Ιόνιο Πανεπιστήμιο – Τμήμα Πληροφορικής

Εισαγωγή στην Επιστήμη των Υπολογιστών

2019-20

Αναπαράσταση Μη Αριθμητικών Δεδομένων

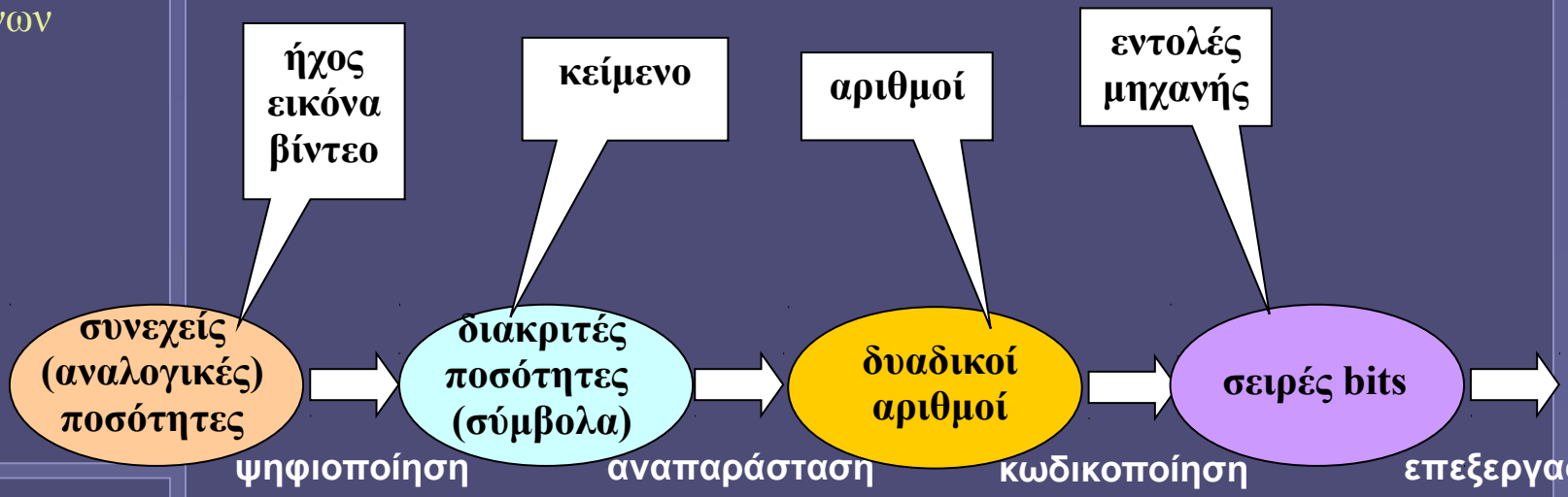
(κείμενο, ήχος και εικόνα στον υπολογιστή)

http://mixstef.github.io/courses/csintro/

Μ.Στεφανιδάκης

Αναπαράσταση δεδομένων

- Αναπαράσταση δεδομένων



!

Δεδομένα: ανεξάρτητα από τύπο και προέλευση, στον υπολογιστή υπάρχουν σε μία μορφή: 0 και 1

- Ψηφιοποίηση
  - Διαδικασία μετατροπής συνεχών τιμών σε διακριτά σύμβολα
- Αναπαράσταση
  - Διαδικασία αντιστοίχισης συμβόλων σε δυαδικούς αριθμούς
- Κωδικοποίηση
  - Αποθήκευση δυαδικών αριθμών σε σειρές bits

Η ερμηνεία της αναπαράστασης

- Αναπαράσταση δεδομένων

!

Στα ερωτήματα αυτά μπορεί να απαντήσει μόνο ο προγραμματιστής της εφαρμογής που χειρίζεται τα δεδομένα!

- Κάπου στη μνήμη του υπολογιστή…
  - Βρίσκεται αποθηκευμένη η σειρά bits

    0100110111010001
- Πόσα σύμβολα αναπαριστά;
  - Πόσα bits ανά σύμβολο;
- Ποιος ο τύπος των δεδομένων;
- Ποια συγκεκριμένη ποσότητα συμβολίζει;
- Πώς θα το χειριστεί ο υπολογιστής;

Αναπαράσταση με δυαδικούς αριθμούς

- Αναπαράσταση δεδομένων

- **Σειρά *n* bits**
  - Δυαδικός αριθμός με ***n*** **bits** (n≥1) μπορεί να αναπαραστήσει 2^n διαφορετικά σύμβολα
- **Μη αριθμητικά δεδομένα**
  - Κείμενο, εντολές μηχανής, ήχος, εικόνα…
    - Σύνολο διαφορετικών αντικειμένων (**συμβόλων**)
  - Αντιστοίχιση κάθε συμβόλου σε μοναδικό δυαδικό αριθμό (code point)
    - "Αναπαράσταση"
    - Η ακριβής αντιστοίχιση συνήθως ορίζεται σε ένα **πρότυπο** (standard)

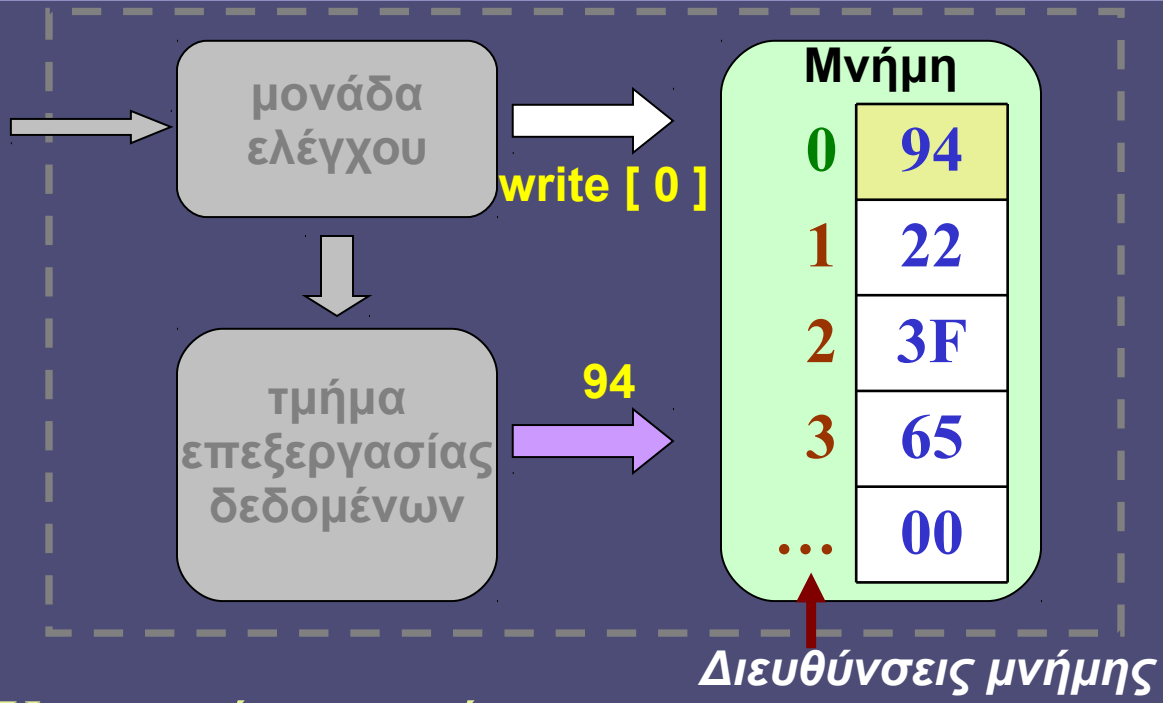
Η επικοινωνία με τη μνήμη

- Αναπαράσταση δεδομένων

- Μοντέλο μνήμης
  - Συστοιχία αποθηκευτικών θέσεων
  - Σε κάθε θέση αποθηκεύεται (συνήθως) 1 byte
  - Κάθε θέση διαθέτει μοναδική διεύθυνση
    - Επιλογή θέσης κατά την προσπέλαση (ανάγνωση-εγγραφή)

Η επικοινωνία με τη μνήμη

- Αναπαράσταση δεδομένων



- Χωρητικότητα μνήμης
  - Εκφράζεται σε πολλαπλάσια του byte
  - 1 KByte (KB) = 1024 Bytes (2^{10})
  - 1 MByte (MB) = 1024 KBytes (2^{10})
    - κλπ

;

Με διεύθυνση των n bits, πόσες διαφορετικές θέσεις μνήμης μπορούμε να προσπελάσουμε;

Θέματα αποθήκευσης δυαδικών αριθμών

- Αναπαράσταση δεδομένων

;

Πώς σχετίζεται η σειρά αποθήκευσης των bytes με τα "Ταξίδια του Γκιούλιβερ";

- Όταν
  - Ένας δυαδικός αριθμός χρειάζεται περισσότερα από ένα byte για να αποθηκεύσει τα ψηφία του
- Παράδειγμα: 3FC (hex) = 11 1111 1100

  Χρειάζονται 2 bytes!

  0000 0011 1111 1100

  περισσότερο σημαντικό byte ***λιγότερο σημαντικό byte***
- Προφανώς σε συνεχόμενες θέσεις μνήμης

  Αλλά: ποιο byte αποθηκεύεται πρώτο;

Θέματα αποθήκευσης δυαδικών αριθμών

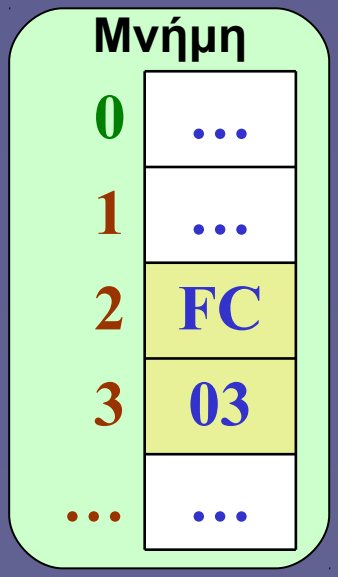
- Αναπαράσταση δεδομένων

αποθηκεύοντας το 03FC

00000011 11111100

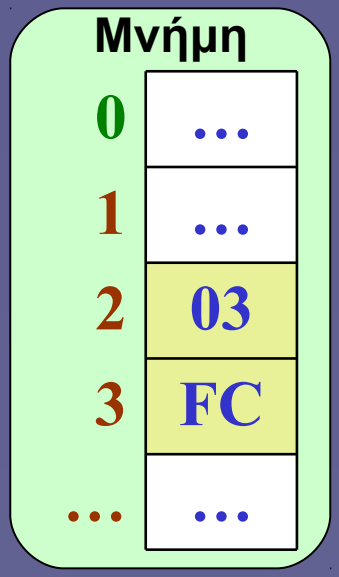
i

Στο Διαδίκτυο τα δεδομένα διακινούνται σε big-endian μορφη



"little-endian"

Το λιγότερο σημαντικό byte στη θέση μνήμης με μικρότερη διεύθυνση



"big-endian"

Το περισσότερο σημαντικό byte στη θέση μνήμης με μικρότερη διεύθυνση

Αρχικές αναπαραστάσεις κειμένου

- Αναπαράσταση δεδομένων
- Κείμενο

- **Οι πρώτες αναπαραστάσεις κειμένου**
  - Στον υπολογιστή
  - 6-7 bits ανά χαρακτήρα
    - Πόσοι διαφορετικοί χαρακτήρες;
- **Μη εκτυπώσιμοι χαρακτήρες**
  - Χαρακτήρες ελέγχου
    - Ιδιαίτερα χρήσιμοι για τις συσκευές εξόδου της εποχής (εκτυπωτές, τηλέτυπα…)
    - Νέα γραμμή (LINE FEED – LF)
    - Επιστροφή κεφαλής εκτύπωσης (CARRIAGE RETURN – CR)
    - Καμπανάκι (BELL) κλπ

Κώδικας ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

i

ASCII: American Standard Code for Information Interchange

- Βασικό αρχικό πρότυπο αναπαράστασης κειμένου
  - 7 bits ανά χαρακτήρα

STANDARD ASCII ΚΩΔΙΚΑΣ

hex	char	hex	char	hex	char
20		40	@	60	`
21	!	41	A	61	a
22	"	42	B	62	b
23	#	43	C	63	c
24	$	44	D	64	d
25	%	45	E	65	e
26	&	46	F	66	f
27	'	47	G	67	g
28	(	48	H	68	h
29	)	49	I	69	i
2A	*	4A	J	6A	j
2B	+	4B	K	6B	k
2C	,	4C	L	6C	l
2D	-	4D	M	6D	m
2E	.	4E	N	6E	n

Κείμενο σε κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

;

Με 7 bits ανά χαρακτήρα και χρήση bytes, 1 bit μένει αχρησιμοποίητο. Πόσοι επιπλέον χαρακτήρες με το bit αυτό;

- 7 bits ανά χαρακτήρα
  - 128 χαρακτήρες
  - Αναπαράσταση με τους αριθμούς 0…127
- Κανονικοί χαρακτήρες (εκτυπώσιμοι)
  - 32…64, 91…96, 123…126 = σημεία στίξης κ.ά. (32 = SPACE!)
  - 65…90 = κεφαλαία λατινικά (A-Z)
  - 97…122 = πεζά λατινικά (a-z)
- Χαρακτήρες ελέγχου (μη εκτυπώσιμοι)
  - 0…31, 127 – επιζούν τα: 9 (TAB), 13/10 (CR/LF, σήμανση “νέας γραμμής”)

Κείμενο σε κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

!

Εφόσον η κωδικοποίηση είναι με 1 byte ανά χαρακτήρα, δεν τίθεται θέμα "little-" ή "big-endian"

- Παράδειγμα

H	a	v	e		a		n	i	c	e		d	a	y	!
72	97	118	101	32	97	32	110	105	99	101	32	100	97	121	33

- Γλώσσες προγραμματισμού
  - Συμβολοσειρά (string)
  - Σε γλώσσες όπως η C, το 0 (αριθμητικό) συμβολίζει το τέλος της συμβολοσειράς
  - Ο υπολογιστής μπορεί να κάνει πράξεις (π.χ. σύγκριση) με τη συμβολοσειρά

Επεκτάσεις κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

!

Χρησιμοποιώντας τον ISO-8859-1 δεν είναι δυνατή η αναπαράσταση των ελληνικών!

- Χρήση του 1 επιπλέον bit του byte
  - 128 + 128 χαρακτήρες, αριθμοί 0…255
  - 0…127 αντιστοιχούν στον αρχικό ASCII
  - 127…255: επεκταμένα αλφάβητα
- Επέκταση αλφαβήτων (πρότυπα)
  - Χαρακτήρες που δεν υπάρχουν στον ASCII
  - Διαφορετικά ανά γλώσσα! Π.χ.:
    - ISO-8859-1: Δυτική Ευρώπη (Å, Ñ, Æ,ä, ø κλπ)
    - ISO-8859-7: Νέα Ελληνικά
    - …και πολλά άλλα πρότυπα για τις υπόλοιπες γλώσσες
  - Επίσης: μη πρότυπες λύσεις
    - Για Windows, Mac ..

Κώδικας ISO-8859-7

- Αναπαράσταση δεδομένων
- Κείμενο

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*unused*															
1x																
2x	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*unused*															
9x																
Ax	NBSP	‘	’	£	€	₯	¦	§	¨	©	ͺ	«	¬	SHY		―
Bx	°	±	²	³	΄	΅	Ά	·	Έ	Ή	Ί	»	Ό	½	Ύ	Ώ
Cx	ΐ	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
Dx	Π	Ρ		Σ	Τ	Υ	Φ	Χ	Ψ	Ω	Ϊ	Ϋ	ά	έ	ή	ί
Ex	ΰ	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο
Fx	π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	

[Wikipedia]

Κείμενο σε κώδικα ISO-8859-7

- Αναπαράσταση δεδομένων
- Κείμενο

!

Οι αναπαραστάσεις αλφαβήτων με 1 byte ανά χαρακτήρα τείνουν να καταργηθούν!

- Παράδειγμα

Γ	ε	ι	α		σ	ο	υ	!
195	229	233	225	32	243	239	245	33

- Επέκταση κώδικα ASCII
  - 0…127 όπως στον ASCII
  - 128…159 πρόσθετοι χαρακτήρες ελέγχου
  - 160…255 ελληνικά και σχετικά σύμβολα

Πρότυπο Unicode

- Αναπαράσταση δεδομένων
- Κείμενο

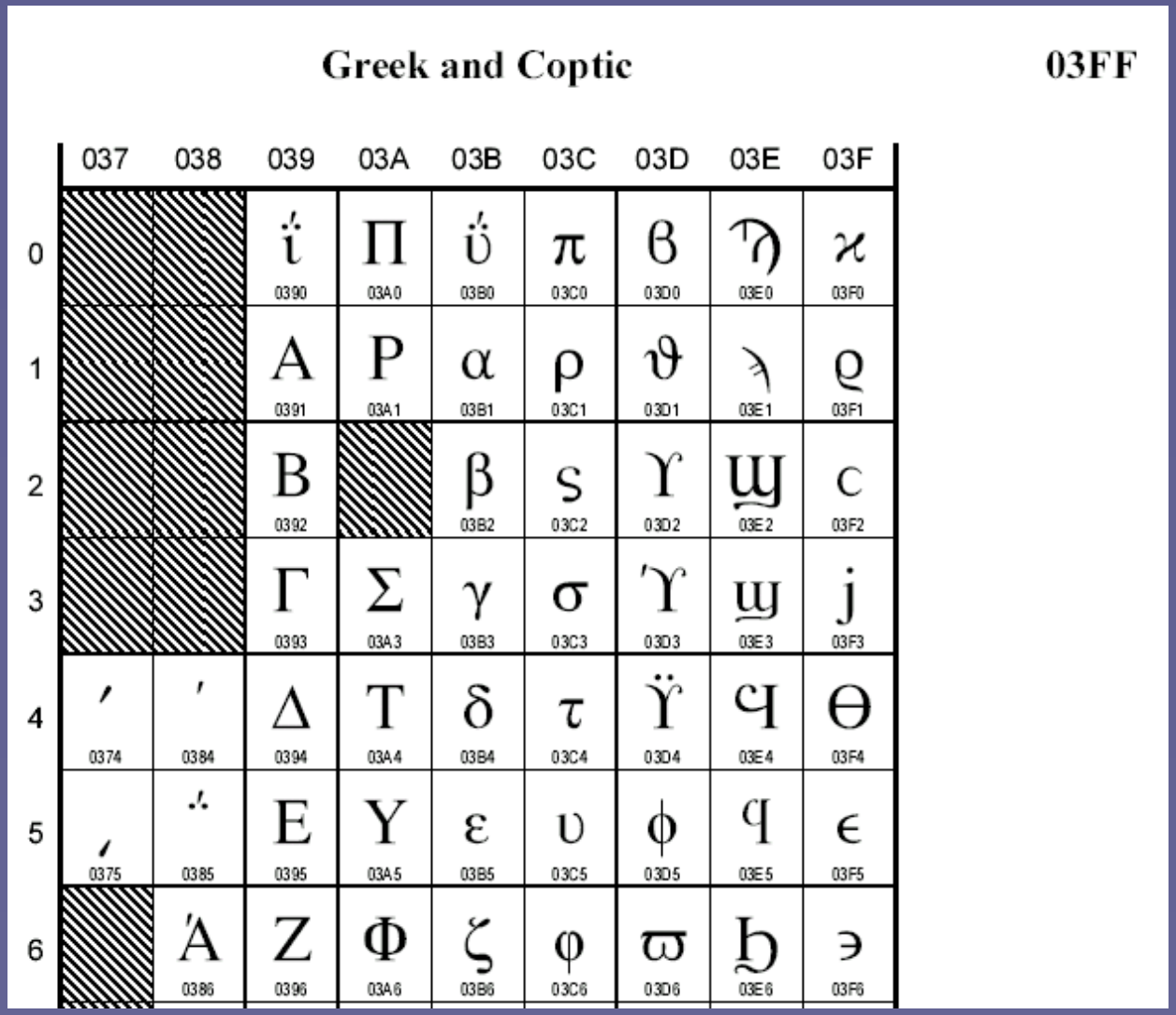
!

Με περισσότερα από 1 bytes ανά χαρακτήρα τίθεται θέμα σειράς αποθήκευσης των bytes!

- **Για την αναπαράσταση όλων των αλφαβήτων!**
  - Έχουν οριστεί σχεδόν 100.000 χαρακτήρες
  - Καλύπτει ιδεογράμματα, φωνητικές αναπαραστάσεις κλπ
  - Θα μπορούσε να καλύψει πάνω από 1 εκ. χαρακτήρες! (0 ... 10FFFF)
  - Κάθε χαρακτήρας αναπαρίσταται με **περισσότερα από ένα bytes**
    - Συνήθεις κωδικοποιήσεις: UCS-2 (ή UTF-16) και UTF-8
  - Το πρότυπο Unicode περιέχει επίσης
    - πληροφορία ισοδύναμων ή παρόμοιων χαρακτήρων
    - οδηγίες συνδυασμών τόνων/διακριτικών και γραμμάτων

Ελληνικά και Unicode

- Αναπαράσταση δεδομένων
- Κείμενο



Κείμενο σε Unicode

- Αναπαράσταση δεδομένων
- Κείμενο

- Παράδειγμα

Γ	ε	ι	α		σ	ο	υ	!
915	949	953	945	32	963	959	965	33
0393	03B5	03B9	03B1	0020	03C3	03BF	03C5	0021

δεκαεξαδικό

Κωδικοποίηση UCS-2 (big-endian)

03	93	03	B5	03	B9	03	B1	00	20	03	C3	03	BF	03	C5	00	21

Κωδικοποίηση UCS-2 (little-endian)

93	03	B5	03	B9	03	B1	03	20	00	C3	03	BF	03	C5	03	21	00

Unicode σε κωδικοποίηση UTF-8

- Αναπαράσταση δεδομένων
- Κείμενο

!

Η κωδικοποίηση UTF-8 τείνει να επικρατήσει σε όλα τα προγράμματα που χειρίζονται κείμενα Unicode!

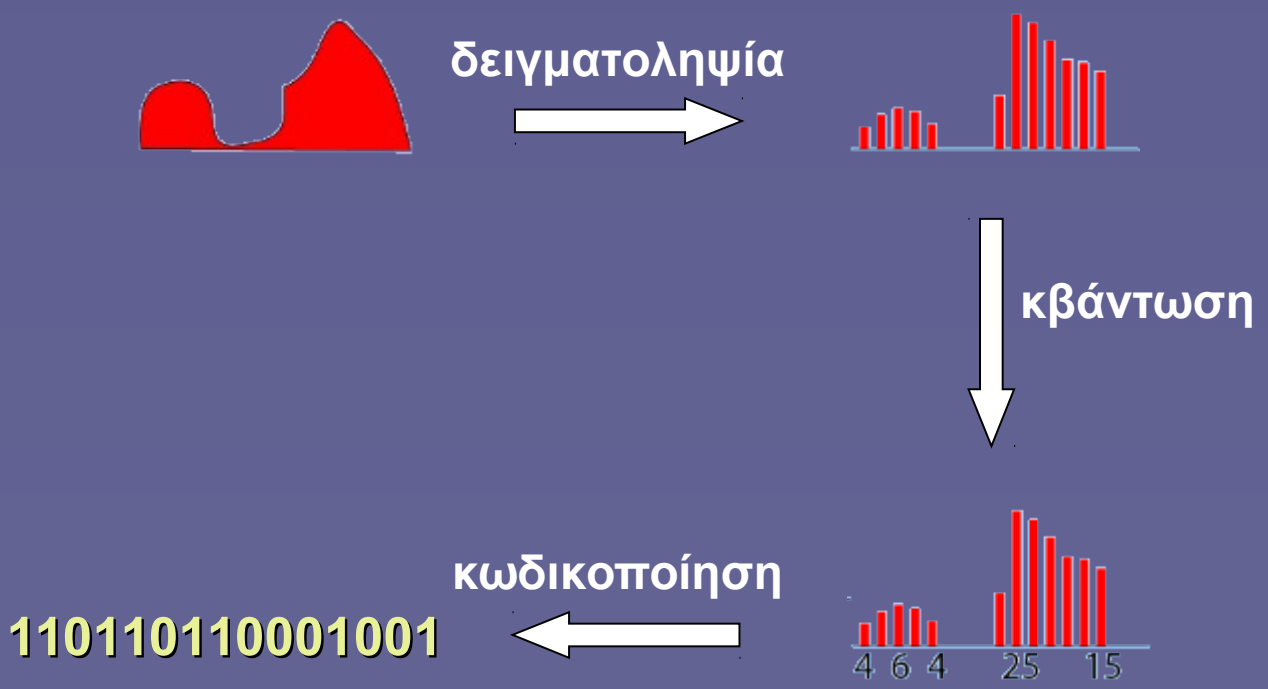
- Αναπαράσταση μεταβλητού μήκους

Unicode	Κωδικοποίηση UTF-8
00…7F	0xxxxxxx
80…7FF	110xxxxx 10xxxxxx
800…FFFF	1110xxxx 10xxxxxx 10xxxxxx
10000…10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

  - Το βασικό λατινικό αλφάβητο (ASCII) χρησιμοποιεί 1 byte ανά χαρακτήρα
    - Προς τα πίσω συμβατότητα
  - Τα ελληνικά, 2 bytes
    - Ποια η κωδικοποίηση κατά UTF-8 του τελευταίου παραδείγματος;

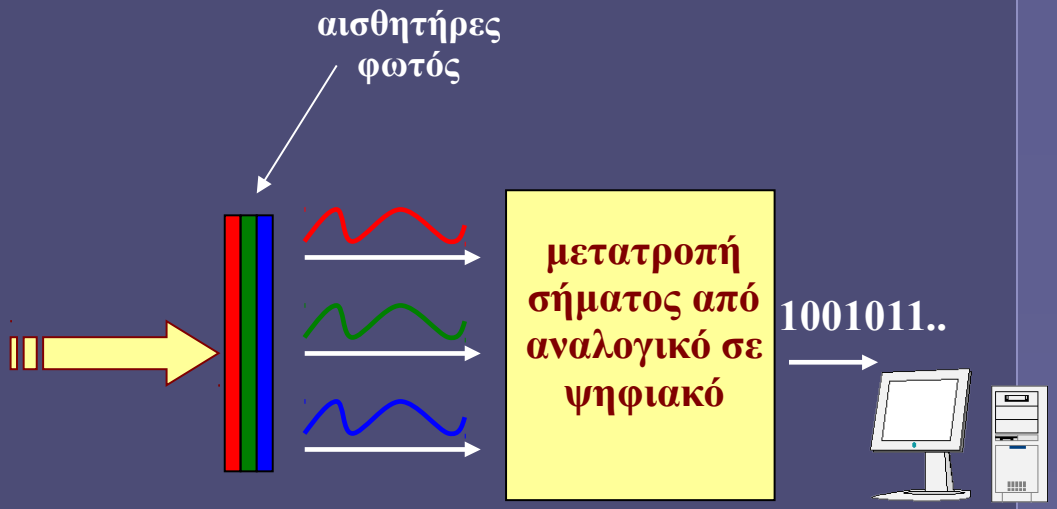
Ήχος: Ψηφιοποίηση και Αποθήκευση

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος



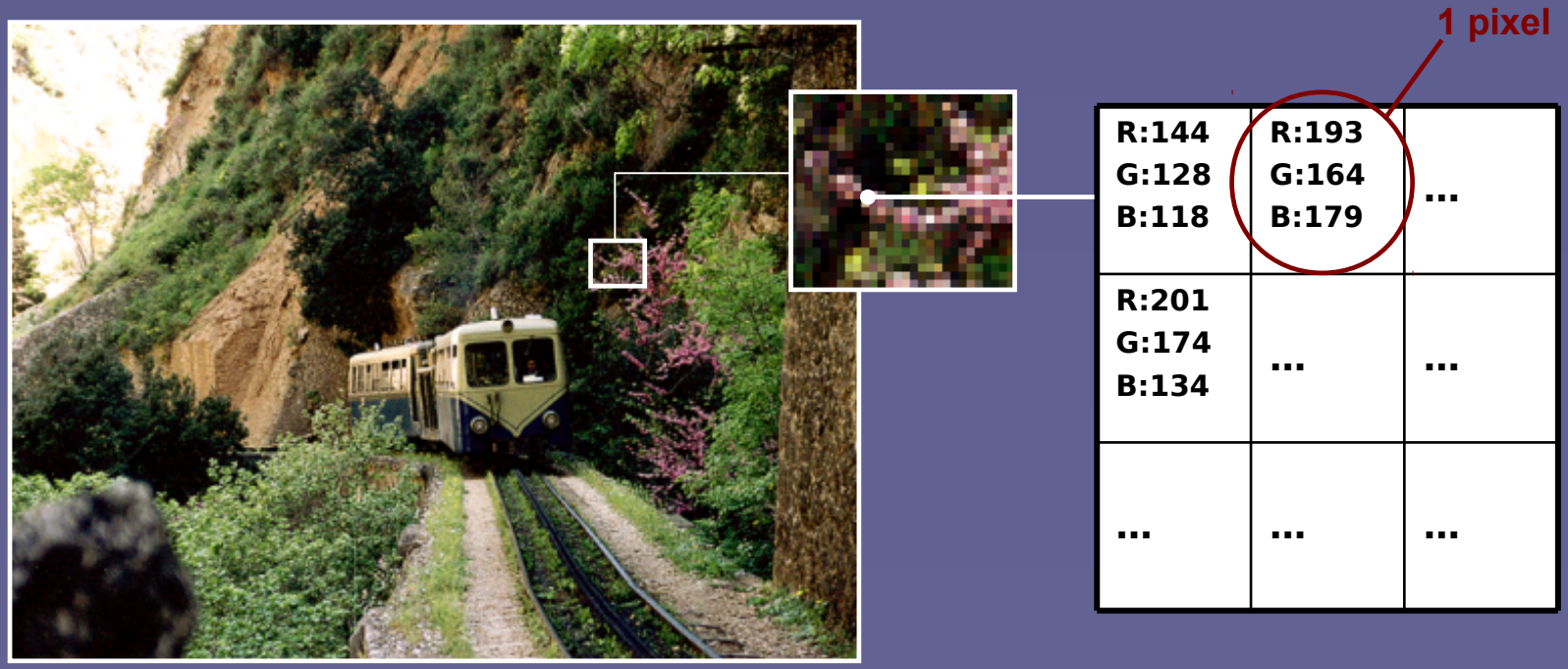
Εικόνα: από τον αναλογικό στον ψηφιακό κόσμο

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος
- Εικόνα



- Φωτοευαίσθητα κύτταρα
  - για τρία χρώματα (κόκκινο-πράσινο-μπλε)
- Μετατροπή σήματος σε ψηφιακή πληροφορία

Παράδειγμα: απλή αναπαράσταση pixels με 16,7 εκ. χρώματα



- 3 bytes/pixel (24bits): R(ed) G(reen) B(lue)
  - 256 στάθμες ανά συνιστώσα χρώματος
    - 256x256x256 = 16.777.216 χρώματα
  - εικόνες με μεγαλύτερο βάθος χρώματος
    - 32 έως 48 bits

Εναλλακτικά: διανυσματικά γραφικά

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος
- Εικόνα

- **Περιγραφή σχημάτων**
  - Ως σύνολο ευθύγραμμων και καμπύλων τμημάτων
  - Με συντεταγμένες
  - Εύρεση σημείων μέσω μαθηματικού τύπου
- **Εύκολη αλλαγή μεγέθους γραφικών**
  - Χωρίς παραμόρφωση των σχημάτων

Αναπαράσταση βίντεο

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος
- Εικόνα
- Βίντεο

- "Κινούμενη εικόνα" (καρέ)
  - όπως αναπαριστούμε τις απλές εικόνες
  - αλλά: με χρήση συμπίεσης
    - Για μείωση όγκου δεδομένων
    - Γειτονικά καρέ έχουν πολλές ομοιότητες

Κωδικοποίηση εντολών μηχανής

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος
- Εικόνα
- Βίντεο
- Εντολές Μηχανής