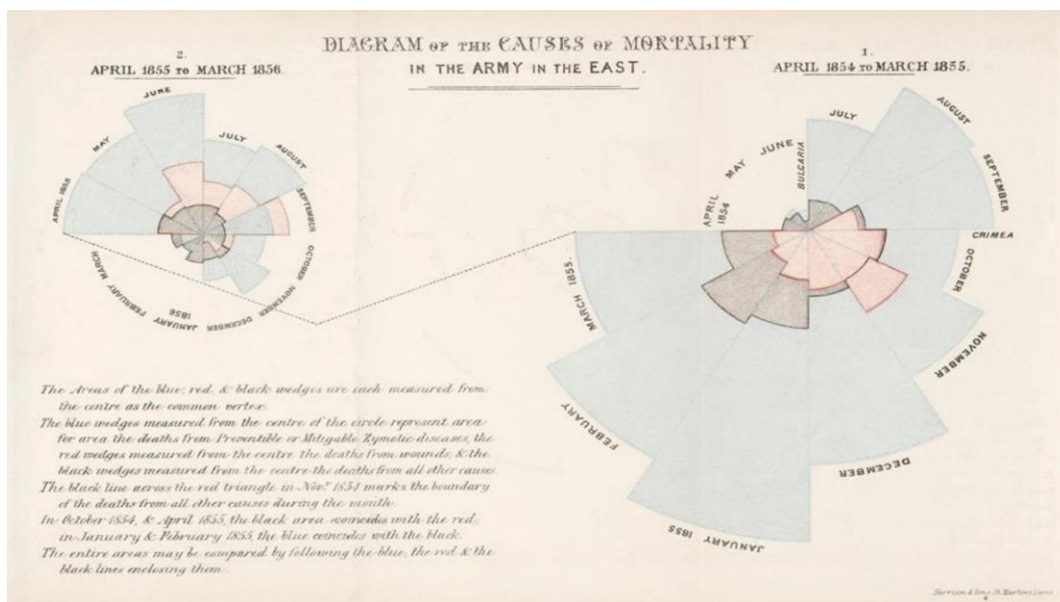


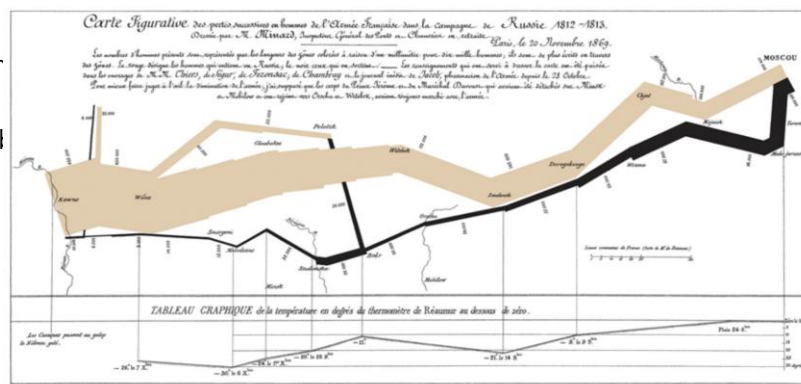
## Diagram of the Causes of Mortality in the Army in the East

- vytvořila Florence Nightingale během Krymské války (1853–1856)
- zjistila, že drtivá většina úmrtí nebyla způsobena zraněními z boje, ale infekcemi
- tzv. „coxcomb diagram“ (někdy se mu říká polární plošný graf)
- jednotlivé segmenty (měsíce) mají barevné klíny:
  - modrá: úmrtí na preventabilní nemoci (tyfus, cholera atd.)
  - červená: úmrtí na zranění
  - černá: úmrtí na jiné příčiny
- čím větší plocha klínu, tím více lidí zemřelo
- názorná ukázka toho, že více než 2/3 úmrtí mohly být eliminované
- graf považován za jeden z prvních příkladů „evidence-based policymaking“



## Charlese Minard Graph

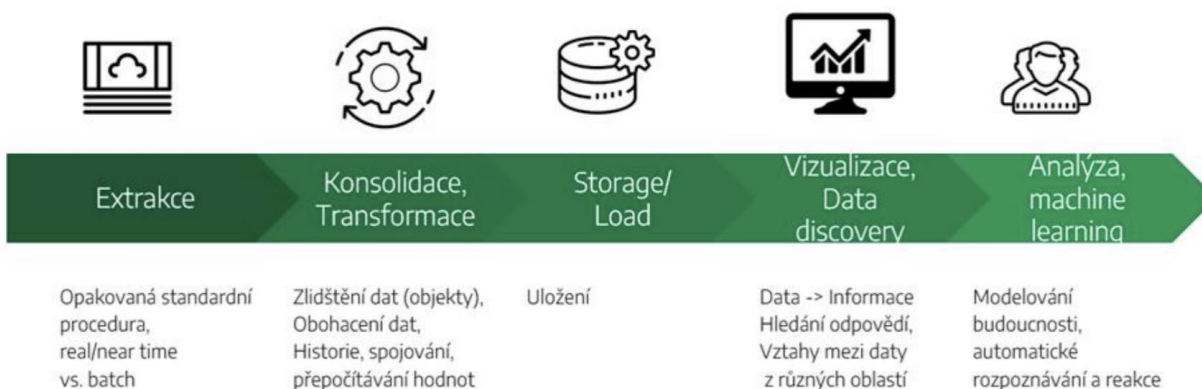
- Napoleonova armáda vyrazila s více než 400 000 vojáků, ale po katastrofálním ústupu se domů vrátilo jen kolem 10 000 (tažení je béžové, návrat černý)
- graf je považován za nejlepší všech dob; kombinuje více dimenzí dat v jednom obraze:
  - čas
  - geografii
  - velikost armády
  - teplotu
  - směr pohybu



## nejčastější chyby v datové analýze

- chybné implikace a kauzality
- chybná interpretace dat
- zbytečně složité řešení
- chybně postavené grafy a vizualizace
- chybná vstupní data nebo předpoklady
- chybný koncept (špatně definovaná myšlenka projektu)

## životní cyklus dat



- formáty přenosu dat (ve fázi Storage/Load) jsou CSV, JSON a XML

## Entity Relationship Model

- datové modelování = návrh datového modelu ve fázi konsolidace/transformace
- postup vytváření
  - vymezit informace, které budou uchovány
  - určit vlastnosti, které jsou důležité a dostupné
  - přemýšlet o tom, jaké jsou mezi objekty vztahy
- **entita** reprezentuje určitou skupinu objektů reálného světa
  - je popsána množinou atributů (např. zákazník banky)
- **atribut** je vlastností entity která nás v kontextu problému zajímá
  - každý atribut má definovaný **datový typ** (např. datum narození, jméno)

<b>Integer</b> (#whole number)	47	16b
<b>Float</b> (#decimal)	47,25	32b
<b>Date &amp; Time</b>	YYYY-MM-DD hh:mm:ss	
<b>String</b> (#character)	'Mr. Mon' / "Mr. Mon"	8b
<b>Boolean</b> (#binary)	True (pravda, 1) nebo False (nepravda, 0)	1b

## DATOVÉ TYPY

- **Číselné**
  - tinyint [0,255]
  - smallint [-32768,32767]
  - int [-2<sup>31</sup>, 2<sup>31</sup> - 1]
  - bigint [-moc, +moc]
  - integer
  - numeric
- **Řetězcové**
  - char, varchar
  - nchar, nvarchar
  - text

- **relace** je vztah mezi několika entitami (např. zákazníkův vlastní účet)
- **kardinalita** je max. počet vztahů daného typu, ve kterých může být jedna entita
  - jeden zákazník může mít několik účtů 1:N
  - jeden zákazník může mít pouze jeden účet 1:1
- **klíč** je atribut, jehož hodnota jednoznačně identifikuje řádek v relační tabulce
- **primární klíč (PK)** jednoznačně identifikuje řádek
- **cizí klíč (FK)** je původní PK použitý v dalším výskytu k vyjádření vazeb

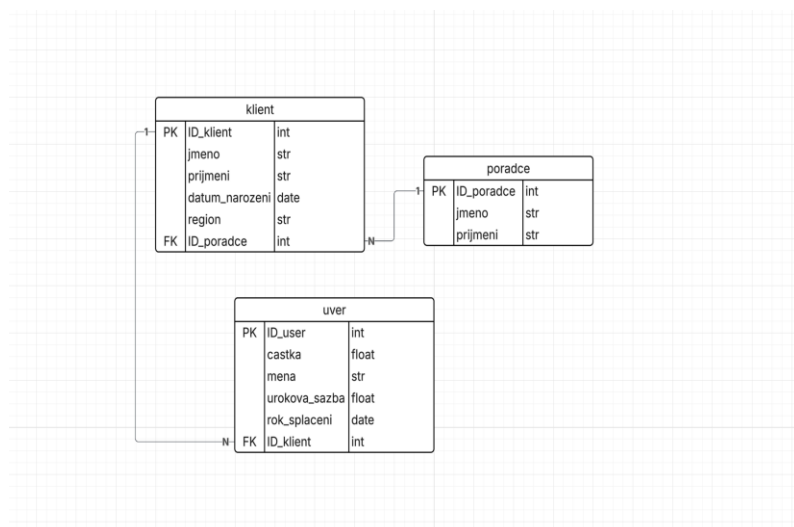
### Jedinečného identifikátor (id)

### Metrika (hodnotový údaj)

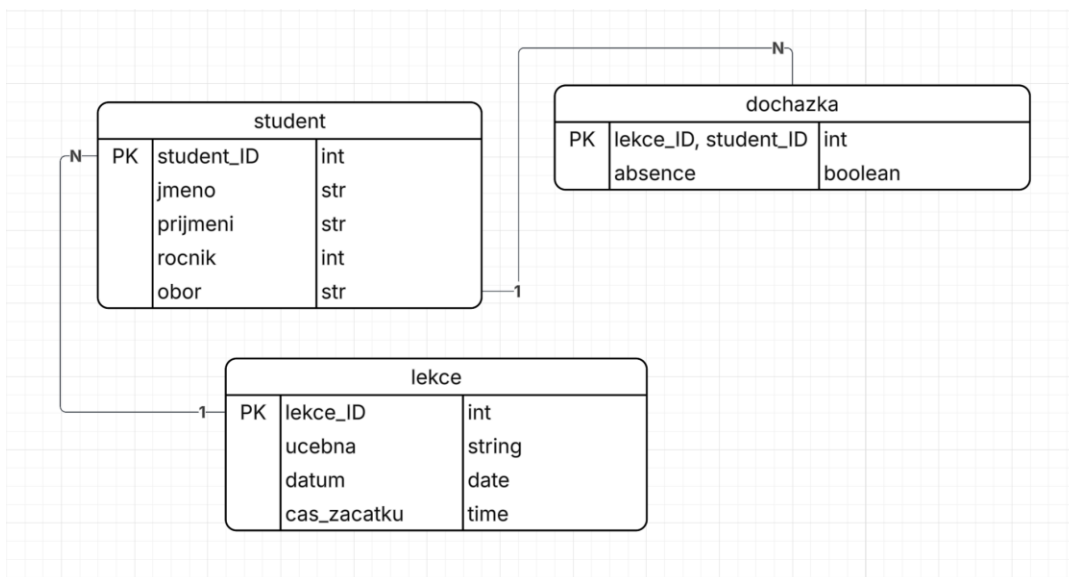
id_klienta	klient	segment	oblast	rok	zisk	obrat	naklady
1	Automyčka s.r.o.	MSE	Praha	2018	50	100	50
2	Spol a.s.	TOP MSE	Brno	2018	70	150	80
3	Pekárna s.r.o.	MSE	Praha	2018	40	90	50
4	Truhlářství s.r.o.	MSE	Brno	2018	50	70	20
5	Group a.s.	TOP MSE	Praha	2018	60	130	70

### Dimenze (popisný údaj)

příklad datového modelu



můj první datový model

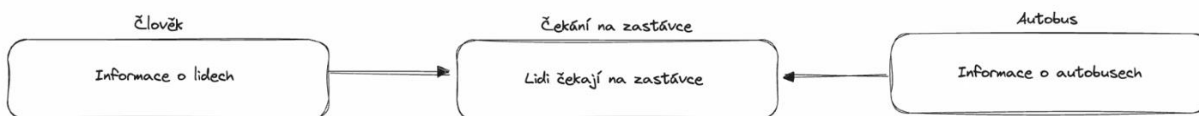


relační model dat

- většinou u vazební tabulky (např. klient/uver) máme N
- typy datových modelů
  - koncepční (jak to přibližně bude)
  - logický (co všechno entity obsahují a jaké mají mezi sebou vztahy)
  - fyzický (to, co máme v Lucidcart)

vývoj modelu

- **konceptuální model** má největší míru abstrakce (jen vypíšu, co potřebuju)
- výběr konceptů, volba rozlišení (jak moc jít do detailu)
  - z existujících informací vybíráme to důležité



- **logický model** identifikuje vlastnosti (převod koncept entita), vazby
- zde rozhodujeme o typu databáze (např. objektová vs. relační)
- **fyzický model** zahrnuje konkrétní názvy entit

datový model

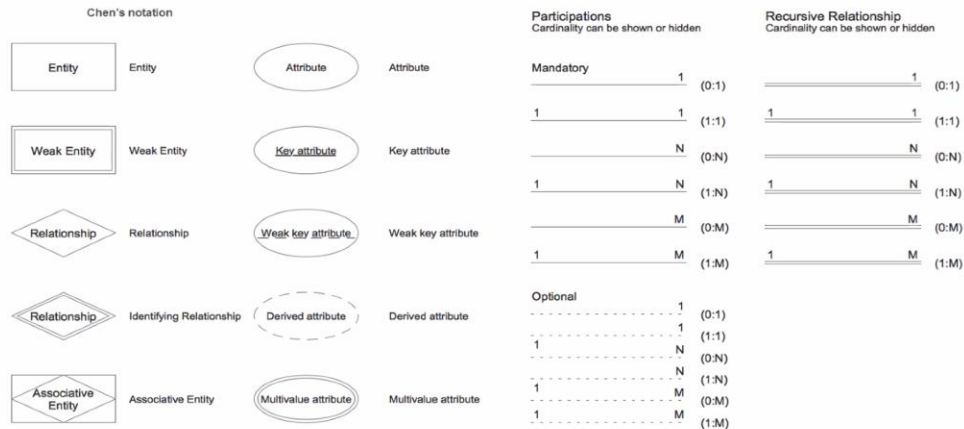
- vazební tabulka

zápis datového modelu

- OMT (objektově orientované modelování)
- IDEF (modelování procesů)
- Bachman (používá šipky mezi entitami)

- Martin notace (nejpoužívanější pro ERD, 1:N, M:N, apod.)
- UML notace (např. v business analýze)
- Crow`s foot notace (vrání nožka místo čísel/písmenek)
- Chen`s notace (používání oválů a kosočtverců)

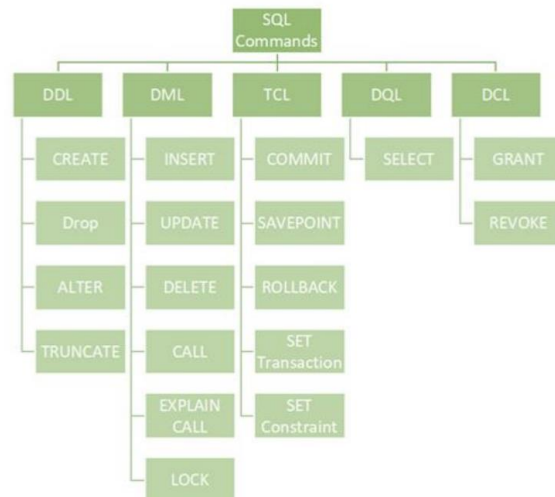
## Chen's notation



## SQL

- structured query language
- standardizovaný strukturovaný dotazovací jazyk
- nejobvyklejší typy MS SQL, MySQL, Oracle, PostgreSQL, no SQL, SQLite

## druhy příkazů



- DQL (Data Query Language) dotazování na data
- DDL (Data Definition Language) vytváření a změna objektů v databázi
- DML (DATA Manipulation Language) manipulace s daty (vkládání, mazání, změna)
- TCL (Transaction Control Language) práce s transakcemi
- DCL (Data Control Language) řízení přístupu k datům