Combining Multiple Learners

- many different algorithms/learners
- NO FREE LUNCH THEOREM $\Rightarrow$ no single algorithm is always the BEST one.

- several algorithms
- several hyperparameters
  - $\rightarrow$ k-NN (3-NN, 5-NN, 7-NN)
  - $\rightarrow$ MLP (# of hidden nodes, activation functions)

MAIN IDEA $\Rightarrow$ DIVERSITY

(1) How do we generate base-learners that complement each other? if they produce the same predictions, they do not complement each other
very similar

	f_1		f_2	f_1+f_2	$\min(f_1, f_2)$	$\max(f_1, f_2)$
x_1	+	—	+	+		
$\vdots$	−	—	−	−	$\vdots$	$\vdots$
			$\vdots$	$\vdots$		
x_N	+	—	+	+		

(2) How do we combine the outputs of base-learners, for getting the maximum accuracy?

Generating Diverse Learners

① Different Algorithms:

→ MLP + k-NN
→ MLP + k-NN + SVM + DT
→ one parametric + one non-parametric

② Different Hyperparameters:

knn(3) ↓ local

knn(17) ↓ global.

MLP(50) ↳ hidden units

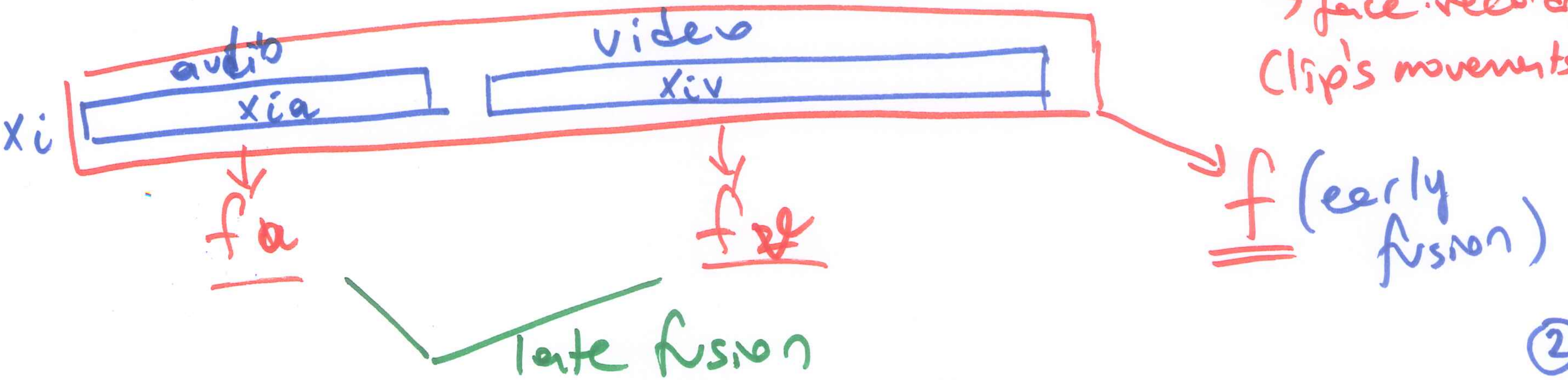
MLP(500) ↳ hidden units

③ Different Input Representations:

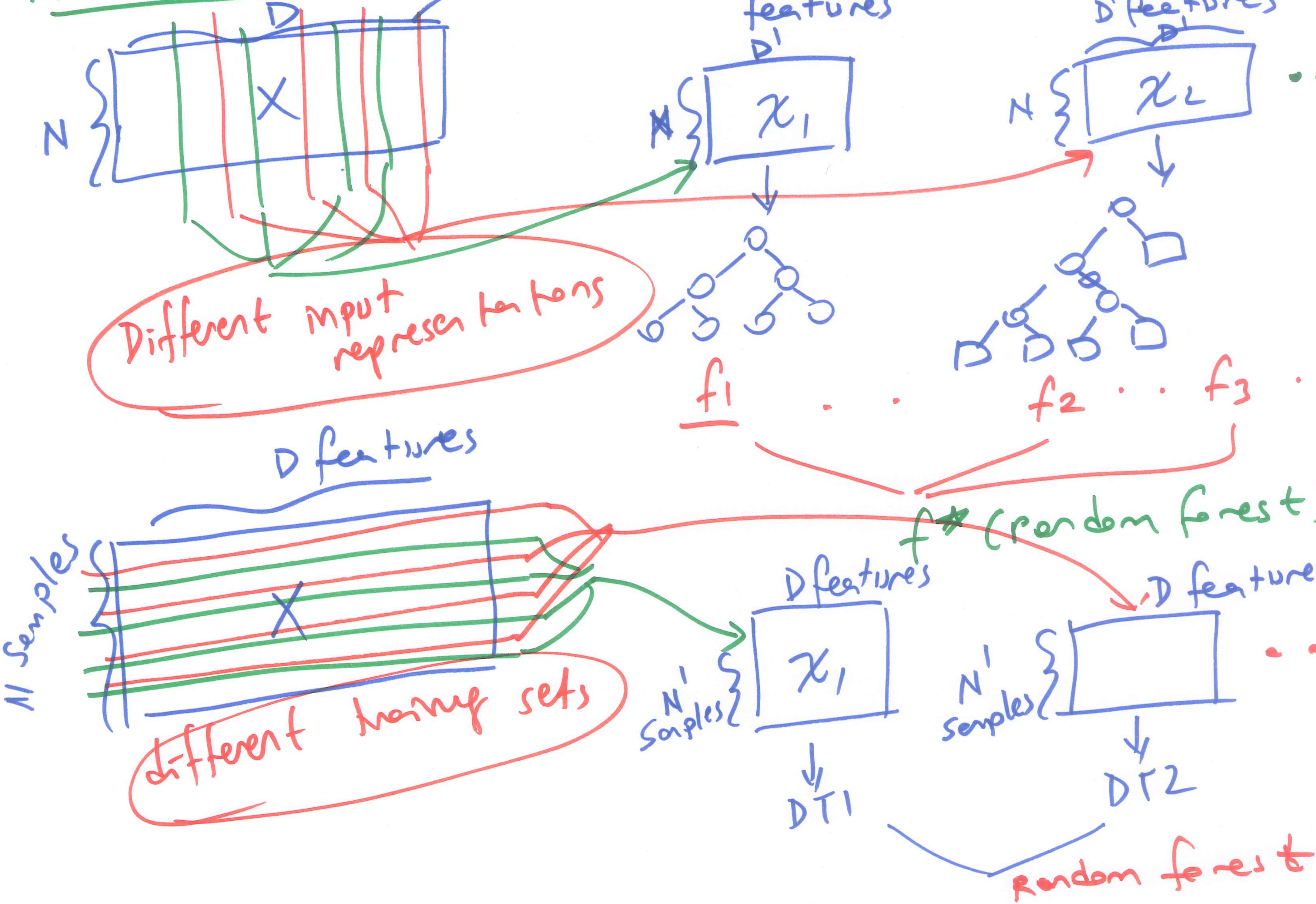
multi-view learning / multi-modal learning

different types of sensors / measurements / modalities

sensor fusion ⇒ speech recognition ⇒ audio → speech/voice recording
video → face recording (lip's movements)



RANDOM FOREST



Model Combination Strategies

multiple expert combination → global (learner fusion)
↘ local (learner selection)

$$f_1 \quad f_2 \quad \cdots \quad f_L$$

$$x^* \Rightarrow f_1(x^*) \quad f_2(x^*) \quad \cdots \quad f_L(x^*)$$ → L : # of base-learners.

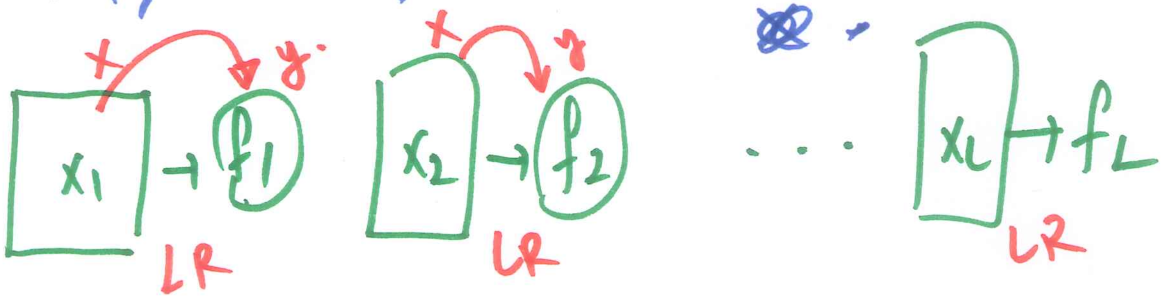
$$x^{**} \quad w_1 \cdot f_1(x^*) + w_2 f_2(x^*) + \cdots + w_L f_L(x^*)$$

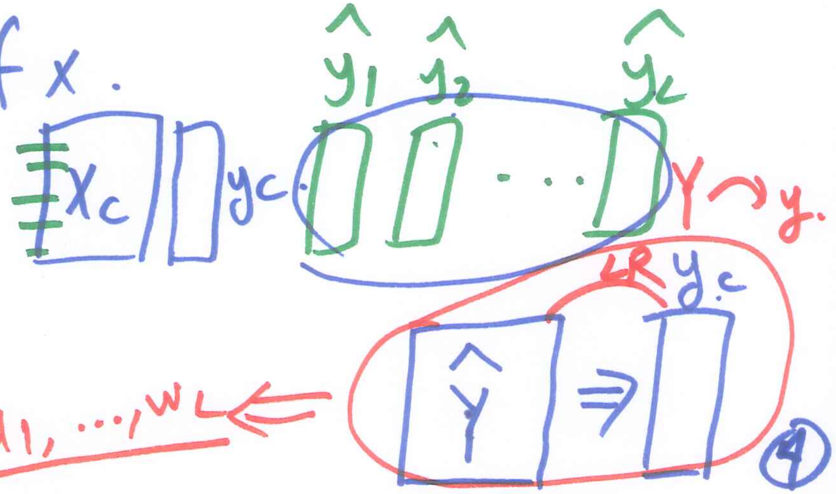
majority voting ⇒ $\underline{w_1 = w_2 = \cdots = w_L} = 1$.

We can learn $w_1, \ldots, w_L$ using another learner.

GLOBAL FUSION

$w_1, \ldots, w_L$ is not a function of x.



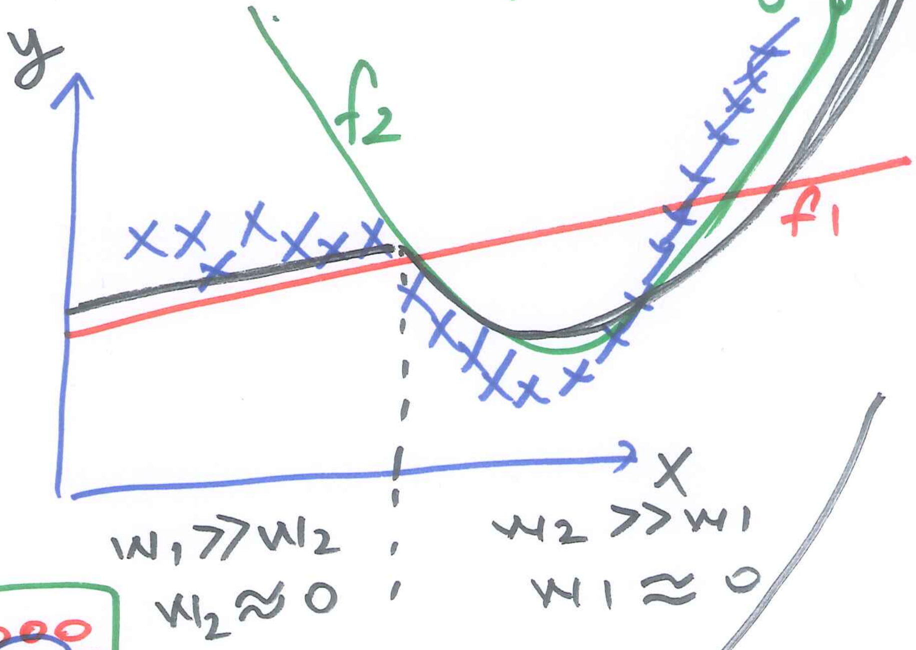


$$
x^* \to f_1 \to \hat{y}_1 \xrightarrow{w_1},\quad x^* \to f_2 \to \hat{y}_2 \xrightarrow{w_2},\quad \dots,\quad x^* \to f_L \to \hat{y}_L \xrightarrow{w_L}
$$

$$
f^* = \underline{w_1} \cdot \hat{y}_1 + \underline{w_2}\, \hat{y}_2 + \dots + \underline{w_L}\, \hat{y}_L
$$

$$
\boxed{w_1 = w_2 = \dots = w_L = \frac{1}{L}}
$$

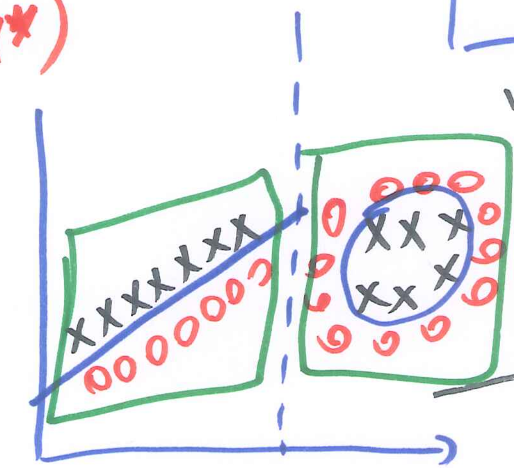
simple averaging.



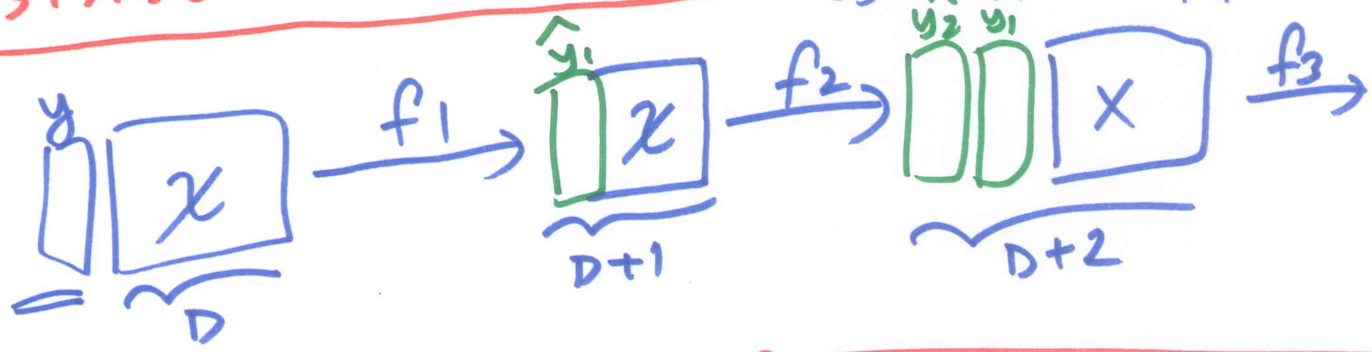
LOCAL FUSION

$$
x^* \to f_1 \to \hat{y}_1 \xrightarrow{w_1(x^*)},\quad x^* \to f_2 \to \hat{y}_2 \xrightarrow{w_2(x^*)},\quad \dots,\quad x^* \to f_L \to \hat{y}_L \xrightarrow{w_L(x^*)} \quad f^*
$$

"mixture of experts"



MULTISTAGE COMBINATION (serial approach)



Let us say we have L base-learners.

$d_j(x)$ $\quad d_1 \quad d_2 \; \ldots \quad d_L.$

$$y = \underbrace{f}_{\text{combination function}}\big(\overbrace{d_1, d_2, \ldots, d_L}^{\text{base-learners.}} \mid \underbrace{\Phi}_{\text{combination parameters.}}\big)$$

VOTING $\Rightarrow$ $\underbrace{y_i = \sum_j w_j d_{ji}}_{\text{ensembles}}$ where $w_j \geq 0$, $\sum_{j=1}^{L} w_j = 1$

ensembles

linear opinion models.

convex combination

6

$$w_j = \frac{1}{L} \Rightarrow y_i = \frac{1}{L}\sum_{j=1}^{L} d_{ji} \Rightarrow \text{sum-rule}$$

$$y_i = \underset{j}{\text{median}}\, d_{ji} \Rightarrow \text{median-rule}$$

$$y_i = \underset{j}{\text{minimum}}\, d_{ji} \Rightarrow \text{minimum-rule}$$

$$y_i = \underset{j}{\text{maximum}}\, d_{ji} \Rightarrow \text{maxmum-rule}$$

$$y_i = \prod_{j=1}^{L} d_{ji} \Rightarrow \text{product-rule.}$$

	c_1	c_2	c_3		
d_1	0.2	0.5	0.3	1.0	✓
d_2	0.0	0.6	0.4	1.0	✓
d_3	0.4	0.4	0.2	1.0	✓
sum	0.2	0.5	0.3	1.0	✓
median	0.2	0.5	0.3	1.0	?
min	0.0	0.4	0.2	0.6	✗
max	0.4	0.6	0.4	1.4	✗
product	0.	0.12	0.024	0.144	✗

Three classifiers.

d_1	d_2	d_3
$8/10$	$8/10$	$8/10$

20%

$P(\text{misclassification}) = P(\text{three of them are wrong}) + P(\text{two of them are wrong})$

$$10\% \simeq \frac{13}{125} = \frac{1}{125} + \frac{12}{125} = \frac{1}{5}\cdot\frac{1}{5}\cdot\frac{1}{5} + 3\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}$$