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

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.	Vyroková logika, složitost
7	28.10.	PŘEDNÁŠKA NENÍ - STÁTNÍ SVÁTEK
8	4.11.	Správné tabulky, klíče a kardinalita, ERA model
9	11.11.	Datové struktury, 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

Co bylo minule a kam jdeme

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

- Data: bit, bajt, kódování, metadata, úložiště
- Algoritmy: postup krok za krokem, logika (podmínky), složitost (cena)
- Šifry: symetrická, asymetrická, Enigma, RSA, důvěryhodnost zprávy a integrita posílaných dat

Dnes: jak data uložit, aby se v nich nedalo ztratit, nedala se zkazit a dalo se z nich učit.
(Na konci hodiny z vaší databáze uděláme dataset pro strojové učení.)

Excel z laboratoře: kolik chyb najdete?

Takhle vypadá evidence zajištěných drog vedená „rychle v Excelu“:

Případ	Datum	Vzorek	Nalezené látky	Množství	Analytik
KR-12/26, Praha 4	3.2.2026	bílý prášek	kokain, levamisol	0,52 g	Novák J.
KR-12/26, Praha 4	2026-02-03	bílý prášek 2	kokain	480 mg	J. Novák
KR-15/26, Brno	5. 2. 26	tableta	MDMA	120	Dvořáková
KR-15/26, Brno	5. 2. 26	tableta	MDMA	120	Dvořáková
KR-17/26, Praha	7.2.2026	rostlina	THC 12 %	3,1 g	

Najděte co nejvíc problémů. Které z nich by vadily počítači a které u soudu?

Excel z laboratoře: co je špatně

- „kokain, levamisol“: **dvě hodnoty v jedné buňce** → nejde vyhledat všechny vzorky s levamisolem
- „KR-12/26, Praha 4“: **případ a místo v jednom poli**; Praha × Praha 4
- tři **formáty data**; **jednotky v textu** (g, mg, nic) → nejde sčítat ani porovnávat
- „THC 12 %“: **koncentrace schovaná v názvu látky**
- „Novák J.“ × „J. Novák“: je to jeden člověk? **Opakovaná data se rozjedou**
- **duplicitní řádek** a žádný jednoznačný identifikátor vzorku
- **chybí analytik** → přerušený řetězec úschovy, důkaz je zpochybnitelný

Každou chybu dnes vyřešíme: atomické hodnoty (1NF), datové typy, rozdělení tabulek (2NF, 3NF), klíče, integrita.

integrita (pocházející z latinského integritas) obecně znamená celistvost, neporušenost, nedotknutelnost nebo vnitřní jednotu

Jak uložit evidenci jednoho případu?

Stejná data (případ KR-12/26, vzorky, osoby, telefony) lze uložit několika způsoby.

Připomeňme si:

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

$$= \text{data i software}$$

DBMS - DataBase Management System , SŘBD - Systém Řízení Báze Dat je softwarové vybavení, které zajišťuje práci s databází.

- Definovat bázi
- Konstruovat databázi
- Manipulovat s databází

Kde databáze potkáte

- V laboratoři: LIMS (laboratorní informační systém) eviduje vzorky, analýzy, přístroje i řetězec úschovy
- V chemii: PubChem (přes 100 milionů látek), spektrální knihovny NIST a SWGDRUG, Reaxys, SciFinder
- Ve forenzní praxi: národní databáze DNA, AFIS (otisky prstů), databáze drog a nových psychoaktivních látek (EUDA)
- V kapse: SQLite v každém mobilu (kontakty, zprávy) $\rightarrow$ i to je forenzní stopa

Systém Řízení Báze Dat: SQLite, PostgreSQL, MySQL, Oracle, MS SQL Server, MS Access

Kterou databázi jste dnes už použili, aniž byste o tom věděli?

Data versus informace

Data - konkrétní hodnoty (údaje), syntaktická složka

Informace - znalost, vyplývající z hodnot uložených v databázi, sémantická složka

databázový model - způsob, jakým jsou data strukturována, propojena a spravována.
Nebo-li: jak spolu data souvisí?
$\Rightarrow$ efektivní ukládání, dotazování, sdílení a integrita dat.

Např: DNA vs. AFIS - jiná data, jiná otázka, jiný model

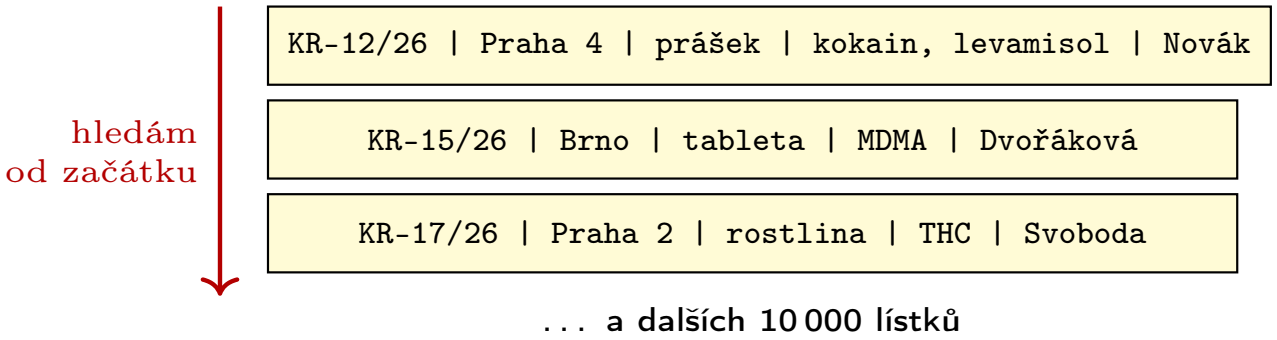
Jak uložit evidenci jednoho případu?

Stejná data (případ KR-12/26, vzorky, osoby, telefony) lze uložit několika způsoby.

Jak byste evidenci vedli, kdybyste neměli počítač?

- sešit nebo kartotéka → seznam
- šanon se složkami → hierarchický model
- nástěnka s fotkami a provázky jako v detektivkách → síťový model
- tabulky s odkazy mezi sebou → relační model
- krabice, kam se hází, co přijde → dokumentový model (NoSQL)

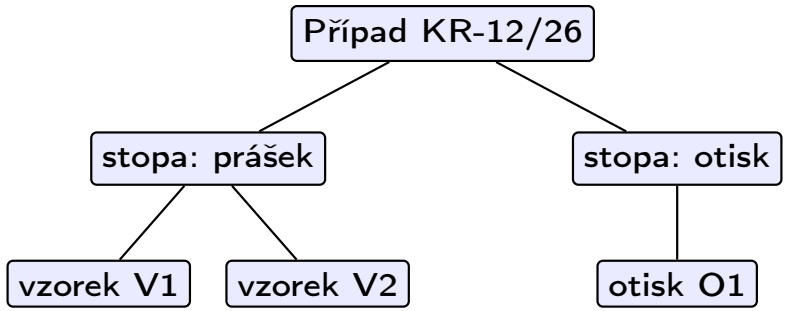
1. Seznam: sešit a kartotéka



- záznam za záznamem; přesně tak vypadá i Excel z úvodu
- kartotéka Carla Linného (kolem 1767), děrné štítky H. Holleritha (sčítání lidu USA 1890)
- $+$ jednoduchý zápis; $-$ hledá se postupně, $O(n)$; opakování; žádné vazby

Jak v kartotéce najdete všechny případy s levamisolem?

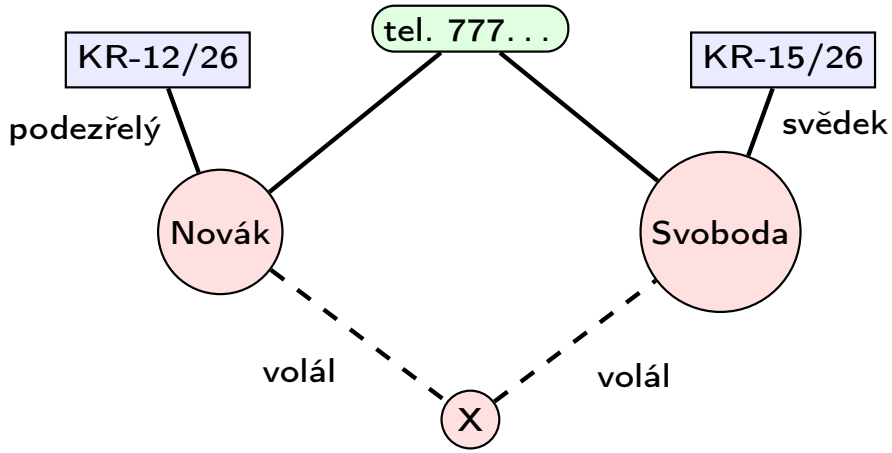
2. Hierarchický model: šanon se složkami



- každý podřízený má právě jednoho nadřízeného (strom)
- IBM IMS (1966) pro program Apollo: raketa → stupeň → díl
- + rychlé procházení shora dolů; − Novák ve dvou případech = dvě kopie

Kde v počítači používáte hierarchickou databázi každý den?
Složky a soubory na disku.

3. Síťový a grafový model: nástěnka s provázky



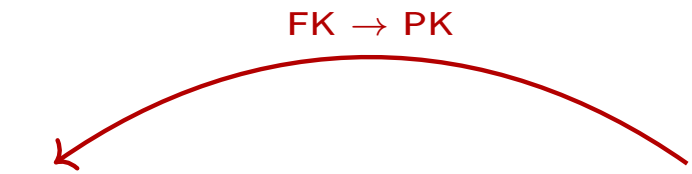
- každý uzel může mít libovolně mnoho vazeb (graf)
- dnes grafové databáze (Neo4j); novináři jimi rozplétali Panama Papers

`https://cs.wikipedia.org/wiki/Panama_Papers`

- + vztahy jsou hlavní informací; − složité dotazy typu „všechno o vzorku“

Jak zjistíte, že se Novák a Svoboda znají přes prostředníka?
Hledáním cesty v grafu: Novák – X – Svoboda.

4. Relační model: tabulky propojené klíči



Případ

id_pripad	misto
KR-12/26	Praha 4
KR-15/26	Brno

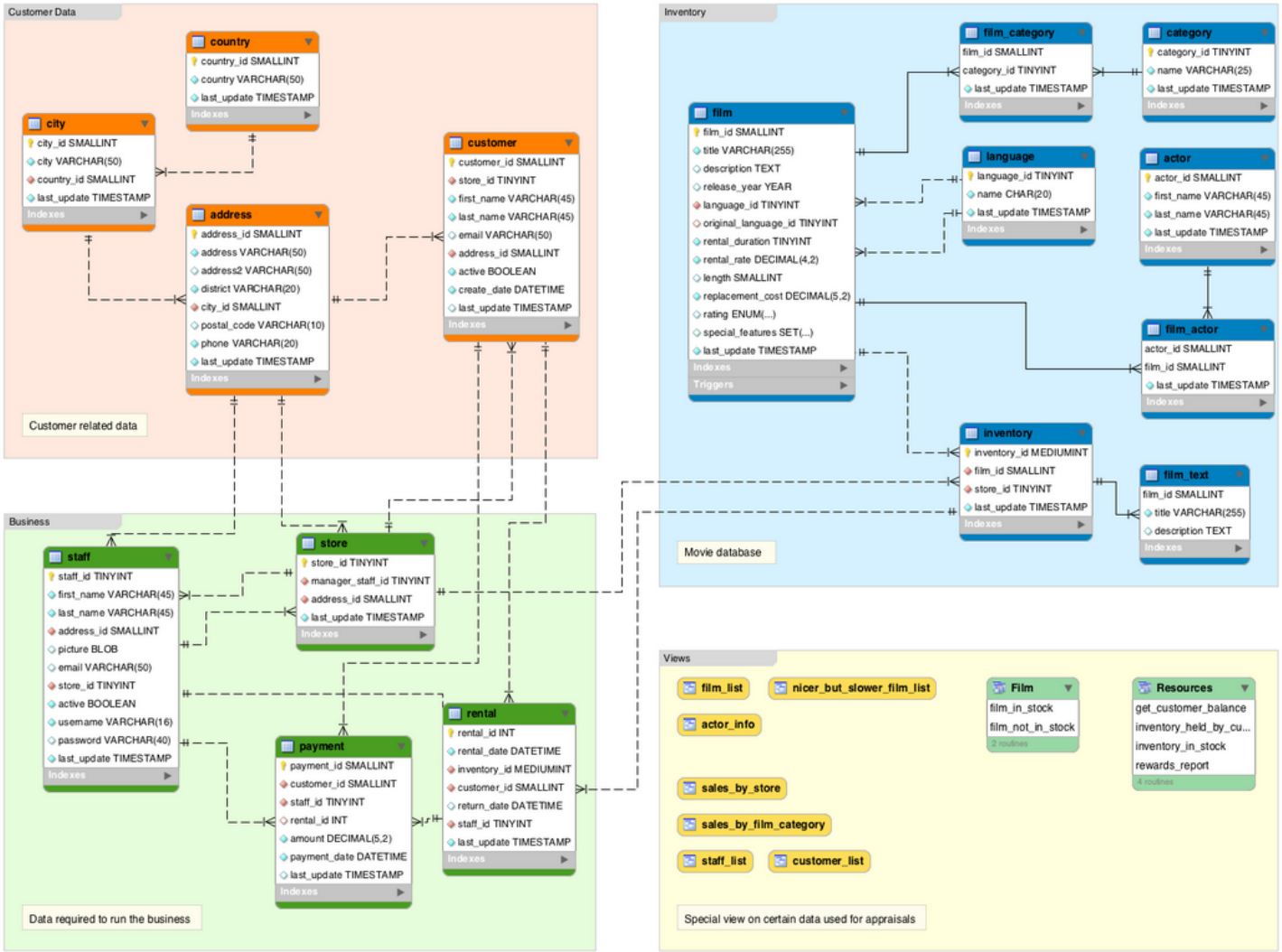
Vzorek

id_vzorek	*id_pripad*	hmotnost_g
V1	KR-12/26	0,52
V2	KR-12/26	0,48
V3	KR-15/26	0,12

- E. F. Codd (IBM, 1970); dotazovací jazyk SQL
- každý fakt v jedné tabulce, vazby přes klíče (primární PK, cizí FK)
- + integrita, standard, univerzální dotazy; − pevná struktura

Kam byste v tabulkách uložili fotografii stopy?
Soubor na disk; do tabulky cestu k souboru a jeho otisk (hash).

Jak vypadá relační databáze



5. Dokumentový model (NoSQL): data ze zabaveného mobilu

```
{ "zarizeni": "telefon Novák",
    "kontakty": [{"jmeno": "X", "cislo": "777..."}],
    "zpravy": [{"komu": "X", "cas": "2026-02-02 23:14",
                "text": "zitra v 10"}],
    "gps": [[50.06, 14.43], [50.07, 14.45]] }
```

- každý záznam může mít jinou strukturu (JSON); MongoDB
- další NoSQL: klíč–hodnota (Redis), sloupcové (Cassandra)
- + pružné, rychlé pro obrovská data; − slabší kontrola integrity

Proč by se data z telefonu špatně ukládala do pevných tabulek?
Každý telefon i aplikace ukládá jiné položky.

Časová osa databázových modelů

Kdy	Model	Kdo / co	Charakteristika
kolem 1767	seznam (kartotéka)	Carl Linné: lístkový katalog	lineární, hledá se postupně
1890	děrné štítky	H. Hollerith, sčítání lidu USA → IBM	strojové třídění
1966	hierarchický	IBM IMS (program Apollo)	strom, pevné vazby
1969	síťový	CODASYL (COBOL)	graf, složitá navigace
1970	relační	E. F. Codd (IBM), SQL	tabulky a klíče; standard dodnes
90. léta	objektový	objektové a objektově-relační DB	složité objekty, multimédia
od 2000	NoSQL, grafové	MongoDB, Neo4j, Cassandra	velká a nestrukturovaná data

Který model zažil nejdelší kariéru a proč asi?

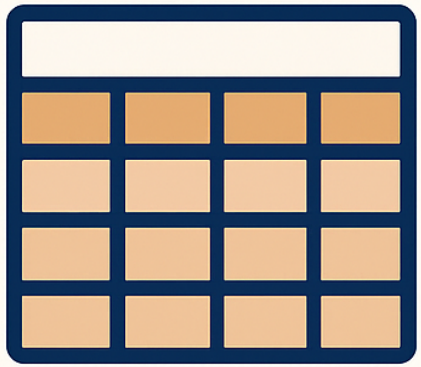
Který model zvolíte?

- periodická tabulka prvků → **relační** (každý prvek má stejné vlastnosti)
- fotodokumentace místa činu ve složkách → **hierarchický**
- síť dealerů a jejich telefonátů→ **grafový**
- data vytěžená ze zabaveného telefonu → **dokumentový (NoSQL)**
- evidence vzorků, analýz a výsledků v laboratoři → **relační** (LIMS)
- databáze DNA profilů a databáze otisků prstů → to je otázka na další snímky

DNA vs. otisk

Proč má DNA databáze jinou strukturu než databáze otisků prstů?







Společná práce

Navrhněte DNA databázi a databázi otisků prstů? V čem by mohli být shodné a v čem se lišit. Jaký model by byl pro tu kterou databázi vhodnější a proč?
DNA databáze a databáze otisků prstů mají odlišnou datovou povahu, způsob porovnávání i požadavky na uložení, jiný cíl...
Proto potřebují odlišnou strukturu databáze, optimalizovanou pro konkrétní typ informací.

DNA vs. Otisk

	DNA profil	**Otisk prstu**
data	alely na STR lokusech (EU: 12 lokusů + pohlaví)	obraz linií → markanty (ukončení, rozdvojení)
příklad	D8S1179: 13, 14 | TH01: 6, 9.3 | ... | AMEL: X, Y	(112; 240; 35°; ukončení) | (150; 198; 120°; rozdvojení) | ...
velikost	desítky čísel	obraz stovky kB, šablona stovky B
porovnání	shoda alel lokus po lokusu	zarovnání markantů (posun, otočení, deformace)
výsledek	shoda/neshoda + pravděpodobnost náhodné shody	seznam kandidátů se skóre → ověří daktyloskop
databázový model	relační: Osoba – Vzorek – Profil – Alela	relační evidence + archiv obrazů + šablony markantů
algoritmy	statistika četností alel; u směsí pravděpodobnostní modely	rozpoznávání obrazu, hledání podobnosti, hluboké učení

AFIS vs. hledání ve spektrální knihovně

	Spektrální knihovna (GC-MS)	**AFIS (otisky prstů)**
neznámý	spektrum neznámé látky	stopa z místa činu
porovnává se	píky (m/z, intenzita)	šablona markantů (poloha, směr, typ)
knihovna	NIST, SWGDRUG	otisky osob, neobjasněné stopy (cBIS)
výsledek	kandidáti s „match factor“	kandidáti se skóre shody
rozhoduje	analytik	daktyloskop

pořadí	**kandidát**	**skóre**
1	karta osoby 2019/0457, pravý ukazovák	**0,94**
2	stopa z případu KR-08/24	0,61
3	karta osoby 2021/1180, pravý prostředníček	0,58

Ukázkový (vymyšlený) seznam kandidátů pro jednu stopu

Proč systém nevrátí rovnou jednu odpověď?

Stopy bývají neúplné a rozmazané: vysoké skóre je jen kandidát k ověření, ne důkaz. Hledání podle podobnosti je typická úloha strojového učení

NÁVRH TABULEK

NÁVRH TABULEK

Jak zapsat měření, aby se v nich dalo počítat?

Tabulka = entita

id_latka	nazev	cas	molarni_hmotnost_g_mol
1	kokain	50-36-2	303,35
2	kofein	58-08-2	194,19
3	MDMA	42542-10-9	193,25

- entita (tabulka) = Látka; atribut (sloupec) = název, CAS, molární hmotnost
- záznam (řádek) = jedna látka; primární klíč = id_latka
- jednotka je v názvu sloupce, v buňkách jsou jen čísla

Proč má tabulka sloupec id_latka, když každá látka má CAS?
Odpověď přijde u klíčů.

Tabulka obsahuje

Atribut: sloupec

- určují vlastnosti objektů (CAS,koncentrace...)
- jedinečný název a - nepoužívat diakritiku a mezery, tedy např.: nazev_odberatele
- určený datový typ (číslo, text, logická hodnota, ...)

Záznam: řádek

- každý záznam je jednoznačně rozlišitelný - jeden řádek reprezentuje např. jednoho zaměstnance s hodnotami daných atributů
- mít svůj jedinečný identifikátor

id_zamestnance	prijmeni	jmeno	plat	datum_nastupu
1	Novák	Adam	32000	2013-05-02
2	Nová	Jana	45000	2013-06-11

Společná práce: měření v triplikátech

Koncentraci benzoylekgoninu (metabolit kokainu) v moči jste změřili třikrát:

- vzorek M1: 152, 148 a 155 ng/ml
- vzorek M2: 98 a 101 ng/ml (třetí nástřik se nepovedl)

Navrhněte tabulku tak, aby se z ní dal spočítat průměr a směrodatná odchylka a aby šlo kdykoli přidat další nástřik.

Měření v triplikátech: dvě častá chybná řešení

Špatně 1: více hodnot v jedné buňce

vzorek	koncentrace
M1	152, 148, 155
M2	98, 101

Špatně 2: opakující se sloupce

vzorek	c1	c2	c3
M1	152	148	155
M2	98	101	

Co uděláte, když přijde čtvrtý nástřik? Jak spočítáte průměr všech měření v laboratoři?

Měření v triplikátech: správně

id_vzorek	opakovani	c_ng_ml
M1	1	152
M1	2	148
M1	3	155
M2	1	98
M2	2	101

- jeden řádek = jedno měření; čtvrtý nástřik = nový řádek
- primární klíč je dvojice (id_vzorek, opakovani)
- každá buňka obsahuje jednu hodnotu jednoho typu → 1. normální forma

Opakování jedné hodnoty: látka v každém řádku

id_vzorek	nazev	cas	M (g/mol)
V1	kokain	50-36-2	303,35
V2	kokain	50-36-2	303,35
V7	kokain	50-63-2	303,35

Najdete chybu? Kolik řádků byste opravovali u 5 000 nálezů kokainu?
Údaje o látce patří do vlastní tabulky **Látka**; nález si jen pamatuje id_latka. Překlep pak opravíte na jednom místě.

Co se dá spočítat, se neukládá

Průměr, směrodatná odchylka, molární hmotnost ze vzorce, stáří vzorku z data odběru...

```
SELECT id_vzorek, AVG(c_ng_ml) AS prumer, COUNT(*) AS n
FROM mereni
GROUP BY id_vzorek;
```

Proč je uložený průměr nebezpečný?

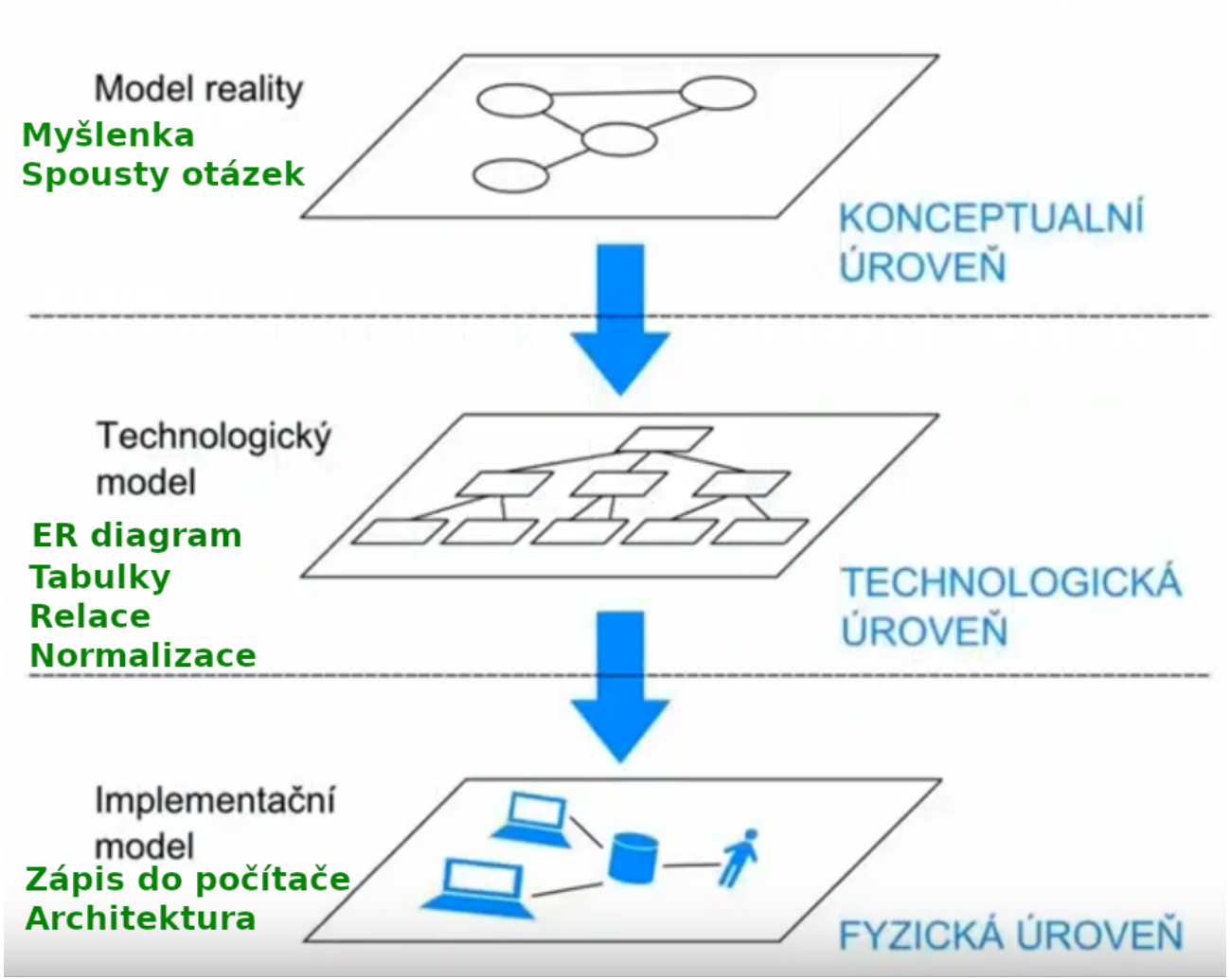
Opravíte jedno měření a uložený průměr zůstane starý → v databázi jsou dvě pravdy.

ERA MODEL

ERA MODEL

Jak si databázi nakreslit dřív, než ji postavíme?

Návrh relační databáze



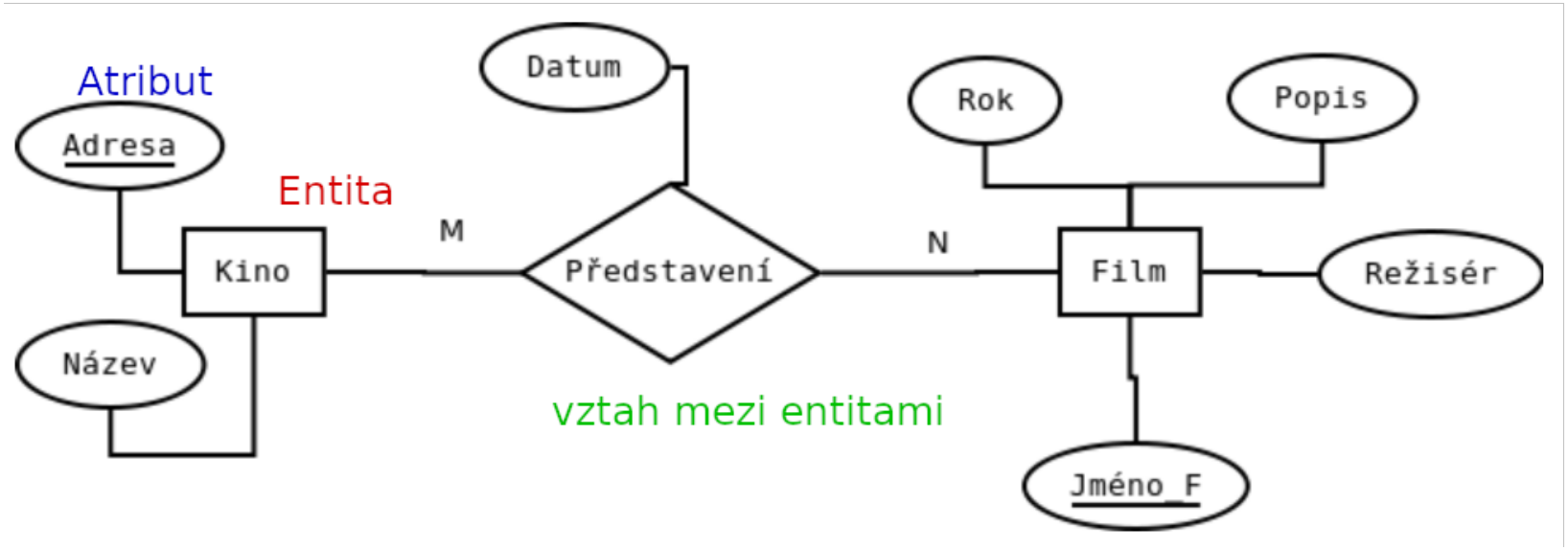
ERA model

- **E**ntity (objekt z reálného světa) → tabulka
- **R**elationship (vztah mezi entitami) → vztahy mezi tabulkami
- **A**ttribute (vlastnosti entit) → sloupeček tabulky

Modeluje „oblast zájmu“ (např. podnik) jako kolekci entit a vztahů (relationships) mezi nimi
- grafická reprezentace struktury databáze
dělá se před vytvářením databáze!! a slouží k rozmyšlení, jak bude databáze vypadat

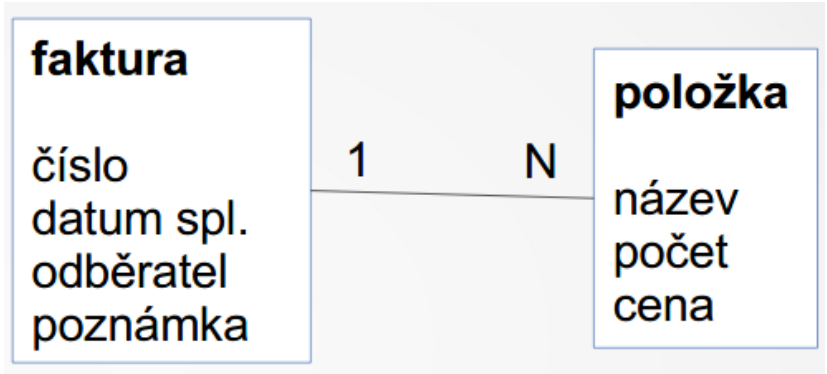
ERA model

Klasický formát

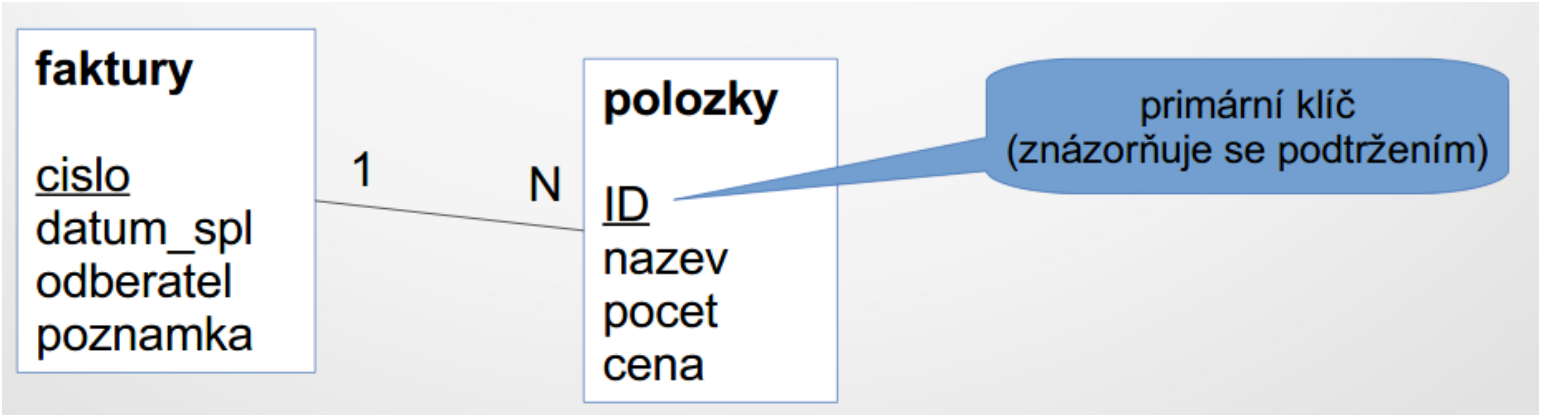


ERA model

tabulková varianta



implementační

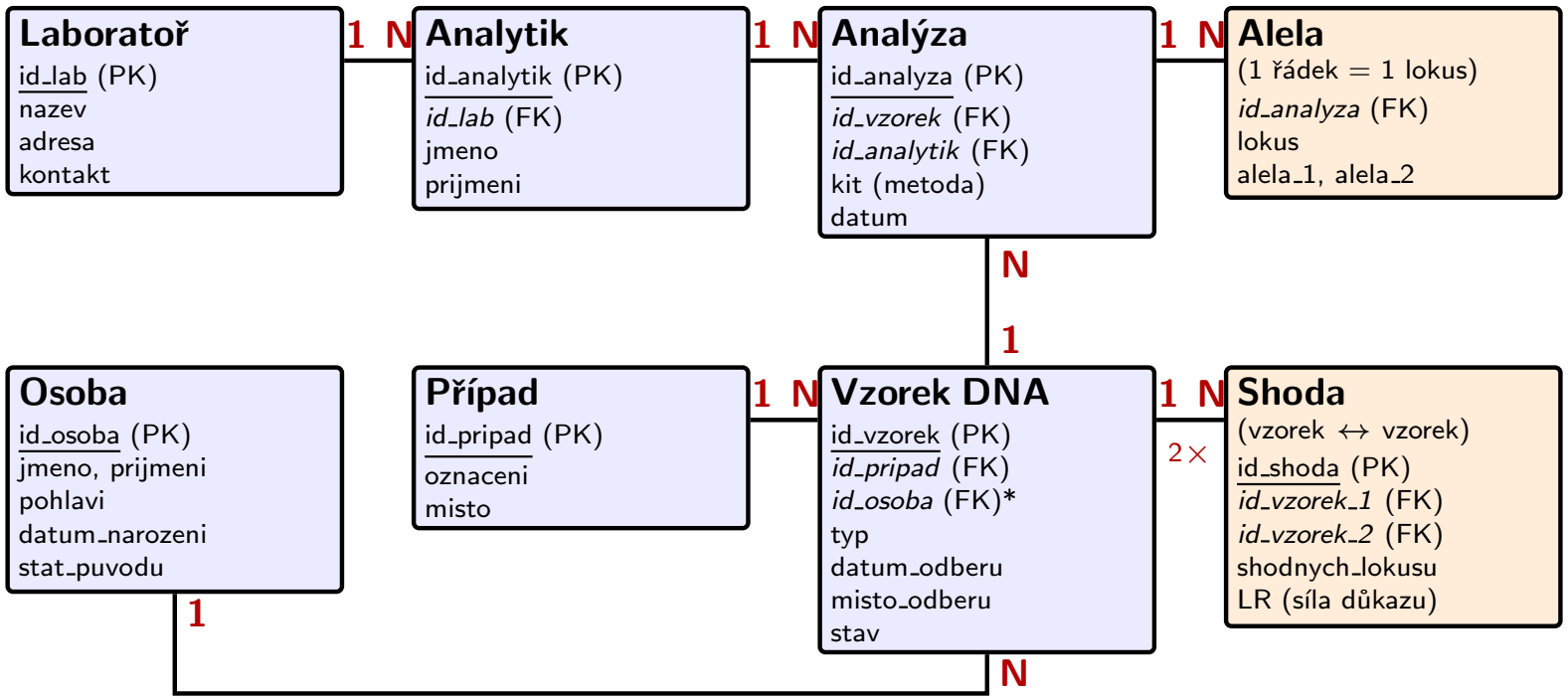


Společná práce

Navrhněte ERA model pro databázi DNA vzorků.
Model by měl umožnit uchovávat informace o:
Osobách (ID, jméno, pohlaví, datum narození, stát původu),
Vzorcích DNA (ID vzorku, datum odběru, místo odběru, stav analýzy),
Laboratořích (název, adresa, kontaktní osoba),
Analýzách (metoda, datum, výsledek, identifikátor osoby),
Shodách (matches) mezi vzorky z různých případů.

Určete entity, jejich atributy a vazby („má vzorek“, „analyzuje“, „obsahuje shodu“).
Co by se stalo, kdyby se změnila struktura pro uložení STR markerů?

Možné řešení: ERA model databáze DNA vzorků



* NULL = dárce stopy zatím neznámý

- Alela: jeden lokus = jeden řádek → nová sada STR markerů nemění strukturu databáze.
- Shoda spojuje dva vzorky (často z různých případů): M:N vzorku se sebou samým.

Kam zapíšete, že shodu ověřil znalec, a kdy?

Klíče a integrita

KLÍČE A INTEGRITA

Udržuje logickou správnost dat

Jak poznat, že jde o tentýž vzorek, a jak zabránit sirotkům?

Je to tentýž vzorek?

popis	hmotnost_g	datum
bílý prášek	0,52	2026-02-03
bílý prášek	0,52	2026-02-03

Jeden vzorek zapsaný dvakrát, nebo dva různé vzorky?

Bez jednoznačného identifikátoru to nepoznáte. Proto má každý řádek **primární klíč** a každá zkumavka čárový kód.

Primární klíč

Integrita Entitní – jedinečná identifikace každé entity - pomocí primárního klíče (PK)

Primární klíč

- zajistí jednoznačnost záznamů v tabulce,
- AUTO INCREMENT - zajistí automatické číslování => hodnota primárního klíče se v tabulce nesmí opakovat
- data nesmí obsahovat duplicitu nebo hodnotu Null.
- teoreticky může být libovolného typu
- může být složený z více sloupců – unikátnost je pak brána ze všech najednou, v jednotlivých sloupcích se tedy mohou opakovat stejné hodnoty

Přirozený, nebo umělý klíč?

- přirozený: rodné číslo, CAS, číslo případu, InChIKey
- umělý: id = 1, 2, 3, ... (AUTO INCREMENT), nemá žádný význam

Je CAS dobrý primární klíč pro tabulku látek?

- kokain 50-36-2 × kokain hydrochlorid 53-21-4: báze a sůl mají různá CAS
- směsi a mnoho nových psychoaktivních látek CAS nemají
- CAS se může změnit nebo zrušit

Proto: umělé id jako primární klíč, CAS jako běžný unikátní atribut.
InChIKey je otisk (hash) struktury, který si spočítá kdokoli.

Referenční klíč

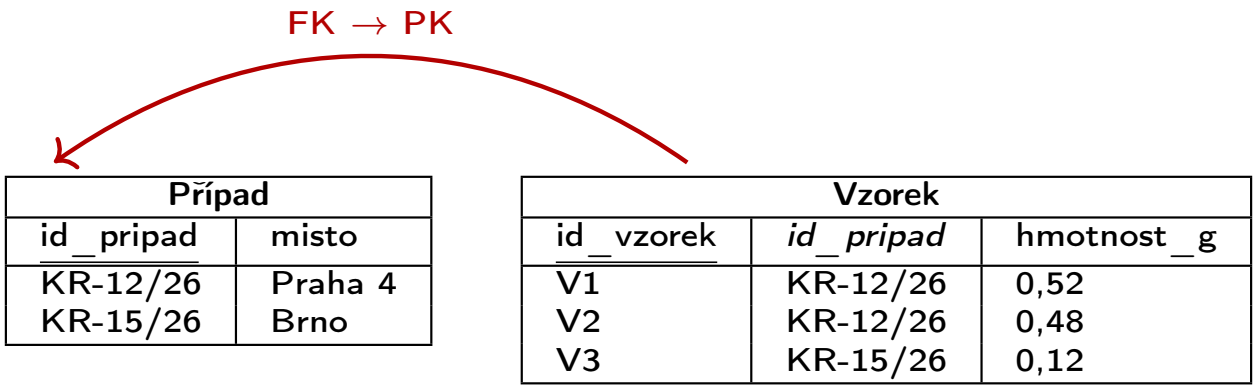
Integrita Referenční – pomocí cizího klíče (FK)

Cizí klíč

- odkazuje se na primární klíč v jiné tabulce – odkazuje se na právě 1 tabulku
- dvojice primární a cizí klíč tvoří vazbu (relaci) mezi tabulkami
- cizí klíč musí být stejného typu a velikosti jako primární

if primární klíč AUTO INCREMENT => cizí klíč dlouhé celé číslo
if primární klíč složený => struktura cizího klíče stejná
Do sloupce s definovaným cizím klíčem nelze vložit hodnotu, která není vložena v nadřízeném sloupci primární tabulky.

Tabulky propojené klíči



Případ	
id_pripad	misto
KR-12/26	Praha 4
KR-15/26	Brno

Vzorek		
id_vzorek	*id_pripad*	hmotnost_g
V1	KR-12/26	0,52
V2	KR-12/26	0,48
V3	KR-15/26	0,12

Cizí klíč v laboratoři: žádní sirotci

id_analyza	id_vzorek (FK)	metoda
A1	V1	GC-MS
A2	V3	HPLC
A3	V9	GC-MS

V tabulce Vzorek jsou jen V1, V2, V3. Analýza A3 odkazuje na neexistující vzorek → databáze ji **odmítne vložit**.

Jako štítek na zkumavce, který odkazuje na protokol o odběru: bez protokolu vzorek u soudu neobstojí.

Kdo by v Excelu z úvodu takovou chybu zastavil?

Nikdo. Databáze ano.

Integrita: co smí databáze smazat?

Mažeme případ, ke kterému patří zajištěné vzorky.

- Cascade: smaže i vzorky → ztráta důkazů
- Set null: vzorky zůstanou „bez případu“ → nikdo neví, k čemu patří
- Restrict: databáze smazání odmítne → **správná volba pro důkazy**

V praxi se důkazy nemažou vůbec: záznam dostane příznak „uzavřeno/archivováno“.

Kde by naopak Cascade dávalo smysl?

Např. smazání dočasného pracovního seznamu a jeho položek.

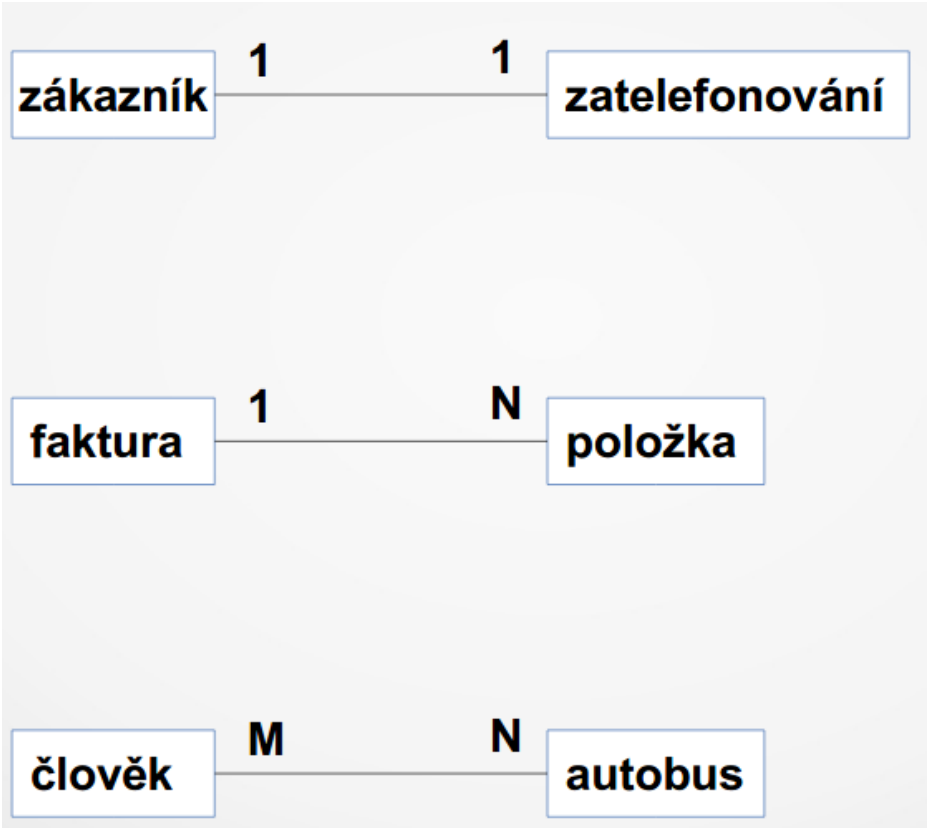
Kardinalita vztahu

mocnost vztahu mezi entitami
neboli
Kolik položek z tabulky A se váže na kolik položek z tabulky B?

3 typy

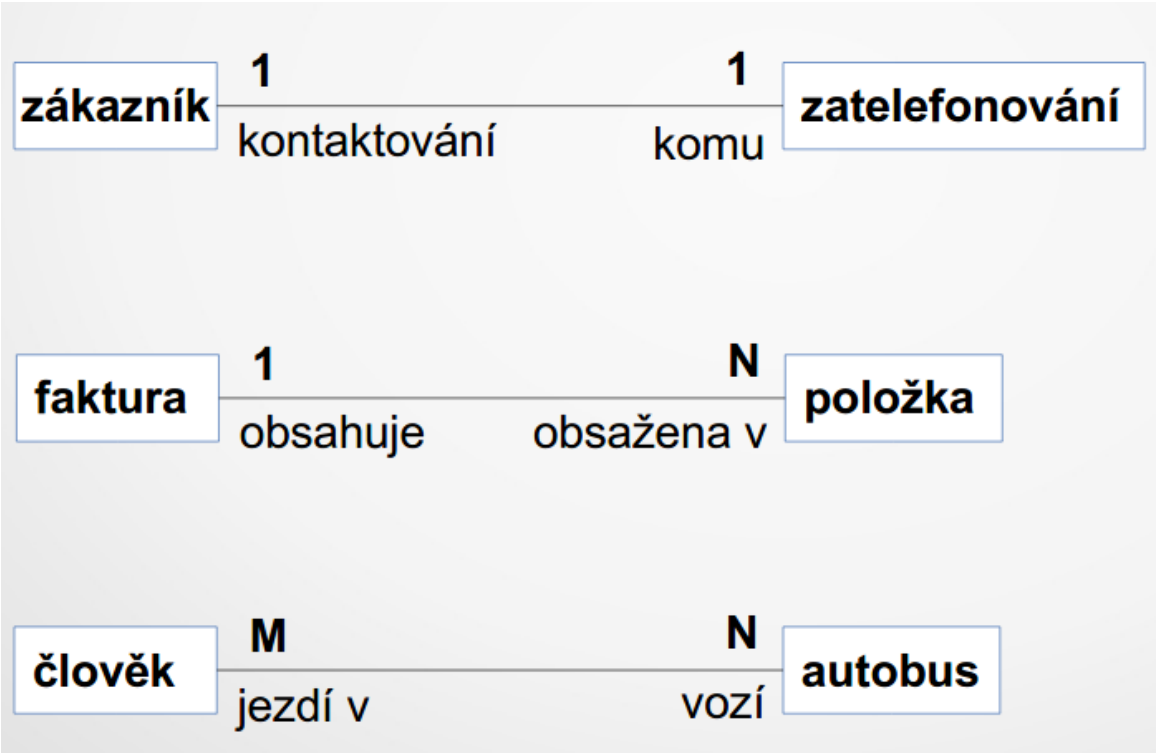
- 1:1 - na 1 položku z tabulky A připadá maximálně 1 položka z tabulky B, ta patří jen té jediné z A, osoba – DNA profil v národní databázi
- 1:N, N:1 - na 1 položku z tabulky A připadá libovolné množství položek z tabulky B, na 1 položku z tabulky B připadá maximálně 1 položka z tabulky A
- M:N - bez omezení množství vazeb v obou směrech, analýza – látka (jedna analýza najde více látek, látka se objeví v mnoha analýzách)

Kardinalita vztahu



Kardinalita vztahu

pomůže lepšímu porozumění ER(A) modelu



Kardinalita 1:N

– jedné entitě odpovídá několik (N) jiných entit
– např. vztah případ – vzorek

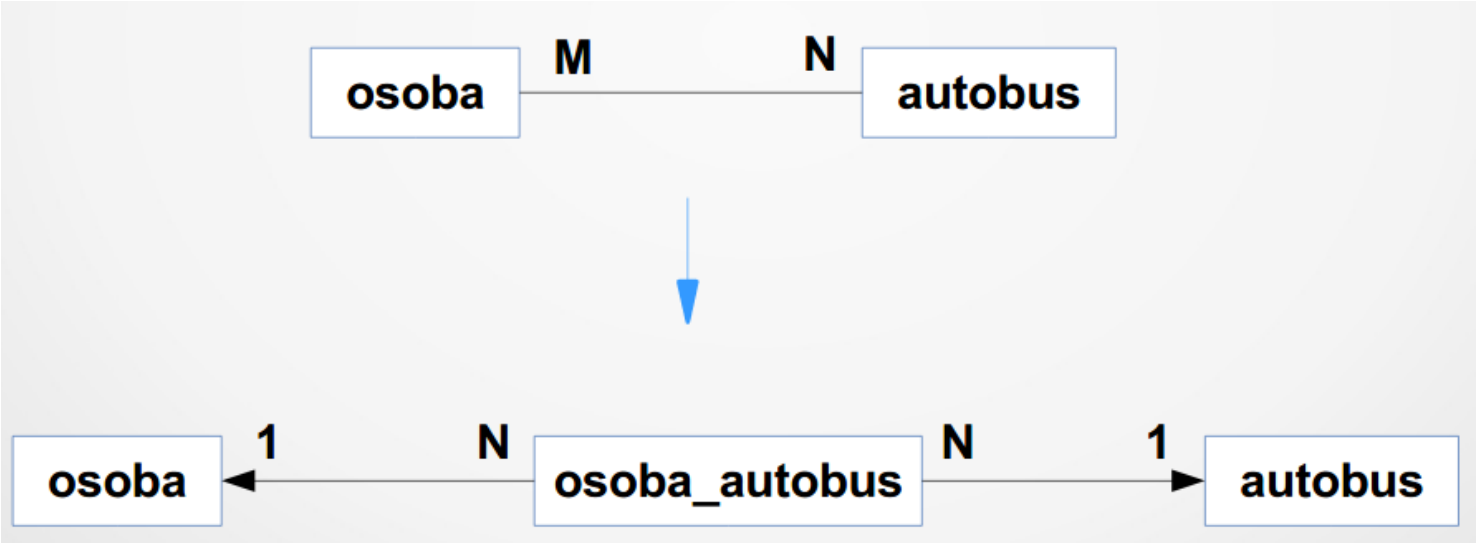
- každý vzorek patří k jednomu případu
- jeden případ má více vzorků

Kardinalita M:N

– několika (M) entitám odpovídá několik (N) jiných entit
– např. vztah analytik – přístroj

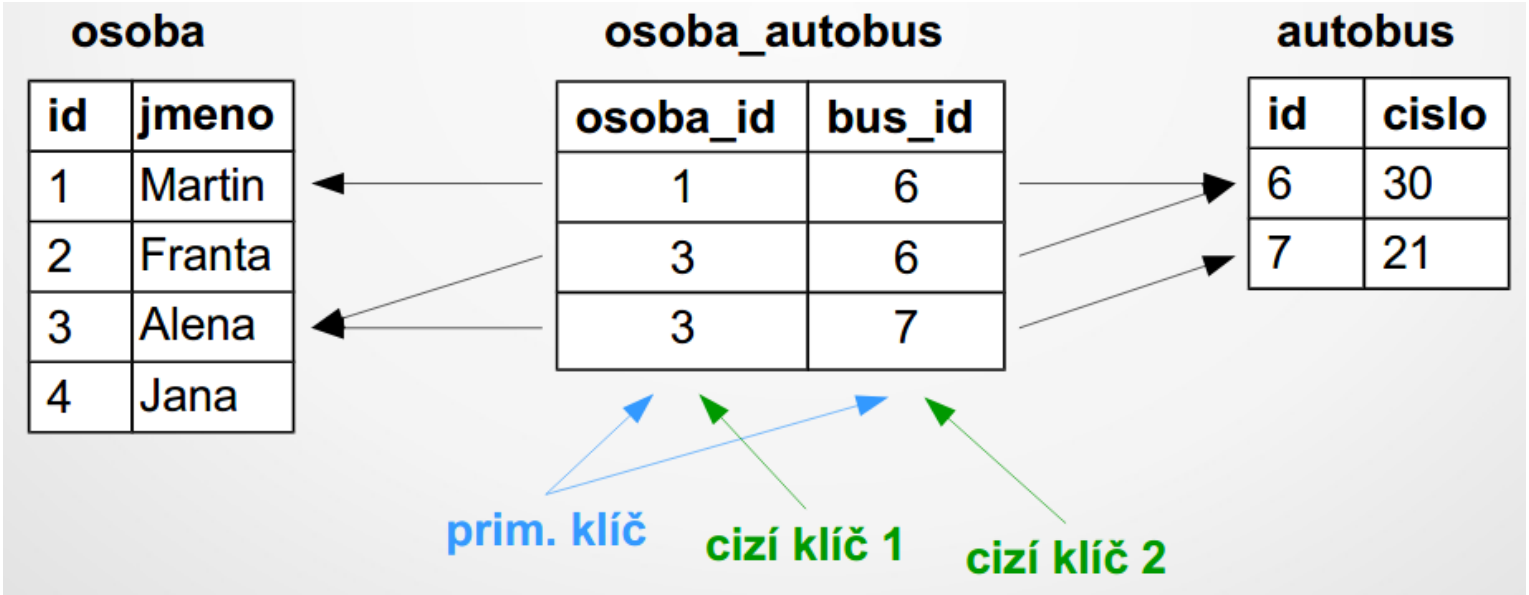
- jeden analytik měří na více přístrojích
  - na jednom přístroji měří více analytiků
  - nejsložitější případ vazby

vazbu M:N nelze vytvořit přímo mezi tabulkami
je potřeba rozložit pomocí nové tabulky na dvě vazby 1:N



Kardinalita M:N

rozkladová tabulka často neobsahuje další informace
primární klíč v rozkladové tabulce se často skládá z obou cizích klíčů (ale může být i vlastní sloupec id)



Kardinalita M:N

spojovací tabulka může obsahovat další informace
– např. v kolik kdo jakým autobusem jezdí
– není nutné kvůli tomu vytvářet další tabulku

osoba

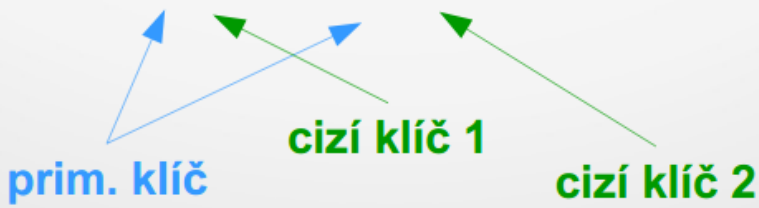
id	jmeno
1	Martin
2	Franta
3	Alena
4	Jana

osoba_autobus

osoba_id	bus_id	cas
1	6	13:30
3	6	7:15
3	7	20:00

autobus

id	cislo
6	30
7	21



M:N v chemii: sumární vzorec je spojovací tabulka

Molekula obsahuje více prvků, prvek je v mnoha molekulách → M:N → rozložíme na dvě vazby 1:N

Molekula	
id_molekula	nazev
1	ethanol
2	kofein

1 N

Složení (spojovací)		
id_molekula	*prvek*	pocet
1	C	2
1	H	6
1	O	1
2	C	8
2	H	10
2	N	4
2	O	2

N 1

Prvek	
prvek	A_r
C	12,011
H	1,008
N	14,007
O	15,999

Ethanol C_2H_6O = tři řádky v tabulce Složení; kofein $C_8H_{10}N_4O_2$ = čtyři řádky.
Jak z tabulek spočítáte molární hmotnost ethanolu?
JOIN a součet počet $\cdot A_r$: $2 \cdot 12{,}011 + 6 \cdot 1{,}008 + 1 \cdot 15{,}999 = 46{,}07$ g/mol.
Neukládá se, počítá se.

Příklad: analytici, přístroje a deník přístroje

Analytici měří na přístrojích (GC-MS, HPLC, ICP-MS) a každé měření zapíšou do deníku přístroje (datum, metoda, poznámka).
Doplňte entity, vazby a kardinality.

- Analytik **1:N** Záznam v deníku **N:1** Přístroj
- mezi analytikem a přístrojem je tedy **M:N**; deník je spojovací tabulka s vlastními atributy

Deník přístroje vyžaduje i akreditace laboratoře: kdo, kdy a na čem měřil.

Kardinalita ve forenzní laboratoři

Určete kardinalitu vztahu:

- případ – vzorek → **1:N**
- vzorek – analýza → **1:N** (vzorek se může analyzovat opakovaně)
- analýza – nalezená látka → **M:N**, spojovací tabulka nese koncentraci
- osoba – případ → **M:N**, spojovací tabulka nese roli (podezřelý, poškozený, svědek)
- osoba – DNA profil v národní databázi → **1:1** (ideálně; hledání duplicit je samo forenzní úloha)
- vzorek – podvzorek (alikvot) → **1:N rekurzivně**: podvzorek je také vzorek

Forenzní M:N: řetězec úschovy (chain of custody)

Vzorek prochází rukama mnoha lidí a každý člověk drží mnoho vzorků → **M:N**.
Spojovací tabulka **Predani**: id_vzorek (FK), od_osoby (FK), komu (FK), cas, ucel, podpis

Proč se z tabulky Predani nikdy nemaže a jak poznáte, že ji nikdo potají nezměnil?

- Mazání by přerušilo řetězec úschovy → důkaz je u soudu nepoužitelný
- Změny se jen připisují (audit log: kdo, kdy, co změnil), záznamy lze zabezpečit otiskem (hash, minulá lekce)
- Požadavky na záznamy a sledovatelnost klade i norma ISO/IEC 17025 pro akreditované laboratoře

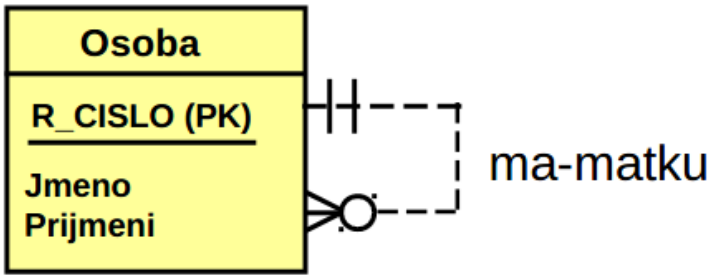
Společná práce

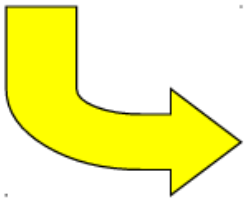
Do vašeho ERA modelu pro databázi DNA vzorků přidejte primární klíče, cizí klíče.
Osoba (ID osoby),
Vzorek DNA (ID vzorku, FK = ID osoby, FK = analytik),
Laboratoř (ID lab),
Analytik (ID analytik, FK = ID lab)
Analýza (ID analyza, FK = ID vzorku, FK = ID analytik),
Shoda (ID shoda, FK = ID vzorku).

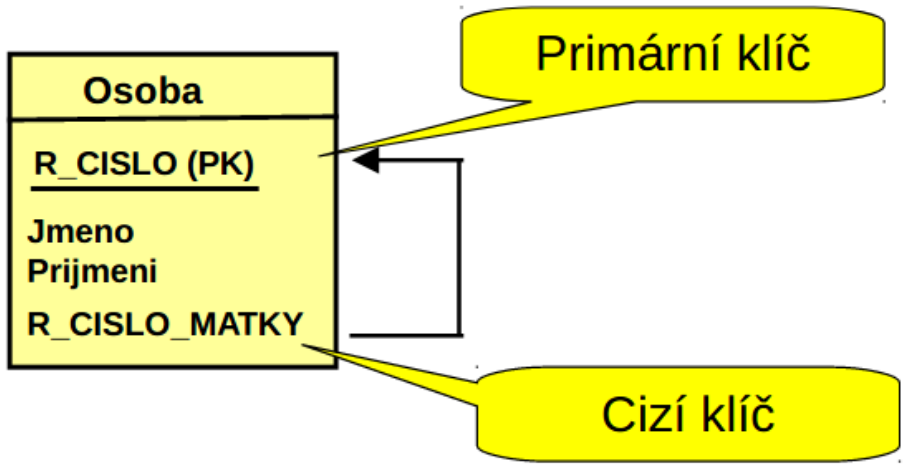
Co se stane, když smažeme osobu, která má přidělené vzorky?
Jak byste zajistili, aby každý vzorek měl přiřazenou laboratorní analýzu?
Co by se stalo, kdyby dva záznamy měly stejný primární klíč?

Rekurze

vztah mezi osobou a matkou dané osoby
Matka osoby je rovněž osobou, nemůže být tudíž v jiné tabulce!







Rekurze v toxikologii: metabolity

Metabolit je také látka → odkaz do **stejné** tabulky:

id_latka	nazev	id_materske_latky (FK)
1	heroin	
2	6-monoacetylmorfin	1
3	morfin	2
4	kokain	
5	benzoylekgonin	4

Proč morfin v krvi nemusí znamenat heroin?

Morfin může pocházet z léku nebo z makového semínka; heroin prokáže až specifický metabolit 6-monoacetylmorfin.

Normalizace

NORMALIZACE

Jak opravit Excel z úvodu?

Co je to normalizace

Normalizace je proces organizace dat v databázi tak, aby:

- se odstranila redundance (zbytečné opakování informací),
- byla zajištěna konzistence dat,
- a zjednodušila se údržba a rozšiřování databáze.

Cílem je, aby každý fakt byl uložen jen jednou a vztahy mezi entitami byly jasně definované.

Proč je to důležité:

- zabraňuje anomáliím při vkládání, mazání a aktualizaci dat
- umožňuje jednoznačné vazby mezi entitami
- dělá databázi přehlednější a efektivnější
- usnadňuje propojování tabulek a tvorbu dotazů

Anomálie ve forenzní databázi

Proč nestačí jedna velká tabulka (jako Excel z úvodu):

- aktualizační: analytička se vdá a změní příjmení → opravit ve 300 řádcích; jeden zapomenu → v datech jsou „dva lidé“
- vkládací: novou látku z varování EMCDDA nelze uložit, dokud ji nenajdeme v nějakém vzorku
- mazací: smažu jediný vzorek s danou látkou → zmizí i její název a CAS

Normalizace = každý fakt uložit právě jednou.

Společná práce: znormalizujte Excel z úvodu

Případ	Datum	Vzorek	Nalezené látky	Množství	Analytik
KR-12/26, Praha 4	3.2.2026	bílý prášek	kokain, levamisol	0,52 g	Novák J.
KR-12/26, Praha 4	2026-02-03	bílý prášek 2	kokain	480 mg	J. Novák
KR-15/26, Brno	5. 2. 26	tableta	MDMA	120	Dvořáková
KR-15/26, Brno	5. 2. 26	tableta	MDMA	120	Dvořáková
KR-17/26, Praha	7.2.2026	rostlina	THC 12 %	3,1 g	

1. Které entity v tabulce vidíte? (nápověda: je jich 5 až 6)
2. Navrhněte tabulky, primární a cizí klíče a kardinality vztahů.
3. Jaký datový typ dostane CAS, hmotnost, datum a koncentrace?
4. Ve které tabulce bude „kokain, levamisol“ a jak?

1. normální forma: v každé buňce jedna hodnota

Relace je v 1NF, pokud každý atribut obsahuje jen atomické hodnoty.

id_pripad	misto	id_vzorek	datum	hmotnost_g	latka	jmeno	prijmeni
KR-12/26	Praha 4	V1	2026-02-03	0,52	kokain	Jan	Novák
KR-12/26	Praha 4	V1	2026-02-03	0,52	levamisol	Jan	Novák
KR-12/26	Praha 4	V2	2026-02-03	0,48	kokain	Jan	Novák
KR-15/26	Brno	V3	2026-02-05	0,12	MDMA	Eva	Dvořáková
KR-17/26	Praha 2	V4	2026-02-07	3,10	THC	Petr	Svoboda

- „kokain, levamisol“ $\rightarrow$ dva řádky; případ a místo, jméno a příjmení zvlášť
- jeden formát data, hmotnost vždy v gramech, duplicitní řádek pryč

Co se v tabulce pořád opakuje?

2. normální forma: vše závisí na celém klíči

Klíč tabulky z 1NF je dvojice (id_vzorek, latka). Datum a hmotnost ale závisí jen na id_vzorek $\rightarrow$ patří do tabulky **Vzorek**.

id_vzorek	id_pripad	misto	datum	hmotnost_g	jmeno	prijmeni
V1	KR-12/26	Praha 4	2026-02-03	0,52	Jan	Novák
V2	KR-12/26	Praha 4	2026-02-03	0,48	Jan	Novák
V3	KR-15/26	Brno	2026-02-05	0,12	Eva	Dvořáková
V4	KR-17/26	Praha 2	2026-02-07	3,10	Petr	Svoboda

id_vzorek	latka
V1	kokain
V1	levamisol
V2	kokain
V3	MDMA
V4	THC

Co se opakuje teď?

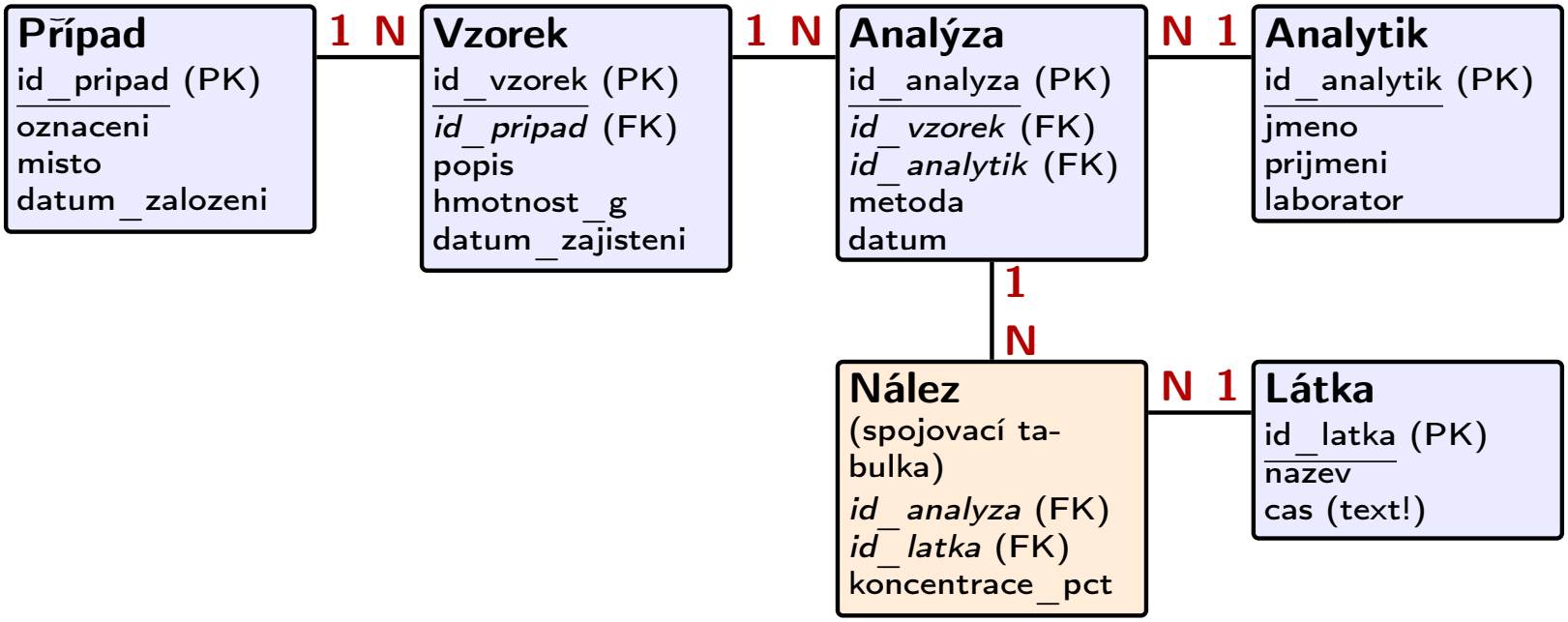
3. normální forma: žádné závislosti přes prostředníka

- místo závisí na případu, ne na vzorku → tabulka **Případ**
- jméno a příjmení závisí na analytikovi → tabulka **Analytik**
- název a CAS závisí na látce → tabulka **Látka**

```
Pripad   (id_pripad, misto)
Vzorek   (id_vzorek, id_pripad -> Pripad, datum, hmotnost_g, id_analytik -> Analytik)
Analytik (id_analytik, jmeno, prijmeni)
Latka    (id_latka, nazev, cas)
Nalez    (id_vzorek -> Vzorek, id_latka -> Latka, koncentrace_pct)
```

Vzorek se může analyzovat opakovaně → přidáme ještě tabulku **Analýza** (další snímek).

Možné řešení: ER model forenzní laboratoře



- „kokain, levamisol“ = **dva řádky** v tabulce Nález (M:N mezi analýzou a látkou)
- hmotnost vždy v gramech (hmotnost_g), datum jako DATE, CAS jako text

Kam byste přidali tabulku Predani (řetězec úschovy)?

Pravidla normalizace

I. pravidlo: Jedinečnost polí. Každé pole v tabulce by mělo představovat jedinečný typ informace.
II. pravidlo: Primární klíče. Každá tabulka musí mít jednoznačný identifikátor neboli primární klíče, který je vytvořen z jednoho nebo více polí v této tabulce.
III. pravidlo: Funkcionální závislost. Pro každou jedinečnou hodnotu primárního klíče se musí hodnoty v datových sloupcích týkat předmětu tabulky a musí tento předmět úplně popisovat.
IV. pravidlo: Nezávislost polí. Pokud se provede změna dat v libovolném poli (ne v primárním klíči), pak nebudou ovlivněna data v jakémkoliv jiném poli.

Co si odnést

- Tabulka = entita, sloupec = atribut, řádek = záznam s jednoznačným **primárním klíčem**
- Vztahy přes **cizí klíče**; kardinalita 1:1, 1:N, M:N (M:N vždy přes spojovací tabulku)
- **ERA model** se kreslí dřív, než vznikne databáze
- **Normalizace**: atomické hodnoty, každý fakt jen jednou, žádné tranzitivní závislosti
- Ve forenzní praxi: důkazy se nemažou (Restrict), řetězec úschovy je M:N s časem a podpisem
- Dobře navržená databáze = čistý dataset pro AI; klíče brání úniku dat mezi učením a testem

Otázky k zamyšlení

1. Proč nesmí být CAS číslo uloženo jako číslo?
2. Jaká je kardinalita vztahu „látka – spektrum ve spektrální knihovně“?
3. Proč je jedna obří tabulka pro strojové učení výhodná, ale pro ukládání dat nebezpečná?
4. Co se stane s modelem, když polovina hmotností je v mg a polovina v g?
5. Navrhněte, jak v databázi uchovat informaci, že výsledek analýzy byl později opraven.

Domácí úkol: ERA model vaší bakalářky

Příprava na závěrečnou konferenci:

1. Jaké entity vznikají ve vaší práci (vzorek, měření, látka, přístroj, ...)?
2. Nakreslete ERA model s primárními a cizími klíči a kardinalitami.
3. Ověřte, že je v 3NF: je každý fakt uložen jen jednou?

Skripta, stránky a materiály

https://www.fi.muni.cz/~popel/lectures/bak_logika/
https://www.siliconrepublic.com/science/logic-maths-puzzle-george-boole
http://www.di.unito.it/~anselma/pdf/preprintBoole.pdf
Simultaneous Equations and Boolean Algebra in the Analysis of Judicial Decisions
Fred Kort Law and Contemporary Problems Vol. 28, No. 1, Jurimetrics (Winter, 1963), pp. 143-163 (21 pages)
https://en.wikipedia.org/wiki/Truth_table
https://sites.ff.cuni.cz/uisk/wp-content/uploads/sites/62/2016/01/Datov%C3%A9-modely-a-n%C3%A1vrhy-rela%C4%8Dn%C3%ADch-sch%C3%A9mat_Sk%C5%99ivan.pdf
http://www.databaze.chytrak.cz/modely.htm
http://books.fs.vsb.cz/dbacc20/Welcome.htm
http://books.fs.vsb.cz/
http://projekty.fs.vsb.cz/463/edubase/VY_01_042/Z%C3%A1klady%20informatiky.pdf
http://projekty.fs.vsb.cz/463/edubase/VY_01_044/Datab%C3%A1zov%C3%A9%20syst%
learning/Datab%C3%A1zov%C3%A9%20syst%C3%A9my%2005.%20Normalizace%20dat.p
https://bytebytego.com/guides/top-6-database-models/