Natural language processing as a cloud based service for CRIS and beyond

Angus Roberts, Matthew Broadbent, Jyoti Sanyal, Anna Kolliakou, Ian Roberts, Robert Stewart

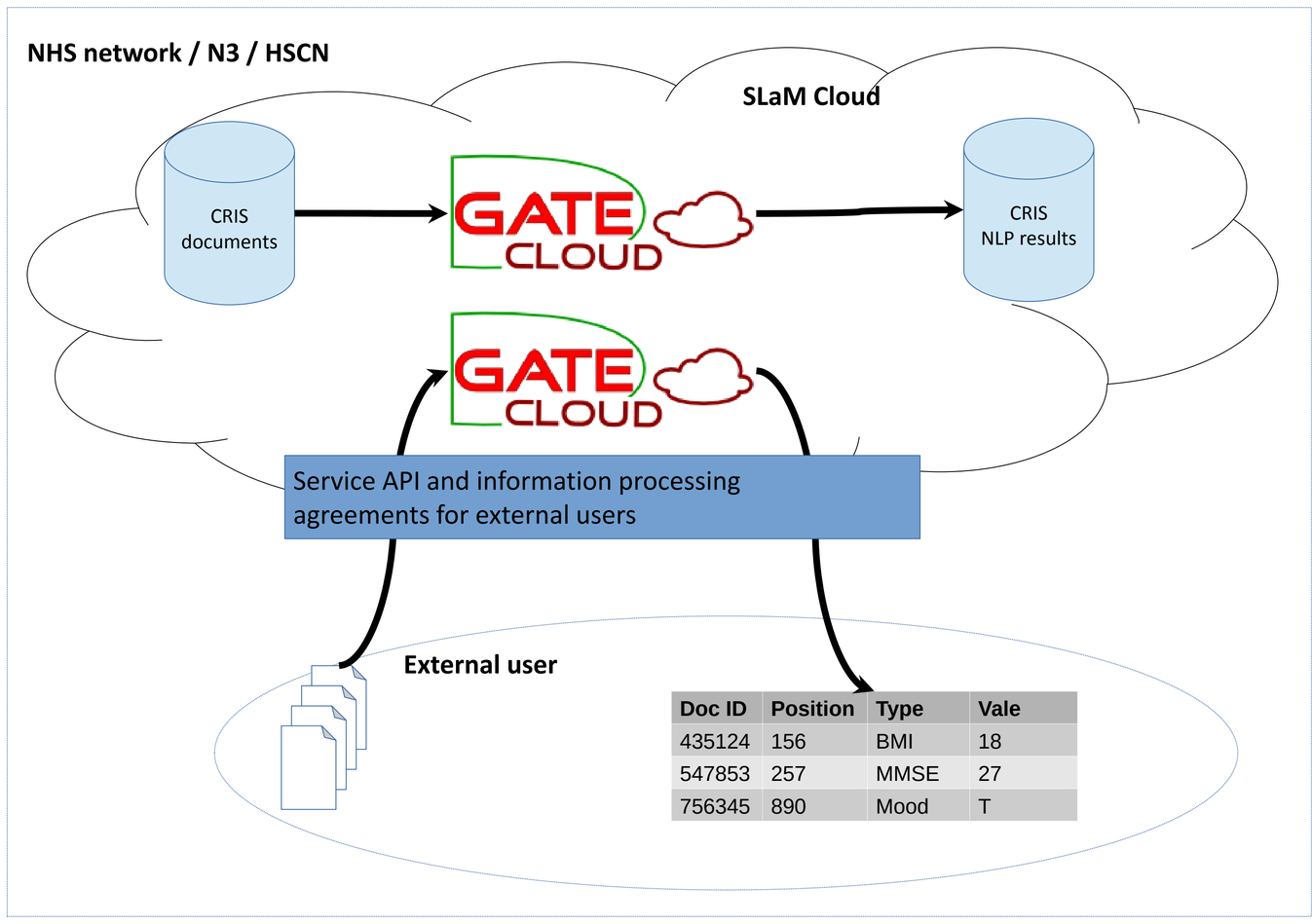
National Institute for Health Research

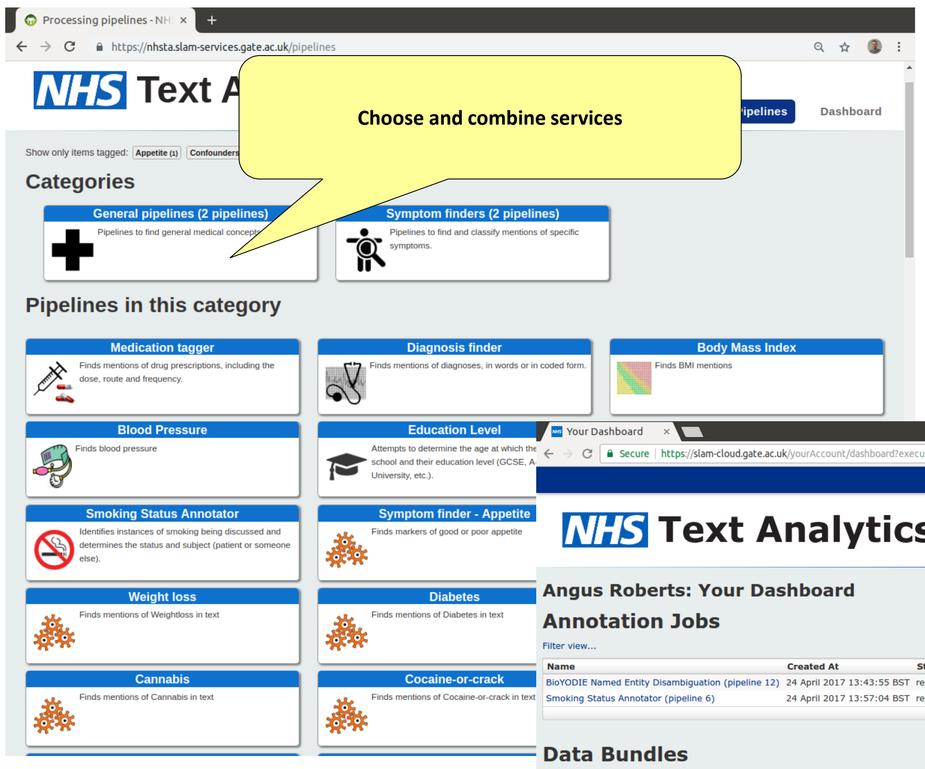
CRIS uses over 70 natural language processing applications to extract information from free text fields, mainly created with the GATE NLP framework

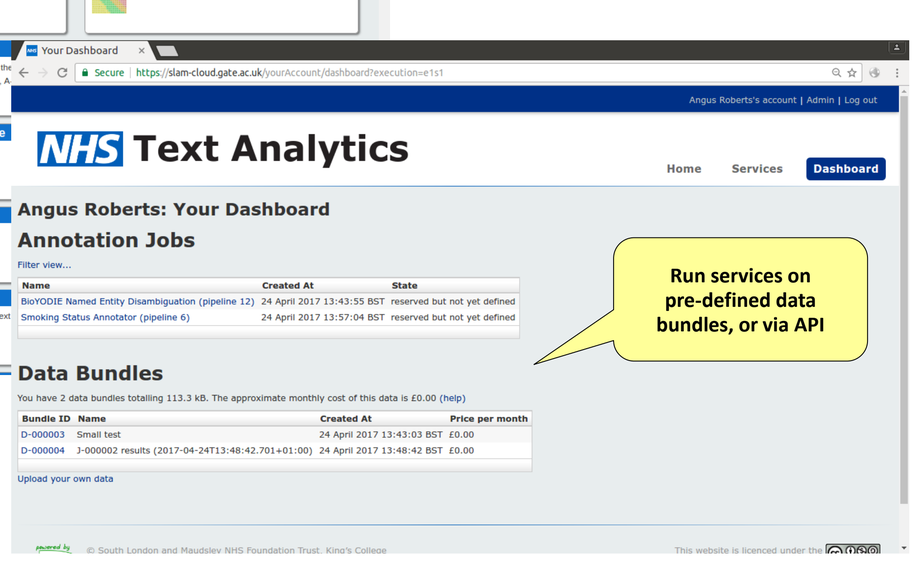
NLP is very compute intensive. It is also highly parallel, with each document being processed through the same “pipeline” of steps.

We have taken advantage of this parallel nature to create a cloud based service for CRIS NLP, using the open source GATE Cloud platform.

The service is available for external users, and resource is available via the MRC Pathfinder project to help external users implement its use.







NLP applications are made available via a “self service” web interface, allowing analysts to upload, make available, and run applications over specific datasets.

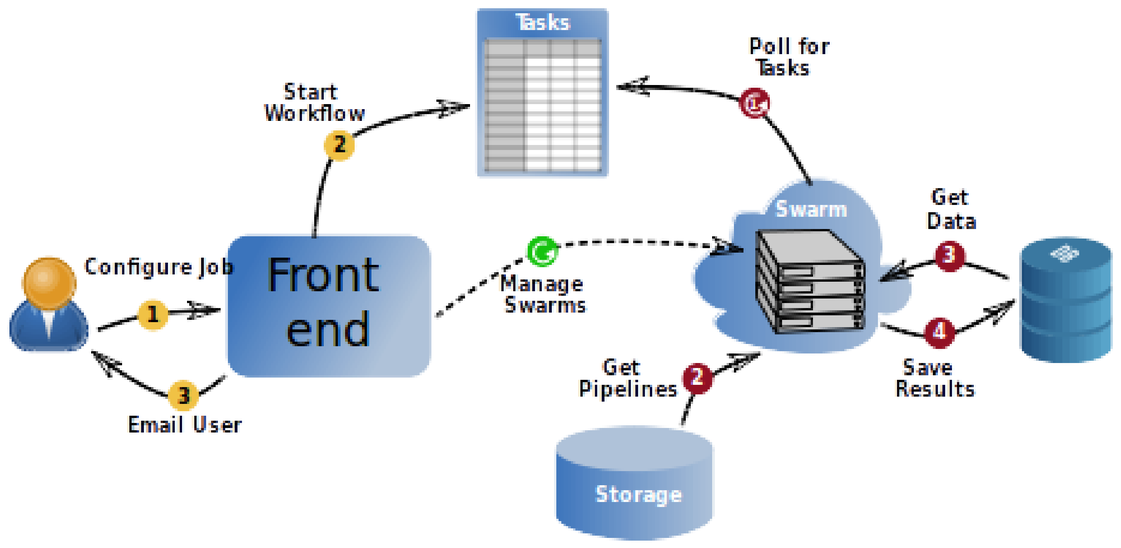
A service API allows the sharing of both applications and compute resources with external NHS users, providing an on-demand NLP service for EHR text.

The API is being further developed in collaboration with the University of Cambridge, as part of the MRC Pathfinder project.

The service is implemented using Microsoft’s Azure cloud platform at the South London and Maudsley NHS Foundation Trust.

The user front end defines tasks and starts the workflow. Compute nodes within a flexibly-sized swarm request new tasks, retrieve the data needed for that task from CRIS, and save results back to CRIS. This means that the user is isolated from the content of CRIS.

The service has a job control API that can be used to create more complex tasks programmatically, such as scheduling regular runs.



Application	Azure machine	Machine x threads	Wall clock time	Documents per machine minute
Medication	D8_V3	15 x 12	2:56	12817
Diagnosis	D8_V3	15 x 12	2:36	15026
Medication + diagnosis	D8_V3	15 x 12	4:26	8515
Bio-YODIE	D8_V3	15 x 12	18:33	1900
Appetite + suicide + low mood	D8_V3	15 x 12	5:27	6790
Nothing	D8_V3	15 x 12	1:43	24554

The table shows indicative processing times for 30 million documents in CRIS.

Applications that were taking several days to run on a desktop workstation are now taking a few hours, and could be speeded up further with a more powerful database and more processing nodes.

The combined medications and diagnosis applications takes 80% of the time of the individual applications, as the platform is able to factor out common parts of the two applications.

The “Nothing” application shows the base time taken to load each document from CRIS. Bio-YODIE is a complex application for labelling documents with multiple terminologies, such as SNOMED, ICD-10 and others.

CLINICAL RECORD INTERACTIVE SEARCH

The University Of Sheffield.

UNIVERSITY OF CAMBRIDGE

Microsoft Azure

MRC Medical Research Council

NHS South London and Maudsley NHS Foundation Trust