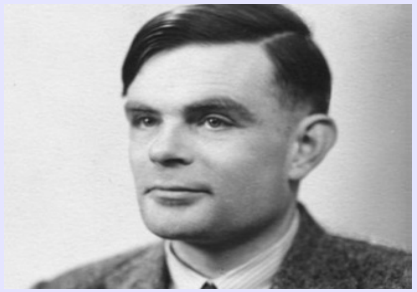
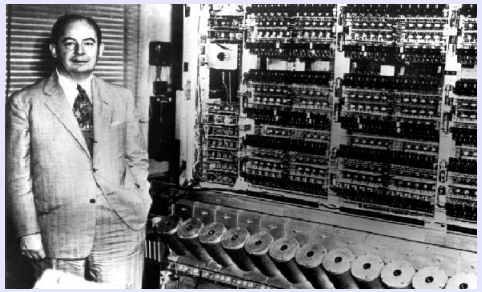




Information, Calcul et Communication

Théorie: Représentation de l'Information (1.4)

R. Boulic

Représentation des symboles

- De l'alphabet aux idéogrammes

Plan

Lien avec les leçons précédentes
- **Rappel des domaines d'applications**
- **Une représentation est une convention**
- **Vers l'unité élémentaire d'information (exercices)**

Manipulation sur les nombres entiers
- **Opérations et domaine couvert**

La virgule flottante: Pourquoi ? Comment ?
- **Un exemple qui pose problème**

Retour à la représentation des symboles
- **De l'alphabet aux idéogrammes**

Comment représenter un alphabet ?

Ensemble fini de signes

Considéré avec des variantes : Majuscule / minuscule

Associé aux symboles des chiffres, de ponctuation

Convention la plus large possible est nécessaire:

Table ASCII de base codifie 2^7 caractères
American Standard Code for Information Interchange
`http://www.asciitable.com/`

Code ASCII

Dec	Hx	Oct	Char	Dec	Hx	Oct	Char	Dec	Hx	Oct	Char	Dec	Hx	Oct	Char
0	0	000	NUL (null)	32	20	040	SPACE	64	40	100	@	96	60	140	`
1	1	001	SOH (start of heading)	33	21	041	!	65	41	101	A	97	61	141	a
2	2	002	STX (start of text)	34	22	042	"	66	42	102	B	98	62	142	b
3	3	003	ETX (end of text)	35	23	043	#	67	43	103	C	99	63	143	c
4	4	004	EOT (end of transmission)	36	24	044	$	68	44	104	D	100	64	144	d
5	5	005	ENQ (enquiry)	37	25	045	%	69	45	105	E	101	65	145	e
6	6	006	ACK (acknowledge)	38	26	046	&	70	46	106	F	102	66	146	f
7	7	007	BEL (bell)	39	27	047	'	71	47	107	G	103	67	147	g
8	8	010	BS (backspace)	40	28	050	(	72	48	110	H	104	68	150	h
9	9	011	TAB (horizontal tab)	41	29	051	)	73	49	111	I	105	69	151	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	74	4A	112	J	106	6A	152	j
11	B	013	VT (vertical tab)	43	2B	053	+	75	4B	113	K	107	6B	153	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	76	4C	114	L	108	6C	154	l
13	D	015	CR (carriage return)	45	2D	055	-	77	4D	115	M	109	6D	155	m
14	E	016	SO (shift out)	46	2E	056	.	78	4E	116	N	110	6E	156	n
15	F	017	SI (shift in)	47	2F	057	/	79	4F	117	O	111	6F	157	o
16	10	020	DLE (data link escape)	48	30	060	0	80	50	120	P	112	70	160	p
17	11	021	DC1 (device control 1)	49	31	061	1	81	51	121	Q	113	71	161	q
18	12	022	DC2 (device control 2)	50	32	062	2	82	52	122	R	114	72	162	r
19	13	023	DC3 (device control 3)	51	33	063	3	83	53	123	S	115	73	163	s
20	14	024	DC4 (device control 4)	52	34	064	4	84	54	124	T	116	74	164	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	85	55	125	U	117	75	165	u
22	16	026	SYN (synchronous idle)	54	36	066	6	86	56	126	V	118	76	166	v
23	17	027	ETB (end of trans. block)	55	37	067	7	87	57	127	W	119	77	167	w
24	18	030	CAN (cancel)	56	38	070	8	88	58	130	X	120	78	170	x
25	19	031	EM (end of medium)	57	39	071	9	89	59	131	Y	121	79	171	y
26	1A	032	SUB (substitute)	58	3A	072	:	90	5A	132	Z	122	7A	172	z
27	1B	033	ESC (escape)	59	3B	073	;	91	5B	133	[	123	7B	173	{
28	1C	034	FS (file separator)	60	3C	074	<	92	5C	134	\	124	7C	174	|
29	1D	035	GS (group separator)	61	3D	075	=	93	5D	135	]	125	7D	175	}
30	1E	036	RS (record separator)	62	3E	076	>	94	5E	136	^	126	7E	176	~
31	1F	037	US (unit separator)	63	3F	077	?	95	5F	137	_	127	7F	177	DEL

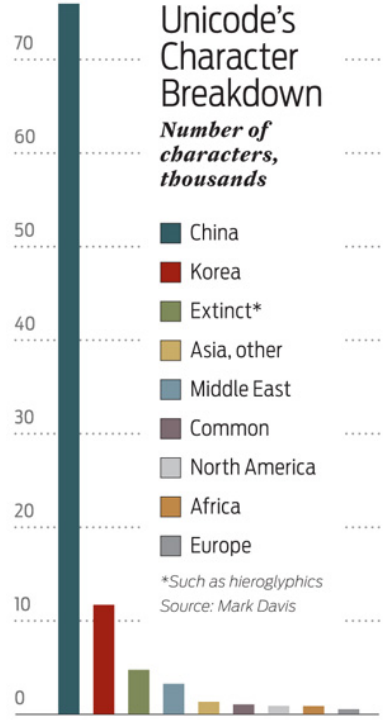
Au-delà du code ASCII de base

ASCII étendu sur 8 bits: Codes 0x80 à 0xFF
Le code étendu ISO 8859 Latin1 offre les caractères accentués minuscules et majuscules des langues occidentales: é è ê à ä ö ü …

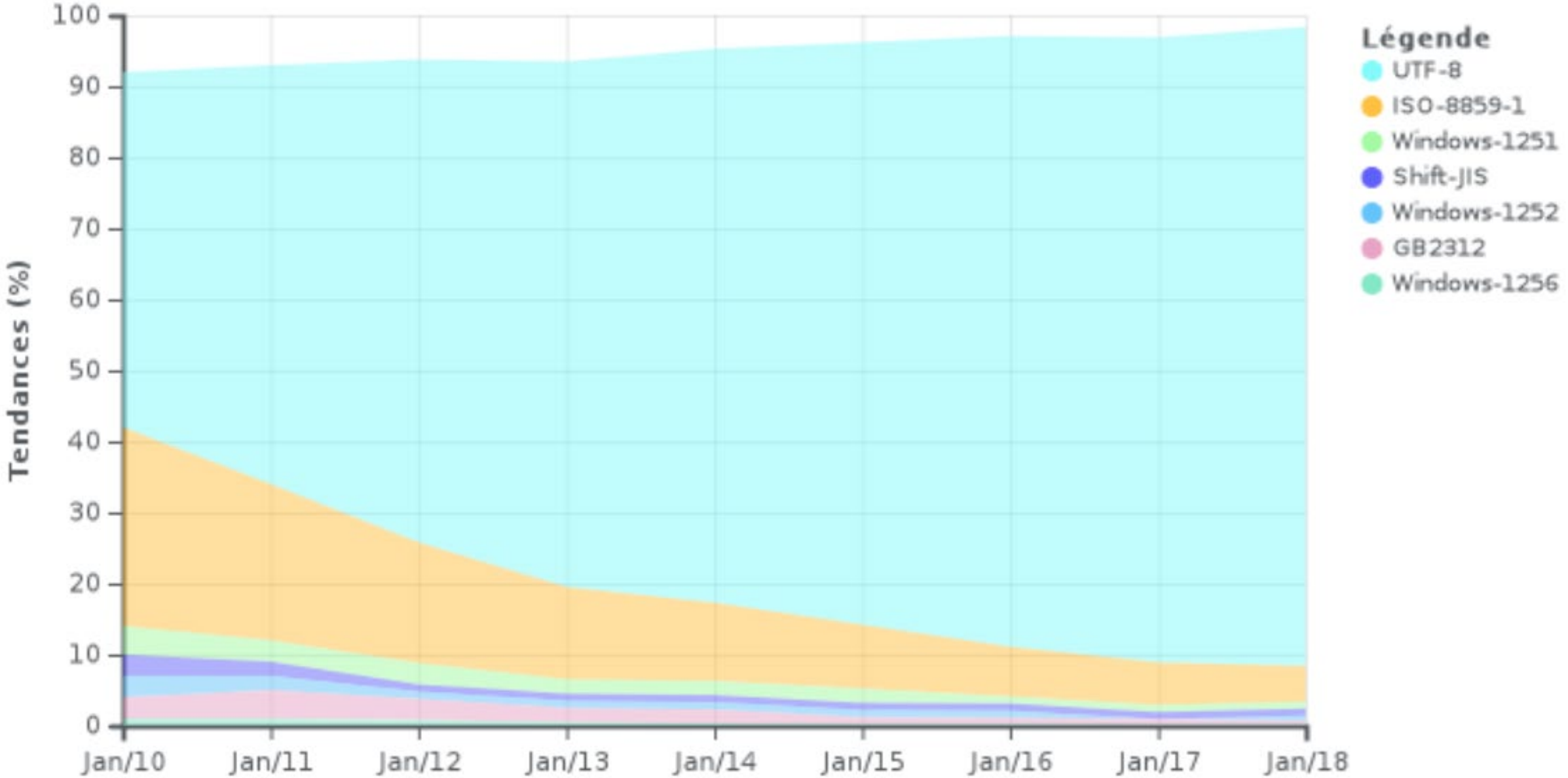
La norme **UNICODE** permet d'intégrer les autres langues, > 109’000 caractères pour 93 écritures dont le chinois.

- les 256 codes d' ISO 8859 sont au début de l'UNICODE

- **UTF-8** est un codage des caractères UNICODE comprenant de 1 à 4 octets. Il est recommandé mais son usage n’est pas encore généralisé.



[IEEE Spectrum 7/12]

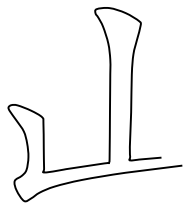


https://fr.wikipedia.org/wiki/UTF-8

Du code multi-byte (idéogramme) à l'image

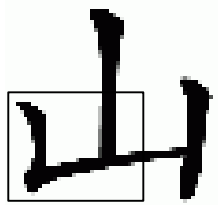
shan = montagne
le symbole peut être codé par 1 à 4 octets en UTF-8

MAIS **la représentation du symbole = son image** demande plus d'information.
Plusieurs approches sont possibles, du plus haut vers le plus bas niveau :



1) Préciser la **police de caractères** = *« style classique »*.

2) Caractériser les **contours** de la forme par un ensemble de **courbes** mathématiques paramétrées (silhouette). C'est la manière dont les polices de caractères sont construites.



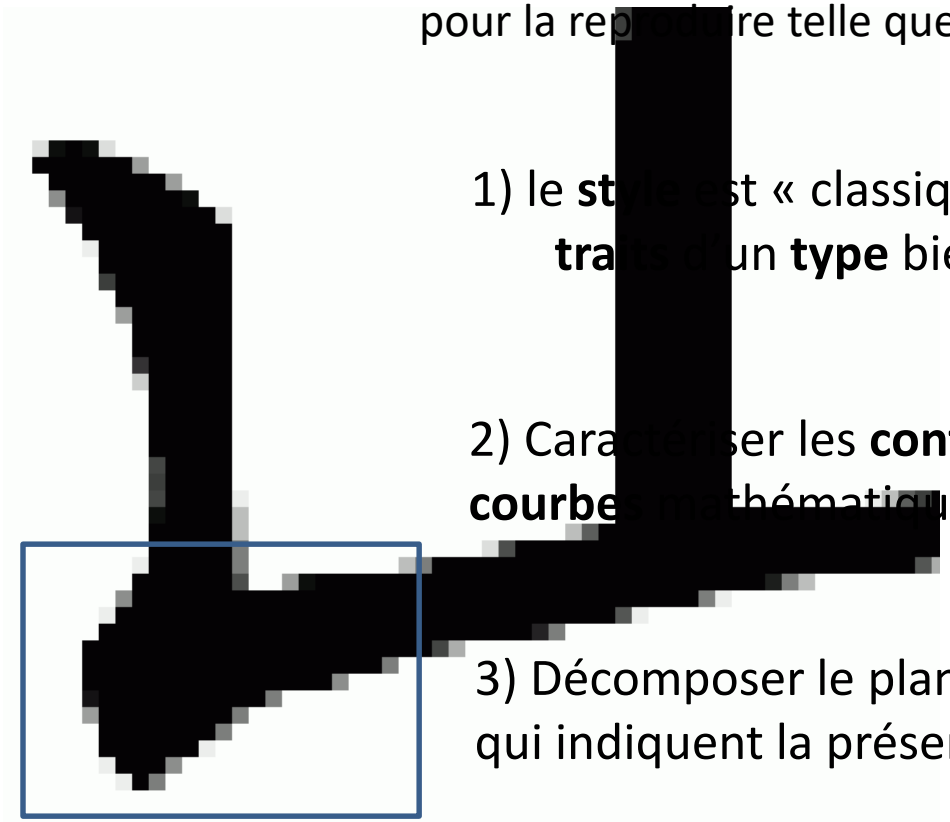
60x60

3) Décomposer le plan de l'image en un **ensemble de cellules** qui indiquent la quantité d'encre (pixel).

Du code multi-byte (idéogramme) à l'image

shan = montagne
le symbole peut être codé par 1 à 4 octets en UTF-8

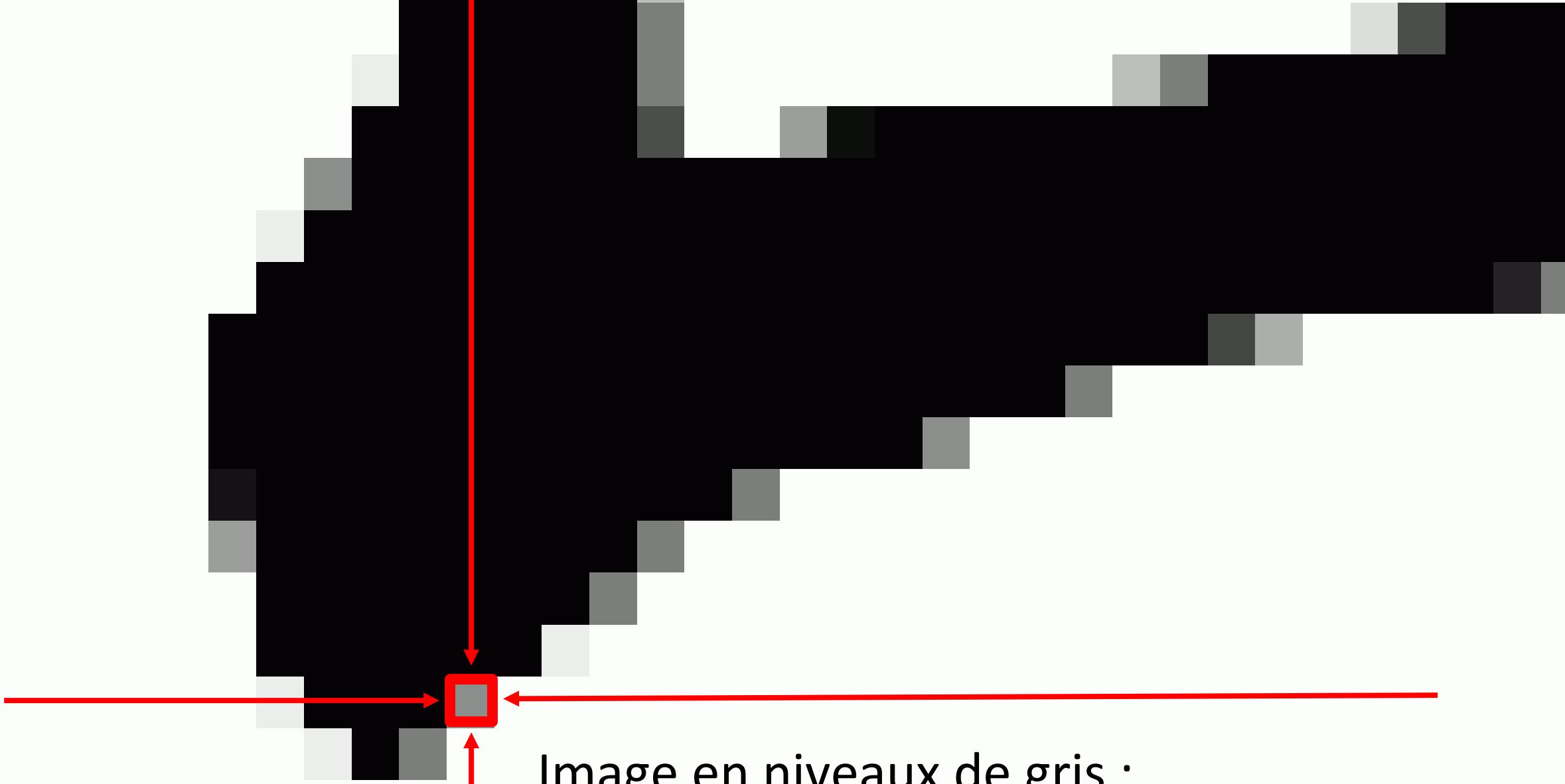
MAIS **l'image du symbole** demande plus d'information pour la reproduire telle quelle. <u>Plusieurs approches:</u>



1) le **style** est « classique ». Il comporte plusieurs **traits** d'un **type** bien précis parmi 37

2) Caractériser les **contours** de la forme par un ensemble de **courbes** mathématiques paramétrées (silhouette).

3) Décomposer le plan de l'image en un **ensemble de cellules** qui indiquent la présence d'encre (pixel).



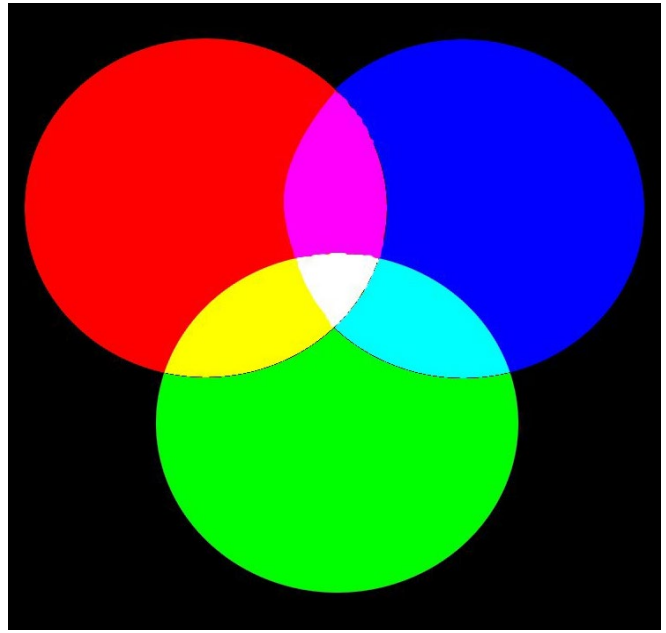
bitmap

Image en niveaux de gris :

chaque pixel mémorise une intensité entre 0 (noir) et le maximum définit par le nombre de bits retenu (blanc)

Image en couleur:

chaque pixel mémorise l'intensité de 3 composantes primaires dont la combinaison permet de restituer un espace de couleurs.



Le codage RGB (Red, Green, Blue)

- synthèse additive des couleurs: Rouge + Vert donne .....

- niveaux de gris lorsque les trois composantes sont égales noir(0,0,0) et le maximum définit par le nombre de bits retenu (blanc)

- parfois complété d'une 4ième composante Alpha (transparence) pour les applications graphiques.

Taille d'une image UXGA 1600x1200 x 3octets/pixel = 5'760'000 octets

Résumé

La représentation des **symboles alphanumériques** est maintenant standardisée à l'échelle de la planète avec UNICODE/UTF8

La représentation d'images requiert de mémoriser au moins une donnée par élément de l'image (pixel), ce qui impacte rapidement la quantité de mémoire nécessaire pour les mémoriser.