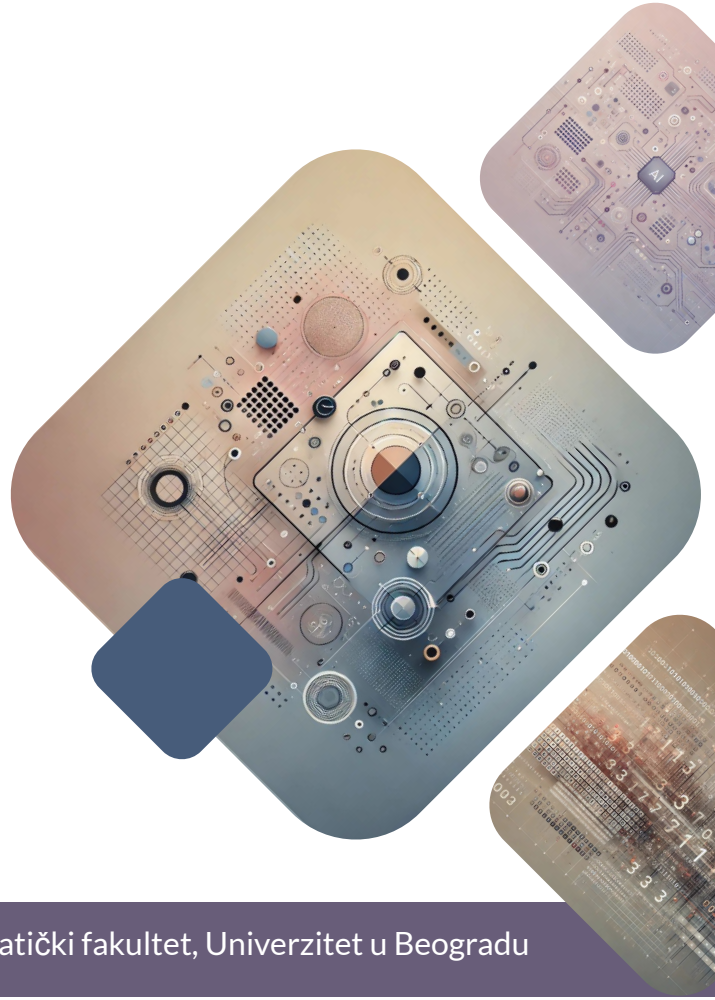
Zapis teksta, zvuka i videa

Danijela Simić



2024 | Matematički fakultet, Univerzitet u Beogradu

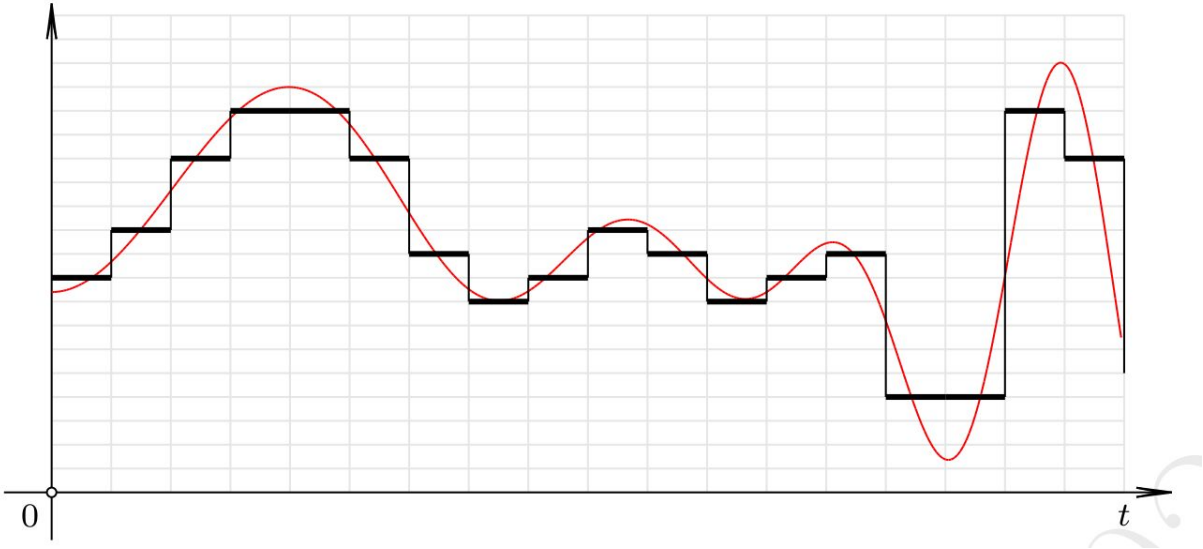
Uvod u digitalni zapis podataka

- Današnji računari su **digitalni** . To znači da su svi podaci koji su u njima zapisani – zapisani kao nizovi brojeva.
- Gubitak dela polaznih informacija.
- Većina podataka koje računari koriste nastaje zapisivanjem *prirodnih signala*.
- **Kontinualna priroda signala –** prirodno se mogu predstaviti *neprekidnim funkcijama*.
- Primer: Zvučni signal predstavlja promenu pritiska vazduha u zadatoj tački i to kao neprekidnu funkciju vremena.

Analogni zapis

- Analogni zapis uspostavlja **analogiju** između signala koji je zapisan i određenog svojstva medijuma na kome je signal zapisan.
- **Prednost**: jednostavno
- **Mane**:
  - nizak kvalitet,
  - teško je napraviti veran zapis,
  - nestalnost medijuma,
  - komplikovana obrada…

Digitalni zapis



- Vrednost signala se izmeri u **određenim vremenskim trenucima** i onda se na medijumu zapišu izmerene vrednosti.
- **Semplovi – niz brojeva** kojim je signal zapisan.
- Koliko često je potrebno vršiti merenje?
- **Najkvist-Šenonova teorema**: *signal je dovoljno meriti dva puta češće od najviše frekvencije koja se u njemu javlja.*
- Na primer: čovekovo uho čuje frekfenciju 20kHz, pa je dovoljno meriti frekfencijom 40kHz.

Digitalni zapis

Prednosti:

- Kvalitet reprodukcije digitalnog zapisa ne zavisi od toga kakav je kvalitet medija na kome su podaci zapisani;
- Kvarljivost medijuma tokom vremena postaje nebitna;
- Omogućava kreiranje apsolutno identičnih kopija što dalje omogućava prenos podataka na daljinu;
- Obrada digitalno zapisanih podataka se svodi na matematičku manipulaciju brojevima i ne zahteva korišćenje specijalizovanih mašina.

Problemi:

- **Neophodno imati veoma razvijenu tehnologiju** da bi se uopšte stiglo do iole upotrebljivog zapisa.
- Na primer, izuzetno je komplikovano napraviti uredaj koji je u stanju da 40 hiljada puta u sekundi izvrši merenje intenziteta zvuka. A onda negde treba čuvati tih 40 hiljada brojeva.
- Zato je digitalni zapis dosta kasnio.

Zapis teksta u računaru

- **Tekst**: „*informaciju namenjenu ljudskom sporazumevanju koja može biti prikazana u dvodimenzionalnom obliku*.”
- U računarima se tekst predstavlja kao **jednodimenzioni (linearni)** niz karaktera koji pripadaju određenom unapred fiksiranom skupu karaktera.
- U zapisu teksta, koriste se specijalni karakteri koji označavaju prelazak u novi red, tabulator, kraj teksta i slično.
- Osnovna ideja koja omogućava zapis teksta u računarima je da se svakom karakteru pridruži određen (neoznačeni) ceo broj i to na unapred dogovoreni način.
- Ovi brojevi se nazivaju **kodovima karaktera**.

Zapis teksta u računaru

- Tehnička ograničenja ranih računara kao i neravnomeran razvoj računarstva između različitih zemalja, doveli su do toga da postoji više različitih standardnih tabela.
- U zavisnosti od broja bitova potrebnih za kodiranje karaktera, razlikuju se 7-bitni kodovi, 8-bitni kodovi, 16-bitni kodovi, 32-bitni kodovi, kao i kodiranja promenljive dužine.
- Tabele koje sadrže karaktere i njima pridružene kodove obično se nazivaju **kodne strane**.

Zapis teksta u računaru

- Tehnička ograničenja ranih računara kao i neravnomeran razvoj računarstva između različitih zemalja, doveli su do toga da postoji više različitih standardnih tabela.
- U zavisnosti od broja bitova potrebnih za kodiranje karaktera, razlikuju se 7-bitni kodovi, 8-bitni kodovi, 16-bitni kodovi, 32-bitni kodovi, kao i kodiranja promenljive dužine.
- Tabele koje sadrže karaktere i njima pridružene kodove obično se nazivaju **kodne strane**.

Zapis teksta u računaru

- Postoji veoma jasna razlika izmedu karaktera i njihove grafičke reprezentacije.
- **Glifovi** – Grafičke reprezentacije pojedinih karaktera
- **Skupovi glifova** – Fontovi
- Korespondencija između karaktera i glifova ne mora biti jednoznačna.
- Fontovi koji se obično instaliraju uz operativni sistem sadrže glifove za karaktere koji su popisani na takozvanoj **WGL4 listi (Windows Glyph List 4)** koja sadrži uglavnom karaktere korišćene u evropskim jezicima, dok je za ispravan prikaz, na primer, kineskih karaktera, potrebno instalirati dodatne fontove.

ASCII

- Mala slova engleskog alfabeta: a, b, ... , z
- Velika slova engleskog alfabeta: A, B, ... , Z
- Cifre 0, 1, ..., 9
- Interpunkcijske znake: , . : ; + * - _ ( ) [ ] { } . . .
- Specijalne znake: kraj reda, tabulator, . . .

ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	STX	SOT	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

- 7 bita
- Prva 32 karaktera – od $(00)_{16}$ do $(1F)_{16}$ – su specijalni kontrolni karakteri.
- Ukupno 95 karaktera ima pridružene grafičke likove.
- Kodovi velikih i malih slova se razlikuju u samo jednom bitu u binarnoj reprezentaciji.

- Cifre 0-9 predstavljene su kodovima $(30)_{16}$ do (39)16, tako da se njihov ASCII zapis jednostavno dobija dodavanjem prefiksa 011 na njihov binarni zapis.
- Slova su poređana u kolacionu sekvencu, u skladu sa engleskim alfabetom.

YU-ASCII, ISO 646

YUSCII	ASCII	kôd	YUSCII	ASCII	kôd
Ž	@	$(40)_{16}$	ž	`	$(60)_{16}$
Š	[	$(5B)_{16}$	š	{	$(7B)_{16}$
Đ	\	$(5C)_{16}$	đ	|	$(7C)_{16}$
Ć	]	$(5D)_{16}$	ć	}	$(7D)_{16}$
Č	^	$(5E)_{16}$	č	~	$(7E)_{16}$

8 bitna proširenja ASCII tabele

ISO-8859-1	Latin 1	večina zapadnoevropskih jezika
ISO-8859-2	Latin 2	centralno i istočnoevropski jezici
ISO-8859-3	Latin 3	južnoevropski jezici
ISO-8859-4	Latin 4	severnoevropski jezici
ISO-8859-5	Latin/Cyrillic	ćirilica većine slovenskih jezika
ISO-8859-6	Latin/Arabic	najčešće korišćeni arapski
ISO-8859-7	Latin/Greek	moderni grčki alfabet
ISO-8859-8	Latin/Hebrew	moderni hebrejski alfabet

Windows-1250	centralnoevropski i istočnoevropski jezici
Windows-1251	ćirilica većine slovenskih jezika
Windows-1252	(često se neispravno naziva i ANSI) zapadnoevropski jezici

Problemi sa 8-bitnim kodnim stranama

- Nije moguće u istom tekstu koristiti i ćirilicu i latinicu
- Za azijske jezike nije dovoljno 256 mesta za zapis svih karaktera
- Nije moguće u istom tekstu koristiti specijalne karaktere (recimo za zapis matematičkih simbola) i pisati tekst na nekom istočnoevropskom jeziku
- …

Problemi sa 8-bitnim kodnim stranama i novi standardi

- Nije moguće u istom tekstu koristiti i ćirilicu i latinicu
- Za azijske jezike nije dovoljno 256 mesta za zapis svih karaktera
- Nije moguće u istom tekstu koristiti specijalne karaktere (recimo za zapis matematičkih simbola) i pisati tekst na nekom istočnoevropskom jeziku
- Universal Character Set — UCS
- *ISO 10646* i projekat *Unicode* (Xerox Parc, Apple, Sun Microsystems, Microsoft, . . .)
- ISO 10646
    - 4 bajta
    - prvih 65536 karaktera koristi se kao osnovni višejezički skup karaktera
    - preostali prostor ostavljen kao proširenje za drevne jezike, naučnu notaciju i slično.

Unicode

- Unicode je za cilj imao da bude:
  - univerzalan (UNIversal) — sadrži sve savremene jezike sa pismom;
  - jedinstven (UNIque) — bez dupliranja karaktera - kodiraju se pisma, a ne jezici;
  - uniforman (UNIform) — svaki karakter sa istim brojem bitova.
- 3 bajta:
  - $(000000)_{16} - (10FFFF)_{16}$

0020-007E	ASCII printable
00A0-00FF	Latin-1
0100-017F	Latin extended A (osnovno proširenje latinice, sadrži sve naše dijakritike)
0180-027F	Latin extended B
...	
0370-03FF	grčki alfabet
0400-04FF	ćirilica
...	
2000-2FFF	specijalni karakteri
3000-3FFF	CJK (Chinese-Japanese-Korean) simboli
...	

UCS-2 UTF8

- **Basic multilingual plane –**
  - u najčešćoj upotrebi
  - sadrži većinu danas korišćenih karaktera (uključujući i CJK — Chinese, Japanese, Korean — karaktere koji se najčešće koriste) čiji su kodovi izmedu $(0000)_{16}$ i $(\mathtt{FFFF})_{16}$.
- **UCS-2**: svaki Unicode karakter osnovne višejezičke ravni jednostavno zapisuje sa odgovarajuća dva bajta.
- **UTF-8**

raspon	binarno zapisan Unicode kôd	binarno zapisan UTF-8 kôd
0000-007F	00000000 0xxxxxxx	0xxxxxxx
0080-07FF	00000yyy yyxxxxxx	110yyyyy 10xxxxxx
0800-FFFF	zzzzyyyy yyxxxxxx	1110zzzz 10yyyyyy 10xxxxxx

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

- Analizirati i uprostiti zadatak.
- **Funkcionalni zahtevi**:
  - Kreirati nalog i mogućnost logovanja
  - Kreirati, menjati i brisati objave (eng. *tweets*)
  - Pratiti druge korisnike
  - Pratiti objave prema vremenu kada su postavljene
  - Pretraga objava
  - Mogućnost da se na objavu odgovori ili da se označi pozitivno (eng. *like*)

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

- **Nefunkcionalni zahtevi**:
  - Svaka objava može imati najviše **280 karaktera**
  - Omogućiti da sistem mo\e da ima **1 milijardu korisnika (10^9 korisnika)**
  - I neka u proseku svaki korisnik ima **200 pratilaca**
  - U proseku neka imamo **100 miliona objava dnevno**
  - Da je uvek dostupan (ili barem 99.99% vremena)
  - Da može da istovremeno u toku dana da radi sa velikim brojem objava
  - Obezbediti privatnost korisnika i njihovih podataka
  - Da brzo radi – korisnik ne treba da čeka dugo na prikaz/učitavanje objava

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

- Jedna objava 200 karaktera. U proseku 1 karakter je 2 bajta (UTF-8 koristimo): 2B * 280 = 560B ~ 1KB ~ 10^3 B
- 100M objava dnevno (100 * 10^6 * 10^3 B) = 10^{11} ~ 10^{12} = 1TB za objave dnevno
- Na godišnjem nivou: 365 * 1 TB ~ 400TB
- Procene:
  - 1 PB ~ 10^{15}
  - 1 TB ~ 10^{12}
  - 1 GB ~ 10^9
  - 1 MB ~ 10^6
  - 1 KB ~ 10^3

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

- Hajde malo da popravimo procenu.
- Trebaju nam i neki podaci o objavama:
  - vreme objave
    - **Kao UNIX timestamp**: U formi broja koji predstavlja sekunde od početka epohe (1. januar 1970.). Timestamp koristi **4 bajta** (32-bitni `int`), ali može biti i **8 bajtova** za `int64` (što omogućava širi opseg datuma).
  - jedinstvena oznaka korisnika koji je objavu napisao: koristimo opet 32-bitni int, odnosno 4B
  - jedinstvena oznaka objave (isto kao i prethodno) ~ 4B
  - Broj “sviđanja” ~ 32-bitni int, odnosno 4B
- Ovo je 20B koje je potrebno dodati na ona 560B za tekst. Praktično ništa ne menja, ne utiče previše.
- Na kraju, možemo reći (još malo proširiti) da je za *objave* potrebno 1PB

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

- Imamo i meta podatke koje čuvamo za svakog korisnika:
  - ime i prezime korisnika
    - neka je ukupno ime + prezime = 20 karaktera
    - opet koristimo UTF-8, 2 bajta
    - 40B za ime i prezime
  - jedinstvena oznaka korisnika (kao jmbg)
    - dovoljan je **int** (32-bitni, odnosno 4B)
  - Ukupno: 44B * 10^9 korisnika ~ 44 GB podataka

- Ali, svaki korisnik ima 200 pratilaca. Ove veze između korisnika je potrebno pamtiti:
  - Uzmimo da pamtimo kao parove (kodKorisnika, kodKorisnikaKogPrati)
  - 200 * 10^9 korisika * 8B = 1.6 * 10^{12} ~ 1.6TB podataka

- Napraviti procenu: Koliko memorije je potrebno za mrežu X (bivši Twitter)?

Ukupni rezultat:

- Twitter postoji od 2006. ~ 20 godina
- 20 * 1PB + 44 GB podataka + 1.6TB = 21 PB prostora
- Javni podatak je da kompanija koristi oko 300PB prostora
- Naša procena nije loša!
- Diskutovati zašto postoji razlika u proceni.

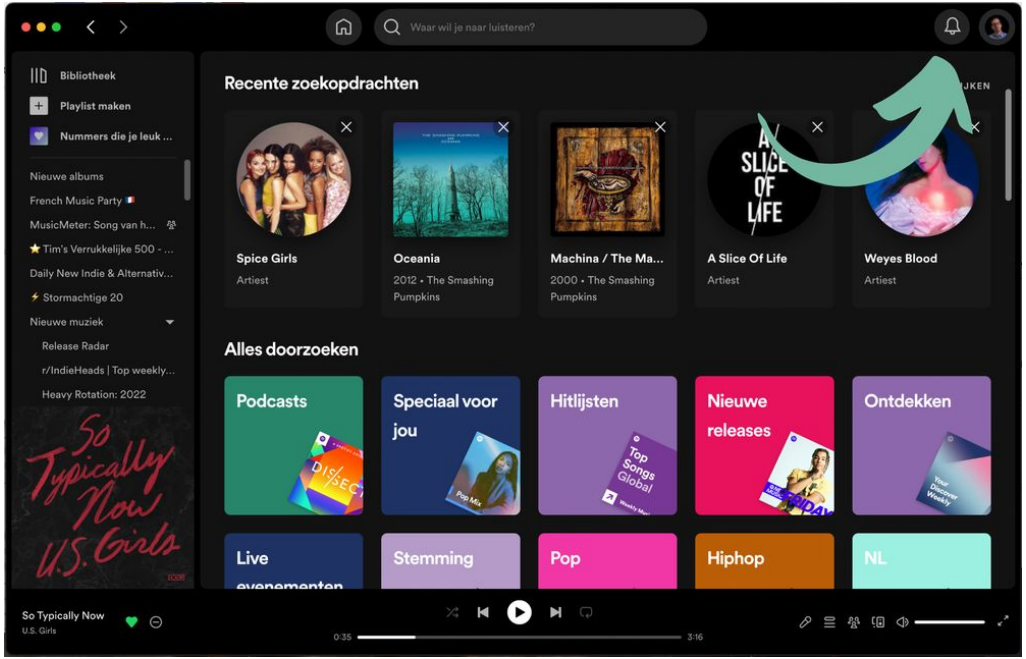
Zapis zvuka

- Zvučni talas predstavlja oscilaciju pritiska koja se prenosi kroz vazduh
- Digitalizacija zvuka se vrši merenjem i zapisivanjem vazdušnog pritiska u kratkim vremenskim intervalima
- Osnovni parametri:
    - **amplituda** (koja odgovara „glasnoći“)
    - **frekvencija** (koja odgovara „visini“).\
- Ljudsko uho obično čuje raspon frekvencija od 20Hz do 20kHz — dovoljno je koristiti frekvenciju odabiranja 40kHz.
- AudioCD standard — 44.1kHz
- miniDV, DVD, digital TV — 48kHz
- U mobilnoj telefoniji — frekvencije odabiranja mogu biti i znatno manje.
- **Broj bitova —** 2 bajta, 65536 različitih nivoa amplitude

Zapis zvuka

- Višekanalnog snimanja zvuka
- **Stereo zvuk** — snimanje zvuka sa dva kanala
- **Surround sistemi** — snimanje sa više od dva kanala (od 3 pa čak i do 10)
- Najpoznatiji takvi sistemi su 5+1 gde se koristi 5 regularnih i jedan bas kanal.
- Jedan minut stereo zvuka u AudioCD formatu zauzima:
    - 2·44100 sample/sec ·60sec·2 B/sample = 10584000B ≈ 10.1M B.
- Tehnike kompresije: **MP3 (MPEG-1 Audio-Layer 3).**

- Napraviti procenu: Koliko memorije je potrebno za Spotify?



Služi za skladištenje i puštanje muzike uživo. Čuva podatke o korisnicima, ali i izvođačima, albumima, pesmama itd…

Zapis video sadržaja

- Video zapis u računaru predstavlja kompleksan proces koji uključuje **efikasno povezivanje slike i zvuka** uz primenu različitih tehnika kompresije i sinhronizacije.
- Video se sastoji od **niza pojedinačnih frejmova**, pri čemu je svaki frejm digitalna slika koja se kombinuje sa odgovarajućim segmentom zvučnog zapisa.
- **Koderi i dekoderi** (eng. codecs)
- **Redundacija podataka** — odnosi na višak podataka koji nisu neophodni za reprodukciju

Zapis video sadržaja – koderi i dekoderi

- **Koderi** — H.264, H.265 (HEVC), AV1 — su algoritmi koji **vrše kompresiju** video i audio podataka tokom snimanja.
- vremenskog (eng. inter–frame) i prostornog (eng. intra–frame) kodiranje
- ***Diskretne kosinusne transformacije*** (eng. Discrete Cosine Transform):
  - podaci rastave na različite frekvencije
  - niske frekvencije - zadržati, visoke frekvencije - delimično ili potpuno ukloniti,
- ***Kompenzacije pokreta*** (eng. Motion Compensation):
  - umesto da se kompletan frejm kodira iznova, kodira se samo razlika između trenutnog i prethodnog frejma, koristeći vektore pokreta za opisivanje kretanja objekata H.265, poznat i kao HEVC, koristi naprednije
- AV1 – kao projekat otvorenog koda

Zapis video sadržaja – koderi i dekoderi

- **Reprodukcija** video sadržaja – **dekoder razdvaja podatke i rekonstruiše originalne frejmove** slike zajedno sa odgovarajućim audio segmentima.
- Sinhronizacija između zvuka i slike postiže se pomoću vremenskih oznaka (eng. timestamp)
- Kompenzacija kretanja zasnova na blokovima (eng. Block-based Motion Compensation) i
- Inverzna kosinusna transformacija (eng. Inverse Discrete Cosine Transform, IDCT)

Zapis video sadržaja – kontejneri sadržaja

- **MP4, MKV, ili MOV**
- Integrišu video i audio, ali i druge, meta podatke (npr. prevode, title)
- **MP4** – široko podržan format; često se koristi za uživo puštanje sadržaja zbog balansa izmedu kvaliteta i veličine fajla.
- **MKV** — pruža veću fleksibilnost i može integrisati više tokova
- **MOV** —razvijen od strane kompanije Apple, nudi visok nivo kvaliteta i često se koristi u profesionalnim okruženjima za uredivanje videa

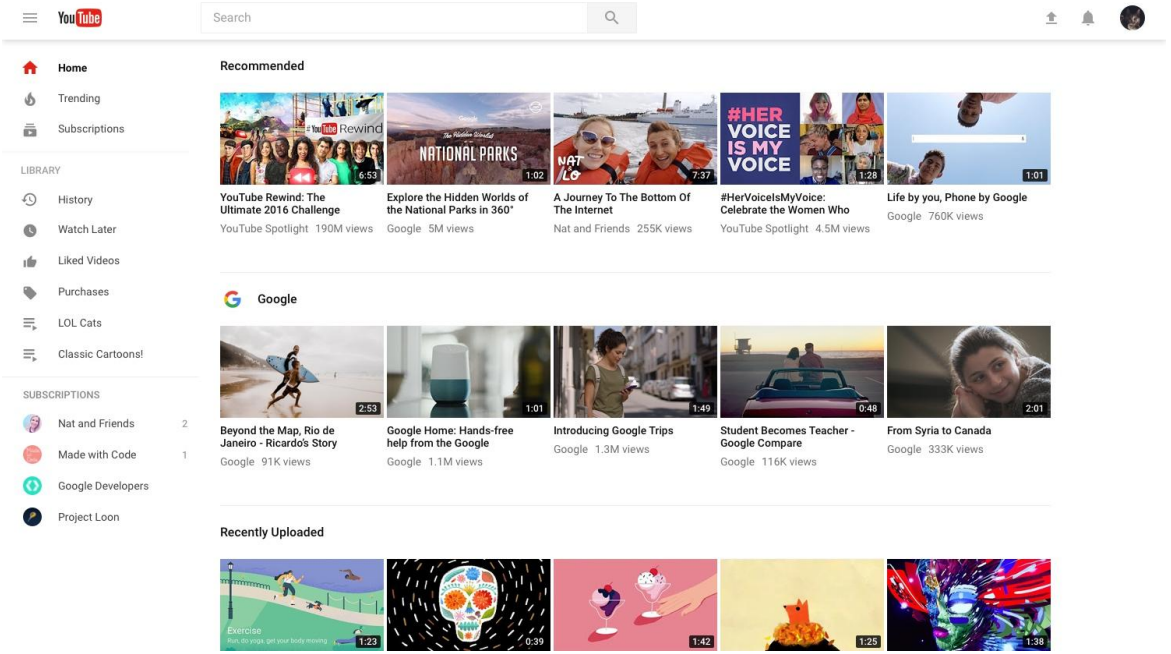
Zapis video sadržaja

- **HD (High Definition)** — odnosi na rezoluciju od 1280 × 720 piksela
- **Full HD** — rezolucija 1920 × 1080 piksela (često kao 1080p)
- **p** označava progresivno skeniranje, što znači da se svaki frejm prikazuje u celosti. U manjim formatima ili u slučaju upletenog (eng. interlaced) skeniranja, frejmovi se ne prikazuju svi odjednom, već se deli na polja, gde se prvo prikazuje polovina slike, a zatim druga polovina, što može dovesti do smanjenja kvaliteta slike ili povećanja **treperenja**.
- 4K — 3840 × 2160 piksela

Zapis video sadržaja - bitrate

- **Brzina bita (brzina protoka, eng. bitrate)** — megabiti po sekundi (Mbps), količina podataka koja se prenosi ili obrađuje u jedinici vremena tokom reprodukcije videa.
- Viša brzina bita – bolji kvalitetom slike, veći kapacitet, šira mrežna propusnost
- Niža brzina bita – gubitak detalja, pojavu vizuelnih nepravilnosti
- Veće rezolucije zahtevaju veću brzinu
- Full HD (1080p) – 8-12 Mbps
- 4K – 35-45 Mbps
- *Moderniji kodeci, H.265 i AV1, omogućavaju bolju kompresiju i kvalitet slike pri nižoj brzini bita u odnosu na starije kodeke*
- H.264 za Full HD (1080p) – 8-12 Mbps
- H.265 za Full HD (1080p) – 4-6 Mbps
- AV1 za Full HD (1080p) – 3-5 Mbps

- Napraviti procenu: Koliko memorije je potrebno za YouTube platformu?



- Napraviti procenu: Koliko memorije je potrebno za YouTube platformu?

- Kolika je cena jednog videa?
  - **1080p (Full HD)**, prosek je 5-10 Mbps, za 5 minuta, veličina videa je **3 GB**.
  - **720p (HD)**, prosek je 2.5-6 Mbps, za 5 minuta, veličina videa je **1 GB**.
  - **480p (SD)**, prosek je 1-2 Mbps, za 5 minuta, veličina je **300 MB**.
  - Prema samom servisu, dozvoljena veličina videa je 256GB
- Neka imamo 1 milijardu korisnika, 10^9
- Neka se dnevno objavi 1 milion videa ~ 10^6
- Ukupno za video dnevno: 10^6 * 1GB ~ 1PB (godišnje je 300PB)
- Prema statističkim podacima, 500 sati videa se postavlja na YouTube svakog minuta:
  - prosek - 7.5 Mbps = 7.7 * 10^6 bita ~ 9 * 10^5 B
  - 9 * 10^5 * 500 * 60 minuta * 60 sekundi ~ 1TB podataka svakog minuta
  - Godišnje: 1TB * 24 * 60 * 365 ~ 100PB podataka

- Napraviti procenu: Koliko memorije je potrebno za YouTube platformu?

- Procena za meta podatke za korisnike je ista kao i kod Twittera.
- Par interesantnih činjenica:
  - svaki video se čuva u više formata i rezolucija
  - video je podeljen na delove (chunk) (paralelizacija, ušteda, brzina, lako menjanje nivoa kvaliteta itd…)
  - Čuvaju se kopije