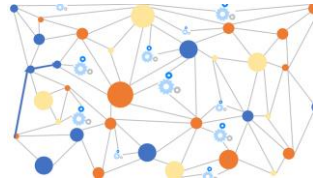
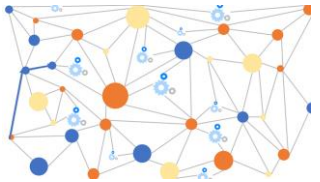
# DATA ENGINEERING - MINTAKA



- **Data Preprocessing**
  **Mintaka dataset – Movies**
  **Splits**
    - **Train**
    - **Validation**
    - **Test**
- Structure of Datapoints
- Data Preparation
- Extracting Entities, Predicates and Triples

| Dataset \ Category | Movies |
|---|---|
| Train | 700 |
| Validation | 107 |
| Test | 198 |

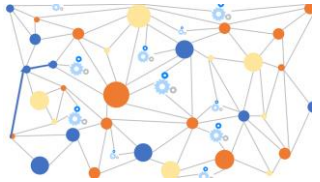
# DATA ENGINEERING - MINTAKA



- Data Preprocessing - Mintaka dataset - Movies
  - Splits
    - Train
    - Validation
    - Test
- **Structure of Datapoints**
- Data Preparation
- Extracting Entities, Predicates and Triples


```
{
  "id":"2723bb1b",
  "question":"Which actor was the star of Titanic and was born in Los Angeles, California?",
  "translations":{
    "ar":"مَنْ الممثل الذي لعب دور البطولة في فيلم \"تيتانيك\" وهو من مواليد لوس أنجلوس بكاليفورنيا؟",
    "de":"Welcher Schauspieler war der Star von Titanic und wurde in Los Angeles, Kalifornien, geboren?",
    "ja":"タイタニックのスターで、カリフォルニア州ロサンゼルス生まれの俳優は誰ですか?",
    "hi":"कौन से अभिनेता Titanic के स्टार थे और लॉस एंजेलिस, कैलिफ़ोर्निया में पैदा हुआ थे?",
    "pt":"Qual ator foi a estrela de Titanic e nasceu em Los Angeles, Califórnia?",
    "es":"¿Qué actor protagonizó Titanic y nació en Los Ángeles, California?",
    "it":"Quale attore è stato protagonista in Titanic ed è nato a Los Angeles in California?",
    "fr":"Quel acteur a joué le rôle principal de Titanic et est né à Los Angeles en Californie ?"
  },
  "questionEntity":[
    {
      "name":"Q44578",
      "entityType":"entity",
      "label":"Titanic",
      "mention":"Titanic",
      "span":[
        28,
        35
      ]
    }
```

# DATA ENGINEERING - MINTAKA



- Data Preprocessing - Mintaka dataset – Movies Splits
  - Train
  - Validation
  - Test
- Structure of Datapoints
- **Data preparation**
- Extracting Entities, Predicates, and Triples

- Example:

**Question**: Which actor was the star of Titanic and was born in Los Angeles, California?

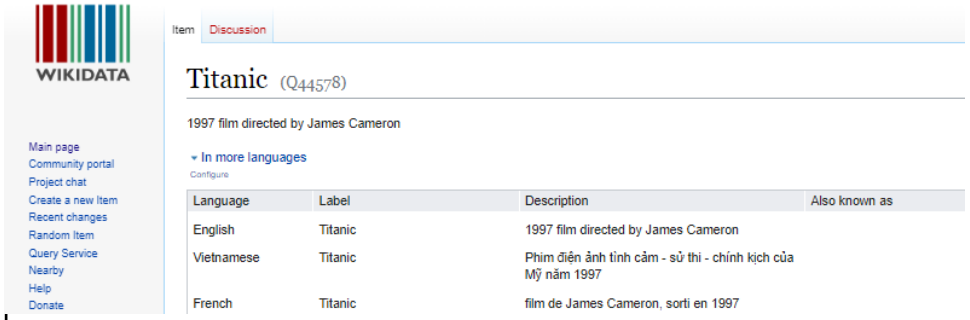
**Answer**: Leonardo DiCaprio

Entities in Question and Answer :

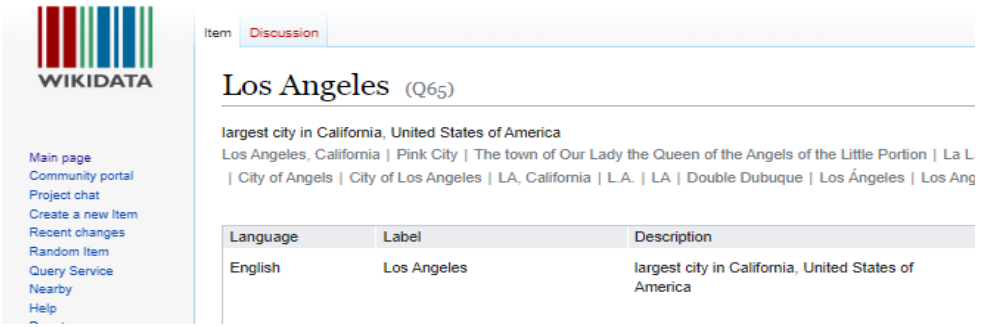
Titanic - **Q44578**
Los Angeles - **Q65**
Leonardo DiCaprio - **Q38111**

# DATA ENGINEERING - MINTAKA

- Data Preprocessing Mintaka dataset – Movies Splits
  - Train
  - Validation
  - Test
- Structure of Datapoints
- **Data preparation**
  - **Wikidata Entities**
- Extracting Entities, Predicates, and Triples




Titanic - **Q44578**

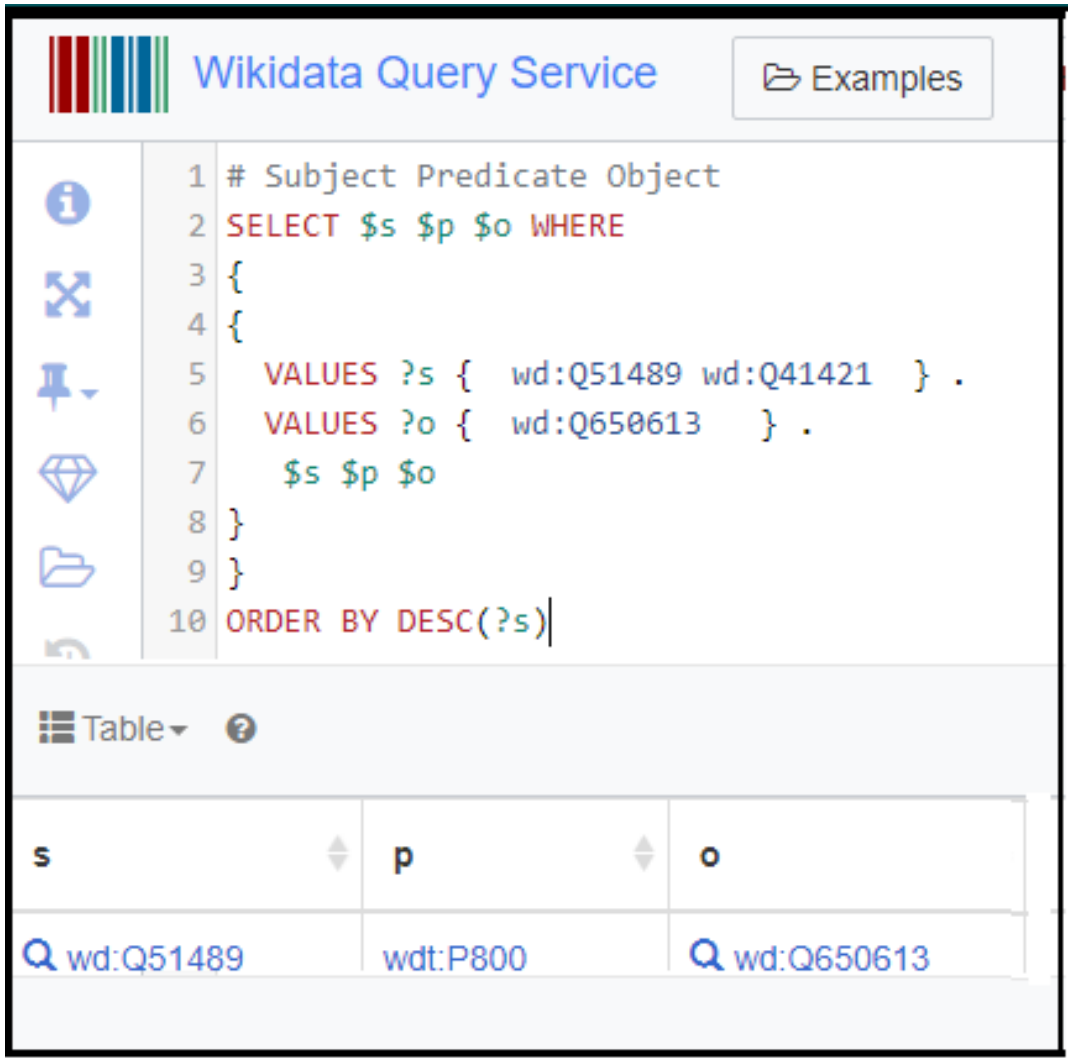



Los Angeles - **Q65**

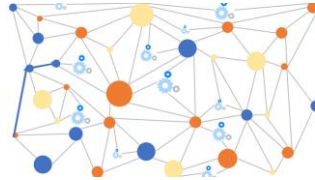

Leonardo DiCaprio - **Q38111**

# DATA ENGINEERING - MINTAKA

- Data Preprocessing - Mintaka dataset – Movies Splits
  - Train
  - Validation
  - Test
- Structure of Datapoints
- **Data preparation**
  - **Wikidata Entities**
  - **SPARQL Queries**
- Extracting Entities, Predicates and Triples



# DATA ENGINEERING - MINTAKA



- Data preprocessing - Mintaka dataset - Movies
  Splits
    - Train
    - Validation
    - Test
- Structure of Datapoints
- Data preparation
    - Wikidata Entries
    - SPARQL Queries
- **Extracting Entities, Predicates and Triples**

- Extract all the entities from Question and Answers from the trimmed dataset

Example : Q51489 , Q656013

- Extract all the Relations or Predicates between the Question QID and the Answer QID

Example: P800

- Form the triples as <S , P , O >

Example : **< Q51489 , P800 , Q656013 >**

# DATA ENGINEERING - MINTAKA

- Data preprocessing - Mintaka dataset - Movies
  Splits
    - Train
    - Validation
    - Test

- Structure of Datapoints

- Data Engineering
    - Wikidata Entries
    - SPARQL Queries

- **Extracting Entities, Predicates and Triples**

| No of Entities | 10118 |
|---|---|
| No of Triples | 28027 |
| No of Predicates | 323 |

**Entity table**

| Entity | Name | Alias |
|---|---|---|
| Q189489 | Zendaya | Zendaya Coleman, Zendaya Maree Stoermer Coleman |
| Q80596 | Arthur Miller | Arthur Asher Miller |
| Q165219 | Robert Down | RDJ, Robert Downey, Robert John Downey Jr. |
| Q23894967 | Tony Stark | Iron Man, Anthony Edward Stark |
| Q172 | Toronto | Toronto, Canada, Hogtown, The Six, City of Toronto, Toronto, Ontario, T-O, The 416, Toronto, ON |
| Q189490 | Jennifer Lawr | J Law, Jennifer Shrader Lawrence, JLaw |

**Property table**

| Property | Name | Alias |
|---|---|---|
| P106 | occupation | work, profession, job, craft, vocation, avocation, career, employ, employment |
| P199 | business division | divisions, has business division, has division |
| P85 | anthem | national anthem, march, team anthem, club anthem, official hymn, official song |
| P47 | shares border with | border, next to, adjacent to, bordered by, borders clockwise |
| P69 | educated at | faculty, education, student of, alma mater, attended, alumna of, alumni of, alumnus of, attended school at, |
| P57 | director | film director, movie director, directed by |
| P19 | place of birth | birth location, birth place, birthplace, location of birth, POB, birth city, born, born at, born in, location born |
| P135 | movement | art movement, artistic movement, school, philosophical movement, literary movement, music scene, artistic school, trend, scientific movement |

**Triplet table**

| Subject | Relation | Object |
|---|---|---|
| Q189489 | P800 | Q27985819 |
| Q188 | P17 | Q39 |
| Q270599 | P162 | Q223992 |
| Q30 | P530 | Q242 |
| Q159846 | P355 | Q7135302 |
| Q6219699 | P3373 | Q51506 |
| Q2266587 | P179 | Q642878 |