

Základy forenzních databází ⇒ Základy AI v chemii

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

Heslo

Tell me and I forget,
teach me and I remember,
involve me and I learn.

Co byste se chtěli dozvědět?

Představte si, že zítra dostanete 10 000 chemických měření.
Co s nimi budete dělat?

Témata jednotlivých lekcí

1	16.9.	Úvod - AI, člověk, myšlení, forenzní, chemie
2	23.9.	Python 1
3	30.9.	Python 2
4	7.10.	Informace, záznam, kodování, hardware
5	14.10.	Algoritmus, kódy, RSA šifra
6	21.10.	Vyrovňovací logika, složitost
7	28.10.	PŘEDNÁŠKA NENÍ - STÁTNÍ SVÁTEK
8	4.11.	Správné tabulky, klíče a kardinalita, ERA model
9	11.11.	Datové struktury, y a řazení
10	18.11.	Vyhledávání dat a Strojové učení
11	25.11.	Strojové učení - algoritmy
12	2.12.	Kvantové počítače pro práci s daty
13	9.12.	PŘEDNÁŠKA NENÍ - spojené 16.12.
14	16.12.	Zkouška - konference - celý den

Jiné předměty o databázích a ML

Databázové systémy v chemii - ing. Sedláková Zuzana, Ph.D.

Databáze v chemické a forenzní analýze - doc. Raich Ivan, Ph.D.

Databázové systémy - ing. Valenta Michal, Ph.D.

Databázové systémy v chemii - doc. Kukul Jaromír, Ph.D.

Metodika analýzy dat: Od základů po aplikace metod strojového učení -
Vrba Jan Ing. Ph.D.

Strojové učení v Pythonu - Hrnčířík Pavel doc. Ing. Ph.D.

Požadavky k úspěšnému ukončení

- aktivní účast
- prezenční účast (min. 9 lekcí)
- závěrečná prezentace 5-10 minut na společné konferenci -> úvod k vaší bakalářské práci, jaká data a kolik dat potřebujete, jaká databáze by se hodila pro vaši práci, jak by taková databáze mohla vypadat, jak by vám ML mohlo pomoci, jaký algoritmus byste použili...
- např.: „Stanovení drog v biologických vzorcích“ - Jaká data vznikají? Kolik jich je? Jak by měla vypadat databáze? Co by mohl predikovat ML model?

Proč jsou data tak důležitá?

chemik dnes nepracuje pouze s: baňkou, spektrometrem, chromatografem, laboratorním zápisem

ale také s: numerickými daty, spektry, chromatogramy, strukturami molekul, vlnovými funkcemi, obrazy, texty, databázemi, výsledky simulací.

experiment: vzorek → přístroj → signál → data → informace → rozhodnutí

jeden vzorek – mnoho dat

Neznámý prášek ...co s ním?

Co vlastně znamená „mít data“?

Nic pokud s nimi neumíme pracovat. Existuje spousta naměřených experimentů a počítačových simulací v šuplíkách.

Správně pochopená a zpracovaná data -> Nobelova cena

Data z telefonů, chatů...Etika

Jaké znáte databáze?

Co mají všechny tyto příklady společného?

určitým způsobem organizované a strukturované údaje.

např: Caffeine

CID: 2519 Formula: $C_8H_{10}N_4O_2$ MW: 194,19 g/mol CAS: 58-08-2

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

Které z těchto údajů jsou vhodné pro člověka a které pro počítač?

Proč lidé začali data ukládat a organizovat?

Když už něco víme, chceme si to pamatovat.
A když toho víme hodně, potřebujeme to uspořádat.

paměť člověka → zápis → seznam → kartotéka → databáze → AI

Chemik v roce 1900:

„Tuto látku jsme už někdy měřili. Kde je ten záznam?“

Chemik dnes:

„Vyhledám látku podle CAS / struktury / spektra.“

Chemik za 100 let?

První databáze?

12st.př.nl. snaha o sběr dat a
vytvoření abecedy - Ugarit
(Sírie) ~ 1500 tabulek
(dopisy, diplomacie, básně,
hebrejská bible...)



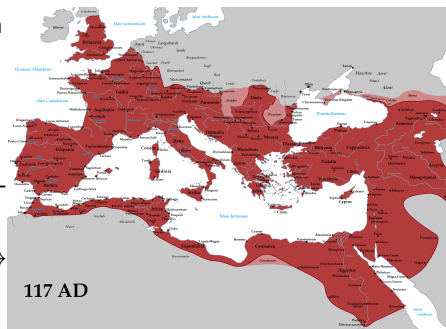
Je hromada tabulek databáze?

Od záznamu k organizaci

475př.n.l. stal Řím republikou - snaha
o třídění společnosti
(Dějiny psané Římem)

...

Nestačí informace pouze uchovat. Potřebujeme je také klasifikovat a najít.
zapsat → uspořádat → klasifikovat →
vyhledat



1729 — kartotéka

1729 první kartotéka - Carl Linnaeus



Periodická tabulka :-)

Když člověk nestačí

1890 prvním elektromechanické zpracování dat - Herman Hollerith -> International Business Machines (IBM)

Potřeba: zpracovat obrovské množství záznamů

Ruční přepis → počítač →
databáze →
automatizované zpracování



Od kartotéky k počítači

- 1951 UNIVAC - první digitální počítač pro komerční využití - Electronic Discrete Variable Automatic Computer (EDVAC)
- 1960 Common business-oriented language (COBOL-60) - programovací jazyk pro databáze; magnetické disky - nesériový přístup k datům
- 1970 relační model E. F. Codd

Máme soubor s 10 miliony chemických záznamů.

Chceme najít CAS 58-08-2.



```

Command ==>
=COLS> ----+----1----+----2----+----3----+----
*****
000100 000100 IDENTIFICATION DIVISION.
000200 000200 PROGRAM-ID. HELOWOLD.
000300 000300 ENVIRONMENT DIVISION.
000400 000400 PROCEDURE DIVISION.
000500 000500 DISPLAY 'HELLO WORLD'.
000600 000600 STOP RUN.
*****

```

Když máme miliony záznamů

proč databázové systémy vznikly...

Počítač už umí data ukládat.

Ale jak je rychle najdeme?

Jak je můžeme měnit?

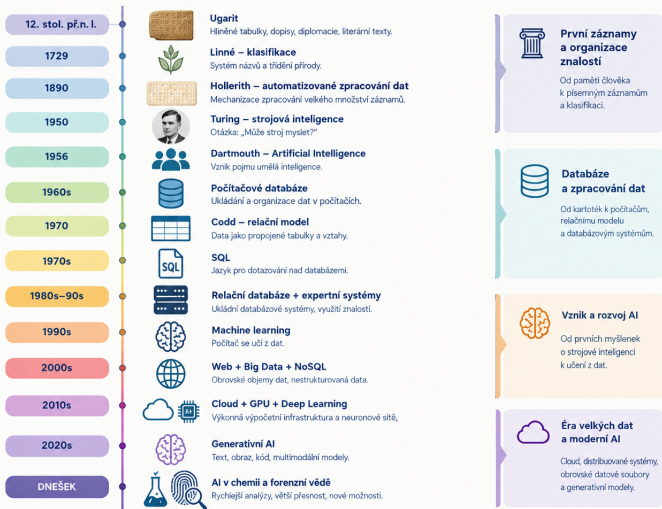
Jak může pracovat více lidí současně?

Jak zabráníme chybám?

→ **databázové systémy**

Od prvních záznamů k umělé inteligenci

Stejný cíl po staletí: jak z velkého množství dat získat užitečnou informaci.



Celou dobu řešíme stejný problém:
jak z velkého množství dat získat užitečnou informaci.

Od ukládání dat k učení z dat

Na začátku jsme data ukládali proto, abychom je **neztratili**.
Potom jsme se naučili data organizovat a **vyhledávat**.
Dnes chceme, aby počítač v datech sám hledal **souvislosti**.

Databáze: „Najdi látky podobné této látce.“

ML: „Na základě tisíců známých látek odhadni vlastnost nové látky.“

Databáze vs. ML vs. AI

Databáze nám umožnily data ukládat a vyhledávat.

Strojové učení nám umožňuje hledat v datech vzory.

AI se snaží tyto vzory využít k řešení problémů.

Databáze: „Najdi všechny molekuly s $MW < 300$.“

ML: „Odhadni jejich rozpustnost.“

AI: „Navrhni molekuly s požadovanými vlastnostmi.“

Co je databáze?

“Definice” Databáze neboli datová základna (Data Base) je místo, kam se ukládají **určitým** způsobem organizované a strukturované údaje.

Databáze = data + způsob jejich organizace + nástroje pro práci s nimi

Data → samotné údaje

Databáze → organizovaná data

Databázový systém → software, který umožňuje databázi vytvářet, spravovat a používat

Co databáze a ML model obsahuje?

co možná kompletní data pro popis dané skutečnosti pro daný účel

Například:

Databáze drog pro forenzní laboratoř:

název, vzorec, strukturu, molární hmotnost, CAS číslo, spektrum, toxicitu, fyzikální vlastnosti, odkazy na zdroje.

barvu obalu, místo činnu, číslo případu, datum nálezu

Co rozhoduje o tom, co do databáze uložíme? → **účel databáze.**

Co rozhoduje o tom, jak ML model vycvičíme? → **účel ML modelu.**

Data × Informace × Znalost

3350 → data

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

vzorek obsahuje alkohol → znalost

vzorek odpovídá očekávanému složení → závěr

Databáze uchovává data. Člověk z nich vytváří informace a znalosti. AI se snaží **pomáhat** s cestou od dat k užitečným závěrům.

Proč databáze?

*Když už se něco dozvíme,
chceme to i později použít.*

Experiment → poznatek → data → uložení → pozdější použití Např:

Máme 100 000 spekter.

Potřebujeme najít všechna spektra: naměřená metodou FTIR, v rozsahu 400–4000 cm^{-1} , pro organické látky, s určitou podobností k našemu vzorku.

Otázka: Jak byste to udělali v obyčejném adresáři souborů?

Databáze a AI jsou všude

Periodická tabulka prvků...telefonní čísla....Archivy počasí...Kartotéka ve zdravotnictví...firmy (Sigma Aldrich)... filmy (Netflix)...hry (Duna)...osobní uložště...ve všech sportech....sociální media...vládní organizace....google...telefony...auta...vysavače...

Databáze nejsou něco, co používají jen programátoři. AI používáme všichni.

A co forenzní věda?

Jaká data byste jako forezníci chtěli mít k dispozici?

DNA, otisky, drogy, léčiva, toxikologická data, spektra, chromatogramy, střelivo, materiály, požáry, dokumenty, fotografie, databáze pachatelů / případů ...

A co forenzní věda?

Jaká data byste jako forezníci chtěli mít k dispozici?

DNA, otisky, drogy, léčiva, toxikologická data, spektra, chromatogramy, střelivo, materiály, požáry, dokumenty, fotografie, databáze pachatelů / případů ...

Forenzní věda je obrovský datový problém.

Př.: Neznámý vzorek má MS spektrum.

Databáze obsahuje 1 000 000 referenčních spekter.

Jak najdeme nejpodobnější?

A: porovnáme se všemi

B: data nejprve uspořádáme

C: vytvoříme index

D: použijeme algoritmus podobnosti

E: naučíme model

Forenzní problém = data + porovnání + rozhodnutí

Neznámý vzorek → změřím → získám data → porovnám s databází → najdu podobné záznamy → vyhodnotím výsledek

Co když databáze obsahuje 100 záznamů?

Vzorek → měření → data → databáze → porovnání → rozhodnutí

Co když jich obsahuje 100 milionů?

Vzorek → měření → data → model → predikce → rozhodnutí

Problém není jen „mít data“. Problém je umět s nimi efektivně pracovat.

Proč databáze místo souborů?

- Potřebujeme najít konkrétní záznam.
- Potřebujeme najít všechny záznamy splňující podmínku.
- Potřebuje pracovat více lidí současně.
- Potřebujeme určit, kdo smí data měnit.
- Potřebujeme pracovat s miliony záznamů.

Právě proto používáme databázové systémy.

Kdo stojí za databází?

- Vědec / forenzní odborník → Co chceme zjistit?
- Analytik / odborník na data → Jaká data k tomu potřebujeme?
- Databázový specialista / programátor → Jak databázi vytvoříme?
- IT architekt → Na jakém systému bude databáze fungovat?

Databáze není primárně programátorský problém. Začíná otázkou odborníka.

Př: Forenzní databáze drog

Forenzní chemik: „Chci identifikovat neznámé vzorky.“

Datový analytik: „Potřebujeme spektra, struktury, názvy, metadata.“

Databázový specialista: „Navrhnou strukturu databáze.“

IT: „Vyřešíme server, zálohy, přístup.“

Správný návrh databáze je velmi abstraktní a velmi složitá záležitost.

Správný návrh databáze je složitý

vzorek | látka | chemik | laboratoř | přístroj | spektrum

Co se stane, když jeden vzorek obsahuje 5 látek?

Co když jedna látka byla měřena 100×?

tabulky, klíče, kardinalita, ER model

Co chceme vědět?



Jaké informace potřebujeme?



Jak spolu souvisejí?



Jak je uložíme?



Jak je budeme hledat?

Např: Které látky byly nalezeny v případech z roku 2025 a zároveň byly identifikovány metodou GC-MS?

Co je potřeba k vytvoření databáze?

- 1 Myšlenka - otázka, účel, odborná znalost, logické propojení informací.
- 2 Počítač - data, hardware, úložiště, databázové prostředí.
- 3 Komunikace - programovací jazyk, syntaxe, sémantika, dotazy, vyhledávání

Př: Které látky se nejčastěji objevují v analyzovaných vzorcích?

Myšlenka: co chceme zjistit

Data: jaká měření máme

Databáze: jak je uložíme

Dotaz: jak je vybereme

Algoritmus: jak je zpracujeme

AI: co můžeme predikovat

Od databází k AI

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

AI nezačíná neuronovou sítí. AI začíná daty.

Př: 10 000 známých spekter → databáze → vyhledávání podobných
spekter → algoritmus podobnosti → ML model → predikce neznámého
vzorku

AI není kouzelná krabička

Tři zásadní myšlenky:

DATA → ALGORITMUS → MODEL

Špatná data → špatný model → špatné rozhodnutí.

Př: Predikce toxicity

Máme 50 000 látek.

Ale:

některé mají špatně změřenou toxicitu,

některé mají chybějící hodnoty,

některé jsou několikrát v databázi,

některé mají různé názvy,

některé byly měřeny za různých podmínek.

Může být model dobrý, když jsou špatná vstupní data?

Co byste před trénováním modelu zkontrolovali?

Co budete za 10 let potřebovat jako chemici?

Laboratoř:

přístroj

vzorek

měření

interpretace

Datová práce:

data

databáze

Python??

algoritmy

ML

AI

Moderní chemik potřebuje obojí.

Co bude předmětem

	Téma		Otázka	
	-----		-----	
	Data		Co je vlastně chemické datum?	
	Úložiště		Kam uložíme 10 TB měření?	
	Přenos		Jak dlouho data potrvají přes síť?	
	Algoritmy		Jak najdeme záznam v milionu položek?	
	Složitost		Proč jeden algoritmus škáluje a jiný ne?	
	Řazení		Jak uspořádáme milion záznamů?	
	Vyhledávání		Jak najdeme podobnou látku?	
	ML		Jak předpovíme vlastnost nové molekuly?	
	AI		Jak může AI pomoci chemikovi?	
	Kvantové počítače		Co mohou změnit v práci s daty?	

Co bude předmětem

Data v chemii a forenzní vědě,
uložiště, přenosová rychlost,
práce s daty,
algoritmizace, složitost
počítačové ukládání, řazení, vyhledávání,
základy stojového učení,
využívání AI,
kvantové počítače jako hrozba?

Co bude předmětem

Data v chemii a forenzní vědě,
uložiště, přenosová rychlost,
práce s daty,
algoritmizace, složitost
počítačové ukládání, řazení, vyhledávání,
základy strojového učení,
využívání AI,
kvantové počítače jako hrozba?

chemické informace → data → datové struktury → algoritmy → strojové učení → AI → budoucnost výpočetní chemie

Témata jednotlivých lekcí

1	16.9.	Úvod - AI, člověk, myšlení, forenzní, chemie
2	23.9.	Python 1
3	30.9.	Python 2
4	7.10.	Informace, záznam, kodování, hardware
5	14.10.	Algoritmus, kódy, RSA šifra
6	21.10.	Vyrovková logika, složitost
7	28.10.	PŘEDNÁŠKA NENÍ - STÁTNÍ SVÁTEK
8	4.11.	Správné tabulky, klíče a kardinalita, ERA model
9	11.11.	Datové struktury, y a řazení
10	18.11.	Vyhledávání dat a Strojové učení
11	25.11.	Strojové učení - algoritmy
12	2.12.	Kvantové počítače pro práci s daty
13	9.12.	
14	16.12.	Zkouška - konference - celý den