SOCRATex Project

Development of a Scalable Search Engine for Clinical Narrative Text in OMOP-CDM

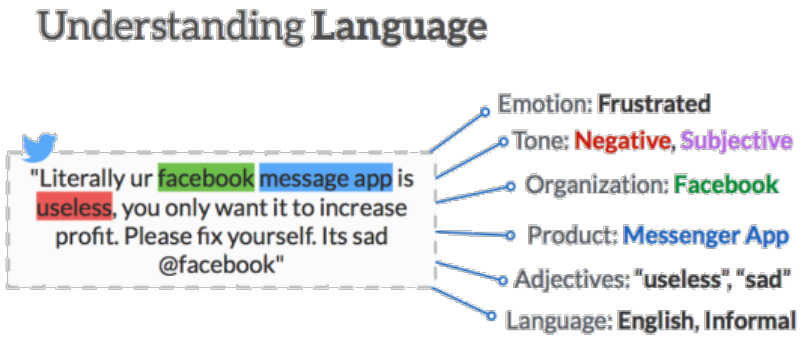
Jimyung Park[*], Seng Chan You M.D. M.S.[**,], Jin Roh M.D. Ph.D[†], Dongsu Park[**], Kwang Soo Jeong[**], Rae Woong Park M.D. Ph.D[*,**]

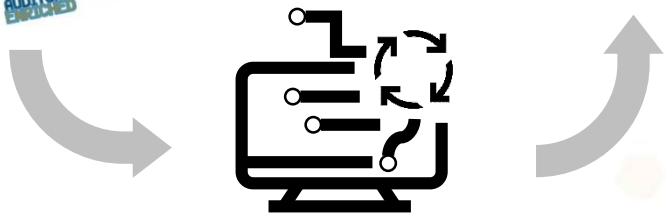
[*] Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon, 16499

[**] Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, 16499

[†]Dept. of Pathology, Ajou University Hospital, Yeongtong-gu, Suwon, 16499

Background





Natural Language Processing

- To unlock the information in narrative text, annotation and processing using Natural Language Processing (NLP) is necessary
- Recent NLP techniques use labeled and unlabeled data result from NLP tools

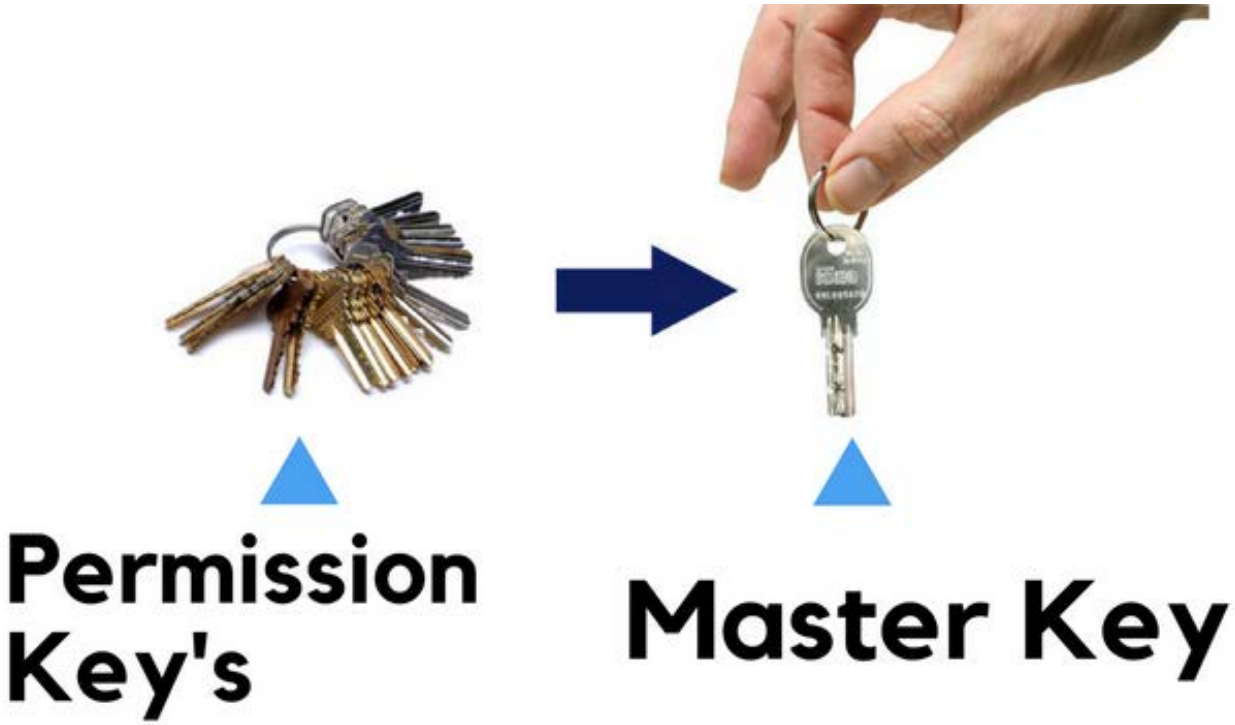
Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017;73:14-29.

Background

- Most of existing NLP tools are developed based on EHR system
- Since the EHR systems are diverse according to the institutions, reusing the NLP tools is challenging

Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77:34-49

Background

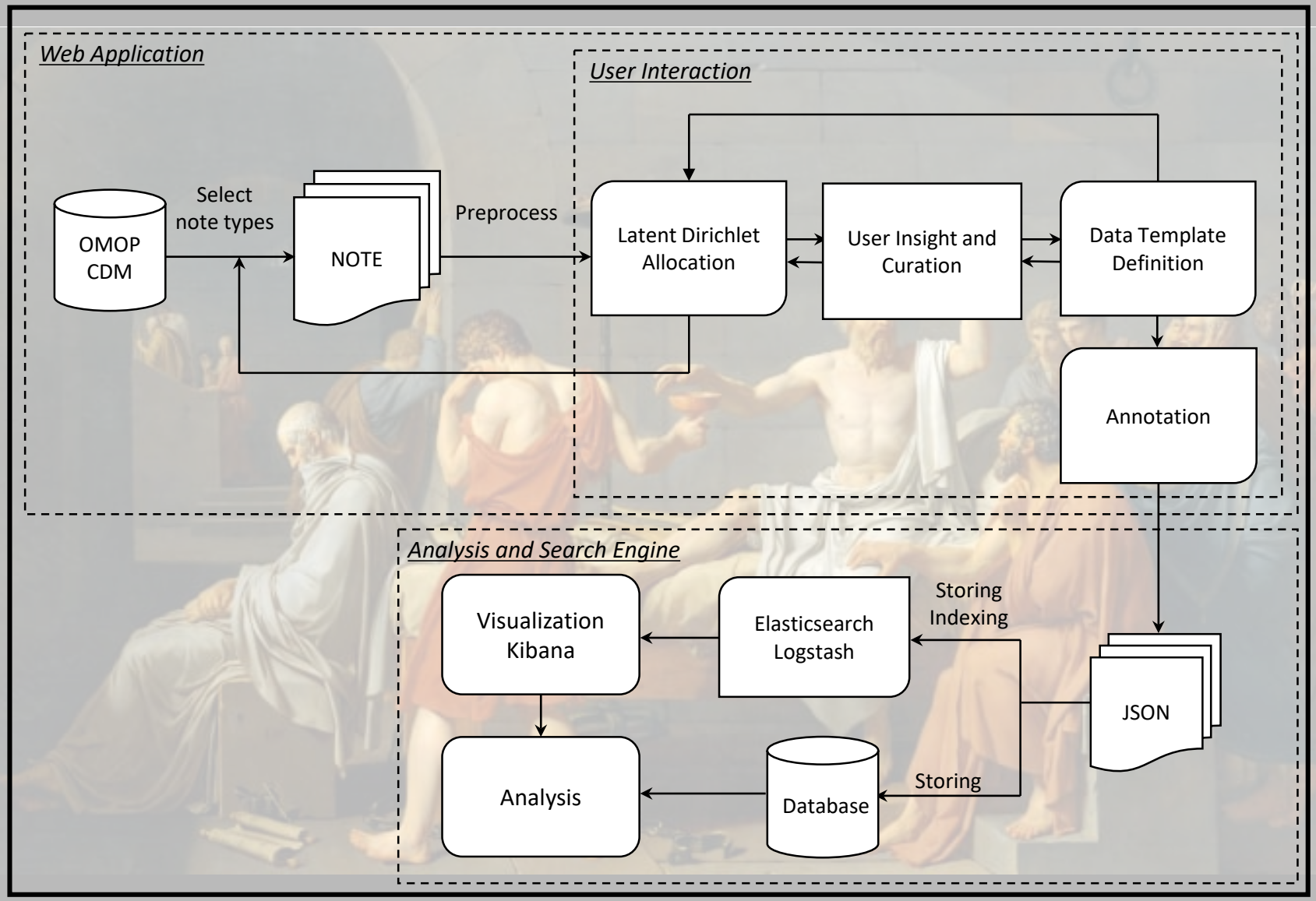


- ✓ Can a single process be applied to several databases?
- ✓ Is it possible to build one single data standard for narrative text?
- ✓ Can NLP tool generated labeled or unlabeled data be utilized for analysis?

- **SOCRATex**

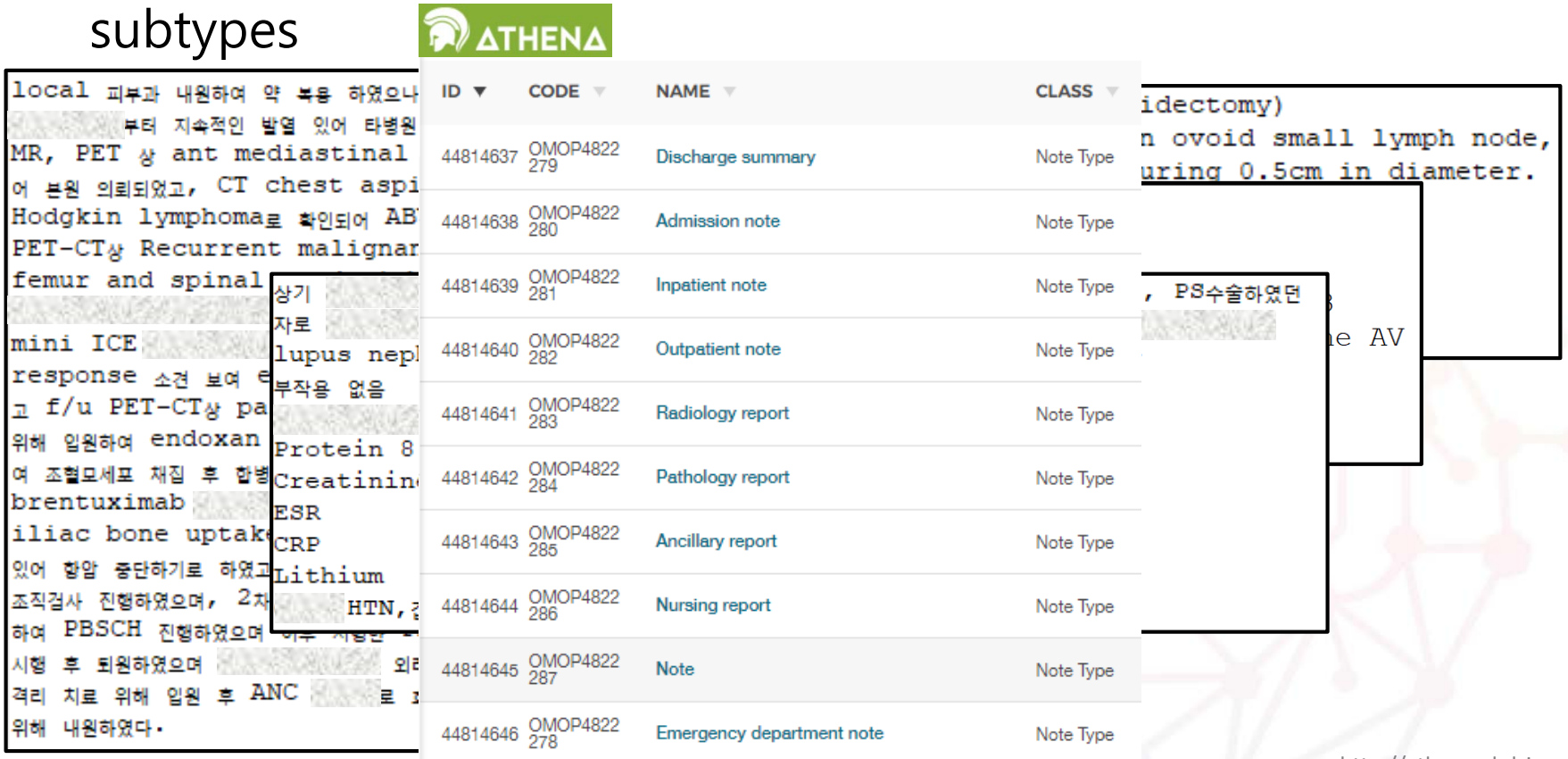
Staged **O**ptimization of **C**uration, **R**egularization, **A**nnotation of clinical **Tex**t

- A system of processing NOTE documents in OMOP-CDM
  - ✓ Topic model (Latent Dirichlet Allocation)
  - ✓ JSON schema and annotation
  - ✓ ELK stack (Elasticsearch, Logstash, Kibana)
  - ✓ Analysis using annotated JSON data
- **SOCRATex** aims to build a scalable and extensible NLP system based on OMOP-CDM


Web Application
User Interaction
OMOP CDM
Select note types
NOTE
Preprocess
Latent Dirichlet Allocation
User Insight and Curation
Data Template Definition
Annotation
Analysis and Search Engine
Visualization Kibana
Elasticsearch Logstash
Storing Indexing
JSON
Analysis
Database
Storing

Method

- NOTE contains diverse types of documents and contents
- ATHENA provides essential concept_ids for note type classification, yet subtypes are not provided
- Reviewing the whole reports is necessary to detect subtypes

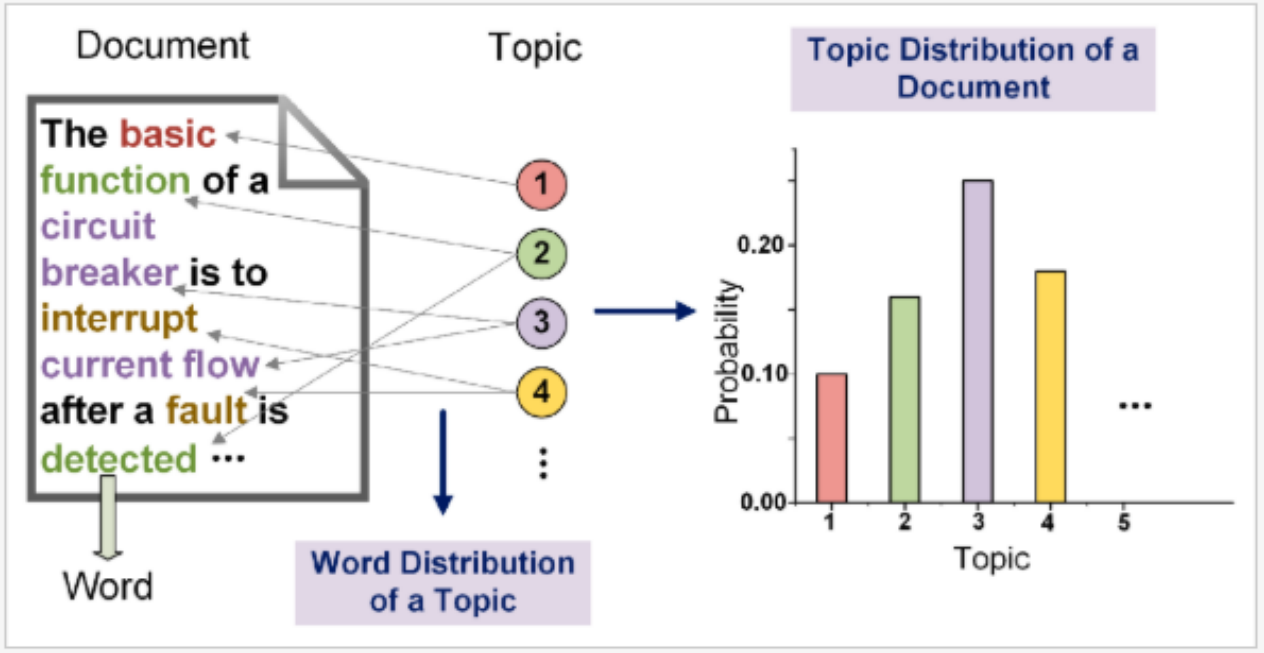


http://athena.ohdsi.org/

Method : LDA clustering

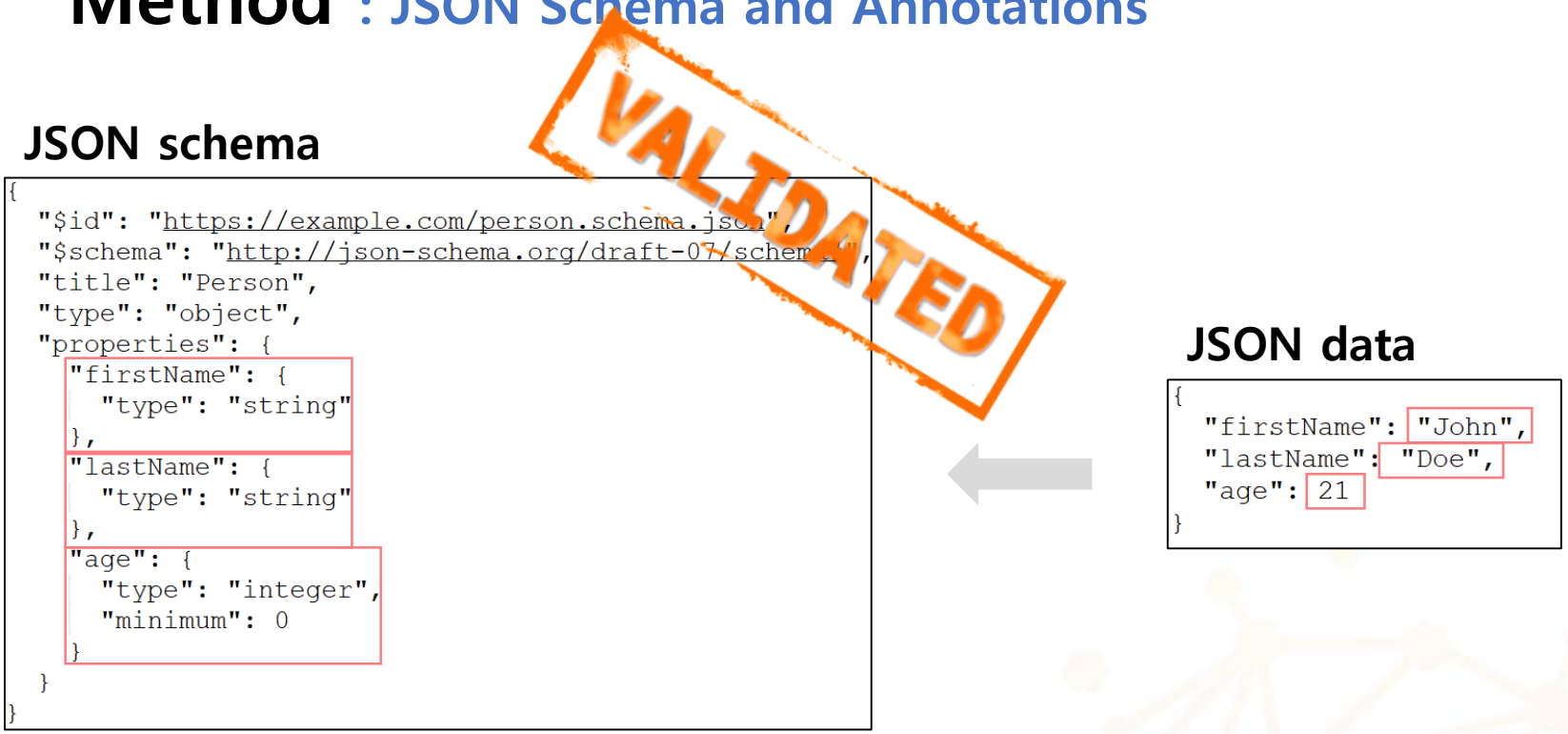
- To avoid reviewing all clinical reports, clustering is needed to see through into the documents
- Latent Dirichlet Allocation (LDA) can detect and classify the topics among the documents
- The topics can describe the contents of the documents

Figure 3. An illustrating example of Latent Dirichlet Allocation (LDA).



Li G, Wang X, Yang A, Rong M, Yang K. Failure Prognosis of High Voltage Circuit Breakers with Temporal Latent Dirichlet Allocation. 2017;10(11):1913.

Method : JSON Schema and Annotations



- With clustering result, user can get an information to define a JSON schema which contains the documents
- JSON schema defines the structure of JSON which ensures the quality of data
- After defining JSON schema, user need to annotated JSON based on the template

https://json-schema.org/understanding-json-schema/about.html

Method : ELK stack

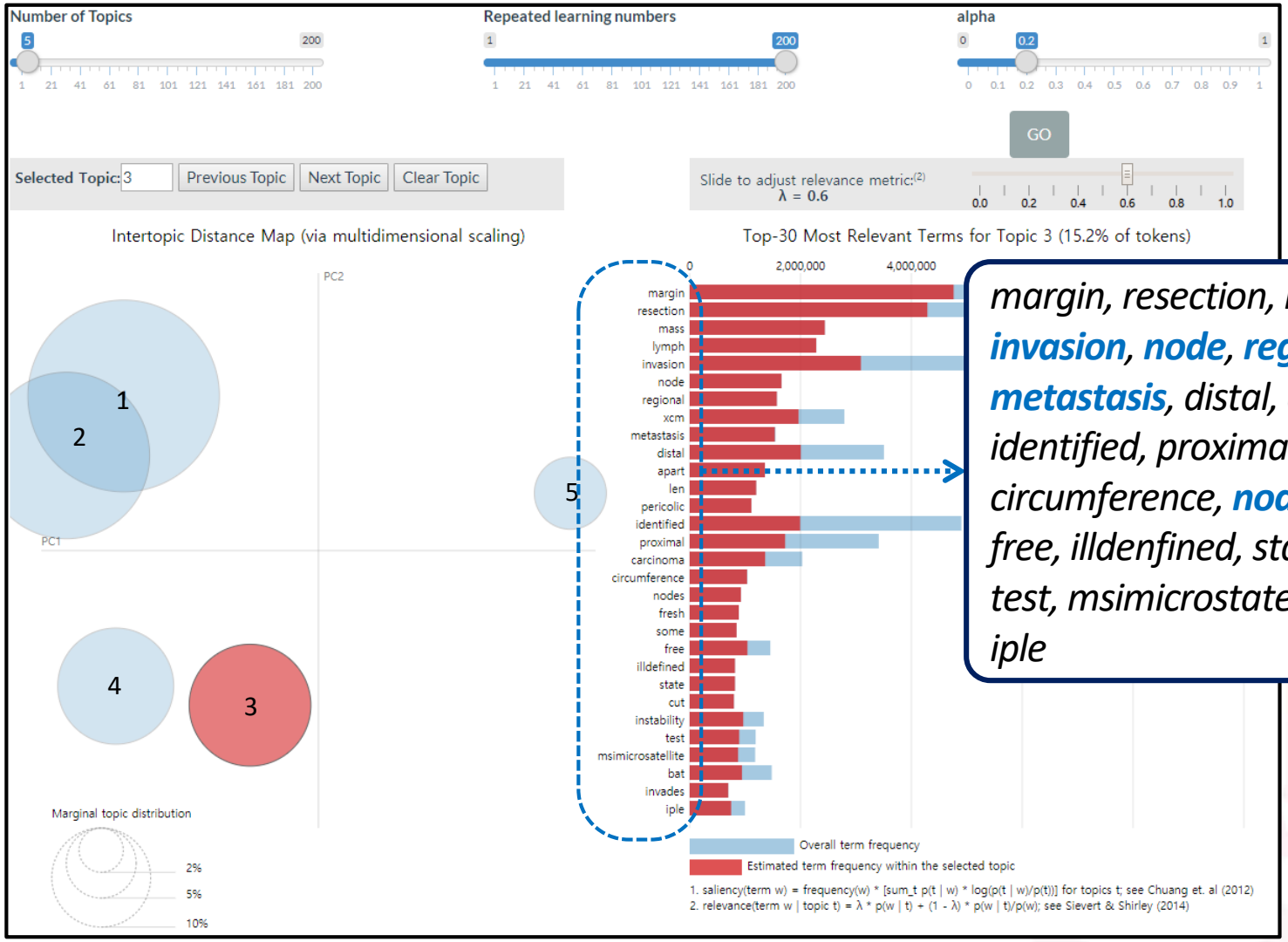
- The annotated JSON can be sent to Elasticsearch
- Elasticsearch is a full text search engine with a schema-free JSON documents
- By building ELK stack with annotated JSON data, clinical text search engine can be constructed

https://github.com/elastic/elasticsearch

Method : Validation

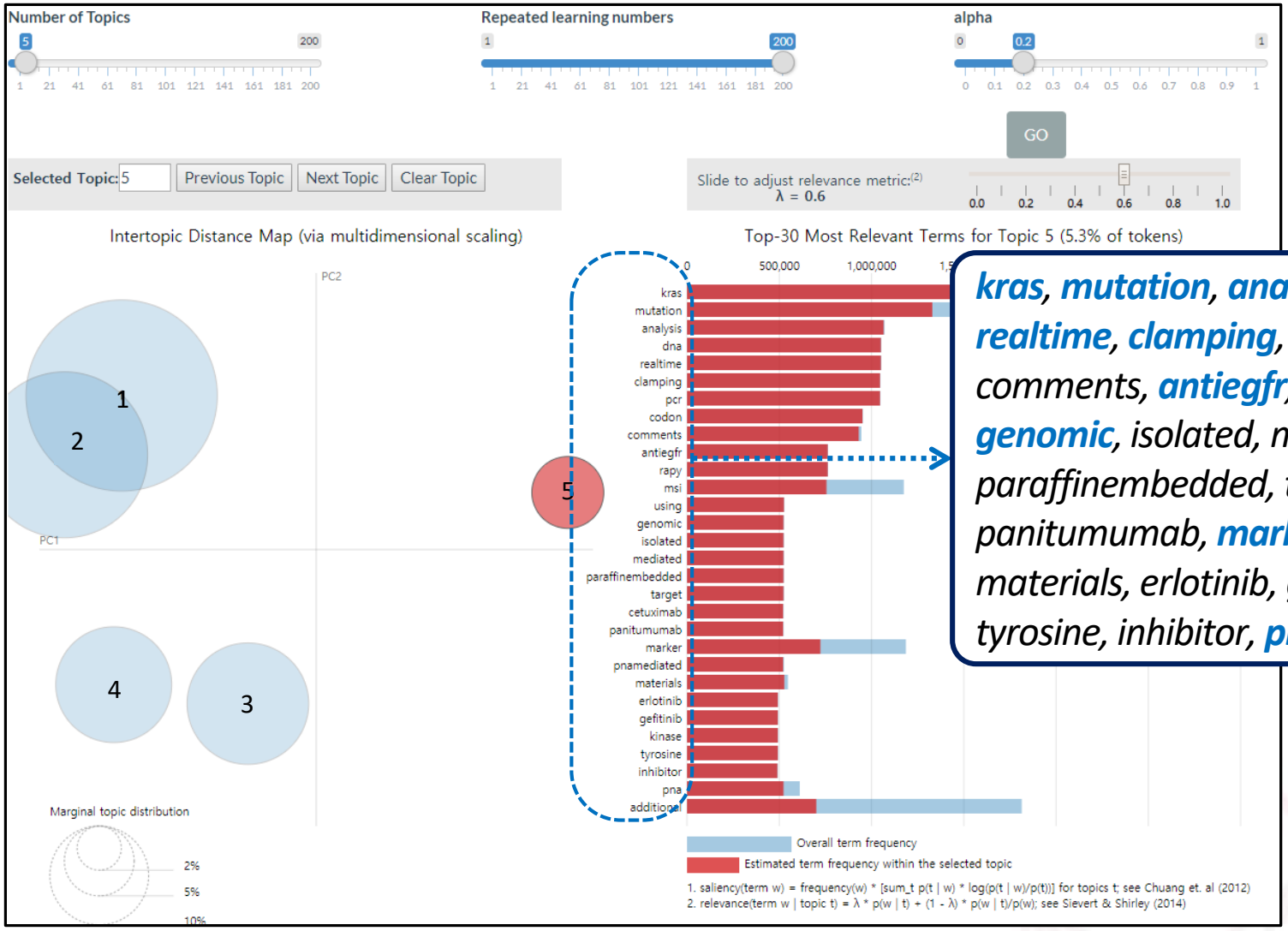
- Ajou University Medical Center
- ICD-10th C18-20 diagnosed from 2014-2017 were included
  - Malignant neoplasm of colon, rectosigmoid junction and rectum
- 1,989 pathology reports on colorectal cancer of 1,929 patients were included

Results : LDA clustering



*margin, resection, mass, **lymph**, **invasion**, **node**, **regional**, xcm, **metastasis**, distal, apart, len, **pericolic**, identified, proximal, **carcinoma**, circumference, **nodes**, fresh, some, free, illdenfined, state, cut, instability, test, msimicrostatellite, bat, **invades**, iple*

Results : LDA clustering

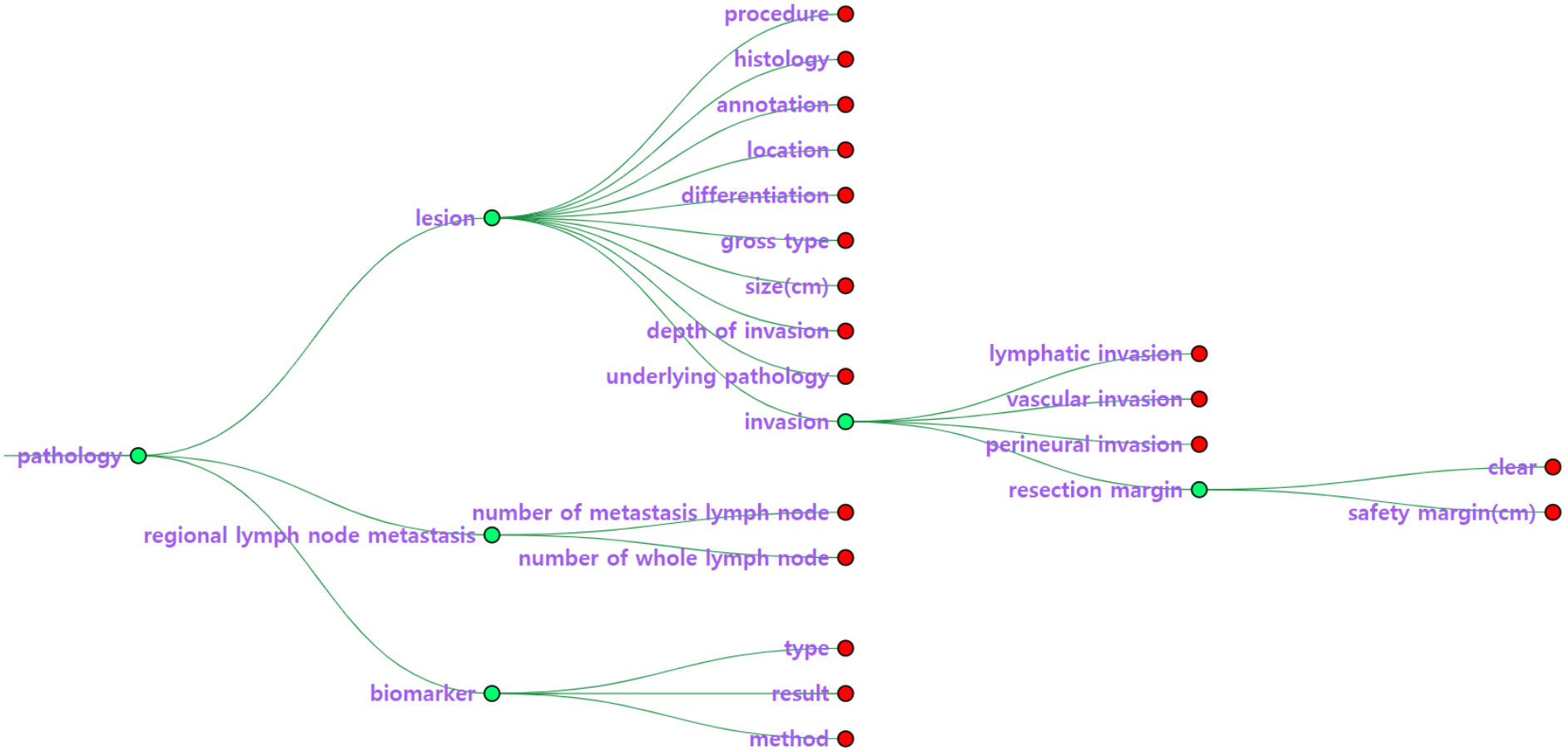


***kras*, *mutation*, *analysis*, *dna*, *realtime*, *clamping*, *pcr*, codon, comments, *antiegfr*, rapy, *msi*, using, *genomic*, isolated, mediated, paraffinembedded, target, cetuximab, panitumumab, *marker*, pnamediated, materials, erlotinib, gefitinib, *kinase*, tyrosine, inhibitor, *pna*, additional**

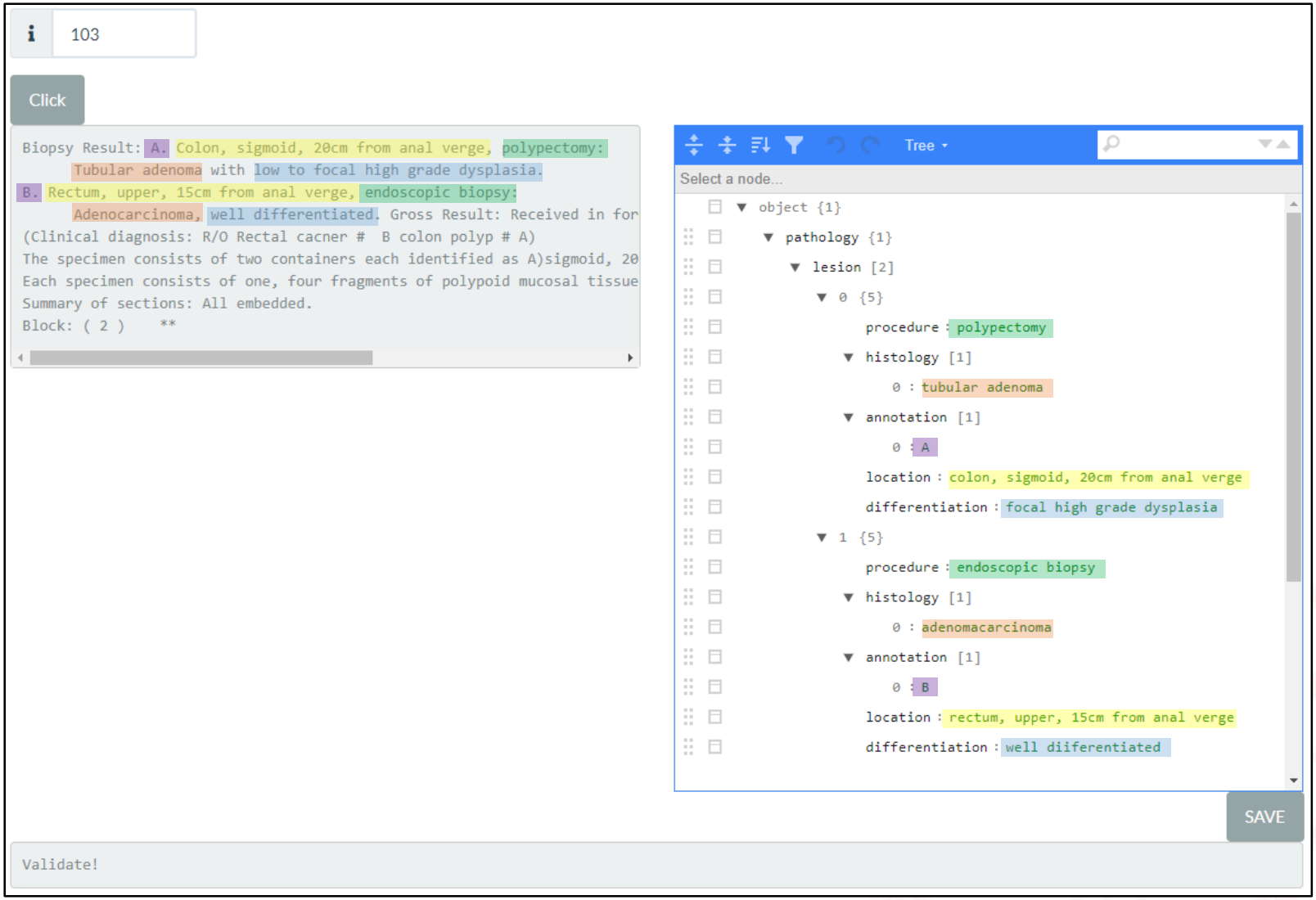
Results : LDA clustering

Topic		Terms
Topic1	Malignant, biopsy	***biopsy***, *all, consists, xxcm, embedded,* ***mucosal***, *received, measuring, diagnosis,* ***sections***, *tissue, pie ces, labelled, gross,* ***biopsied***, ***adenocarcinoma***, ***cancer***, ***differentiated***, *colon,* ***moderately***, *rectal, v erge, rectum, anal, four, sigmoid,* ***endoscopic***, *largest, five, one*
Topic2	Benign, biopsy	*anal, verge, colon, one,* ***tubular***, ***adenoma***, ***low***, ***grade***, ***dysplasia***, ***biopsy***, *transverse,* ***polypectomy*** *, containers, each, ascending, identified, consists, two, polyp, largest, sigmoid, descending, polypoid , hyperplastic,* ***mucosal***, *proximal,* ***endoscopic***, *polyps, xxcm, three*
Topic3	Lymph node invasion, surgery	*margin, resection, mass,* ***lymph***, ***invasion***, ***node***, ***regional***, *xcm,* ***metastasis***, *distal, apart, len,* ***peric olic***, *identified, proximal,* ***carcinoma***, *circumference,* ***nodes***, *fresh, some, free, illdefined, state, cut, i nstability, test, msimicrosatellite, bat,* ***invades***, *iple*
Topic4	Cancer, surgery	***invasion***, ***adenoma***, *resection, margin, submitted, consu, ation, hampe,* ***grade***, *histopathologic, sta ined, size,* ***adenocarcinoma***, ***dysplasia***, ***high***, ***tumor***, ***tublovillous***, *type, depth, low, biopsy, gross,* ***w ell***, ***tubular***, *labelled,* ***differentiated***, *polypectomy, colon, endoscopic, whitish*
Topic5	Gene mutation analysis	***kras***, ***mutation***, ***analysis***, ***dna***, ***realtime***, ***clamping***, ***pcr***, *codon, comments,* ***antiegfr***, *rapy,* ***msi***, *usin g,* ***genomic***, *isolated, mediated, paraffinembedded, target, cetuximab, panitumumab,* ***marker***, *pna mediated, materials, erlotinib, gefitinib,* ***kinase***, *tyrosine, inhibitor,* ***pna***, *additional*

Results : JSON Schema

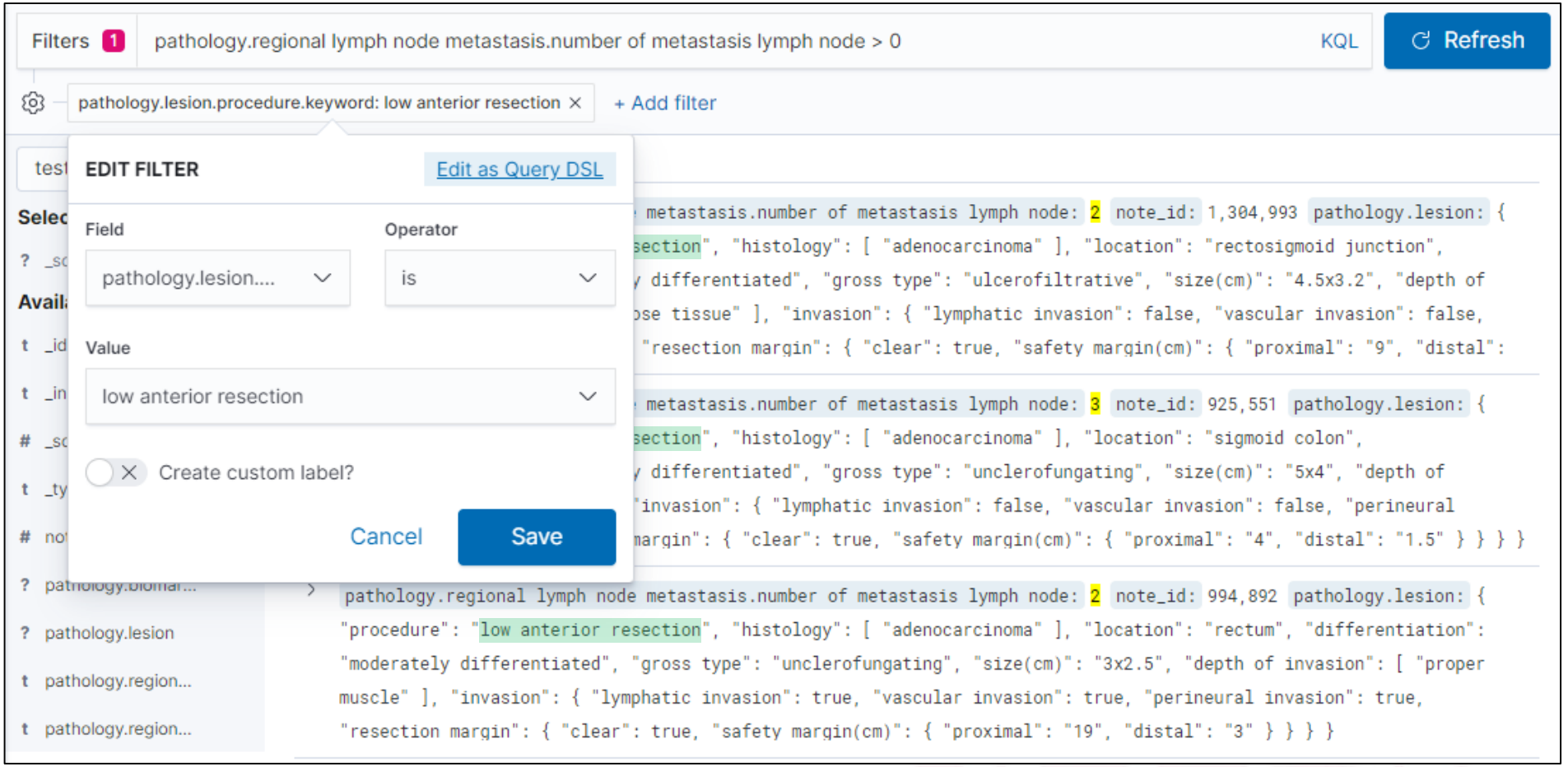


Results : JSON Annotation

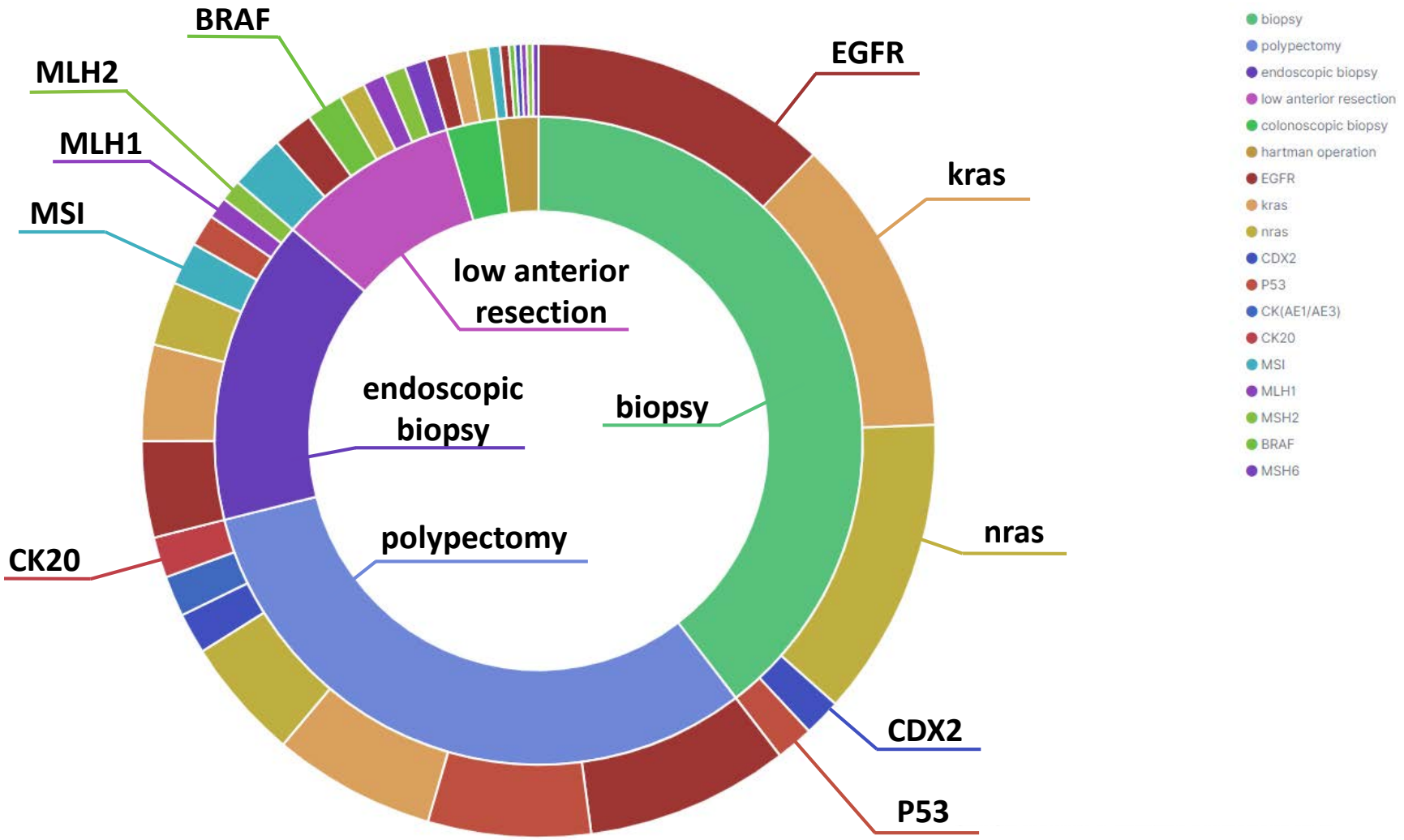


Results : ELK stack

Results : ELK stack

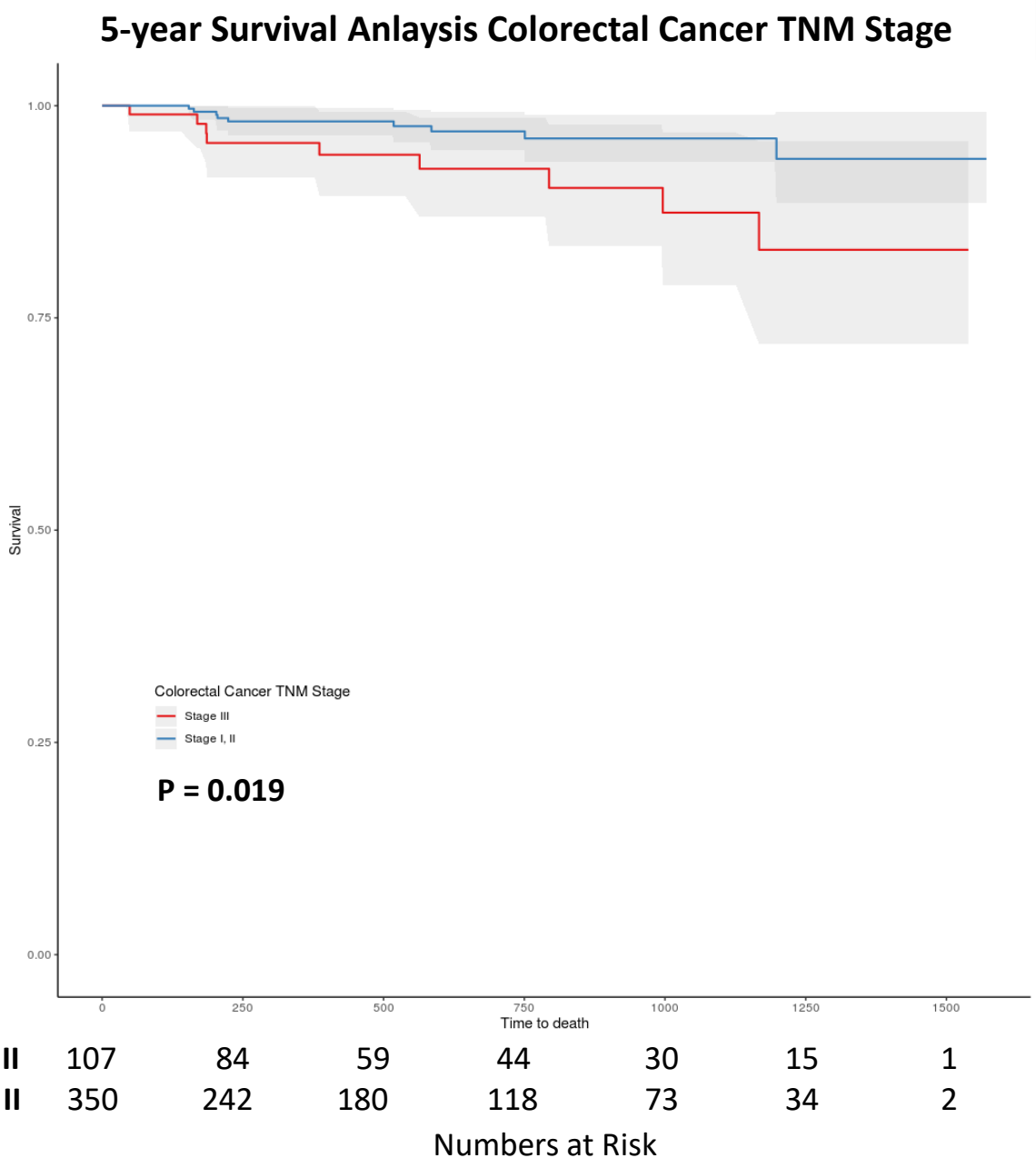


Results : ELK stack



Results : Analysis

- Using annotated JSON data, TNM stage was extracted using '**depth of invasion**' and '**metastasis lymph node**' keys
- 5-year survival analysis was conducted
- As a result, it shows similar survival rates to previously known survival rates
- This analysis indicates that SOCRATex combined with OMOP-CDM can result in significant clinical findings



https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html

Conclusions

- SOCRATex can be applied to any type of clinical documents on OMOP-CDM
- It helps users to explore and cluster text data through a topic model and a search engine
- SOCRATex can produce labeled and unlabeled data which can be used for clinical analysis
- However, curation and annotation still needs human intervention and manual annotation

Thank you

Q&A