спецкурс
“СУЧАСНІ КОМП’ЮТЕРНІ МЕТОДИ В ХІМІЇ”

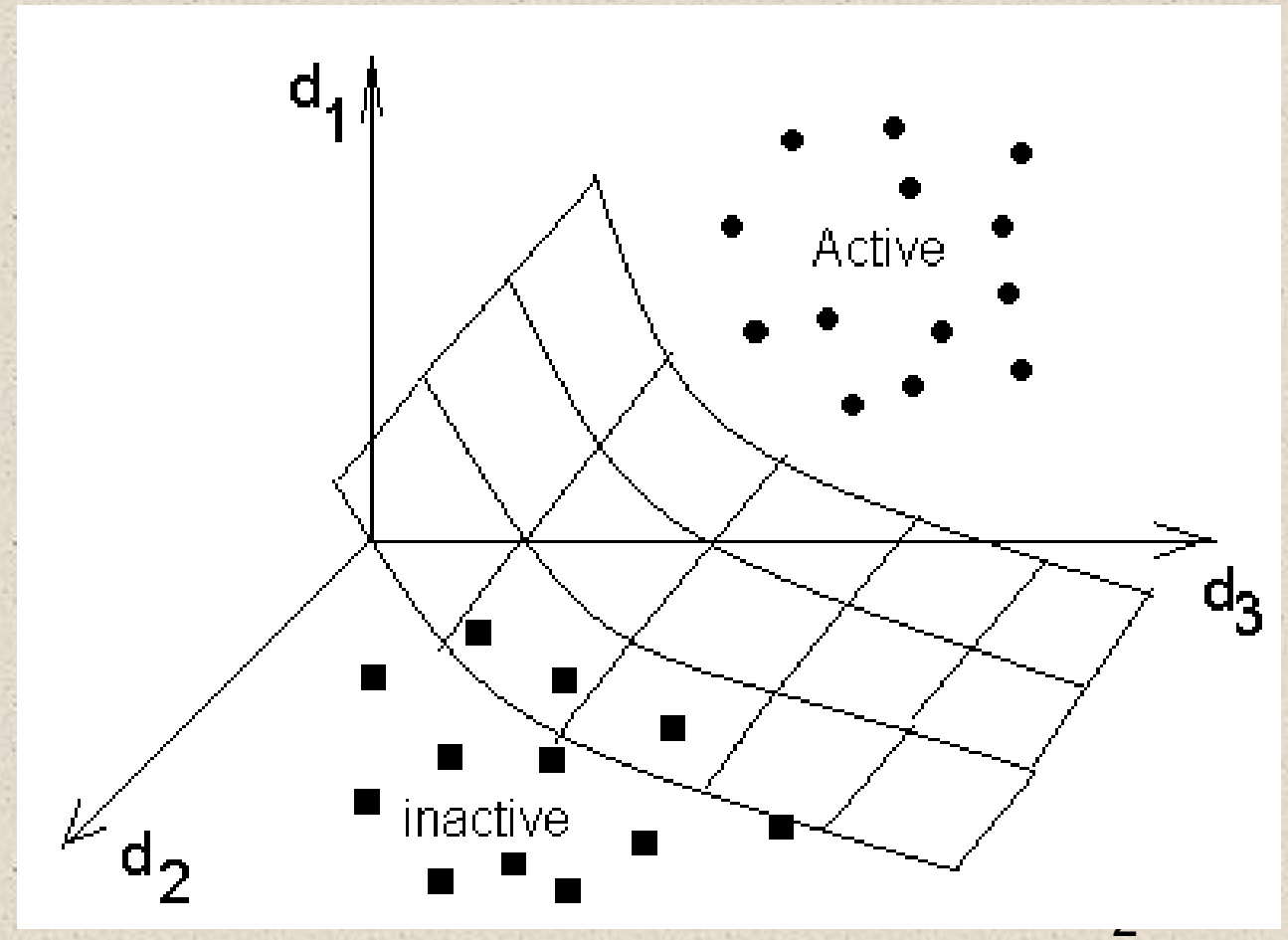
Методи класифікації

В.В.Іванов

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

ДИСКРИМІНАНТНИЙ (дискримінаційний) АНАЛІЗ

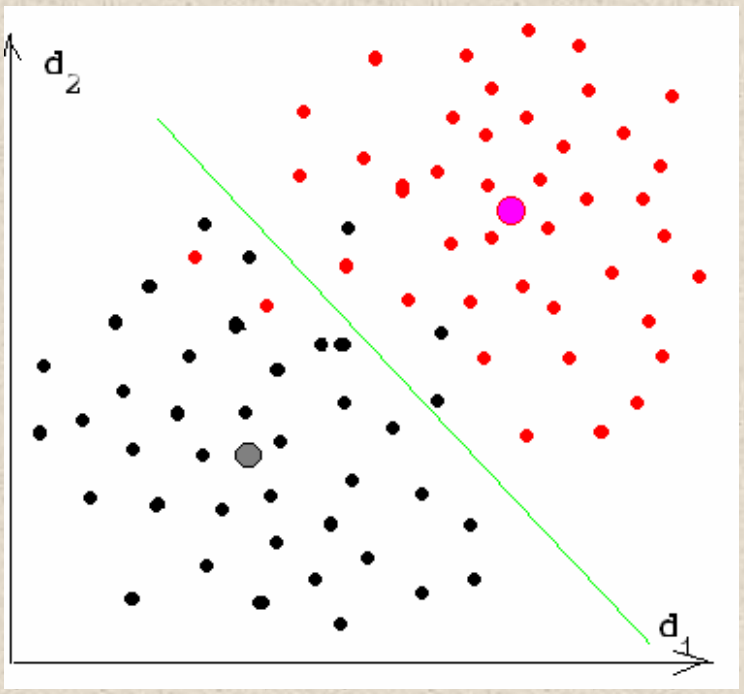
$$D = a_0 + a_1 d_1 + a_2 d_2 + a_3 d_3 + b_{11} d_1^2 + b_{12} d_1 d_2 + b_{13} d_1 d_3 ....$$

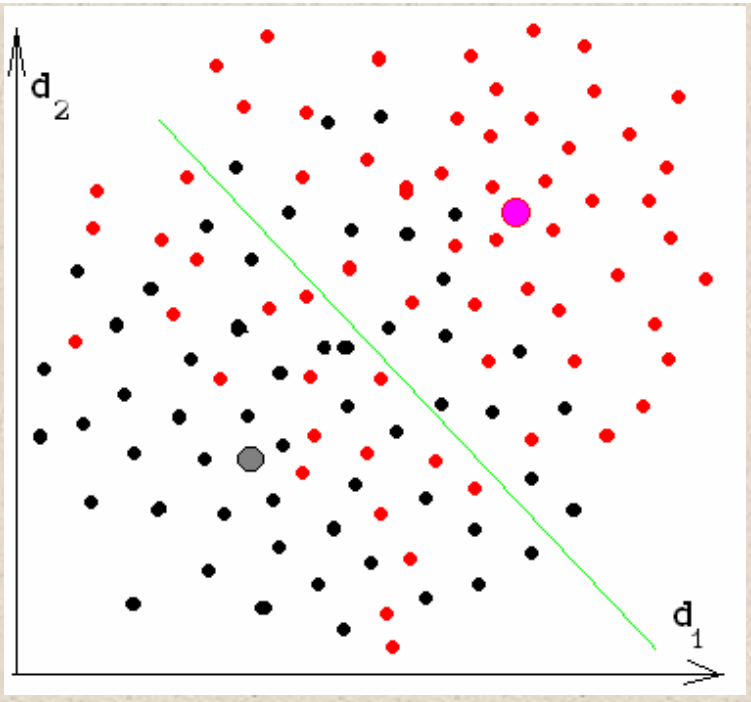


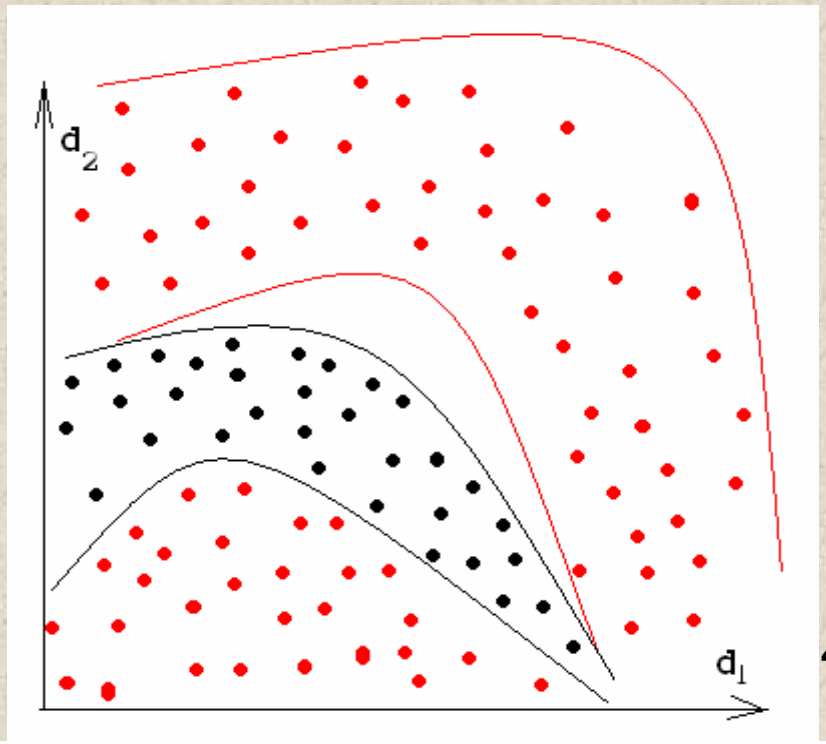
Мета класифікації

- Ідентифікація змінних (дескрипторів), які Найкращім чином розділяють *активні* і *неактивні* молекули

- Створення класифікаційного правила для розділення систем на групи


d2
d1


d2
d1


d2
d1

Метод Фішера
лінійна дискримінація (LDA) по 2 класам

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

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

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

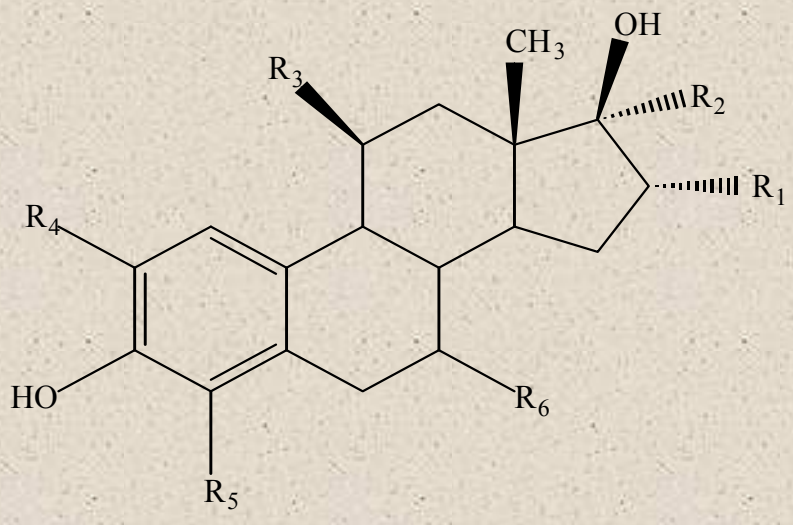
$$C_{ij} = \sum_{\alpha=1}^{2} \sum_{m=1}^{n_\alpha} \left(d_{i\alpha m} - \bar{d}_{i}\right)\left(d_{j\alpha m} - \bar{d}_{j}\right)$$

1938, Рональд Фішер

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

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

Дискримінантний аналіз
Активність естрадіолів



Активність:
Log(RBA) > 1.5

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

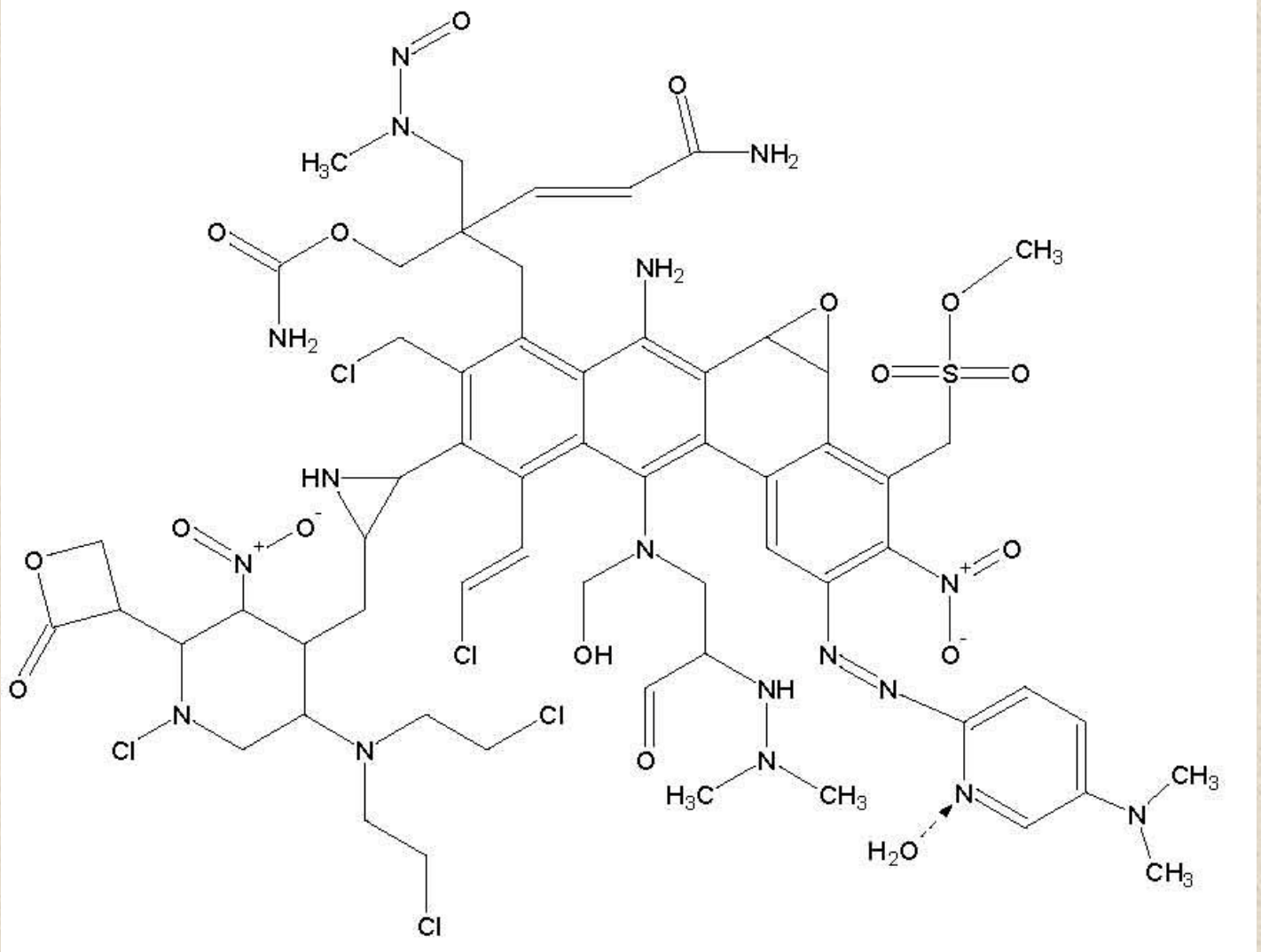
ε_{HBMO} Енергія найнижчої вакантної МО (еВ). розрахунок AM1

$\lg P$ Ліпофільність α - Поляризовність ($Å^3$)

Ефективність (LOO) = 91 %

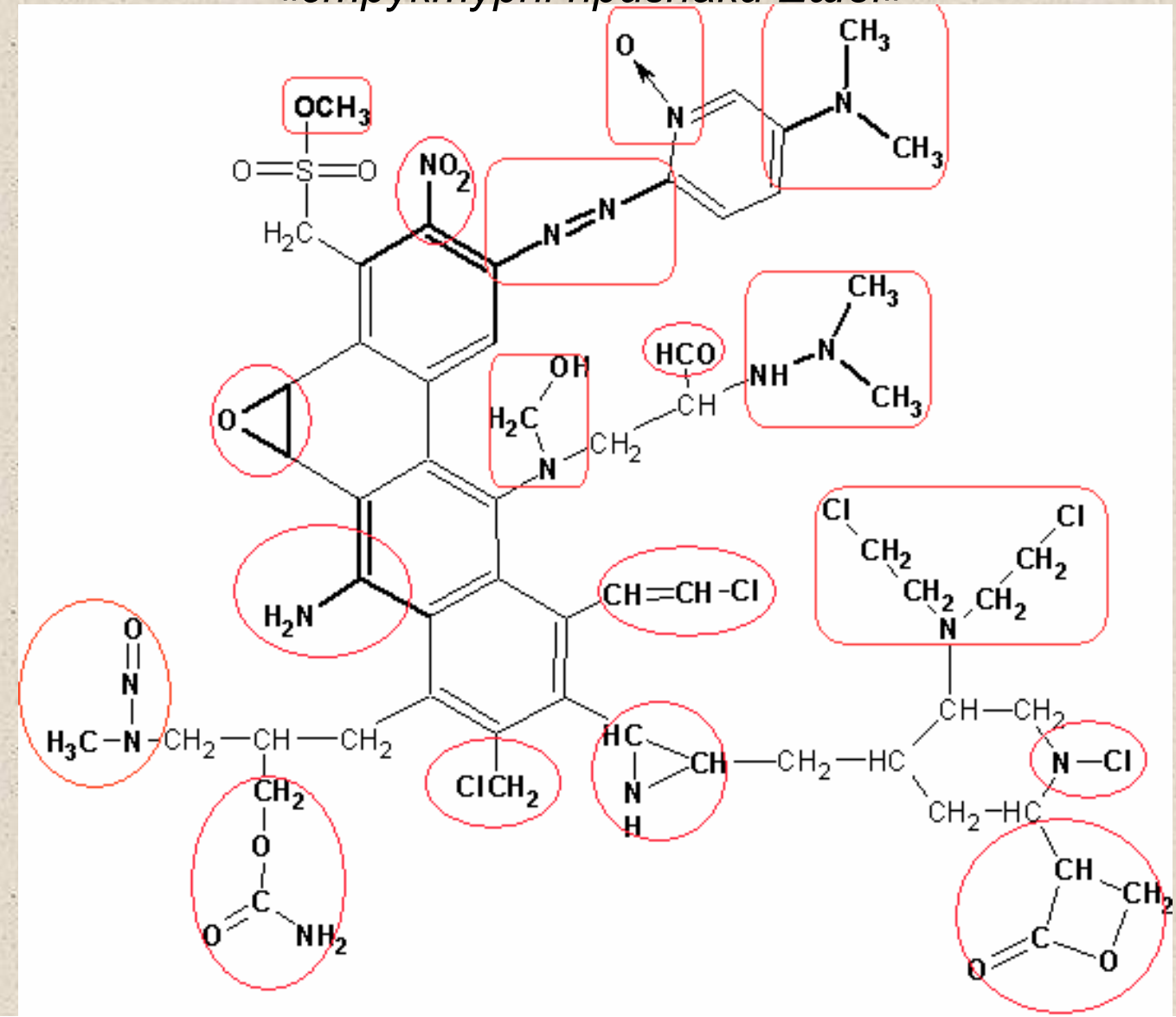
Модель поліканцерогена Ешбі (Ashby)

«*структурні признаки Ешбі*»



Модель поліканцерогена Ешбі (Ashby)

«*структурні признаки Ешбі*»



Розповсюдження РАН

Звідки беруться РАН в довколишньому середовищі ?

- Неповне сгоряння палива
- Дизельні мотори
- Індустрія
- Згоряння дерева

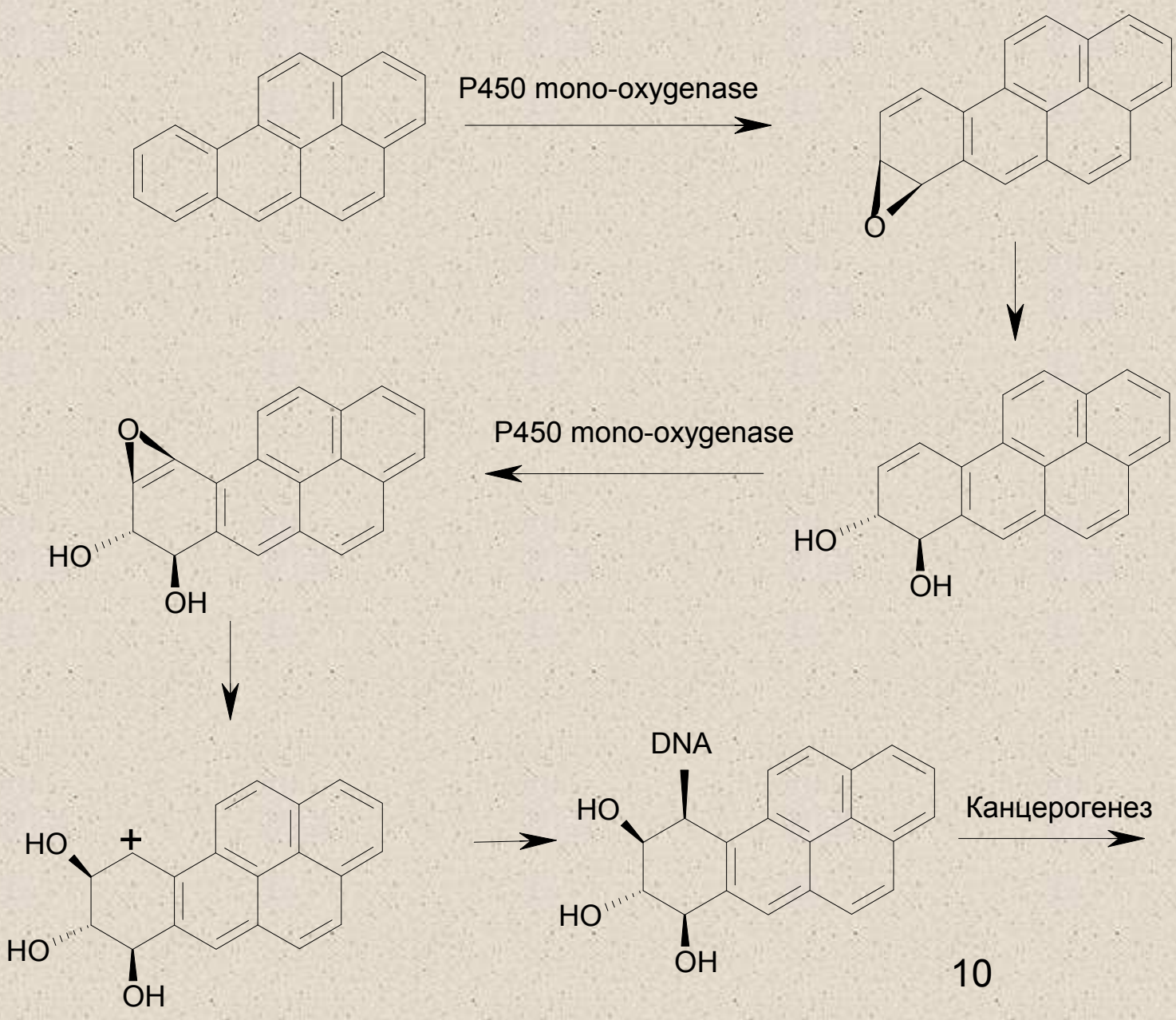
Як потрапляють РАН в організм людини ?

- Їжа
- Повітря

Згідно Стокгольмскої конвенції РАН є постійними органічними забруднювачами:

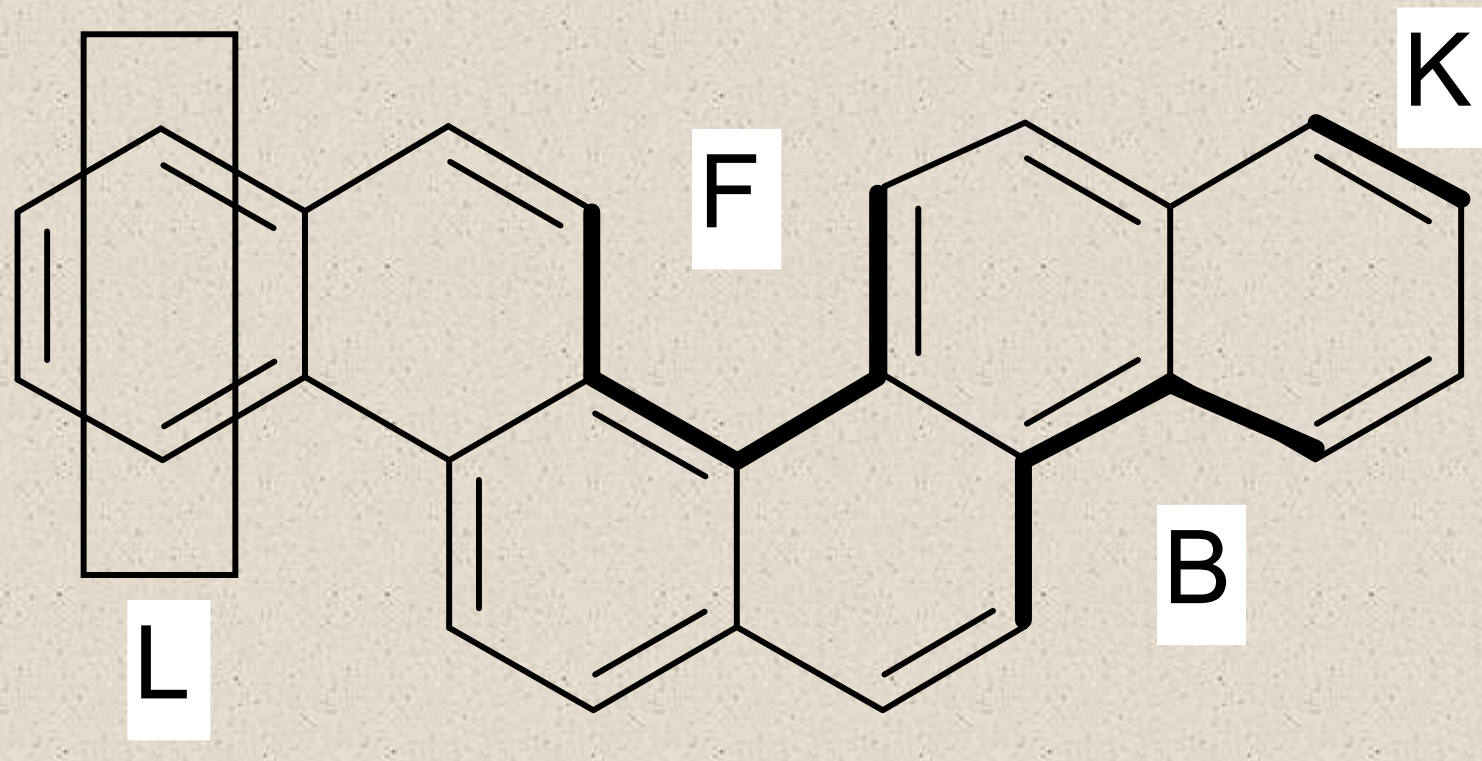
- **Тератогенність**
- **Мутагенність**
- **Канцерогенність**

Метаболізм PAH в організмі



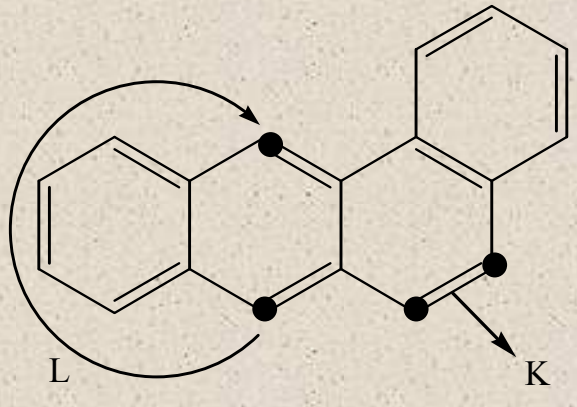
Цитохром P450

Активні області PAH



Дискримінантний аналіз

Канцерогенна активність конденсованих вуглеводнів

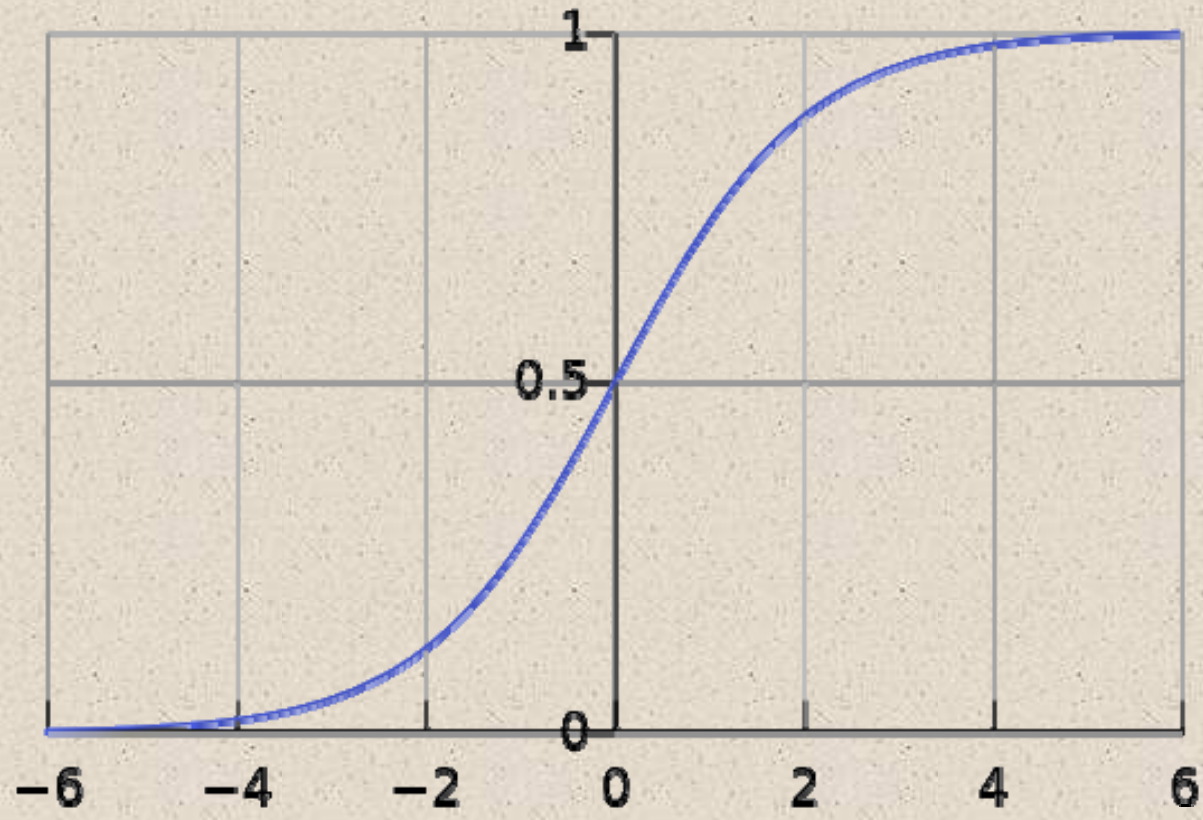


$$-24.1+17.7\,S_e(K)-4.2\,S_e(L)>\ 0$$

S_e – “суперделокалізовність” (індекс Фукуі) К і L області.

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

Логістична регресія



Логістична регресія

- ***Логістична модель*** *найбільш популярна бінарна модель*
- ***Логістична модель***

Відгук системи:

Y = 1 (істина, да, активно)

Y = 0 (не істина, ні, неактивно)

В загальному випадку Y = [0 - 1]

- ***Логістична модель*** може бути узагальнена на випадок більше ніж два відгуки (*multinomial logistic regression model*)

Логістична регресія

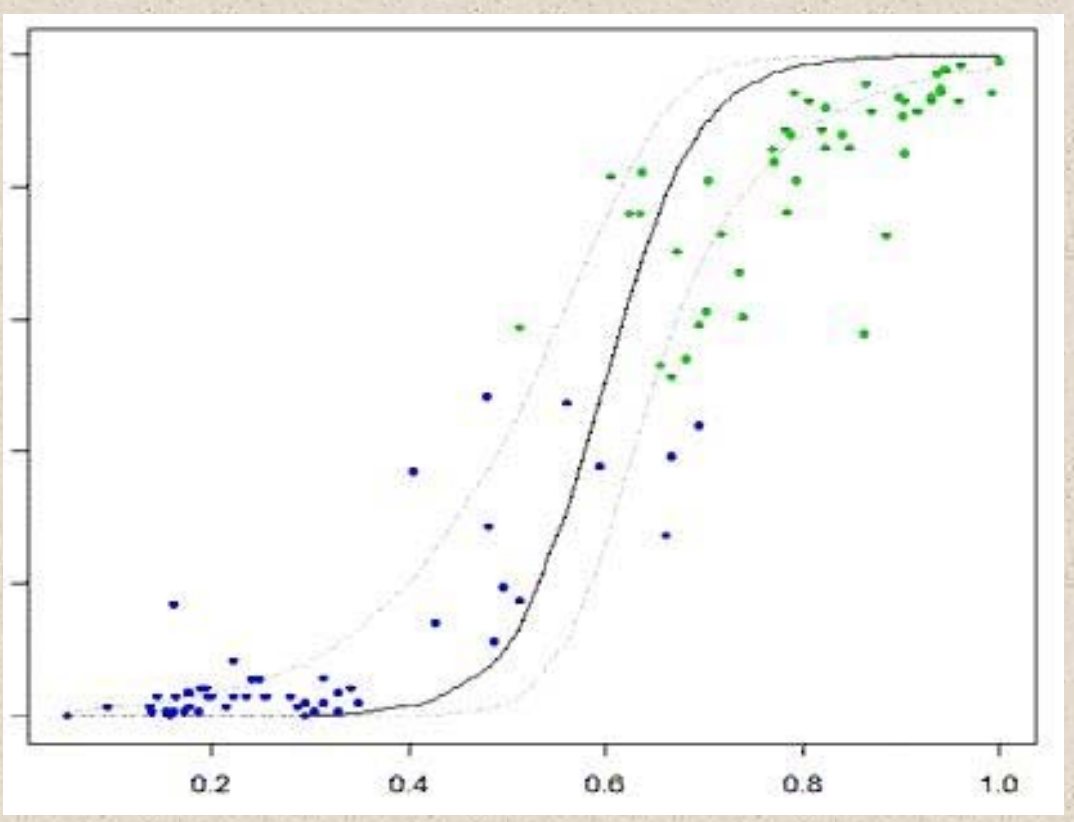
$$y_j = \frac{\exp(t_j)}{1+\exp(t_j)} = \frac{1}{1+\exp(-t_j)}$$

$$t_j = a_0 + \sum_i a_i d_{ji}$$

d_{ji} – дескриптори
(j-номер молекули,
i- номер дескриптору)

$\pi = Y_{(calc)}$

відгук системы
(клас)



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

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

Логістична регресія

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

π_j - розраховані значення y_j

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

Якість апроксимації

1. Процент вірно класифікованих систем, **η**
2. Логарифм функції правдоподібності, **ln(L)**
3. Інформаційний критерій, **AIC=-2(ln(L)-2)**
4. Коефіцієнт детермінації, $\mathbf{R^2}$
5. Урегульований (adjasted), $\mathbf{R^2}$

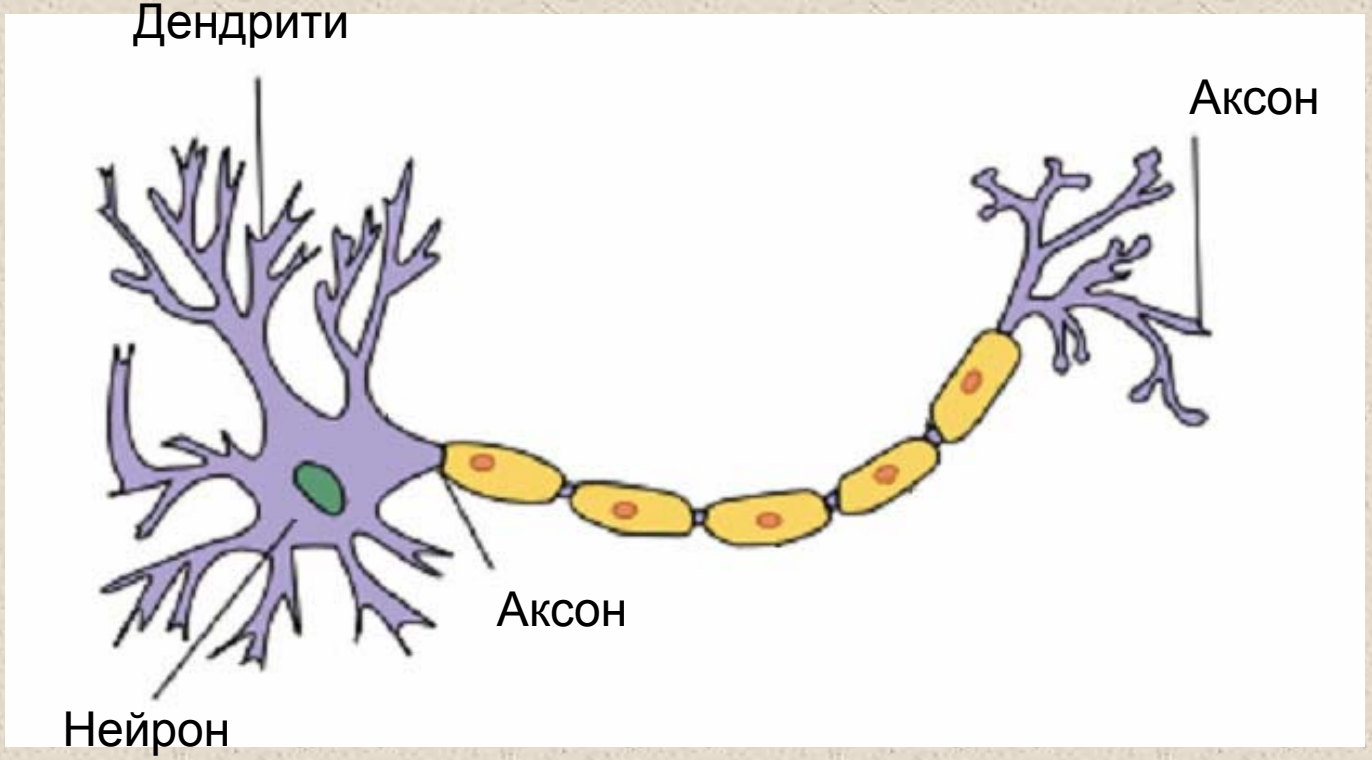
Процедура LOO

N

d_1	d_2	d_3	….	d_p	Class
					1
					0
					…

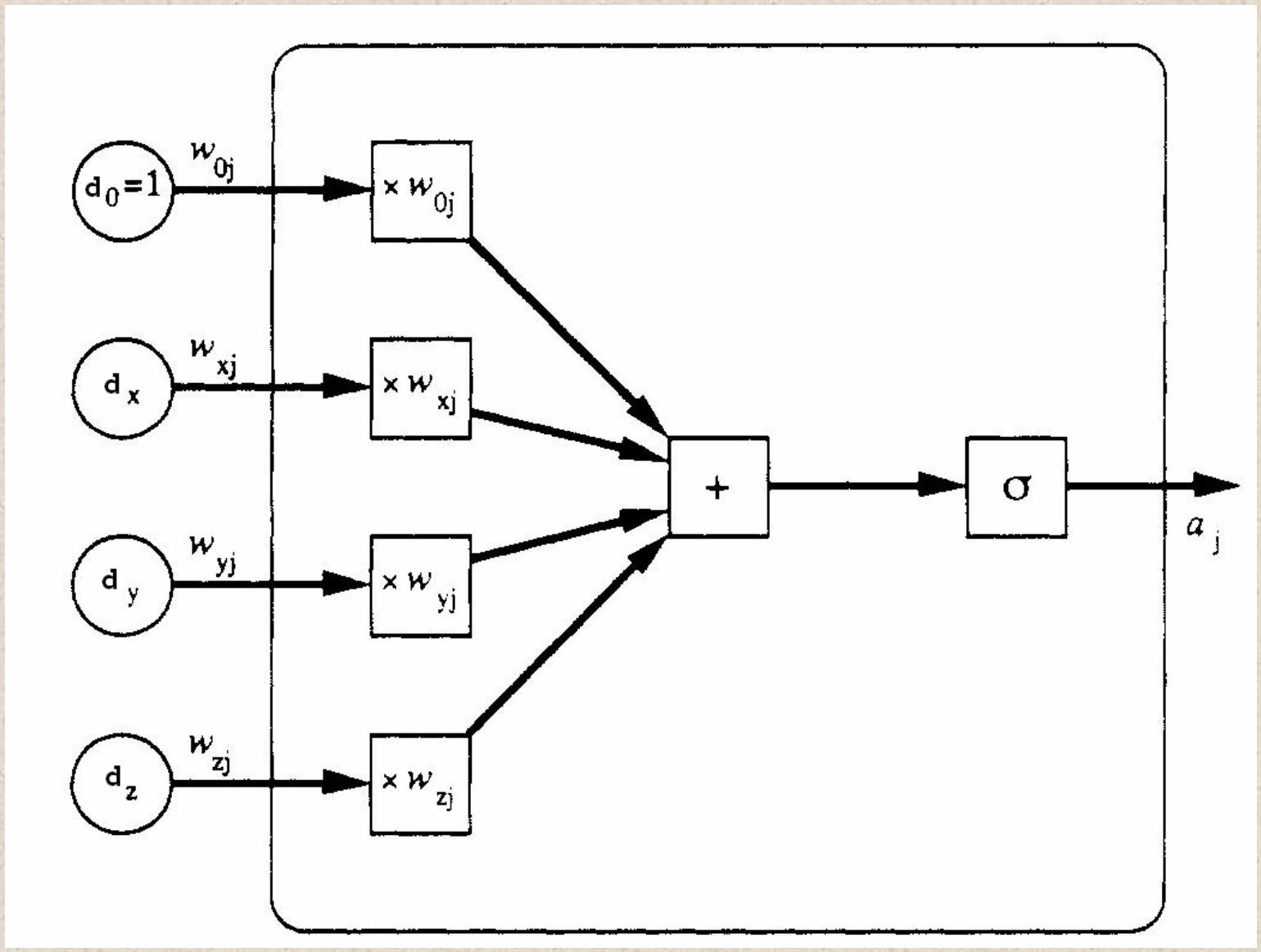
Метод Нейронних мереж

Структура нейрона

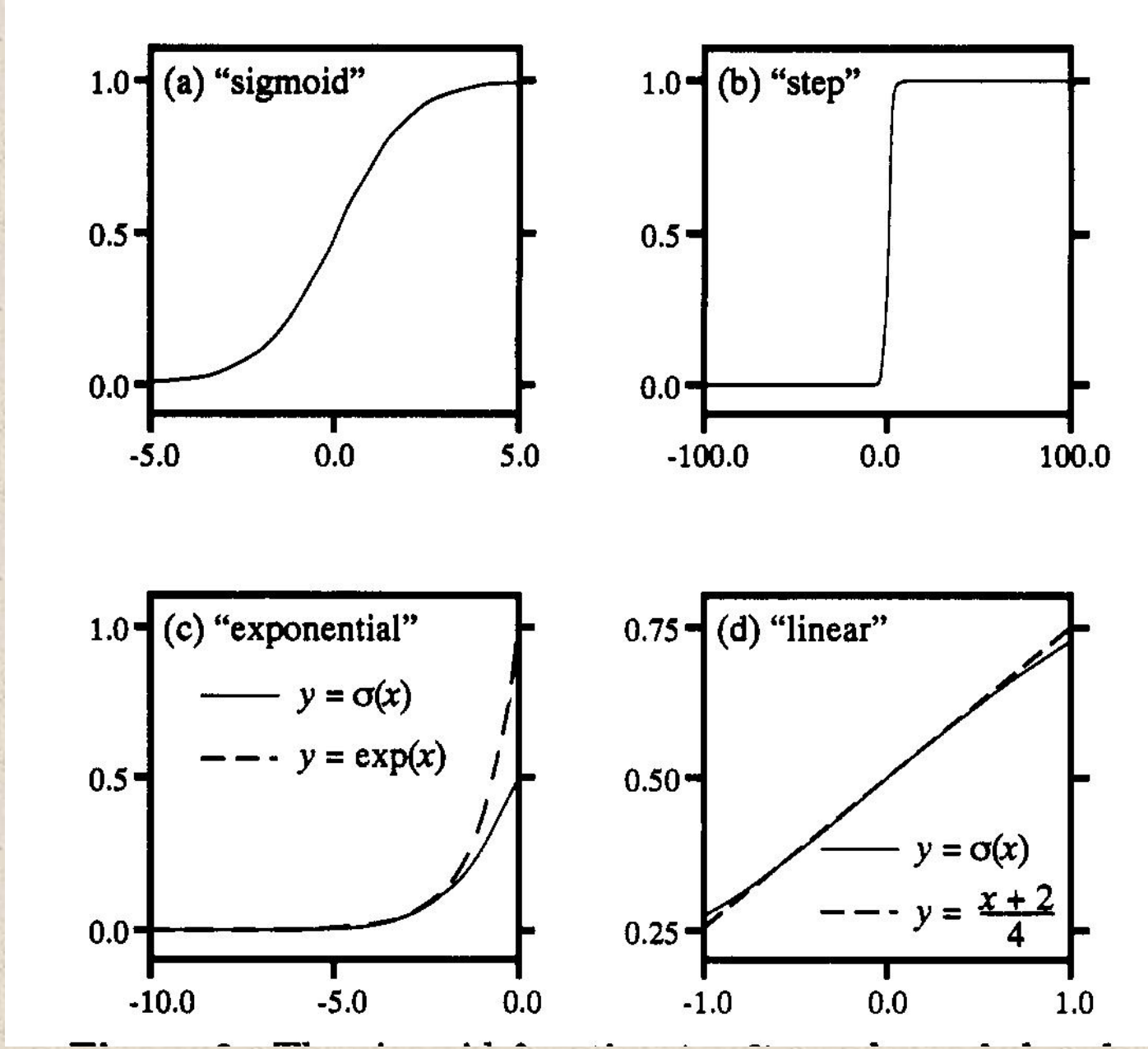


В человеческом организме ~ 10^{15} нейронов

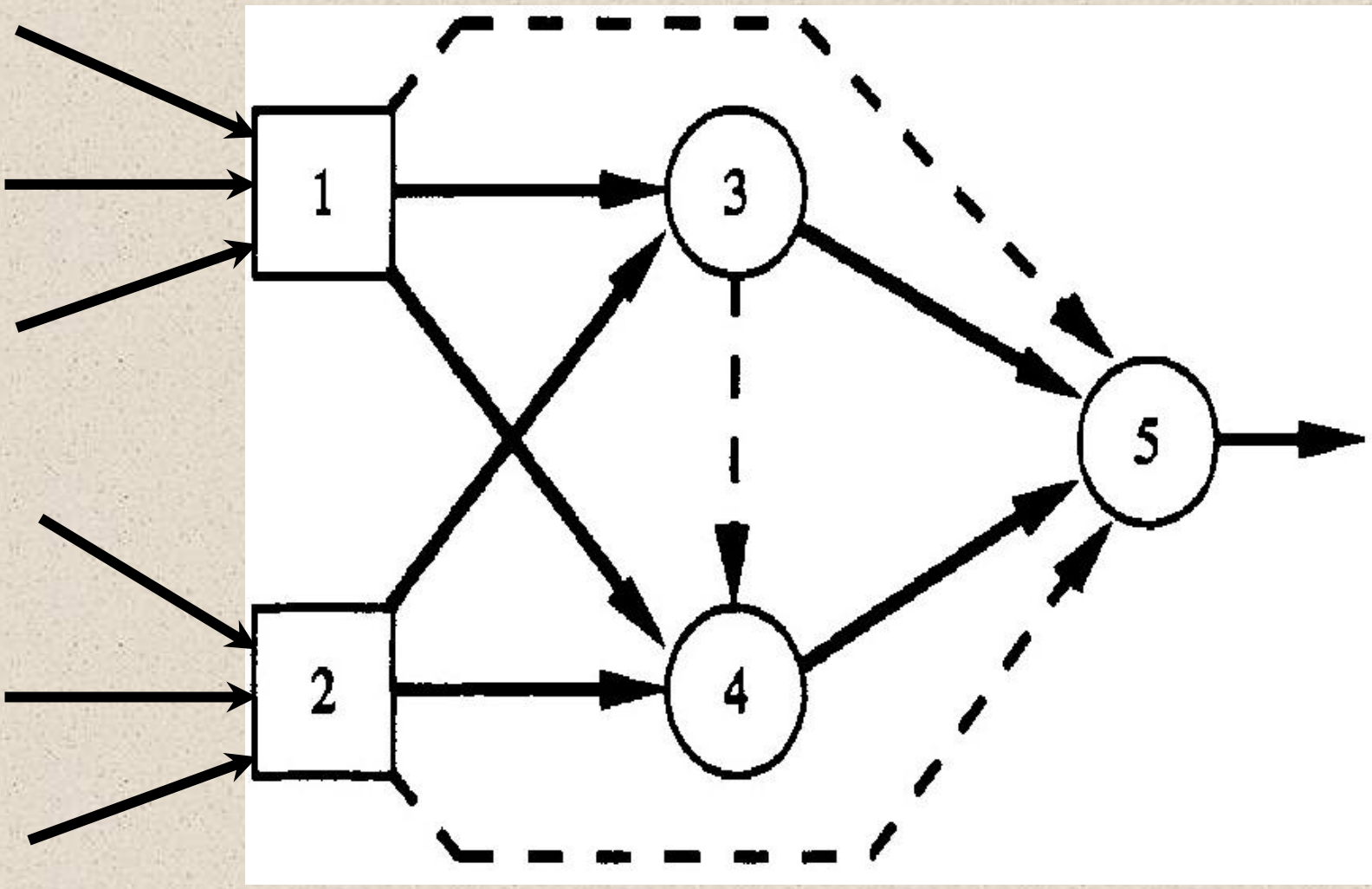
Модель Нейрона



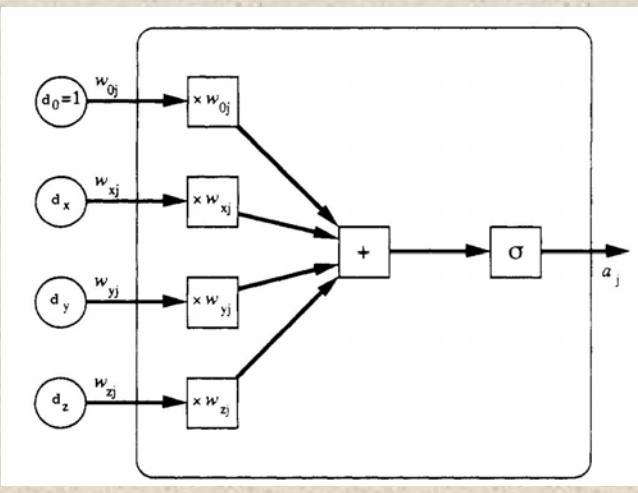
Вихідна функція σ



Модель Нейроної мережі



Навчання Нейроної мережі



Начальне наближенн W

Вычисляем a_j

d_0	d_x	d_y	d_z	a
				a_1
				a_2
				a_3
				a_4

Вычисляем Функцию ошибки F

$$F = \sum_i (a_i - a_i^{(calc)})^2$$

$$\left|\frac{\partial F}{\partial w_{ij}}\right| \le \varepsilon$$

да

STOP

нет

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