

Základy forenzních databází

Tereza Uhlíková

verze y26

Kdo jsem

Tereza Uhlíková

Ústav analytické chemie

Skupina Teoretické Spektroskopie - Broad Peak

místnost B4335

<https://web.vscht.cz/~uhlikovt/>

tereza.uhlikova@vscht.cz



Co bylo minule

- Proč jsou data důležitá
- Proč vznikly databáze
- Co je to databáze
- Proč, kdo, kdy, jak ...
- Něco málo z historie
- **Problém není jen „mít data“. Problém je umět s nimi efektivně pracovat.**
- **Databáze není primárně programátorský problém. Začíná otázkou odborníka.**

Vysvětlete sousedovi rozdíl DB × ML × AI na příkladu z chemie.

O čem budeme mluvit dnes

„Z molekuly do bitů“

Informace, záznam, číslo, kódování, hardware, ukládání, uložště, energie
– a proč to potřebuje AI

DATA → DATABÁZE → VYHLEDÁVÁNÍ → ALGORITMY →
STROJOVÉ UČENÍ → AI

Záhadný soubor

```
1094.905 0.1609871
1095.387 0.1592218
1095.869 0.1578855
1096.351 0.1573563
1096.833 0.1564317
1097.315 0.1551866
1097.797 0.1547755
1098.279 0.1543083
1098.761 0.1532006
1099.244 0.152494
1099.726 0.1528435
1100.208 0.152638
1100.69 0.1518068
1101.172 0.1519589
1101.654 0.1522427
1102.136 0.1515442
```

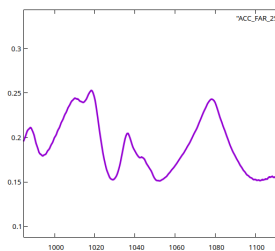
Co potřebuje program vědět, aby tento soubor správně otevřel?
kódování, oddělovač, desetinná čárka/tečka, jednotky, osy...

Záhadný soubor

```

1094.905 0.1609871
1095.387 0.1592218
1095.869 0.1578855
1096.351 0.1573563
1096.833 0.1564317
1097.315 0.1551866
1097.797 0.1547755
1098.279 0.1543083
1098.761 0.1532006
1099.244 0.152494
1099.726 0.1528435
1100.208 0.152638
1100.69 0.1518068
1101.172 0.1519589
1101.654 0.1522427
1102.136 0.1515442

```



```

0001900 3735 3139 3020 312e 3239 3133
0001910 3836 302e 3136 2032 2e30 3833
0001920 0a32 3632 2e38 3435 3333 3020
0001930 3731 3937 320a 3936 302e 3532
0001940 3831 3032 3334 320a 3936 352e
0001950 2e30 3831 3432 3939 0a38 3632
0001960 3739 3020 312e 3438 3339 3439
0001970 342e 3137 2038 2e30 3831 3637
0001980 3732 2e30 3539 3933 3020 312e
0001990 3739 320a 3137 342e 3633 2033
00019a0 3734 3737 0a31 3732 2e31 3139
00019b0 322e 3130 3531 3236 320a 3231
00019c0 2033 2e30 3931 3638 3631 0a32
00019d0 3838 3432 3020 312e 3939 3934
00019e0 3337 332e 3436 2036 2e30 3032
00019f0 0a36 3732 2e33 3438 3736 3020
0001a00 3534 3532 320a 3437 332e 3832
0001a10 3032 3535 3736 0a33 3732 2e34
0001a20 3020 322e 3231 3234 3934 320a
0001a30 3339 2031 2e30 3232 3534 3738
0001a40 2e35 3737 3235 3020 322e 3034

```

Mapa cesty: život jednoho spektra

Vzorek → detektor (analogový signál) → převodník : převést spojitý (analogový) signál na diskrétní (digitální) data
Analog-to-Digital Converter (ADC) vzorkování, kvantizace

Mapa cesty: život jednoho spektra

Vzorek → detektor (analogový signál) → převodník (ADC) → čísla → bajty → soubor (kódování, komprese, metadata) → disk / paměť → CPU / GPU → databáze / ML

Kde na této cestě se podle vás nejčastěji ztrácí nebo poškozují data?

Mars Climate Orbiter (1999) – Záměna jednotek

Web: [https:](https://science.nasa.gov/mission/mars-climate-orbiter/)

[//science.nasa.gov/mission/mars-climate-orbiter/](https://science.nasa.gov/mission/mars-climate-orbiter/)

Příčina: Rozpor v používaných fyzikálních jednotkách mezi dvěma inženýrskými týmy.

Chyba: Firma Lockheed Martin dodala data v libra-silách (lbf-s), ale NASA (JPL) očekávala Newton-sekundy (N-s).

Důsledek: Sonda prováděla manévry s $4,45\times$ větší silou. Po devíti měsících minula bezpečnou dráhu, vletěla příliš hluboko do atmosféry Marsu a shořela.

Škoda: 327 milionů dolarů a ztracená vědecká data.

Ariane 5 (1996) – Přetečení paměti (Integer Overflow)

Web: [https:](https://www.facebook.com/watch/?v=1382704783052999)

[//www.facebook.com/watch/?v=1382704783052999](https://www.facebook.com/watch/?v=1382704783052999)

Příčina: Nekritické převzetí programového kódu ze starší verze rakety Ariane 4.

Chyba: Systém se pokusil zkonvertovat 64bitové číslo (float) (rychlost rakety) na 16bitové celé číslo (integer). Hodnota však byla příliš velká a software zkolaboval.

Důsledek: Selhal primární i záložní počítač. Trysky motorů se natočily do extrému, aerodynamický tlak raketu poškodil a 37 sekund po startu došlo k autodestrukci.

Škoda: 370 milionů dolarů a čtyři vědecké satelity.

Excel a geny (2020) – Příliš chytrý software

Web: <https://www.science.org/content/article/one-five-genetics-papers-contains-errors-thanks-microsoft-excel>

Příčina: Automatická konverze dat v Microsoft Excelu na výchozí nastavení „Obecný“.

Chyba: Excel vyhodnocoval zkratky genů (např. SEPT2, MARCH1) jako kalendářní data a nevratně je přepisoval na 2-Sep nebo 1-Mar.

Důsledek: Znehodnocení dat v 20 % publikovaných genetických studií. Vědci při importu CSV souborů zapomínali ručně zapnout formát „Text“.

Řešení: Microsoft chování nezměnil, takže vědecká komunita musela kapitulovat a oficiálně přejmenovat 27 lidských genů (např. na SEPTIN2).

Data – informace – znalost (a zdroj)

3350 → data

poloha pásu v IČ spektru = 3350 cm^{-1} → informace

vzorek obsahuje alkohol → znalost

vzorek odpovídá očekávanému složení → závěr

Claude Shannon (1916 – 2001) – Definice informace 1948

- **Základní princip:** Informace = snížení nejistoty (odstranění neurčitosti či pochybnosti).
- **Jednotka měření:** Zaveden bit jako matematické měřítko (volba mezi dvěma stejně pravděpodobnými možnostmi).
- **Informační hodnota:** Čím méně je zpráva pravděpodobná, tím více informace (bitů) přináší.
- **Změna paradigmatu:** Nahrazení staršího fyzikálního pojmu „opak entropie“ čistě matematickou teorií pravděpodobnosti.
- **Dopad:** Položení základů pro veškerou moderní digitální techniku, internet a kódování dat.

Informace dnes a kdysi

dnes vs. kdysi

kdysi - problém informace získat a získat je včas

dnes - problém vyhodnotit kvalitu a relevanci informace -> zdroj informace

Jaké vlastnosti by měla mít informace, která ovlivní vaše rozhodování?

- **Správnost** (jsou uvedeny zdroje?)
- **Odbornost autora** (jsou informace o autorovi?)
- **Objektivnost** (je dán prostor více stranám?)
- **Aktuálnost** (jsou informace datovány?)
- **Ucelenost** (je zpráva uvedena v kontextu?)

Záznam informace



Bit = jedna otázka ano/ne

jednotka

- Bit = rozhodnutí mezi dvěma možnostmi; n bitů (otázek) rozliší 2^n možností
- 6 bitů (otázek) → zaznamená 64 možností (znaků); 20 otázek → 1 048 576 možností
- Jde to i obráceně: Půlení množiny možností = základ binárního vyhledávání (později v kurzu)

Otázky:

- Kolik otázek ano/ne stačí k určení jedné látky z 1 000 000? (20)
- A k určení jednoho člověka z 8 miliard? (33)
- Jsou všechny otázky stejně dobré? Která otázka je nejlepší? (příprava na rozhodovací stromy v ML) - Jak postupujete při hře "Myslím si zvíře" :-)?

Bajt: kolik „váží“ kofein?

1 B = 8 bitů = 256 možných hodnot (2 B = 65 536; 4 B $\approx$ 4,3 miliardy)

Písmeno „A“ = 65 = 01000001

SMILES kofeinu CN1C=NC2=C1C(=O)N(C(=O)N2C)C má 28 znaků = 28 B

Číslo 3350: jako text 4 B, jako 32bitové celé číslo 4 B, jako float64 8 B

Jedno IR spektrum, 3 601 bodů $\times$ 8 B $\approx$ 29 kB

Proč jsou barvy v obrázku často zapsané třemi bajty? (R, G, B)

Biologický základ: Lidské oko má tři typy barevných receptorů (čípky pro červenou, zelenou a modrou).

Technický standard: Každá složka (R, G, B) dostane 1 bajt (hodnoty 0 až 255).

Výsledný efekt: Spojením vzniká 24bitová hloubka (3x8) (True Color) umožňující $26^{24} = 16,7$ milionu barev.

Efektivita: Přesahuje rozlišovací schopnost lidského oka (cca 10 milionů barev) a ideálně pasuje do počítačové paměti.

Předpony

1 KiB (kilobajt) = 2^{10} B = 1 024 B

1 MiB(megabajt) = 2^{20} B = 1 048 576 B

1 GiB(gigabajt) = 2^{30} B = 1 073 741 824 B

1 TiB(terrabajt) = 2^{40} B = 1 099 511 627 780 B

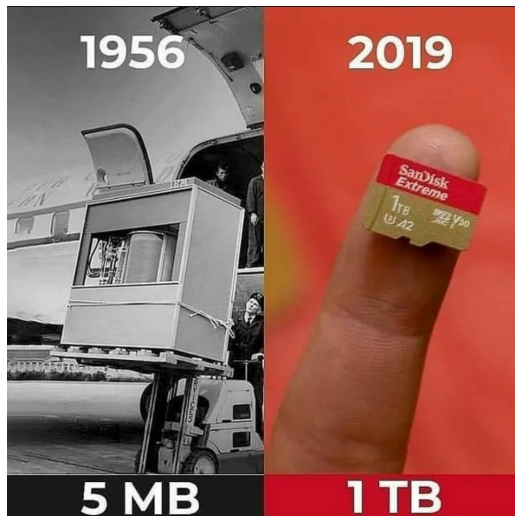
1 PiB(petabajt) = 2^{50} B = 1 125 899 907 000 000 B

1 EiB(exabajt) = 2^{60} B = 1 152 921 505 000 000 000 B

1 ZiB(zettabajt) = 2^{70} B = 1 180 591 621 000 000 000 000 B

1 YiB(yottabajt) = 2^{80} B = 1 208 925 819 614 629 174 706 176 B

Porovnání



<https://www.alza.cz/sandisk-microsdxc-2tb-extreme-rescue-pro-deluxe-ad-edentor-d12720857.htm>

Datová realita v chemii: Od kilobajtů k terabajtům

- Malá data (kB – MB): Chemické struktury, IR spektra, 1D NMR. *Příklad: IR spektrum $\approx$ desítky kB | 1D NMR $\approx$ stovky kB až MB. (Pro srovnání: Fotografie z mobilu $\approx$ 3–5 MB)*
- Střední data (desítky – stovky MB): Rutinní analýzy. *Příklad: 2D NMR nebo GC-MS běžný vzorek $\approx$ desítky MB.*
- Velká data (GB): Moderní high-resolution instrumentace. *Příklad: LC-HRMS (vysoké rozlišení) $\approx$ 1 až 3 GB na jeden běh.*
- Měřitko screeningu (10 TB): 10 000 chemických měření $\approx$ 1 GB = 10 TB.

Důsledek: Moderní chemie už není jen o zkumavkách, ale o IT infrastruktuře (ukládání, zálohování a cloudové zpracování dat).

Datová realita: Výpočetní chemie a simulace

- Malá data (kB – jednotky MB): Inputs & Výsledky

Vstupní souřadnice (.xyz, .gjf) $\approx$ desítky kB.

Textové výstupy DFT optimalizací geometrií $\approx$ jednotky MB.

- Střední data (stovky MB – jednotky GB): Elektronová hustota

Checkpoint soubory s vlnovými funkcemi a orbitaly $\approx$ stovky MB.

Krátké simulace molekulární dynamiky (MD) $\approx$ jednotky GB.

- Velká data (desítky GB – jednotky TB): Makromolekuly & Scratch space

Trajektorie dlouhých MD simulací proteinů (mikrosekundy) $\approx$ stovky GB.

Dočasná data (scratch) pro nejvyspělejší kvantové výpočty [CCSD(T)] $\approx$ stovky GB.

- Měřítko screeningu (TB): Virtuální dokování milionů knihoven léčiv nebo rozsáhlé QM/MM simulace enzymů $\approx$ jednotky až desítky TB.

Fermiho odhad: kolik dat vyprodukuje laboratoř za rok?

Orientační příklad: $20 \text{ vzorků/den} \times 0,5 \text{ GB} \times 220 \text{ dní} \approx 2 \text{ TB/rok}$.

Otázky: Kolik z těch dat budete potřebovat za 5 let? A kdo za jejich uložení zaplatí?

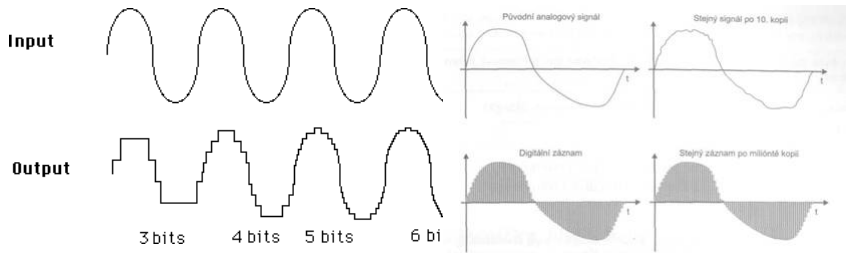
Které z dat jsou surová a která odvozená? Co z toho se dá znovu (a jak rychle) vypočítat?

Od signálu k číslům: vzorkování a kvantizace

strojový záznam

analogový - fyzikální veličina, dochází ke kompresi = ztrátě informace

digitální - vzorkování a převod na binární číslo = minimální ztráta



Vzorkování a kvantizace: pravidla

Vzorkovací frekvence v chemii (Osa X):

Chromatografický pík vyžaduje 15–20 bodů přes pík pro přesný výpočet plochy. Příliš málo bodů = zkreslení plochy a polohy maxima. Příliš mnoho bodů = obří soubory a šum.

Nyquistův teorém (Teoretický limit):

Vzorkovací frekvence musí být $\geq 2 \times$ nejvyšší frekvence v signálu. Nedodržení způsobuje aliasing (vznik falešných signálů).

Kvantizace (Osa Y):

Převod spojitě intenzity signálu na konečný počet digitálních úrovní (zaokrouhlování).

Určuje ji bitová hloubka (např. 24 bitů = 16,7 milionu úrovní citlivosti, jako u RGB).

Zásadní pro dynamický rozsah – schopnost detekovat stopové koncentrace vedle hlavních komponent.

Průběžné shrnutí

- Data jsou jen tak dobrá, jak dobrá jsou jejich pravidla
- Havárie Mars Climate Orbiter i Ariane 5 ukazují, že i ty nejlepší systémy selžou na banálních chybách v komunikaci, rozhraních a nekompatibilitě jednotek či formátů.
- Porozumění digitálním základům mění praxi v laboratoři
- Znalost Shannonovy definice, vzorkování a kvantizace není suchá teorie. Pomáhá nám správně nastavit přístroje, eliminovat šum a vyhnout se nevratnému poškození dat (jako u genů v Excelu).
- Moderní chemie je datová věda
- Éra high-resolution instrumentace a rozsáhlých počítačových simulací posunula chemii od zkumavek k terabajtovým datovým sadám (10 000 měření = 10 TB). Správa, komprese a archivace dat jsou dnes stejně důležité jako samotný experiment.

Číselné soustavy a proč dvojkově

číslo $\times$ číslice = abstraktní entita $\times$ znak pro zápis čísla
číselné soustavy

- poziční
- adiční (římská)

Dvojková (binární; obsahuje pouze číslice 0 a 1)

Osmičková (oktálová; číslice 0-7)

Šestnáctková (hexadecimální; 0-9, A-F)

„Rok 2026 třemi způsoby“: MMXXVI $\times$ 11111101010 $\times$ 7EA.

Proč chceme dvojkovou soustavu?

Proč je záznam informace jako 0 a 1?

Převody

Číslo dělíme 2 a zbytek po dělení zapisujeme zprava.

$$2026 = 11111101010_2$$

Dvojková soustava na desítkovou:

$$110: \text{Zprava: } 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 = 4 + 2 + 0 = 6_{(10)}$$

$$\text{Kontrola zpětně: } 1024 + 512 + 256 + 128 + 64 + 32 + 8 + 2 = 2026$$

$$\text{Hexadecimálně: } 7EA_{(16)} = 2026_{(10)} = 7 \cdot 16^2 + 14 \cdot 16^1 + 10 \cdot 16^0$$

Převod hexadecimálního čísla na dekadické:

$$\text{Číslo } 7EA_{(16)} \text{ zapíšeme ve tvaru: } 7 \cdot 16^2 + E \cdot 16^1 + A \cdot 16^0$$

Kolik bitů potřebuji na číslo 1 000 000?

$$11110100001001000000 \text{ (20)}$$

Zajímavost - Hex ve forenzní praxi: soubor poznáme podle obsahu

Hexdump: vypíše data pouze v hexadecimálním formátu

dvě šestnáctkové číslice = 1 bajt

„Magic bytes“ na začátku souboru: PDF = 25 50 44 46 („%PDF“), PNG = 89 50 4E 47, JPEG = FF D8 FF, ZIP = 50 4B 03 04 („PK“)

Přípona souboru lze změnit, začátek obsahu těžko; file carving = hledání souborů v surových datech disku

Další místa, kde potkáte hex: barvy #RRGGBB, MAC adresy, hash
Forenzní odborník nevěří příponě. Soubor „foto.txt“ může být ve skutečnosti JPEG. Podíváme se na první bajty a poznáme, co to je.

Bitová reprezentace reálných čísel

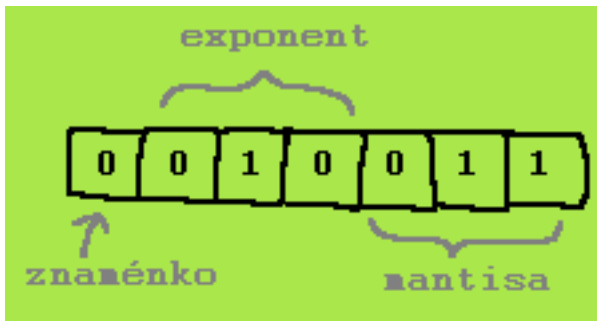
Desetinná čísla lze reprezentovat dvěma způsoby:

- v pevné řádové čárce (5 číslic před a 3 číslice za)
- v plovoucí řádové čárce:

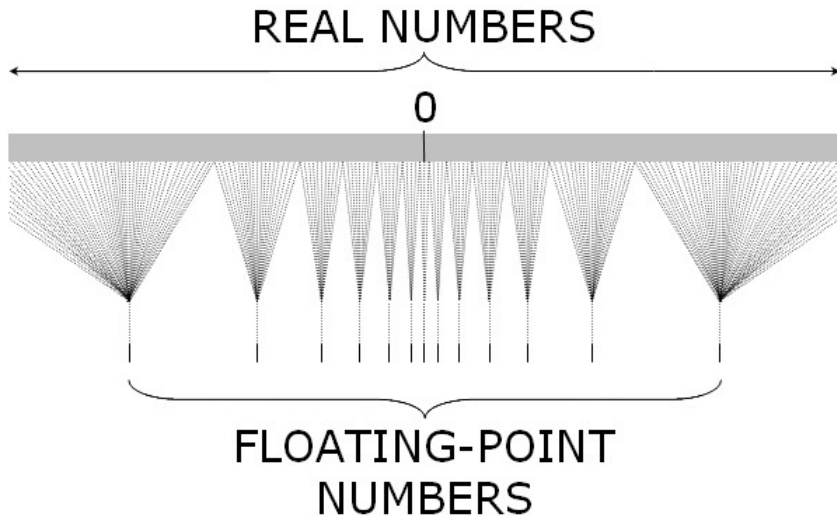
Znaménko: Speciální bit (0 +, 1 -) určuje znaménko čísla

Mantisa: Představuje samotné číslice čísla

Exponent: Udává, o kolik míst se má desetinná čárka v mantise posunout, aby se získalo původní číslo.



Přesnost v desetinných číslech



Bitová reprezentace reálných čísel

- Čím více bitů na mantisu, tím větší je přesnost uložení čísla.
- Čím více bitů na exponent, tím větší je rozsah čísel, která lze uložit.

jednoduchá přesnost (single-precision) - 32 bitů x dvojitá přesnost (double-precision) - 64 bitů

- Kvůli omezenému počtu bitů je binární reprezentace desetinných čísel často nepřesná!

AI trénink používá i 16 a 8 bitů (FP16, BF16, FP8): rychlejší, méně paměti, ale menší přesnost

Kvantová chemie, DFT a kalibrace obvykle double

Číslo není všechno: jednotky, metadata, sémantika

- 3350 bez jednotky, přístroje, data, operátora, kalibrace ... je jen číslo
- Formáty spekter a MS dat nesou metadata (JCAMP-DX, mzML); u CSV pomáhá „datový slovník“
- FAIR principy: Findable, Accessible, Interoperable, Reusable
- Forenzní kontext: řetězec předání (chain of custody) vyžaduje, aby bylo vždy známo, kdo, kdy a čím data vytvořil

Mars Climate Orbiter: čísla byla správná, jen v jiných jednotkách
Které údaje by musel kolega znát, aby za pět let správně použil váš soubor?

Kódování

Morseova abeceda, Braillovo písmo, ...

zakódování X různých znaků = X různých dvojkových čísel

1 znak = 1 B

textová - znaková sada

ASCII (American Standard Code for Information Interchange) - 7bitový,
128 znaků, jeden znak jeden byte

UNICODE - UTF-8 - až 4B ($4 \times 8 = 32$, $2^{32} = 4\,294\,967\,296$ znaků)

ASCII

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]	48	30	110000	60	0	96	60	1100000	140	`
1	1	1	1	[START OF HEADING]	49	31	110001	61	1	97	61	1100001	141	a
2	2	10	2	[START OF TEXT]	50	32	110010	62	2	98	62	1100010	142	b
3	3	11	3	[END OF TEXT]	51	33	110011	63	3	99	63	1100011	143	c
4	4	100	4	[END OF TRANSMISSION]	52	34	110100	64	4	100	64	1100100	144	d
5	5	101	5	[ENQUIRY]	53	35	110101	65	5	101	65	1100101	145	e
6	6	110	6	[ACKNOWLEDGE]	54	36	110110	66	6	102	66	1100110	146	f
7	7	111	7	[BEL]	55	37	110111	67	7	103	67	1100111	147	g
8	8	1000	10	[BACKSPACE]	56	38	111000	70	8	104	68	1101000	150	h
9	9	1001	11	[HORIZONTAL TAB]	57	39	111001	71	9	105	69	1101001	151	i
10	A	1010	12	[LINE FEED]	58	3A	111010	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	111011	73	!	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	111100	74	<	108	6C	1101100	154	l
13	D	1101	15	[CARRIAGE RETURN]	61	3D	111101	75	=	109	6D	1101101	155	m
14	E	1110	16	[SHIFT OUT]	62	3E	111110	76	>	110	6E	1101110	156	n
15	F	1111	17	[SHIFT IN]	63	3F	111111	77	?	111	6F	1101111	157	o
16	10	10000	20	[DATA LINK ESCAPE]	64	40	1000000	100	@	112	70	1110000	160	p
17	11	10001	21	[DEVICE CONTROL 1]	65	41	1000001	101	A	113	71	1110001	161	q
18	12	10010	22	[DEVICE CONTROL 2]	66	42	1000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	1000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	1000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	1000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	1000110	106	F	118	76	1110110	166	v
23	17	10111	27	[END OF TRANS. BLOCK]	71	47	1000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	1001000	110	H	120	78	1110000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	1001001	111	I	121	79	1110001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	1001010	112	J	122	7A	1110100	172	z
27	1B	11011	33	[ESCAPE]	75	4B	1001011	113	K	123	7B	1110101	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	1001100	114	L	124	7C	1111000	174	
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	1001101	115	M	125	7D	1111001	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	1001110	116	N	126	7E	1111100	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	1001111	117	O	127	7F	1111101	177	[DEL]
32	20	100000	40	[SPACE]	80	50	1010000	120	P					
33	21	100001	41	!	81	51	1010001	121	Q					
34	22	100010	42	"	82	52	1010010	122	R					
35	23	100011	43	#	83	53	1010011	123	S					
36	24	100100	44	\$	84	54	1010100	124	T					
37	25	100101	45	%	85	55	1010101	125	U					
38	26	100110	46	&	86	56	1010110	126	V					
39	27	100111	47	'	87	57	1010111	127	W					
40	28	101000	50	(	88	58	1011000	130	X					
41	29	101001	51	)	89	59	1011001	131	Y					
42	2A	101010	52	*	90	5A	1011010	132	Z					
43	2B	101011	53	+	91	5B	1011011	133	[					
44	2C	101100	54	,	92	5C	1011100	134	\					
45	2D	101101	55	-	93	5D	1011101	135	]					
46	2E	101110	56	.	94	5E	1011110	136	^					
47	2F	101111	57	/	95	5F	1011111	137	_					

SMILES

- Zjednodušený textový kód: Simplified Molecular Input Line Entry System.
- Jazyk pro počítače: Zapisuje 2D strukturu chemické molekuly jako jeden řetězec běžných textových znaků (ASCII).

Základní pravidla hry

- Atomy: Používají klasické chemické značky (C, O, N, S).
- Vazby: Dvojitá se píše jako =, trojitá jako # (jednoduchá se obvykle vynechává).
- Větvení řetězce: Zapisuje se do kulatých závorek ().
- Cykly (kruhy): Očíslováním atomů, kde se kruh spojuje (např. 1).

SMILES: příklady

Příklady pro rychlé pochopení:

Voda: O

Ethanol (lineární): CCO

Kyselina octová (větvená): CC(=O)O

Benzen (cyklus): c1ccccc1

Proč o tom jako chemici musíte vědět?

- Úspora místa: Místo velkých obrázků ukládáte molekuly jako krátký text do tabulek a databází.
- Strojové učení a AI: Moderní algoritmy a neuronové sítě pracují s textem – SMILES je klíčem k chemické informatice a AI.

Složitější příklad: Kofein a Chirality

Kofein: CN1C=NC2=C1C(=O)N(C(=O)N2C)C

L-Alanin: N[C@@H](C)C(=O)O

PubChem Sketcher - nakreslíte 2D strukturu a web okamžitě vygeneruje její SMILES řetězec (a naopak).

<https://pubchem.ncbi.nlm.nih.gov/sketch/>

MolDraw / SmilesDrawer - čistě webový převodník, kam vložíte SMILES kód a on vám vygeneruje vektorový obrázek (SVG/PNG)

<https://www.moldraw.com/tools/free-chem-tools/smiles-to-structure.html>

National Cancer Institute (CACTVS) - dokáže přeložit běžný textový název chemikálie (např. "Aspirin") přímo na korektní SMILES řetězec.

<https://cactus.nci.nih.gov/chemical/structure>

Kódování obrazu

kodování obrazu



- rastrový obrázek - pixel, 24 bitů = 3 B na pixel

0	000	černá
1	001	modrá
2	010	zelená
3	100	červená
4	110	žlutá
5	101	magenta
6	011	cyan
7	111	bílá

- vektorová grafika - analýza geometrie, fotografie 12 Mpx $\approx$ 36 MB

Proč (by) se do databáze nikdy nevkládá spektrum jako JPEG?

Spektrum je v podstatě vektor (číselná řada);
screenshot spektra je rastr = ztráta dat

Kompresa: bezztrátová × ztrátová

Kompresa - zmenšuje veľkosť súborů a súložek, kompresní poměr
v počítači jsou všechny informace kódovány pomocí nul a jedniček – (např.
0010111010110001010110 atd.)

→ často se opakující úseky (např. čtveřice 0011), lze nahradit kratším
úsekem

Např: AAAAABBBCCCCCCCCC → 5A3B8C

kompresní algoritmy - bzip2, gzip, zip, LZ4, xz, rar, 7z, LHa, ...

- ztrátová - některé informace jsou trvale ztraceny → psychovizuální
nebo psychoakustický model. MP3, OGG, WMA, WMV, MPEG,
DivX, JPEG, ...
- bezztrátová - zip, rar, tiff, png, gif, flac, wmaLossless, ...

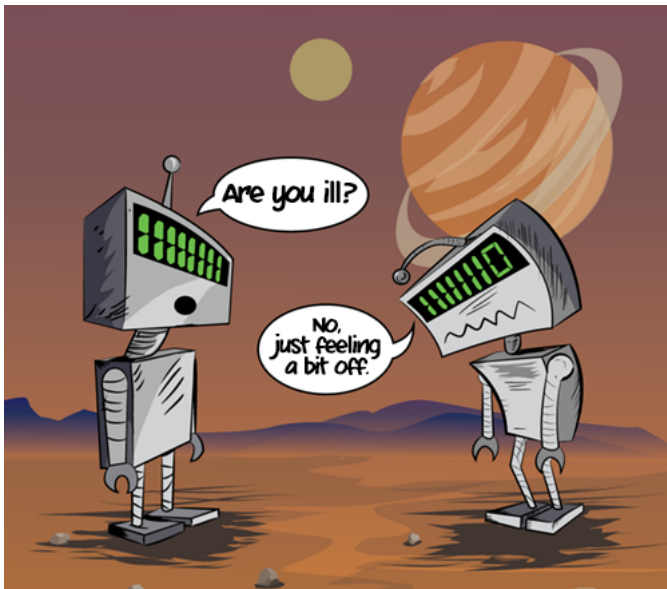
Souvislost umělou inteligencí (AI) a kompresí dat

komprese je inteligence nebo-li Hledání skrytých pravidelností

Náhodná data se stlačit nedají; čím více pravidelností, tím lépe se komprimuje (souvislost s entropií)

- Klasická komprese (RLE/ZIP): Hledá jednoduchá, doslovná pravidla (např. „zopakuj písmeno C 8krát“).
- AI komprese: Hledá komplexní a abstraktní pravidla. AI dokáže pochopit, co je na obrázku nebo v textu. Místo ukládání pixelů si model „zapamatuje“ koncept (např. „zde je lidská tvář s modrýma očima“).
- LLM (např. ChatGPT): Komprimovaný internet. Místo textu ukládá vztahy mezi slovy. Generování textu je pak procesem inteligentní dekomprese.
- Závěr: Dokonalá komprese vyžaduje dokonalé porozumění datům. Proto komprese = inteligence.

a bit off



Moje bakalářka jako datový projekt

- Příprava na závěrečnou prezentaci (5–10 min na konferenci):
 - 1 Jaká data vznikají (typ, formát)?
 - 2 Kolik (velikost/rok)?
 - 3 Jaká metadata?
 - 4 Kam uložím a jak zálohuji?
 - 5 Jak ověřím integritu?
 - 6 Co by mohlo ML predikovat?

Skripta, stránky a materiály

http://projekty.fs.vsb.cz/463/edubase/VY_01_042/Z%C3%A1klady%20informatiky.pdf
<http://home.zcu.cz/~pkral/texty.pdf>
<https://www.konvoj.cz/Nakladatelstvi/Knihy/100289-haluza-rybicka-hala-scriptum-43-uvod-do-informatiky/100289-uvod-do-informatiky.pdf>
http://www.bigyzt.cz/shared/clanky/2893/ICT-Pripravy/IS4-Pripravy_informace.pdf
<http://www.ivt.mzf.cz/informatika-v-kostce/zakladni-pojmy/kodovani-a-sifrovani/>
<https://www.itnetwork.cz/hardware-pc/principy-pocitacu/reprezentace-cisel-v-pocitaci>
<https://www.napocitaci.cz/33/symetricke-a-asymetricke-sifrovani-uniqueidgOkE4NvrWuNY54vrLeM677jX7sp3Lu-ZpLpGVMY1prA/>
https://en.wikipedia.org/wiki/Charles_Babbage
<https://en.wikipedia.org/wiki/FLOPS>
<https://explodingtopics.com/blog/data-generated-per-day>
<https://www.google.com/about/datacenters/>

odkazy na obrázky

obrázky:

<https://comic.browserling.com/tag/hardware>

zaznam <https://www.gym-karvina.cz/ivt/okruhy/informatika.htm>

porovnání <https://www.zmescience.com/feature-post/technology-articles/computer-science/how-big-data-can-get/>

ascii <https://simple.m.wikipedia.org/wiki/File:ASCII-Table.svg>

<https://vektorova-grafika.cz/rastrova-a-vektorova-grafika-jaky-je-rozdil/>

<https://www.fi.muni.cz/usr/jkucera/pv109/xdavidov.html>

<https://commons.wikimedia.org/wiki/File:VNEUMAN.GIF>

<https://leonardo-supercomputer.cineca.eu/hpc-system/>

<https://www.datacenterknowledge.com/supercomputers/lawrence-livermore-national-lab-begins-building-el-capitan-supercomputer>