SOCRATex

Staged Optimization of Curation, Regularization, and Annotation of clinical Text

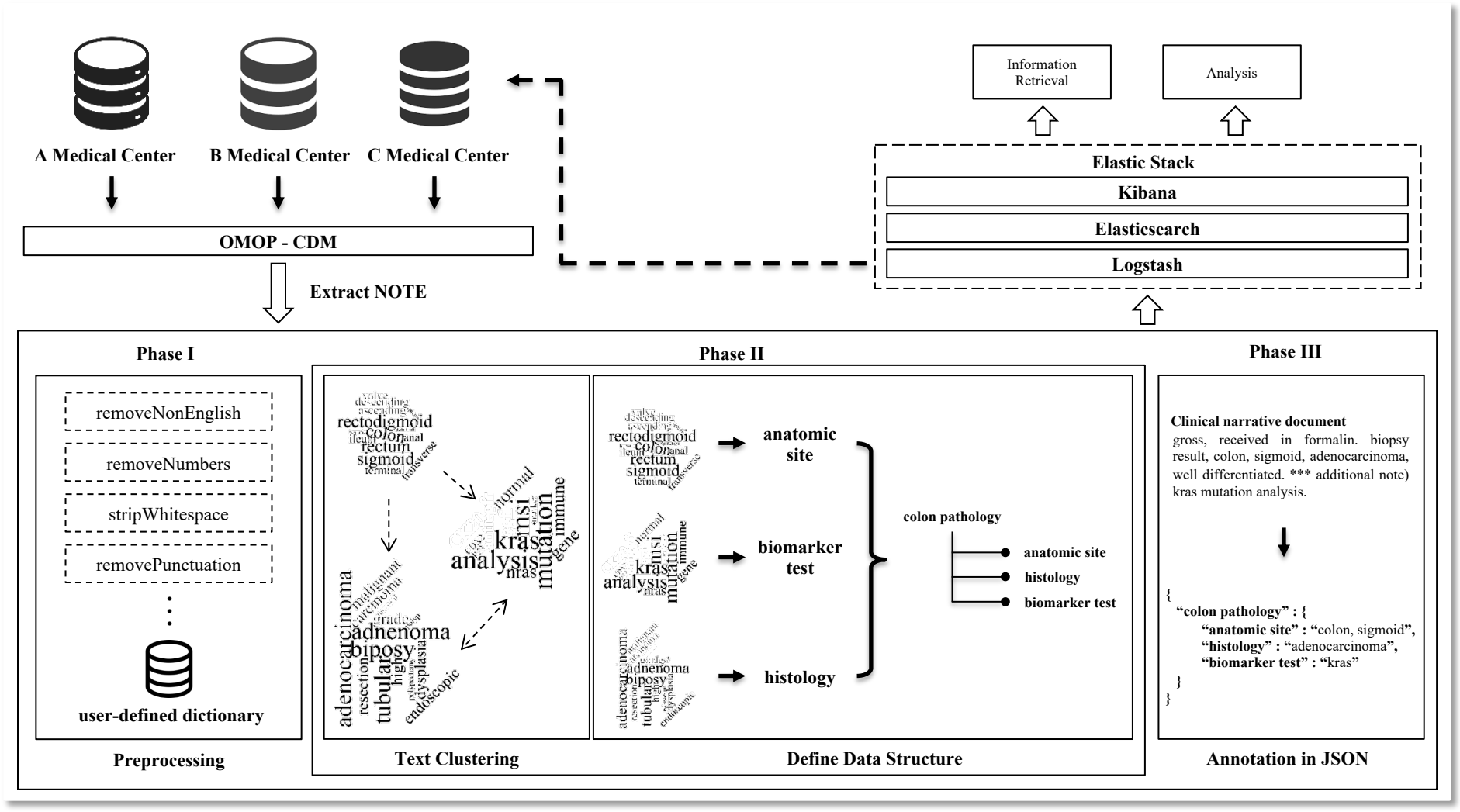
Jimyung Park[1], Seng Chan You M.D. M.S.[2], Jin Roh M.D. Ph.D[4], Rae Woong Park M.D. Ph.D[1,2,3]

1 Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon, 16499
2 Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, 16499
3 FEEDER-NET(Federated E-Health Big Data for Evidence Renovation Network)
4 Dept. of Pathology, Ajou University Hospital, Yeongtong-gu, Suwon, 16499

Background

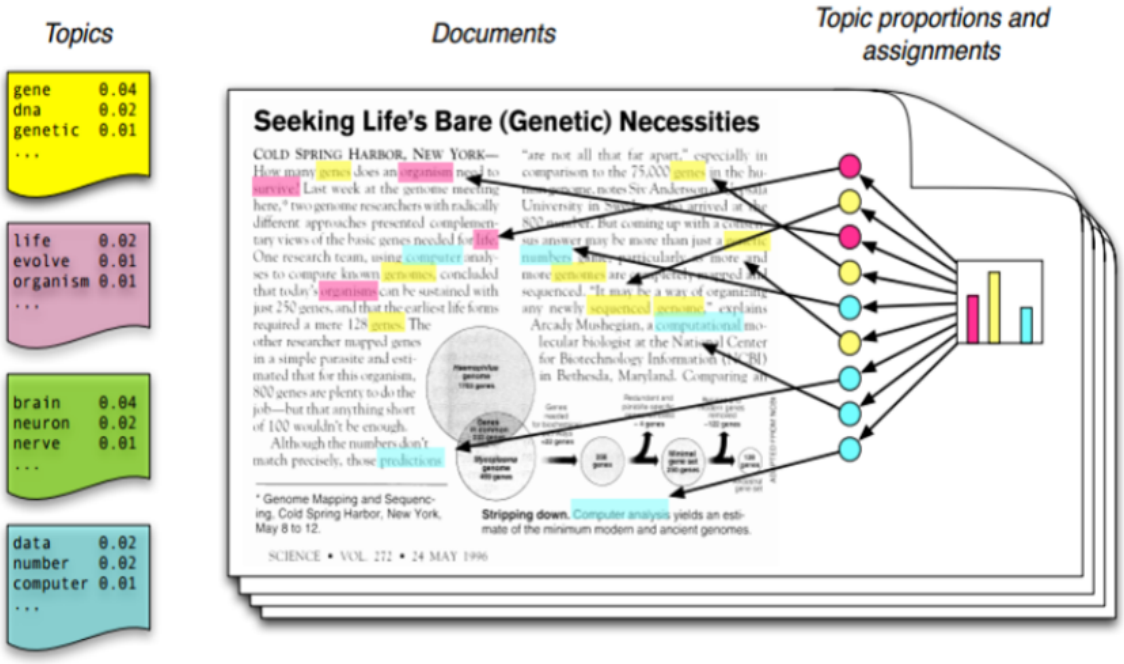
- State-of-the-art (SOTA) methods made great stories on Natural Language Processing (NLP) tasks
- Yet, the SOTA methods usually require **massive amounts of labeled-data** to learn
- SOCRATex is a Natural Language Processing (NLP) system which helps users to **understand**, **annotate**, and **retrieve** their text documents

System Architecture of SOCRATex



Latent Dirichlet Allocation

- Latent Dirichlet Allocation (LDA) is a statistical topic model method which captures **latent topics** in documents

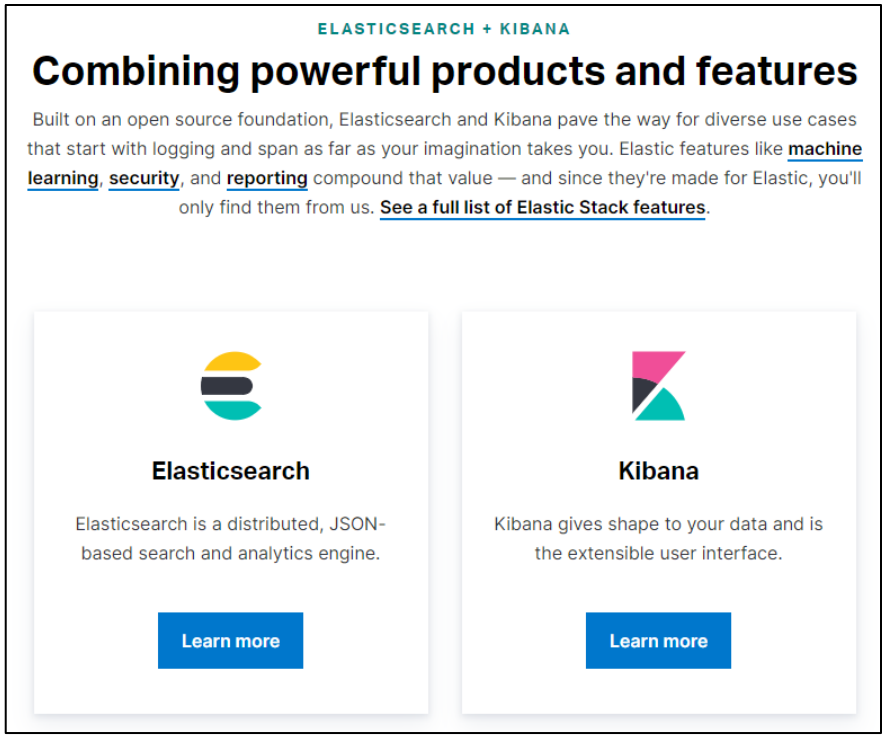


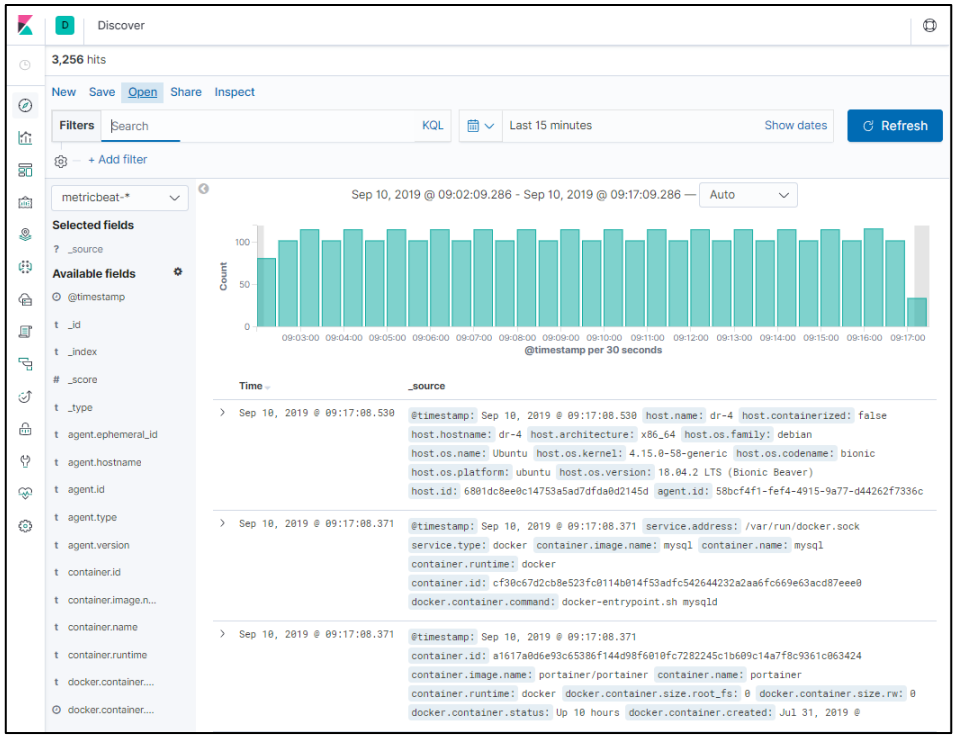
I. LDA is an unsupervised method which can reduce time and cost of reviewing the reports

II. LDA results can give an insight how to annotate and re-construct their text data

Elastic Stack

- **Elasticsearch** is a open-source indexing framework for searching, which can accept tree-based documents such as JSON documents.
- **Kibana** helps to visualize indexed data stored in Elasticsearch



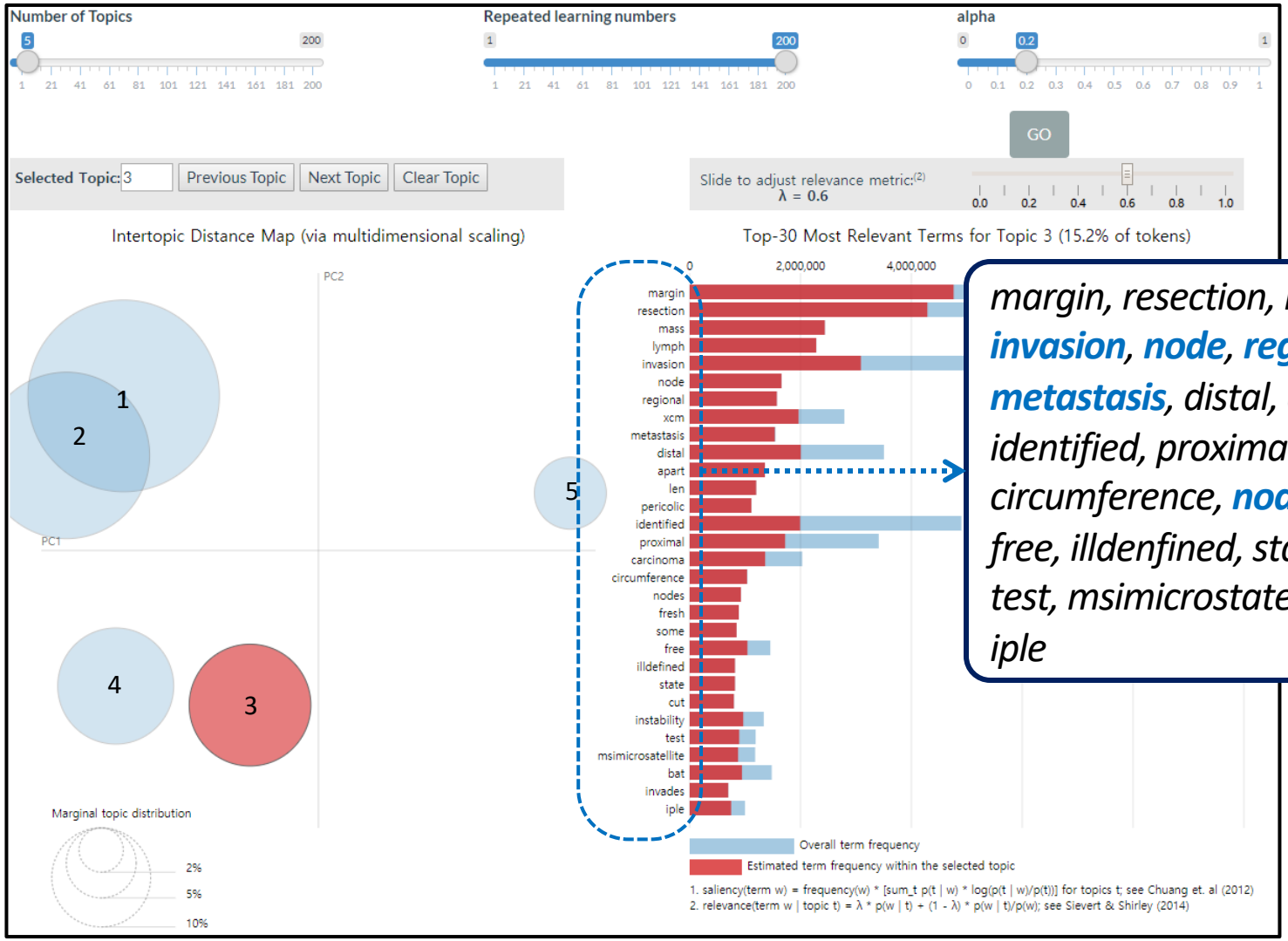


https://www.elastic.co/ ; McEwan R, Melton GB, Knoll BC, Wang Y, Hultman G, Dale JL, et al. NLP-PIER: A Scalable Natural Language Processing, Indexing, and Searching Architecture for Clinical Notes. AMIA Jt Summits Transl Sci Proc. 2016;2016:150-9

Data Source

- Ajou University Medical Center
- ICD-10th C18-20 diagnosed patients from 2014-2017 were included
- 1,989 pathology reports on colorectal cancer of 1,929 patients were included

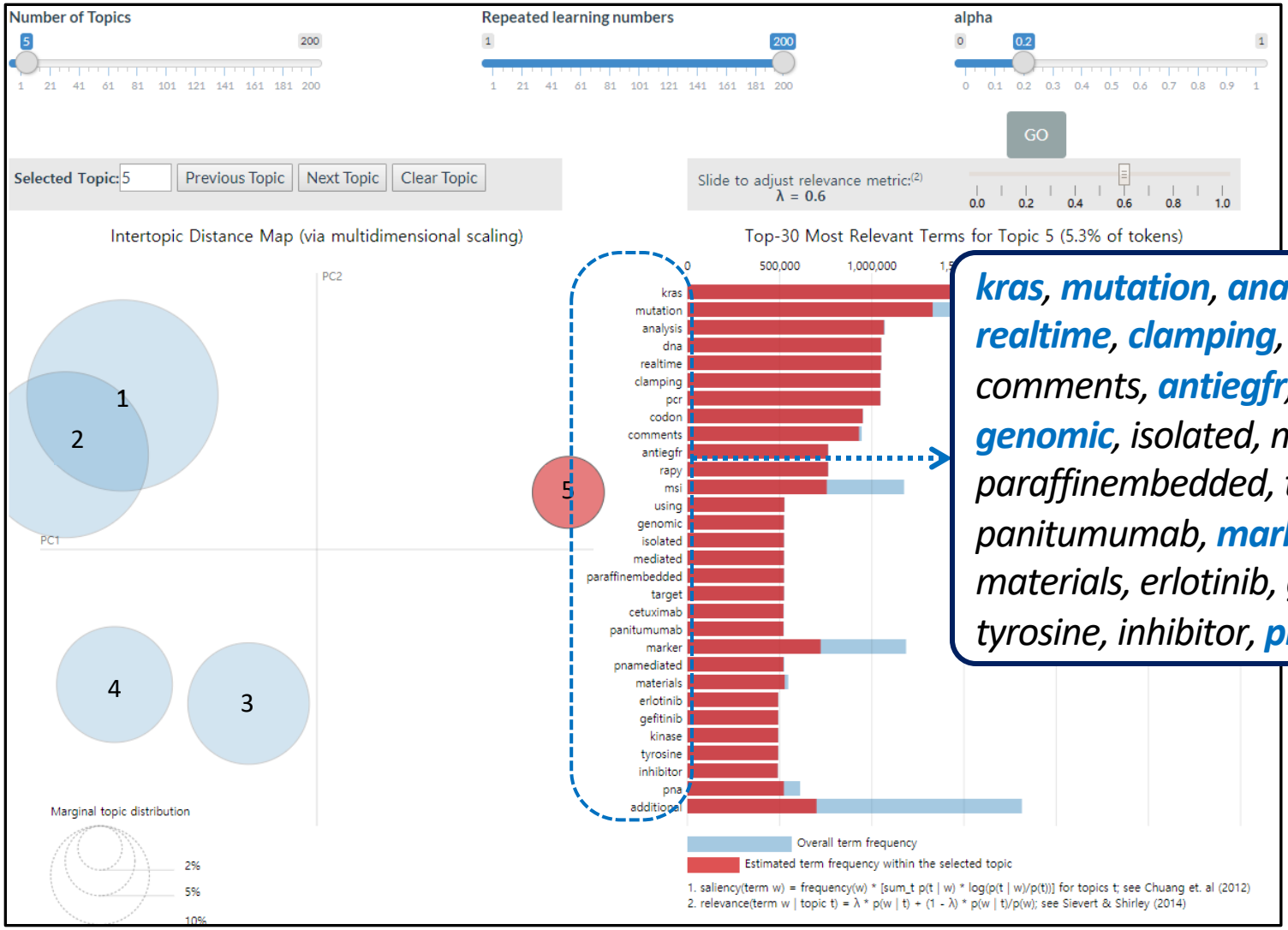
Results of LDA



*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*

Sievert C, Shirley K, editors. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces; 2014

Results of LDA



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*

Sievert C, Shirley K, editors. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces; 2014

Results of LDA

Topic		Terms
Topic num.	Expertise Annotation	
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*

Defining JSON structure based on LDA Results

Topic1

Rectum, endoscopic biopsy: tubulovillous adenoma with high grade dysplasia and adenocarcinoma

Topic2

Colon, proximal transverse, polypectomy: Hyperplastic polyp. Tubular adenoma with low grade dysplasia and clear resection margin.

Topic3

Colon, radical sigmoid colectomy: Adenocarcinoma

…

Regional lymph node metastasis: no metastasis in all 17 regional lymph node nodes (pN0) (pericolic 0/17) Lymphatic invasion: not identified

Topic4

Colon and rectum, Hartman's operation: Adenocarcinoma, moderately differentiated. 1. Location: rectum 2. Gross type: ulcerofungating 3.Size: 7.5x6cm 4.invased perirectal adipose tissue

Topic5

…

Additional report. Kras Mutation Analysis Report; Kras mutation is not detected by PNA mediated real-time PCR clamping method. Materials: Genomic DNA isolated from paraffin-embedded tissue

- Endoscopic biopsy
- Tubulovillous adenoma
- High grade dysplasia
- Adenocarcinoma

- Polypectomy
- Hyperplastic polyp
- Clear resection

- Colectomy
- Regional lymph node metastasis
- Lymphatic invasion

- Hartman's operation
- Adenocarcinoma

- Kras mutation analysis
- Not detected
- PNA mediated real-time PCR clamping method

- Procedure
- Histology
- Annotation
- Location
- Differentiation
- Gross type
- Size(cm)
- Depth of invasion
- Underlying pathology

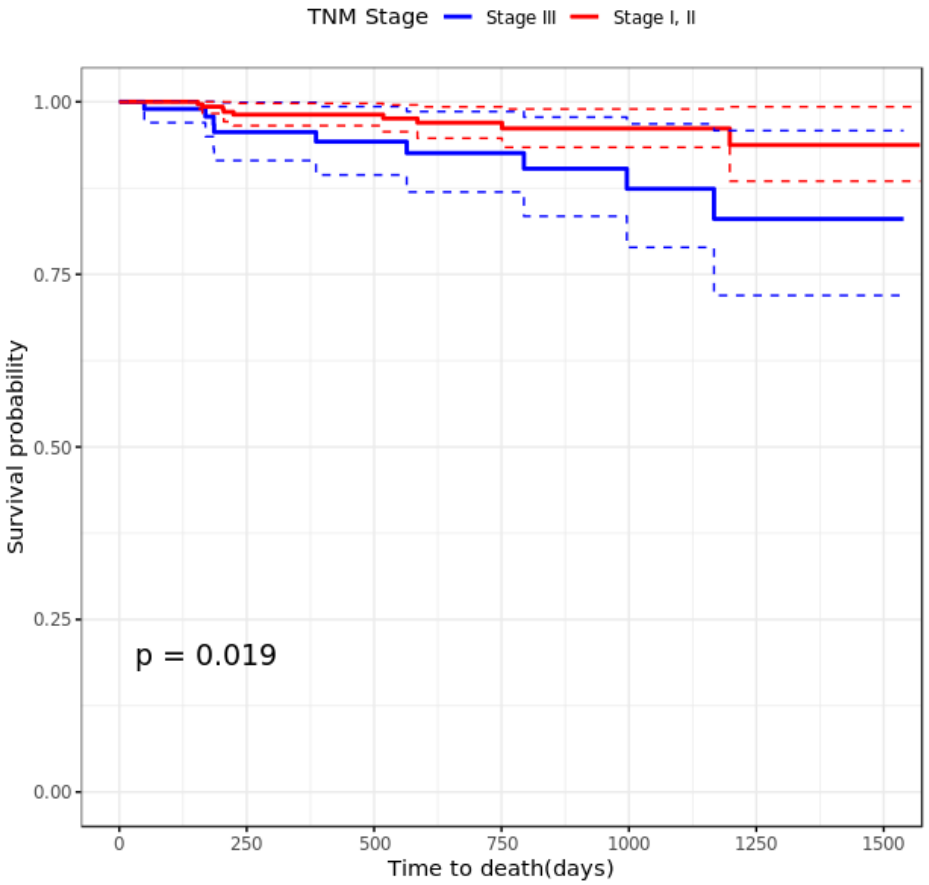
- Number of metastasis lymph node
- Number of whole lymph node
- Invasion
  - Lympathic invasion
  - Vascular invasion
  - Perineural invasion
  - Resection margin
    - Clear
    - Proximal
    - Distal
    - Radial

- Biomarker test
- Biomarkaer method
- Biomarker result

Clinical Analysis using the Annotated JSON Documents

- Using two keys of JSON structure, **TNM stage** was easily extracted, and **5-year survival analysis** was conducted
  - depth of invasion
  - metastasis lymph node
- This research shows **SOCRATex** can actually be combined with **OMOP CDM** and help to conduct **clinical analysis**
- **SOCRATex** is available at

  https://github.com/ABMI/SOCRATex



Numbers at risk

Stage I, II	350	242	180	118	73	34	2
Stage III	107	84	59	44	30	15	1

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

Thank You