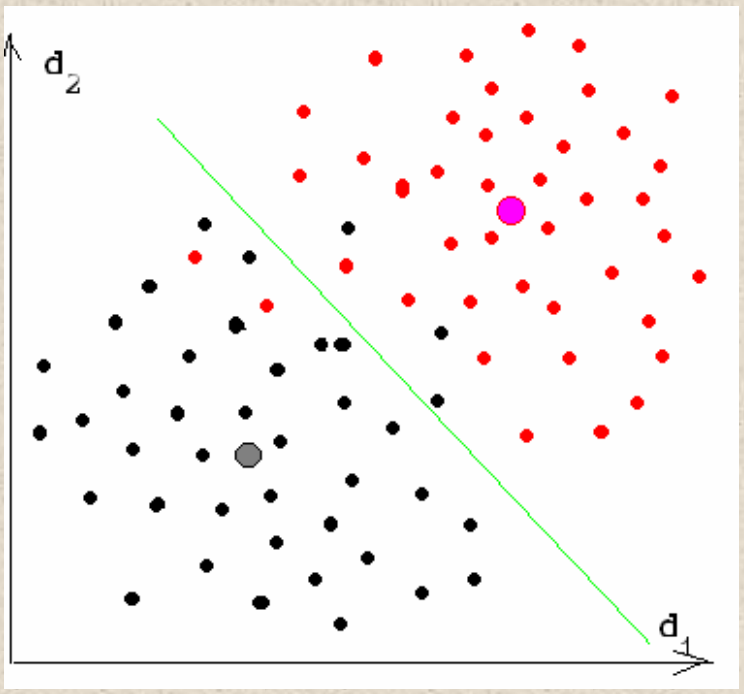
$$D = a_0 + a_1 d_1 + a_2 d_2 + a_3 d_3 + b_{11} d_1^2 + b_{12} d_1 d_2 + b_{13} d_1 d_3 ....$$

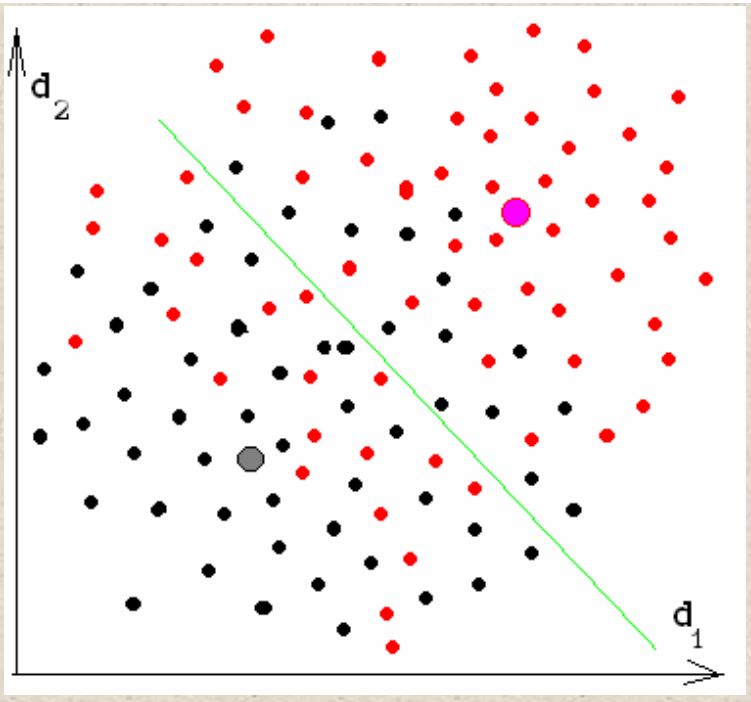


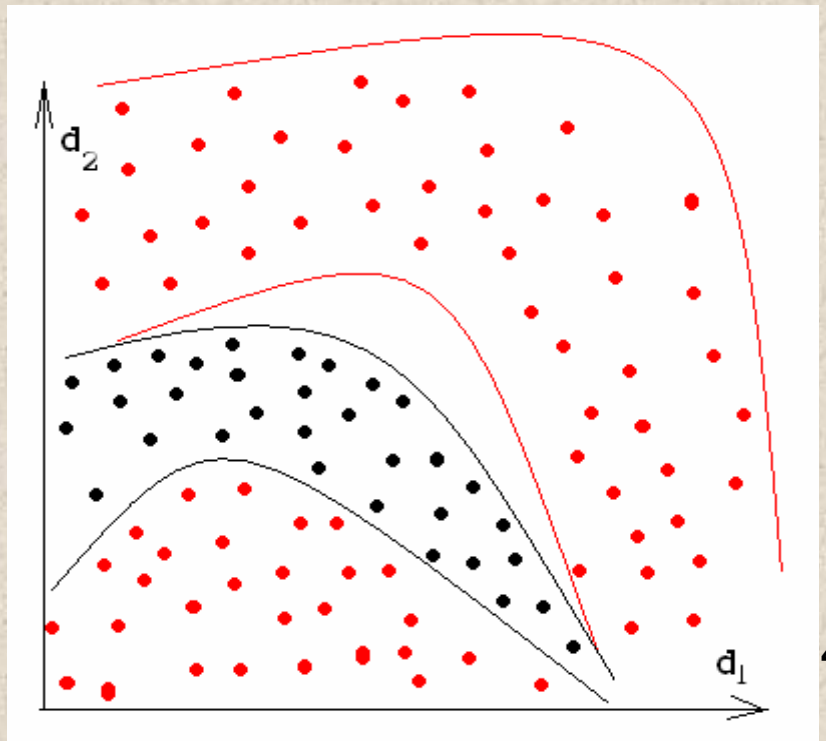
Цели классификации

- Идентификация переменных (дескрипторов), которые наилучшим образом разделяют *активные* и *неактивные* молекулы

- Создание классификационного правила для разделения систем на группы


d2
d1


d2
d1


d2
d1

Метод Фишера (линейная дискриминация (**LDA**) по 2 классам)

$$x=(d_1,d_2,..,d_p)$$

$$D = a_1 \cdot d_1 + a_2 \cdot d_2 + \dots$$

$$W_{ij} = \sum_{\alpha=1}^{2} \sum_{m=1}^{n_\alpha} \left(d_{i\alpha m} - \bar{d}_{i\alpha}\right)\left(d_{j\alpha m} - \bar{d}_{j\alpha}\right)$$

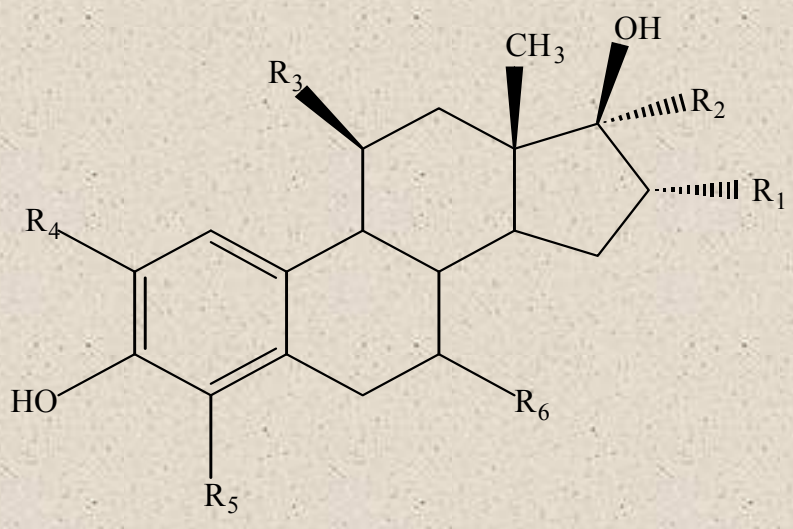
$$C_{ij} = \sum_{\alpha=1}^{2} \sum_{m=1}^{n_\alpha} \left(d_{i\alpha m} - \bar{d}_{i}\right)\left(d_{j\alpha m} - \bar{d}_{j}\right)$$

1938, Рональд Фишер

$$T_{ij} = C_{ij} - W_{ij}$$

$$\max \ \lambda(\vec{a}) = \frac{\sum_{i,j} a_i T_{ij} a_j}{\sum_{i,j} a_i W_{ij} a_j}$$

Активность эстрадиолов



Активность:
Log(RBA) > 1.5

$$1.032 - 1.49\,\varepsilon_{HBMO} - 0.089\,\alpha + 1.41\,\lg P - 0.17(\lg P)^2 > 0$$

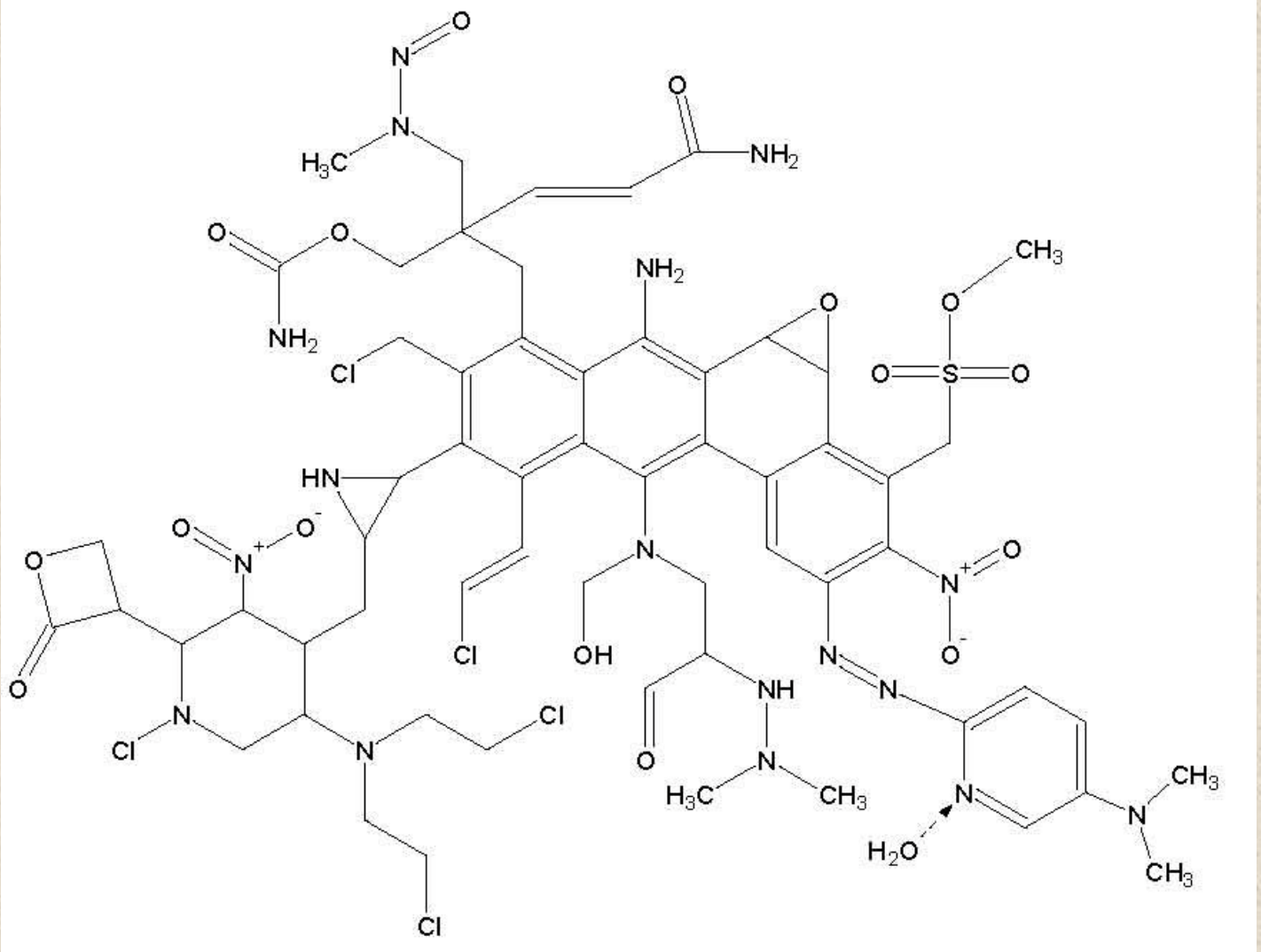
ε_{HBMO} Энергия нижайшей вакантной МО (эВ). расчет AM1

$\lg P$ Липофильность α - Поляризуемость ($Å^3$)

Эффективность (LOO) = 91 %

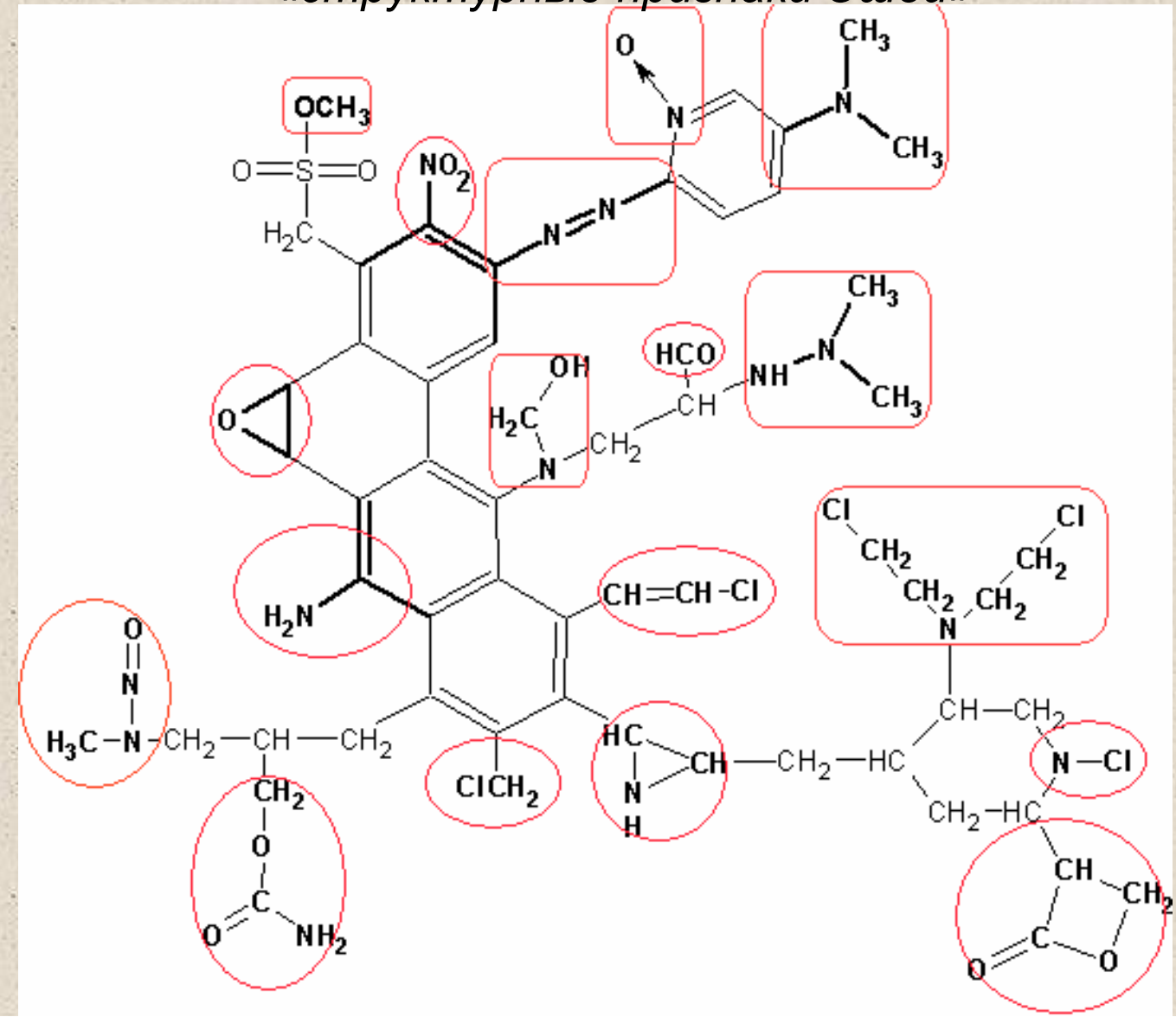
Модель поликанцерогена Эшби (Ashby)

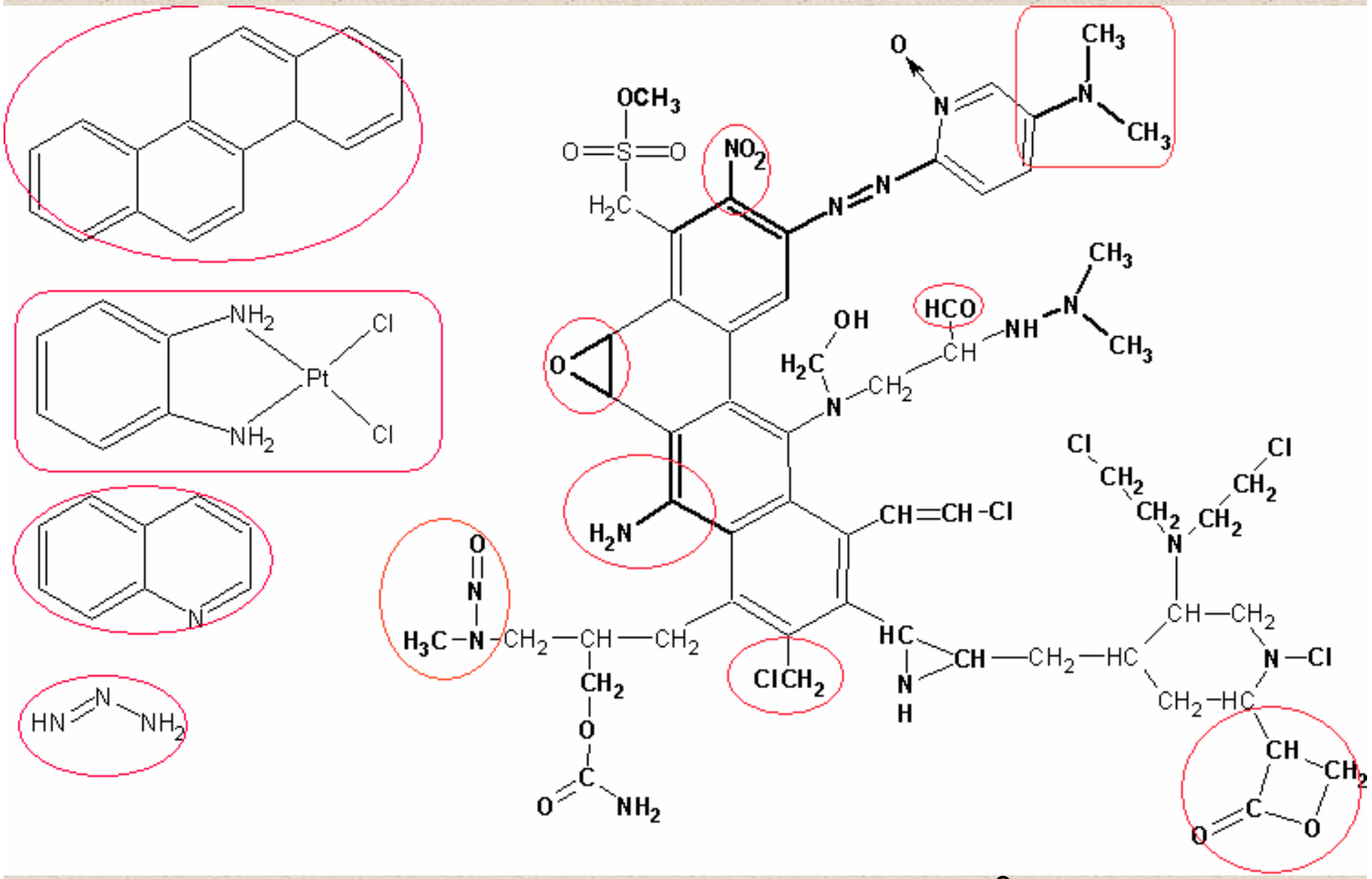
«*структурные признаки Эшби*»



Модель поликанцерогена Эшби (Ashby)

«*структурные признаки Эшби*»





Распространение PAH

Откуда берутся PAH в окружающей среде ?

- Неполное сгорание топлива
- Дизельные моторы
- Индустрия
- Сгорание дерева

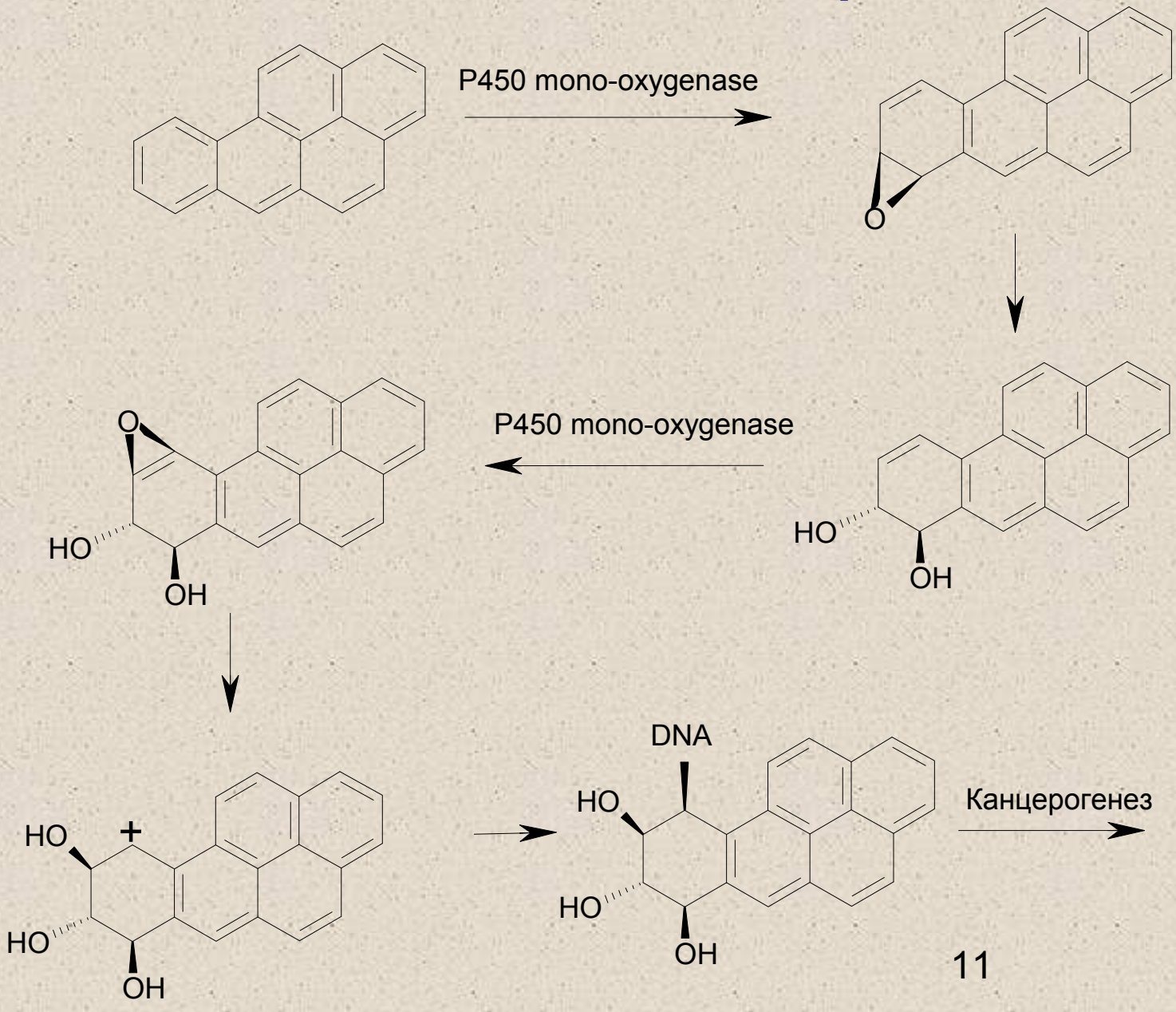
Как попадают PAH в организм человека ?

- Питание
- Воздух

Согласно Стокгольмской конвенции PAH являются постоянными органическими загрязнителями:

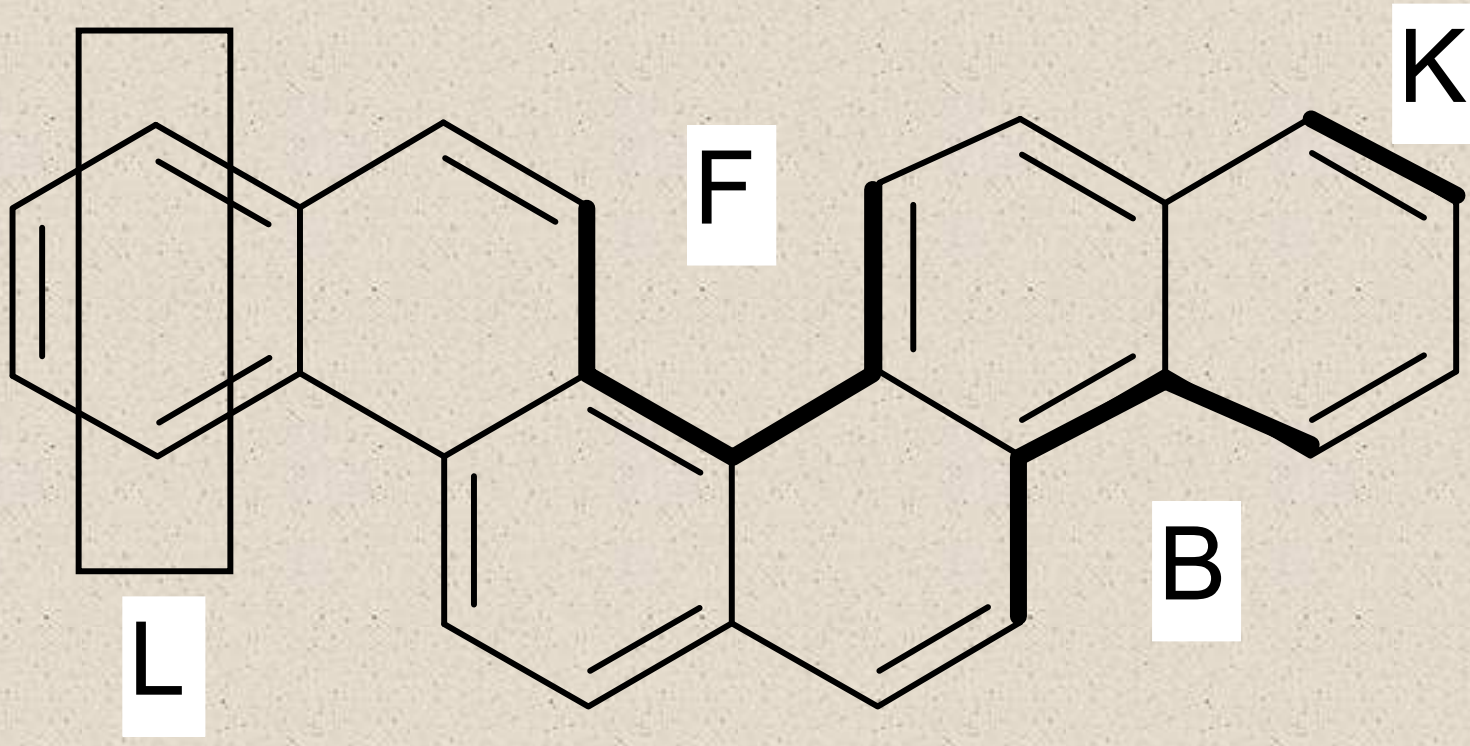
- **Тератогенность**
- **Мутагенность**
- **Канцерогенность**

Метаболизм PAH в организме

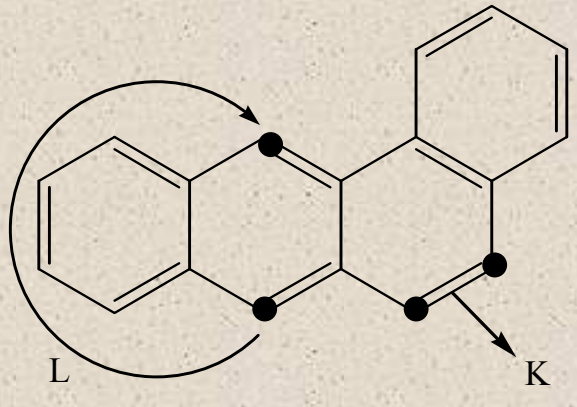


Цитохром P450

Активные области PAH



Канцерогенная активность конденсированных углеводородов

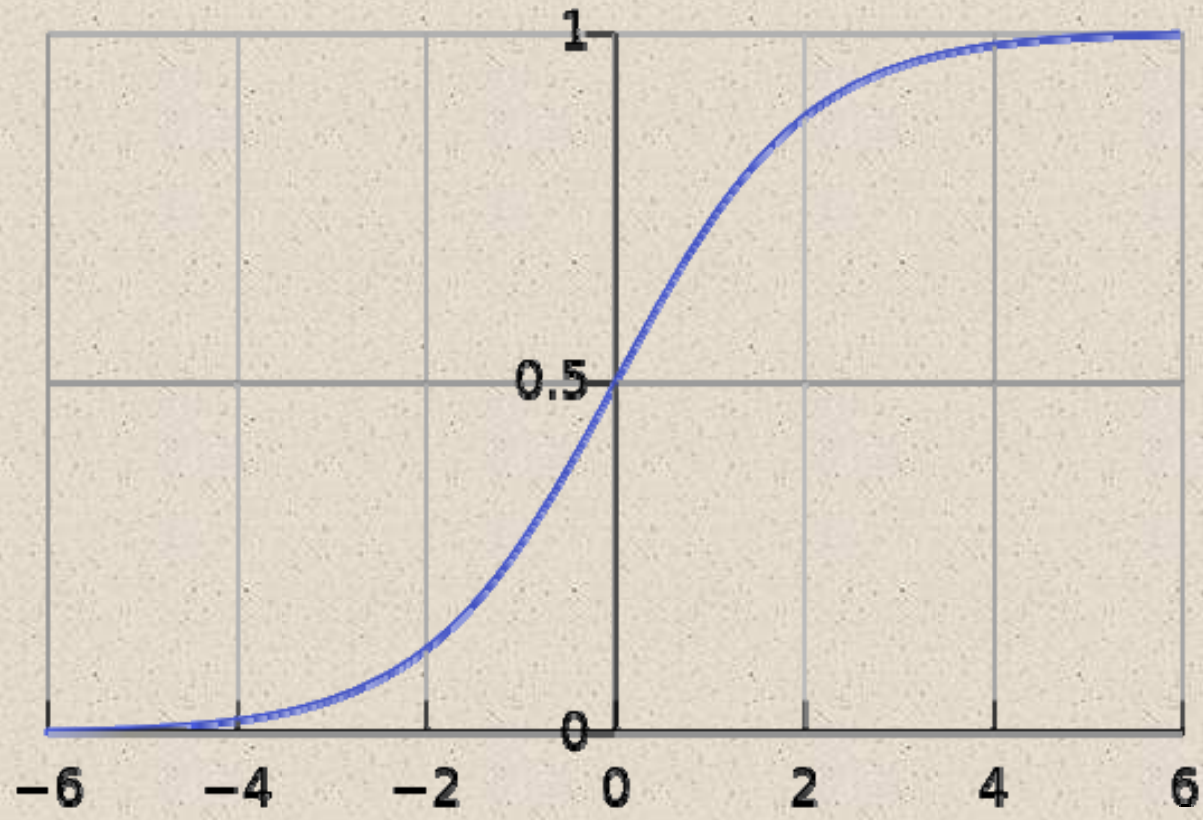


$$-24.1+17.7\,S_e(K)-4.2\,S_e(L) > 0$$

S_e – “сверхделокализуемость” (индекс Фукуи) К и L областей.

Местечкин М.М., Сивякова, Канцерогенная активность полициклических Углеводородов (препринт Института Углехимии, Донецк.)

Логистическая регрессия



•***Логистическая модель*** *наиболее популярная бинарная модель*

•***Логистическая модель*** обычно используется для изучения связи между бинарным откликом и группой предикторов, которые могут быть либо континуальными, либо бинарными.

Отклик системы:

Y = 1 (истина, да, активно)

Y = 0 (ложь, нет, неактивно)

В общем случае Y = [0 - 1]

•***Логистическая модель*** может быть обобщена на случай более, чем два отклика (*multinomial logistic regression model*)

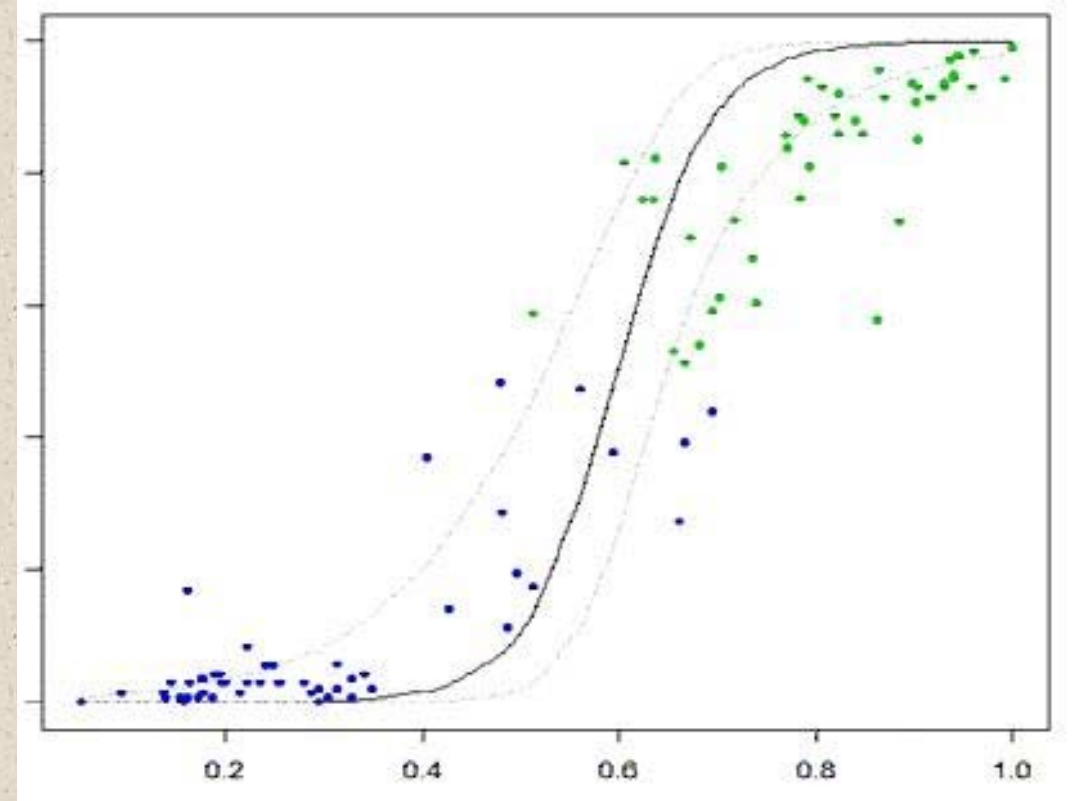
Logistical Regression

$$\pi_j = \frac{\exp(Z_j)}{1+\exp(Z_j)} = \frac{1}{1+\exp(-Z_j)}$$

$$Z_j = b_0 + \sum_i b_i d_{ji}$$

d_{ji} – дескрипторы
(j-номер молекулы,
i- номер дескриптора)

π- отклик системы
(рассчитанный класс)



максимизация $$L = \prod_{j=1}^{N} \pi_j^{y_j} (1 - \pi_j)^{1-y_j}$$

$$\ln(L) = \sum_{j=1}^{n} \{ y_j \ln[\pi_j] + (1-y_j) \ln[1-\pi_j] \}$$

Logistical Regression

$$L = \prod_{j=1}^{N} \pi_j^{y_j} (1-\pi_j)^{1-y_j}$$

π	Y	$\pi^y(1-\pi)^{(1-y)}$	L	ln(L)
1	1	$1^1 (1-1)^{(1-1)}=1$	$1 \cdot 1=1$	0
0	0	$0^0(0-0)^{(0-0)}=1$		
½	1	$(1/2)^1 (1-1/2)^0=0.5$	$0.5 \cdot 0.5 = 0.25$	-1.386
½	0	$(1/2)^0 (1-1/2)^1=0.5$		
0	1	$0^1 (1-0)^{(1-1)}=0$	$0 \cdot 0 = 0$	$-\infty$
1	0	$1^0 (1-1)^{(1-0)}=0$		

Качество аппроксимации

1. Процент верно классифицированных систем, **η**
2. Логарифм функции правдоподобия, **ln(L)**
3. Информационный критерий, **AIC=-2(ln(L)-2)**
4. Коэффициент детерминации, **R^2**
5. Урегулированный (adjasted), **R^2**

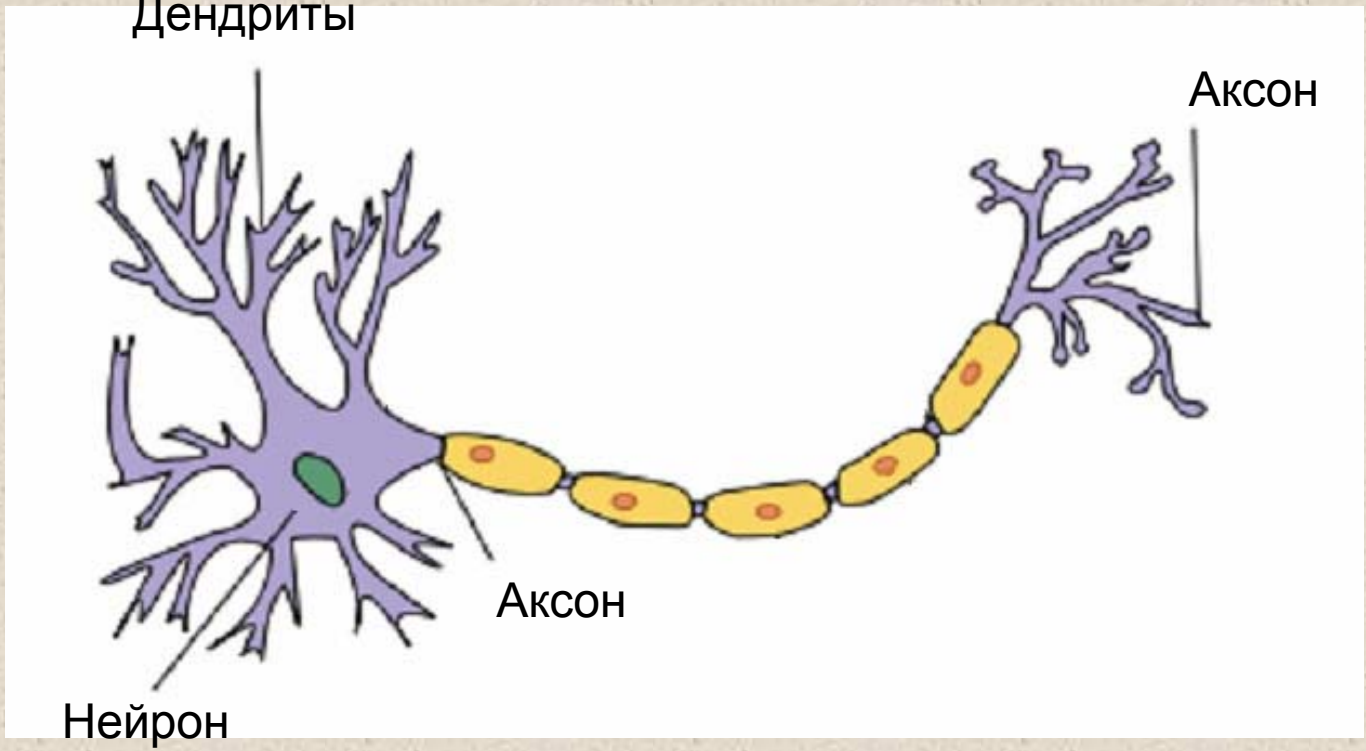
Процедура LOO

N

d_1	d_2	d_3	….	d_p	Class
					1
					0
					…

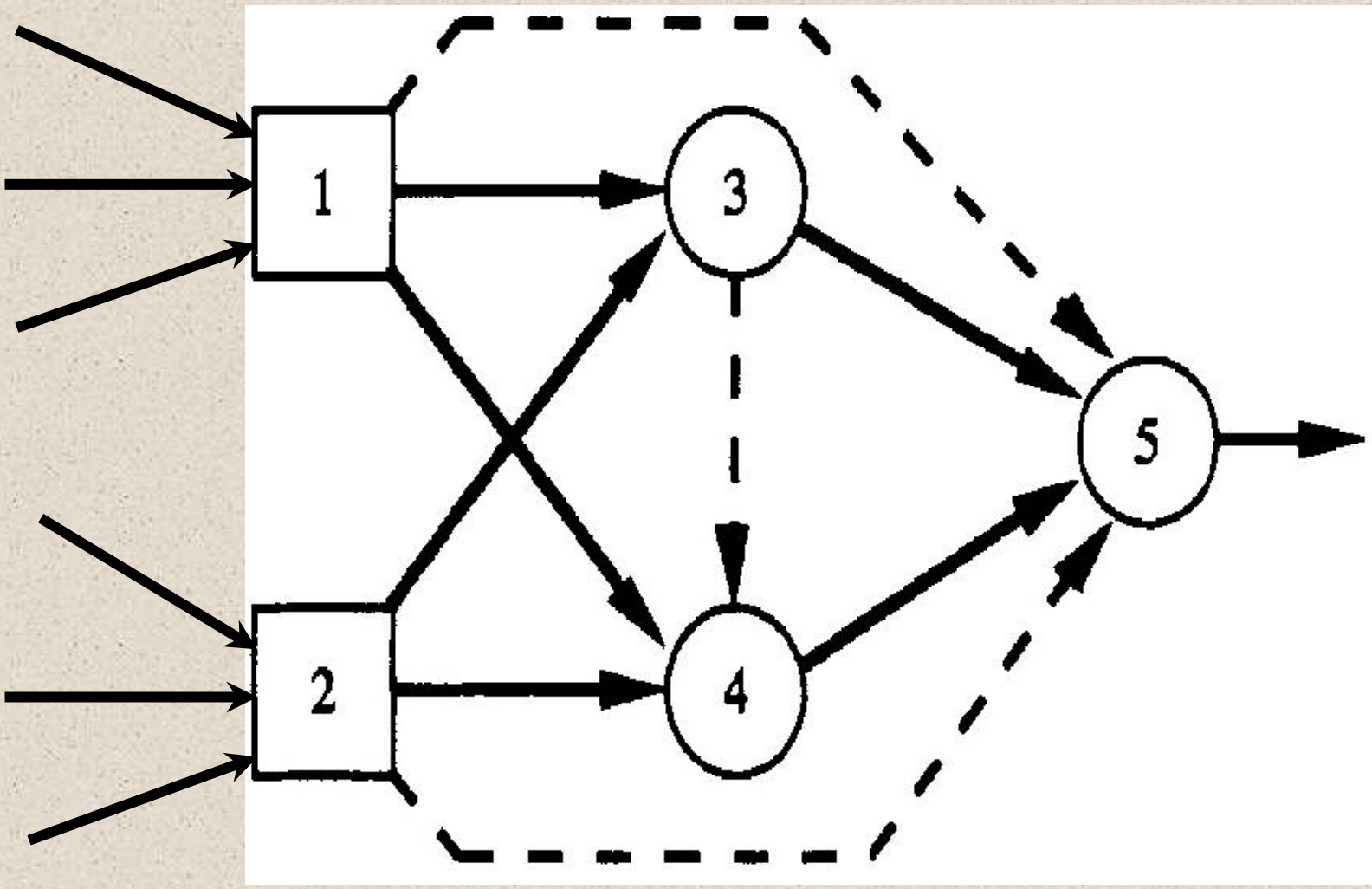
Метод Нейронных сетей

Структура нейрона

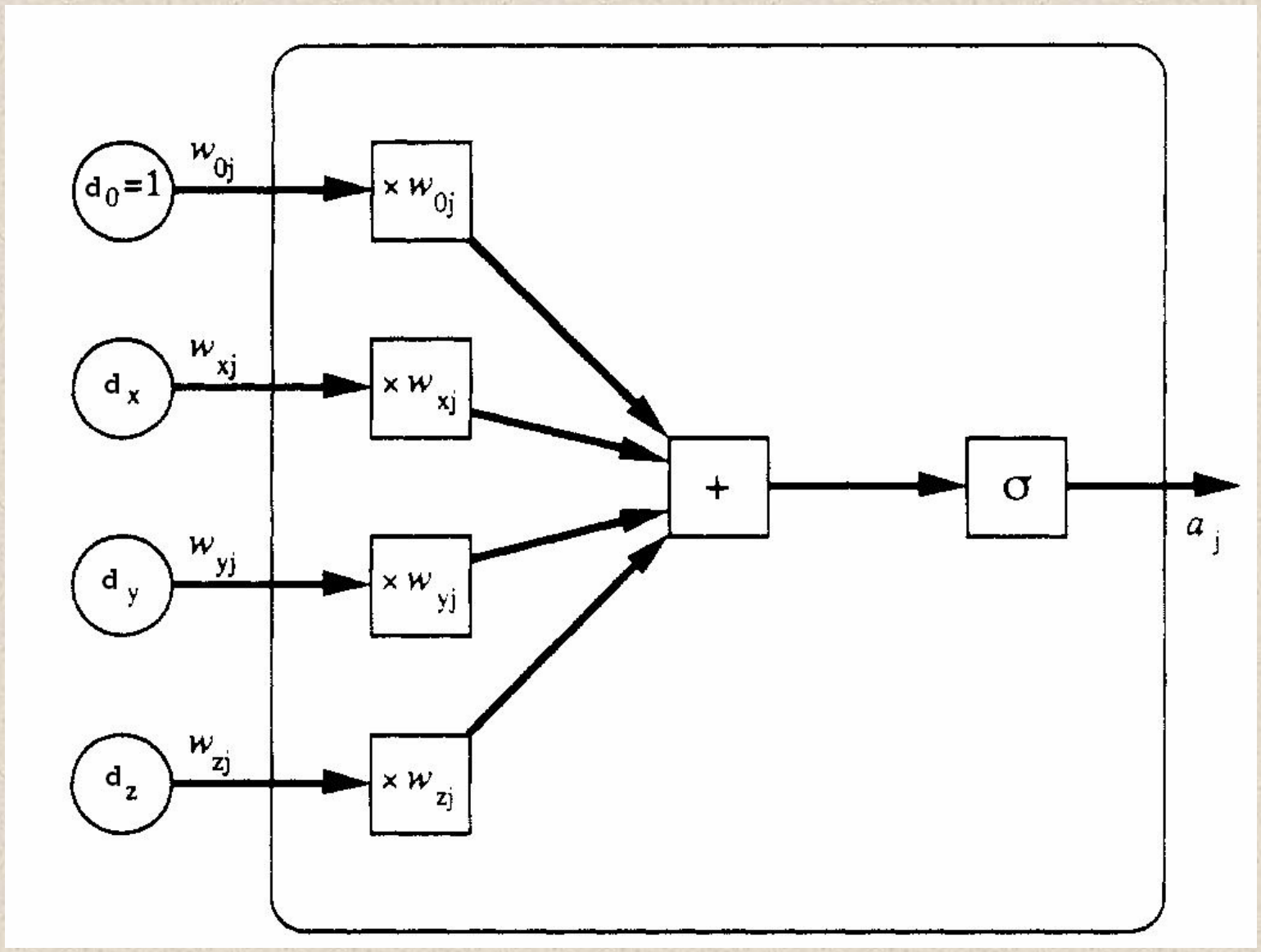


В человеческом организме ~ 10^{15} нейронов

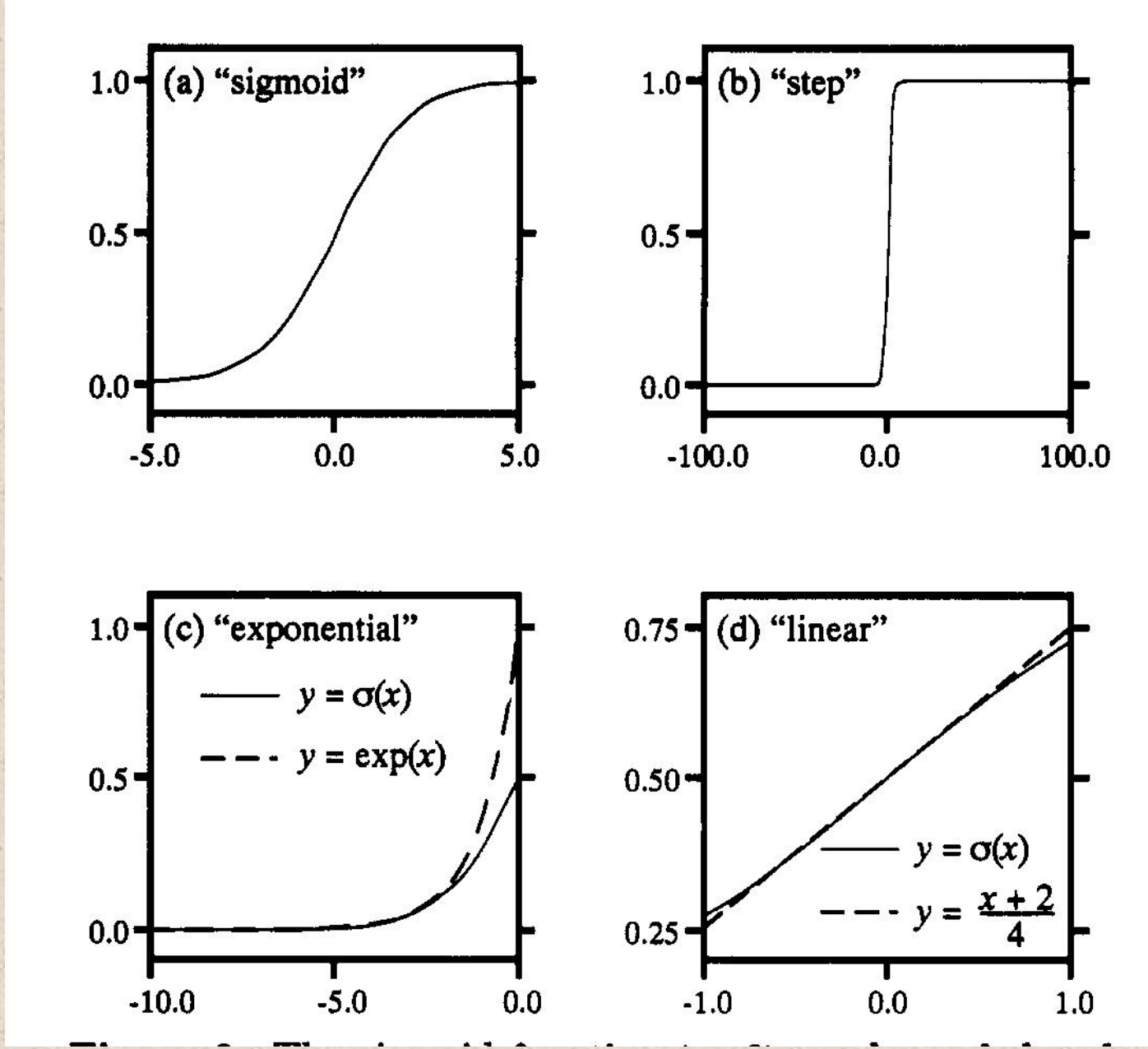
Модель Нейронной сети



Модель Нейрона



Выходная функция σ



$$S_j = b_j + \sum_{i=1}^{j-1} d_i w_{ij} \qquad a_j = \sigma(S_j) = \frac{1}{1+\exp(-S_j)}$$

$$F = \sum_{\text{output}} (\hat{y}_n - a_n)^2 \quad - \quad \min$$

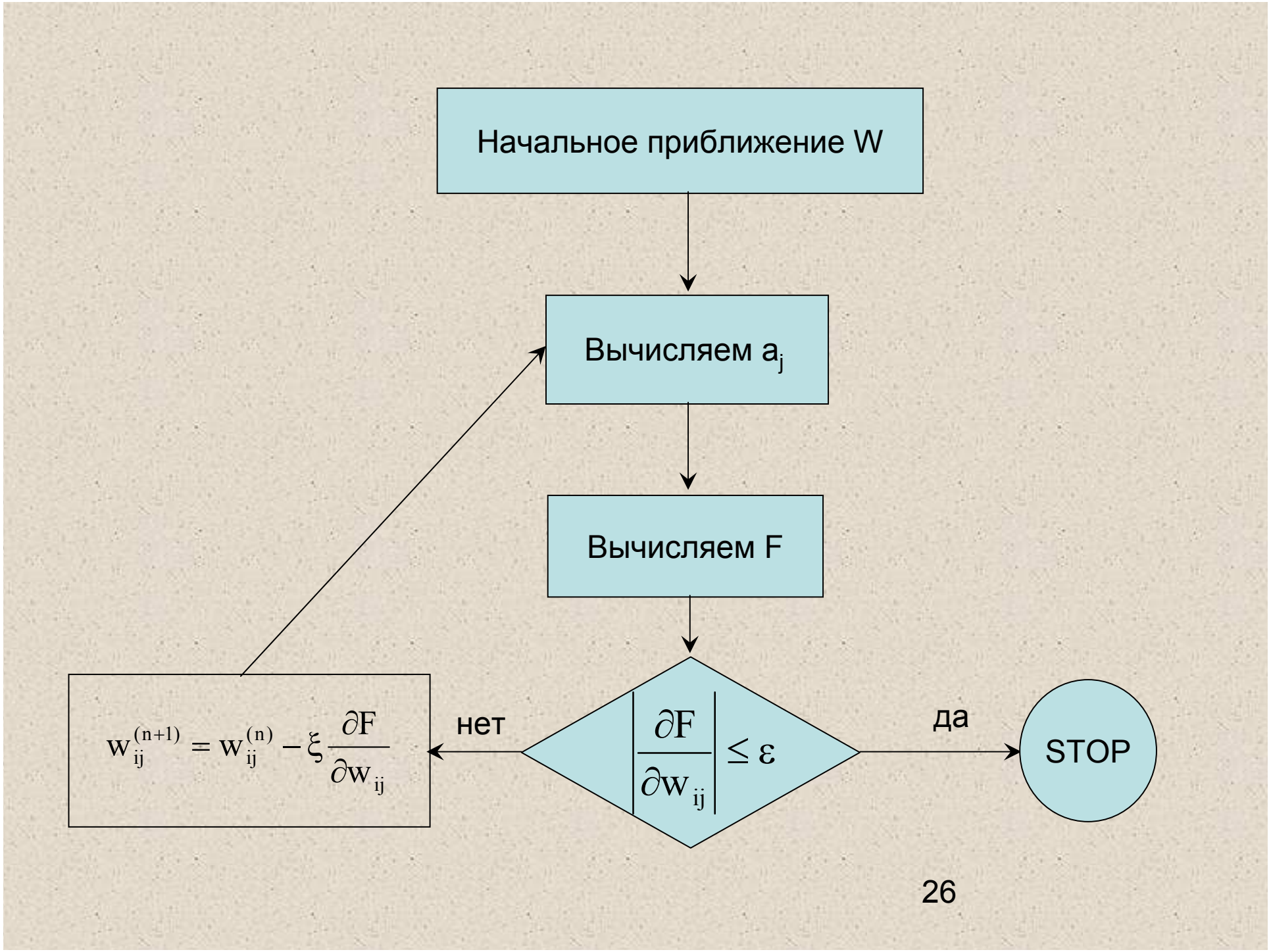
$$\frac{d\sigma(x)}{dx} = \sigma(x)[1-\sigma(x)] \qquad \frac{dS_j}{dw_{ij}} = d_i$$

$$\frac{\partial a_j}{\partial w_{ij}} = \frac{\partial a_j}{\partial S_j}\frac{\partial S_j}{\partial w_{ij}} = a_j(1-a_j)a_i$$

$$w_{ij}^{(n+1)} = w_{ij}^{(n)} - \xi \frac{\partial F}{\partial w_{ij}}$$

$$\xi = 0.001 \div 1$$

Итерируем, пока $\left| \frac{\partial F}{\partial w_{ij}} \right| \leq \varepsilon$


Начальное приближение W
Вычисляем a_j
Вычисляем F
$\left|\frac{\partial F}{\partial w_{ij}}\right| \le \varepsilon$
нет
да
STOP
$w_{ij}^{(n+1)} = w_{ij}^{(n)} - \xi \frac{\partial F}{\partial w_{ij}}$