Forecasting Patient Enrolment for Clinical Trials

2nd Team Project Sprint Review
April 17th, 2020

01

MongoDB

How we set up the MongoDB and the clinical trials collection

02

Country Data

Additional data to enrich the country attribute

03

Mesh Data

Data we scraped to enhance the medical conditions

04

Hospital data

Data we scraped from a hospital db and the Places API

05

Next Steps

Plans on how to proceed in this project

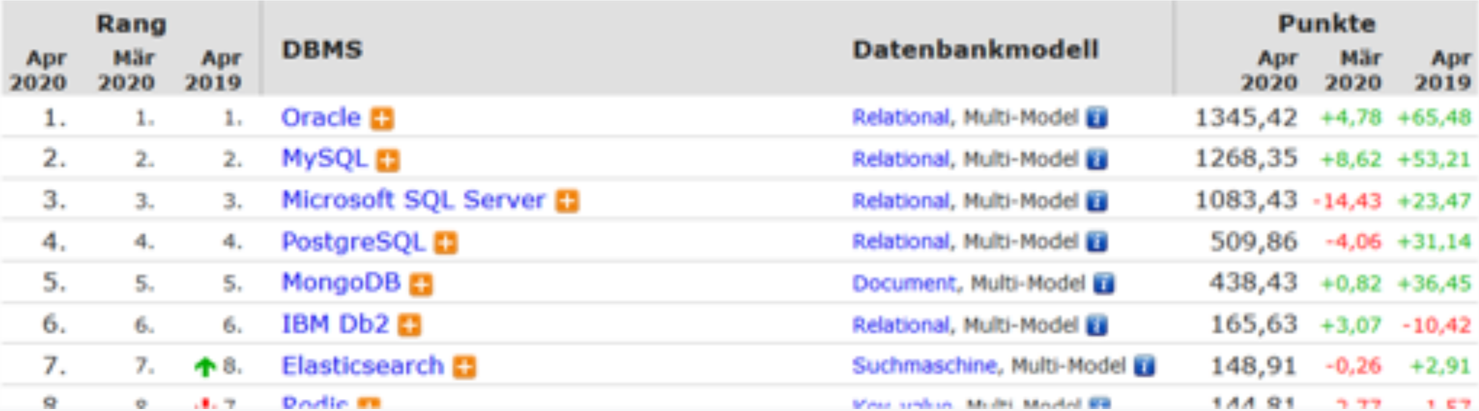
MongoDB

- Why MongoDB?
- Our Approach
- Idea

Why MongoDB?

Some Advantages:

- flexible schema
- scalability & performance
- object-oriented
- documents look like JSON objects

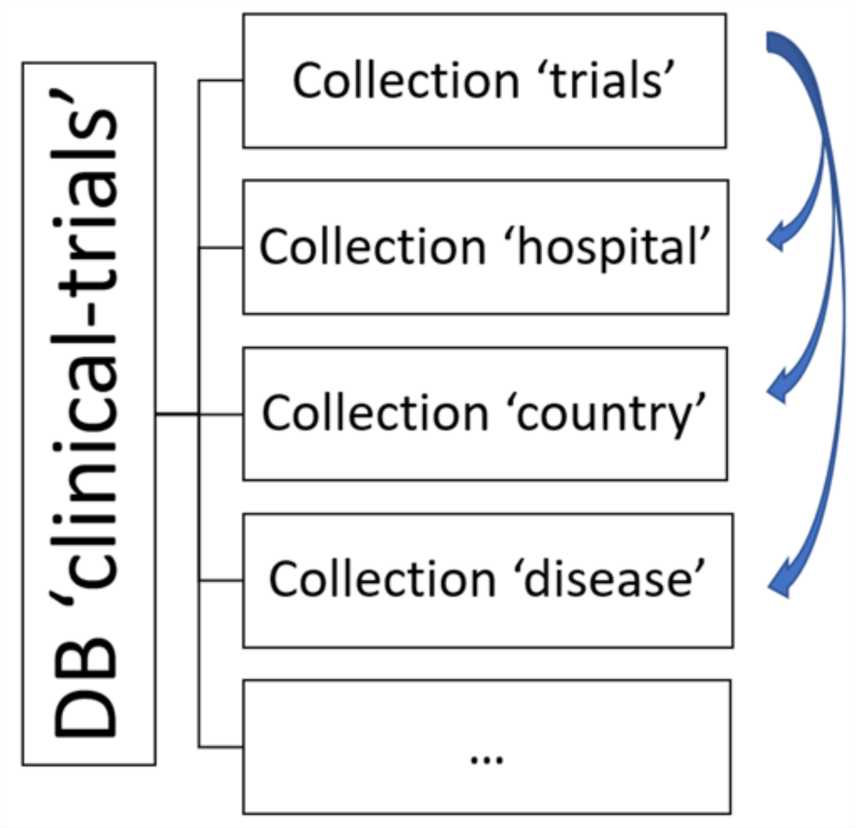


Rang			DBMS	Datenbankmodell	Punkte		
Apr 2020	Mär 2020	Apr 2019			Apr 2020	Mär 2020	Apr 2019
1.	1.	1.	Oracle	Relational, Multi-Model	1345,42	+4,78	+65,48
2.	2.	2.	MySQL	Relational, Multi-Model	1268,35	+8,62	+53,21
3.	3.	3.	Microsoft SQL Server	Relational, Multi-Model	1083,43	-14,43	+23,47
4.	4.	4.	PostgreSQL	Relational, Multi-Model	509,86	-4,06	+31,14
5.	5.	5.	MongoDB	Document, Multi-Model	438,43	+0,82	+36,45
6.	6.	6.	IBM Db2	Relational, Multi-Model	165,63	+3,07	-10,42
7.	7.	↑ 8.	Elasticsearch	Suchmaschine, Multi-Model	148,91	-0,26	+2,91
8.	8.	↓ 7.	Redis	Key-value, Multi-Model	144,81	-2,77	-1,57

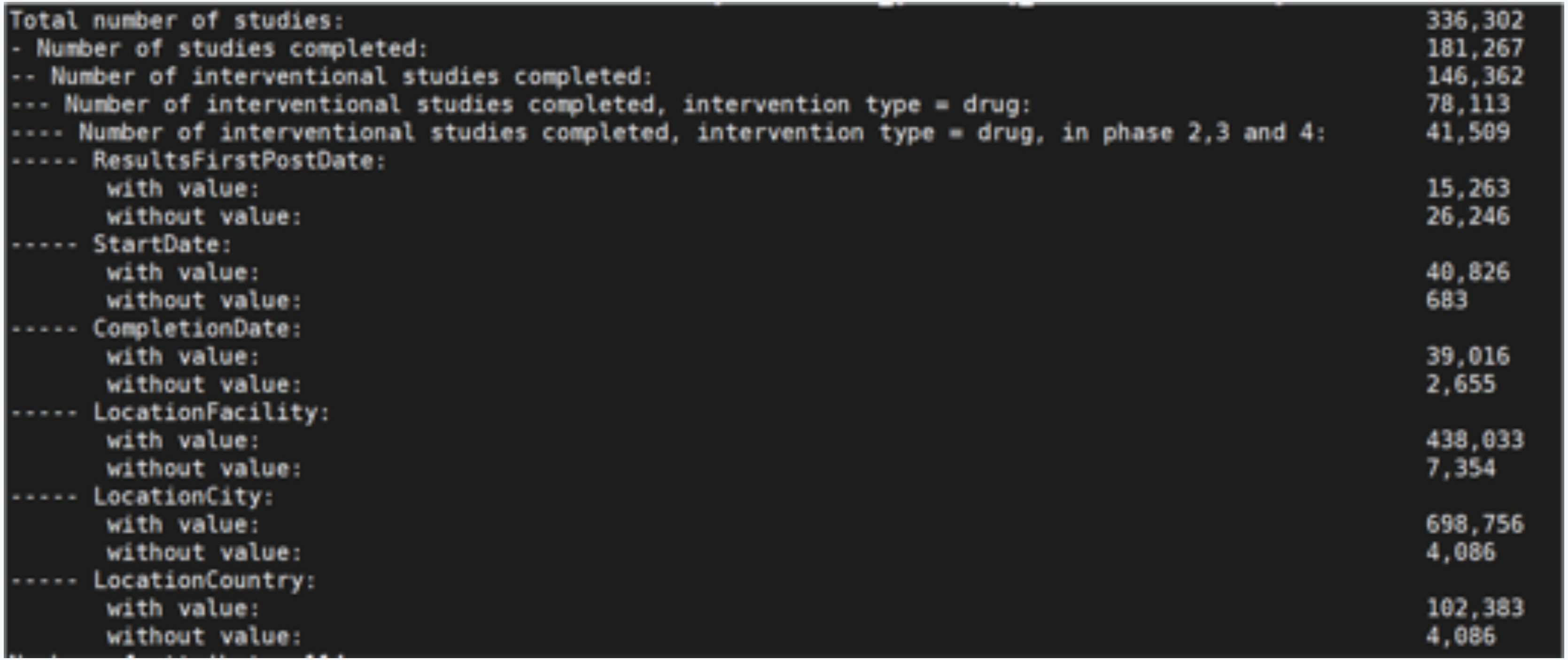
Our Approach

1. Training: Fundamentals, Queries, Filters, Aggregations
2. Installation and Configuration of MongoDB on our server
3. Adaptation of the script to save the trials from clinicaltrials.org to MongoDB
4. Test of the script → Data Consistency
5. Execution (still running):
   - Status: ~ 27.000 studies (out of ~ 41.000)
   - 114 Attributes → 6 requests for each study

Idea


DB ‘clinical-trials’
Collection ‘trials’
Collection ‘hospital’
Collection ‘country’
Collection ‘disease’
...

Insights



```
Total number of studies:                                                                           336,302
- Number of studies completed:                                                                     181,267
-- Number of interventional studies completed:                                                     146,362
--- Number of interventional studies completed, intervention type = drug:                          78,113
---- Number of interventional studies completed, intervention type = drug, in phase 2,3 and 4:     41,509
----- ResultsFirstPostDate:
       with value:                                                                                 15,263
       without value:                                                                              26,246
----- StartDate:
       with value:                                                                                 40,826
       without value:                                                                              683
----- CompletionDate:
       with value:                                                                                 39,016
       without value:                                                                              2,655
----- LocationFacility:
       with value:                                                                                 438,033
       without value:                                                                              7,354
----- LocationCity:
       with value:                                                                                 698,756
       without value:                                                                              4,086
----- LocationCountry:
       with value:                                                                                 102,383
       without value:                                                                              4,086
```

Insights (current status)

Reliable Attributes (<10% missing)	
NCTId OrgFullName OrgClass BriefTitle BriefSummary StudyType OverallStatus Phase StatusVerifiedDate Condition LeadSponsorName LeadSponsorClass StdAge InterventionType InterventionName	EligibilityCriteria Gender DesignPrimaryPurpose HealthyVolunteers ConditionBrowseBranchAbbrev ConditionBrowseLeafName ConditionBrowseLeafRelevance OfficialTitle ConditionMeshId ConditionMeshTerm StartDate ConditionAncestorId ConditionAncestorTerm EnrollmentCount MinimumAge

Insights (current status)

Poor Attributes (>80% missing)		
BaselineMeasureDenomCountGroupId	BaselineMeasurementComment	FlowAchievementComment
BaselineMeasureDenomCountValue	BaselineMeasureCalculatePct	BaselineMeasurementLowerLimit
BaselineMeasureDenomUnits	DesignInterventionModelDescription	BaselineMeasurementUpperLimit
BaselineMeasureDenomUnitsSelected	BaselineClassDenomCountGroupId	FlowPreAssignmentDetails
BaselineTypeUnitsAnalyzed	BaselineClassDenomCountValue	FlowRecruitmentDetails
BioSpecDescription	BaselineMeasurePopulationDescription	EventsTimeFrame
DesignTimePerspective	BaselineClassDenomUnits	BaselineMeasurementSpread
TargetDuration	FlowMilestoneComment	FlowDropWithdrawType
StudyPopulation	AvailIPDType	FlowReasonGroupId
LocationStatus	AvailIPDURL	FlowReasonNumSubjects
FlowAchievementNumUnits	IsFDARegulatedDevice	BaselineClassTitle
FlowDropWithdrawComment	IsFDARegulatedDrug	
FlowReasonComment	BaselinePopulationDescription	
FlowReasonNumUnits	BaselineMeasureDescription	
FlowTypeUnitsAnalyzed	StartDateType	

Country Data

- An overview
- First source - implementation in Python and MongoDB
- Second source - attributes and merging possibilities
- Challenges and next steps

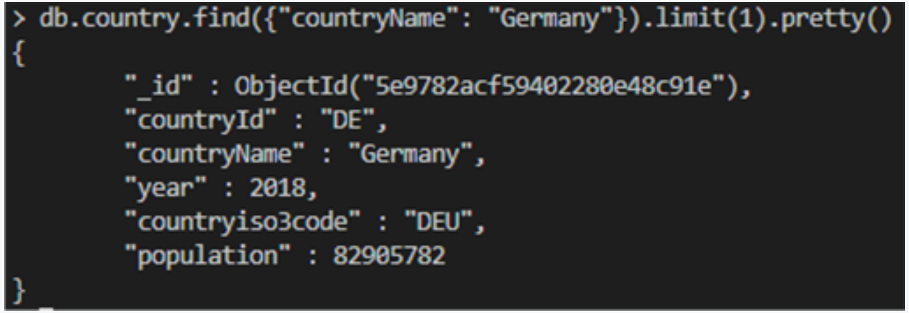
Data crawling and feature selection

- We are collecting data from multiple sources, focusing on overall information and health-related data (e.g. World Bank, UN, Wikidata, DBpedia...)

- Features that could affect to clinical trials: population (in total and in density), hospital beds, life expectancy, geographical location, GDP, human development index, median income...

Population data

- The first attribute about countries is population data which from World Bank.
- The total population is the number of all residents regardless of legal status or citizenship and are shown are mid-year estimates[1]
- Last updated on 09/04/2020
- Python and MongoDB is applied for scripting and storing data. The schema is designed for scaling easily



```
> db.country.find({"countryName": "Germany"}).limit(1).pretty()
{
        "_id" : ObjectId("5e9782acf59402280e48c91e"),
        "countryId" : "DE",
        "countryName" : "Germany",
        "year" : 2018,
        "countryiso3code" : "DEU",
        "population" : 82905782
}
```

Example of a document

[1]: https://data.worldbank.org/indicator/sp.pop.totl

Population data

- The second data source is the wikidata API
- Can be retrieved via SPARQL requests
- Every country can be retrieved and merged with existing data via Country Code
- Example of attributes: GDP, Average life expectancy, population, median Income, Age of majority ...
- Goal: Find “golden-record” of data by merging different sources

https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial/de

To-dos in next sprint(s)

- Merging data from other sources
- Evaluating and pre-processing

Condition Mesh Data

- An overview
- Implementation in Python and MongoDB

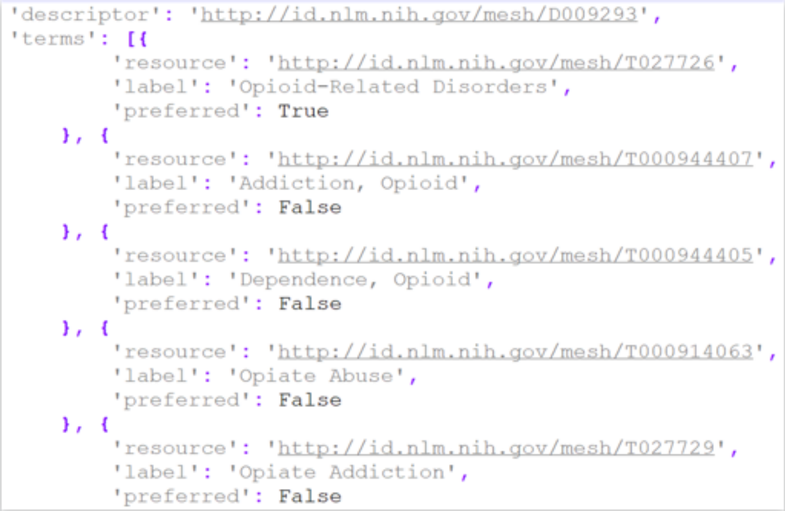
Mesh data overview

- MESH = Medical Subject Headings
- Crawl data from US National Library of Medicine
- Using URI data requests via public API http://id.nlm.nih.gov/ to retrieve related condition terms

https://www.nlm.nih.gov/

Implementation & Merge

- For each entry of Clinical Trials & for each “ConditionMeshId” create URI request
- Retrieve related terms & split into “preferred” and “non preferred”
- Enrich Clinical Trials data set



```
'descriptor': 'http://id.nlm.nih.gov/mesh/D009293',
'terms': [{
        'resource': 'http://id.nlm.nih.gov/mesh/T027726',
        'label': 'Opioid-Related Disorders',
        'preferred': True
    }, {
        'resource': 'http://id.nlm.nih.gov/mesh/T000944407',
        'label': 'Addiction, Opioid',
        'preferred': False
    }, {
        'resource': 'http://id.nlm.nih.gov/mesh/T000944405',
        'label': 'Dependence, Opioid',
        'preferred': False
    }, {
        'resource': 'http://id.nlm.nih.gov/mesh/T000914063',
        'label': 'Opiate Abuse',
        'preferred': False
    }, {
        'resource': 'http://id.nlm.nih.gov/mesh/T027729',
        'label': 'Opiate Addiction',
        'preferred': False
```

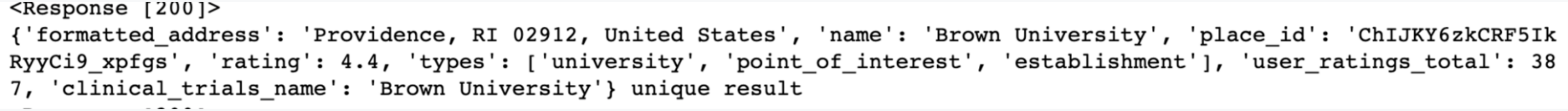
https://www.nlm.nih.gov/

Hospital Data

- The Google Places API
- Attributes of the Hospital Data from Google Places
- The "Ranking Web of World Hospitals“ Hospital data

Hospital Data

- We are currently using find place search, the data looks like this, in JSON format, we



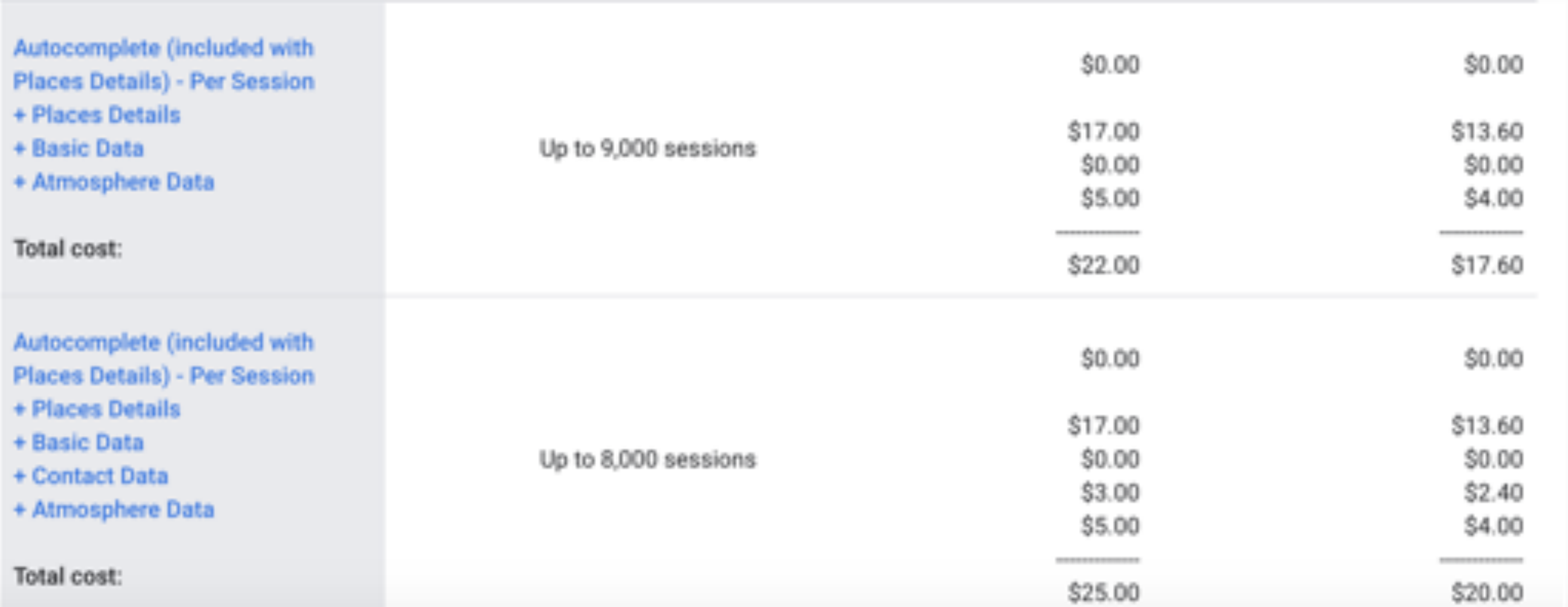
```
<Response [200]>
{'formatted_address': 'Providence, RI 02912, United States', 'name': 'Brown University', 'place_id': 'ChIJKY6zkCRF5Ik
RyyCi9_xpfgs', 'rating': 4.4, 'types': ['university', 'point_of_interest', 'establishment'], 'user_ratings_total': 38
7, 'clinical_trials_name': 'Brown University'} unique result
```

- Problems fixed

Problems	Solution
Cannot retrieve more than 60 data at a time	Change text search to find place search and add credit card to have $200 free amount
There were some irrelevant data and some data with more than 1 result	Adding country information makes the error significantly low
Duplicate data	If formatted_address exists, we compare zipcode, choose the one that has exact the same

Hospital Data

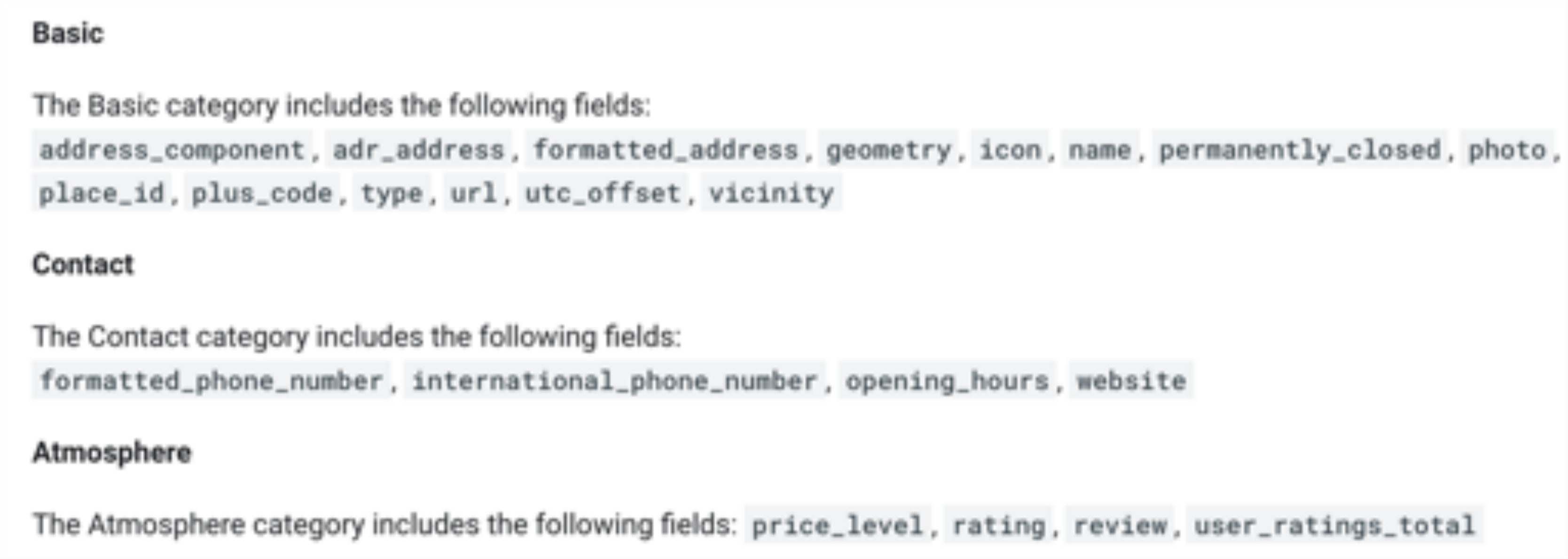
- For Google API, we can have free retrieval on Basic data. However, we want information about ratings, and the paying is listed following, and we have 200$ free quota per month, so we can have about 9000 requests per month, pricing listed here



Autocomplete (included with Places Details) - Per Session + Places Details + Basic Data + Atmosphere Data	Up to 9,000 sessions	$0.00 $17.00 $0.00 $5.00	$0.00 $13.60 $0.00 $4.00
Total cost:		$22.00	$17.60
Autocomplete (included with Places Details) - Per Session + Places Details + Basic Data + Contact Data + Atmosphere Data	Up to 8,000 sessions	$0.00 $17.00 $0.00 $3.00 $5.00	$0.00 $13.60 $0.00 $2.40 $4.00
Total cost:		$25.00	$20.00

Hospital Data

- To be discussed:
  Which attributes do we want to crawl?



Basic

The Basic category includes the following fields:
`address_component`, `adr_address`, `formatted_address`, `geometry`, `icon`, `name`, `permanently_closed`, `photo`, `place_id`, `plus_code`, `type`, `url`, `utc_offset`, `vicinity`

Contact

The Contact category includes the following fields:
`formatted_phone_number`, `international_phone_number`, `opening_hours`, `website`

Atmosphere

The Atmosphere category includes the following fields: `price_level`, `rating`, `review`, `user_ratings_total`

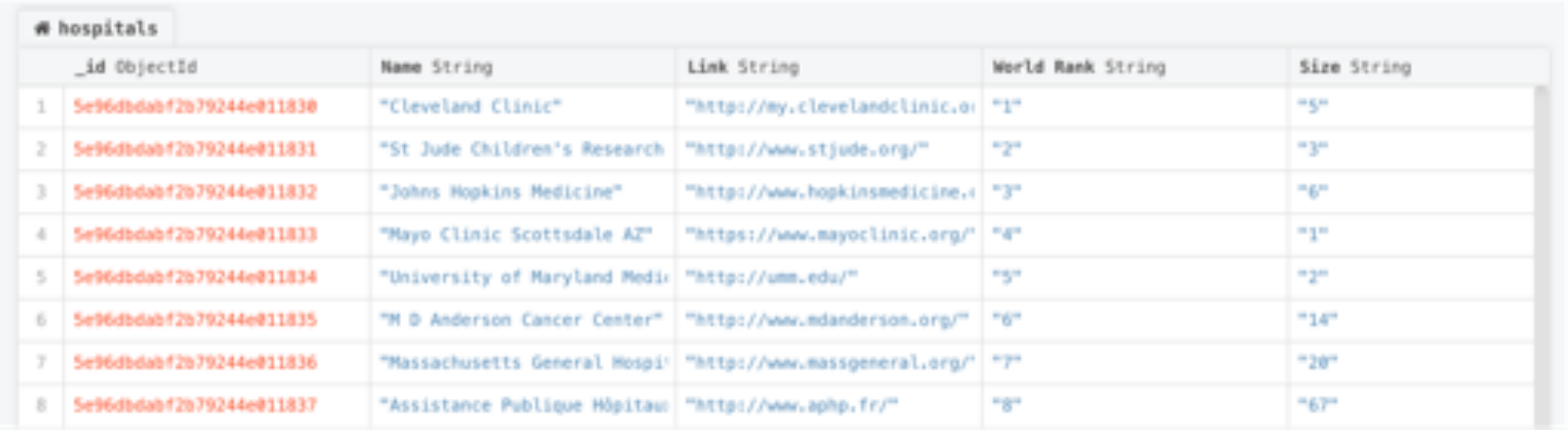
Hospital Data

- Scraped the Data of 12 000 hospitals all over the world
- Data was collected by a spanish research institute
- https://hospitals.webometrics.info/en/World

Name	Website	World Rank	Country	Size

- Primarily gives information about the research activities of a hospital
- Try to combine with the hospitals from the Google Places API via Name or Link

Hospital Data



	_id ObjectId	Name String	Link String	World Rank String	Size String
1	5e96dbdabf2b79244e011830	"Cleveland Clinic"	"http://my.clevelandclinic.o	"1"	"5"
2	5e96dbdabf2b79244e011831	"St Jude Children's Research	"http://www.stjude.org/"	"2"	"3"
3	5e96dbdabf2b79244e011832	"Johns Hopkins Medicine"	"http://www.hopkinsmedicine.	"3"	"6"
4	5e96dbdabf2b79244e011833	"Mayo Clinic Scottsdale AZ"	"https://www.mayoclinic.org/"	"4"	"1"
5	5e96dbdabf2b79244e011834	"University of Maryland Medi	"http://umm.edu/"	"5"	"2"
6	5e96dbdabf2b79244e011835	"M D Anderson Cancer Center"	"http://www.mdanderson.org/"	"6"	"14"
7	5e96dbdabf2b79244e011836	"Massachusetts General Hospi	"http://www.massgeneral.org/"	"7"	"20"
8	5e96dbdabf2b79244e011837	"Assistance Publique Hôpitau	"http://www.aphp.fr/"	"8"	"67"

Next Steps

Next Steps

- Possibly crawl data for disease burden from WHO
- Merging data from different sources → find suitable merging strategy
- Preprocessing of existing data
- Creation of additional features from existing ones (i.e. duration)
- Find alternative storage

Appendix