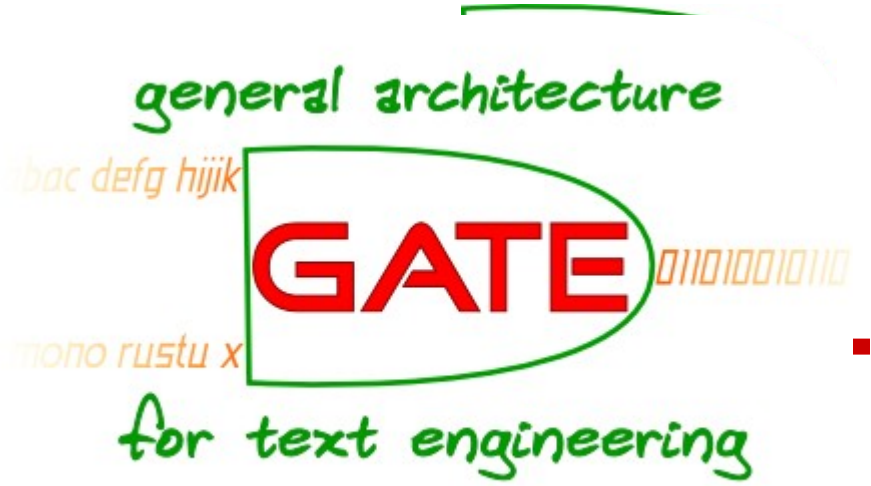


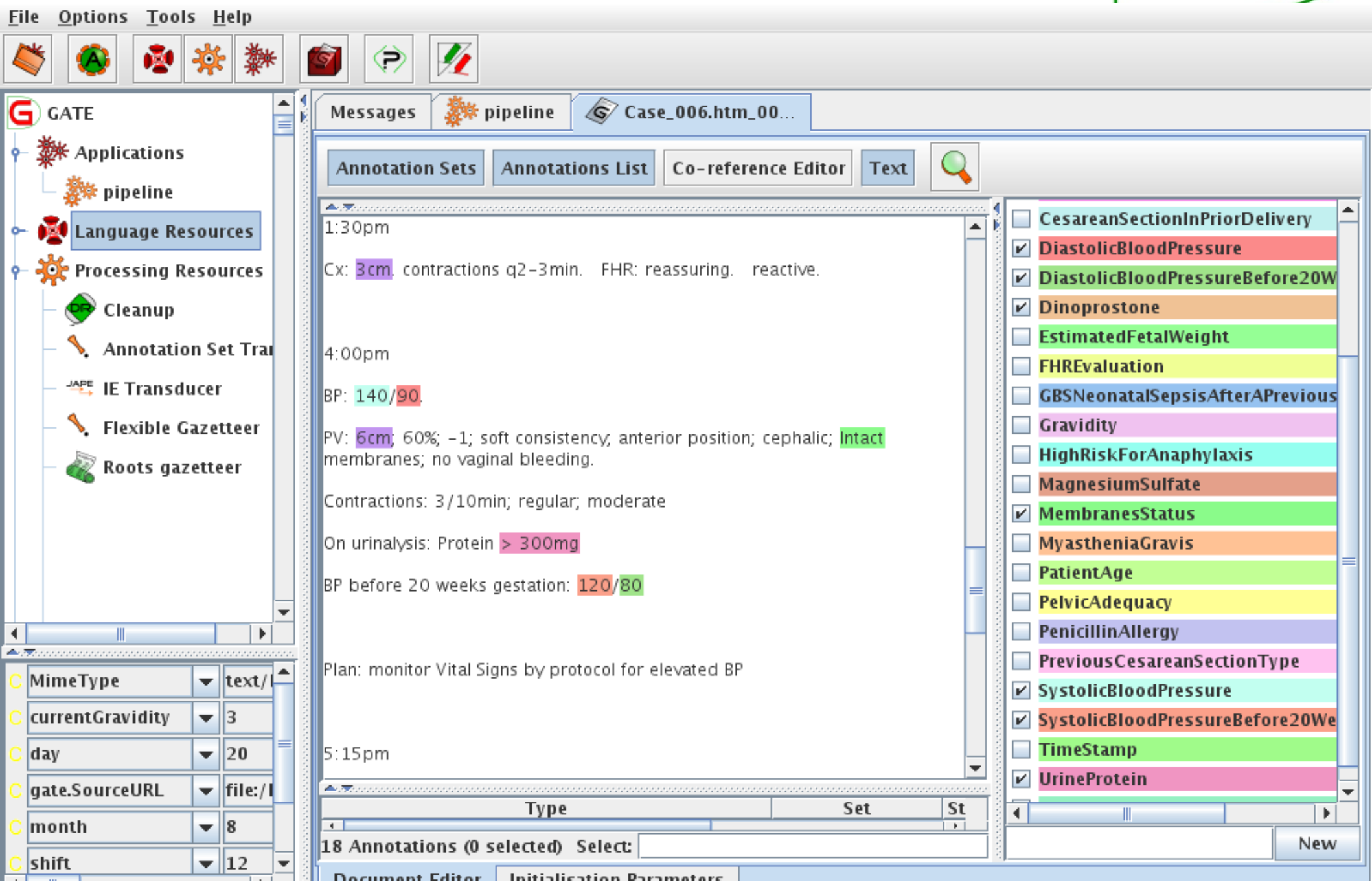
GATE – an introduction

GATE Team
University of Sheffield

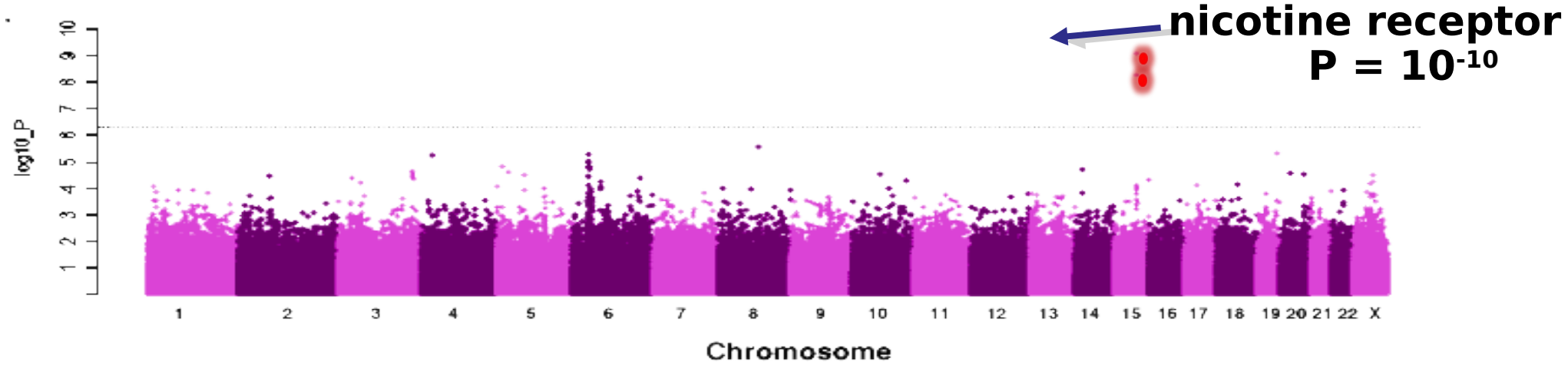
The challenge

- There is a lot of free text in medical records and in the literature
- Medical records - text is:
  - “Convenient to express complex ideas”
  - “Tolerant of ambiguity”
  - “More accurate and reliable”
- Literature:
  - Medline is how big now?
  - And growing
  - Grey literature

Obstetrics use case



WHO Genome Wide Association Studies



- $1x10^5$ to $1x10^6$ SNP markers, 3 to $7x10^3$ cases and controls
- Association with disease expressed as p-values
- The top ranked SNPs are selected for replication
- p-values must be very small to claim significance

WHO Genome Wide Association Studies

- Use a Bayesian model to combine the experimental probability of a SNP with a prior based on the occurrence of text and concepts in literature about proximal genes
- E.g. for lung cancer, we might search for smoking terms

Gene	PubMed article (don't filter keywords)
1138	18783506: A non-synonymous coding SNP in CHRNA5 was associated with case status and, in Caucasians, with experiencing a pleasurable rush or buzz during the first **cigarette**; these sensations were associated highly with current **smoking** (OR = 8.2, P = 0.0001). 18618000: In the 2,827 long-term smokers examined, common susceptibility and protective haplotypes at the CHRNA5-A3-B4 locus were associated with **nicotine dependence** severity. 18759969: Correlated SNPs in the **cholinergic** nicotinic receptor gene cluster CHRNA5-CHRNA3-CHRNB4, in a case-control study of cocaine **dependence** composed of 504 European-American and 583 African-Americans. 18450646: Environmental tobacco smoke and **nicotine** up-regulated the levels of alpha5 and alpha7 expression in a time-dependent fashion 18559515: Increased levels of **cholinergic** receptor nicotinic alpha 5 is associated with squamous cell lung carcinoma 18957677: The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for **nicotine dependence** and for lung cancer.(18227835: A common haplotype in the CHRNA5/CHRNA3 gene cluster on chromosome 15 contains alleles, which predispose to **nicotine dependence**. 19029397: Two distinct variant groups in the CHRNA5-CHRNA3-CHRNB4 gene cluster are strongly associated with heavy **smoking**. The snp rs16969968 alters the coding sequence of these genes.

WHO Genome Wide Association Studies

- Ranked significance of known positive lung cancer SNPs

SNP ID	Old method	New method
rs8034191	1	1
rs1051730	2	2
rs4324798	4	4
rs401681	74	6
rs2736100	77	8
rs3117582	124	10

WHO Genome Wide Association Studies

- Ranked significance of known positive lung cancer SNPs

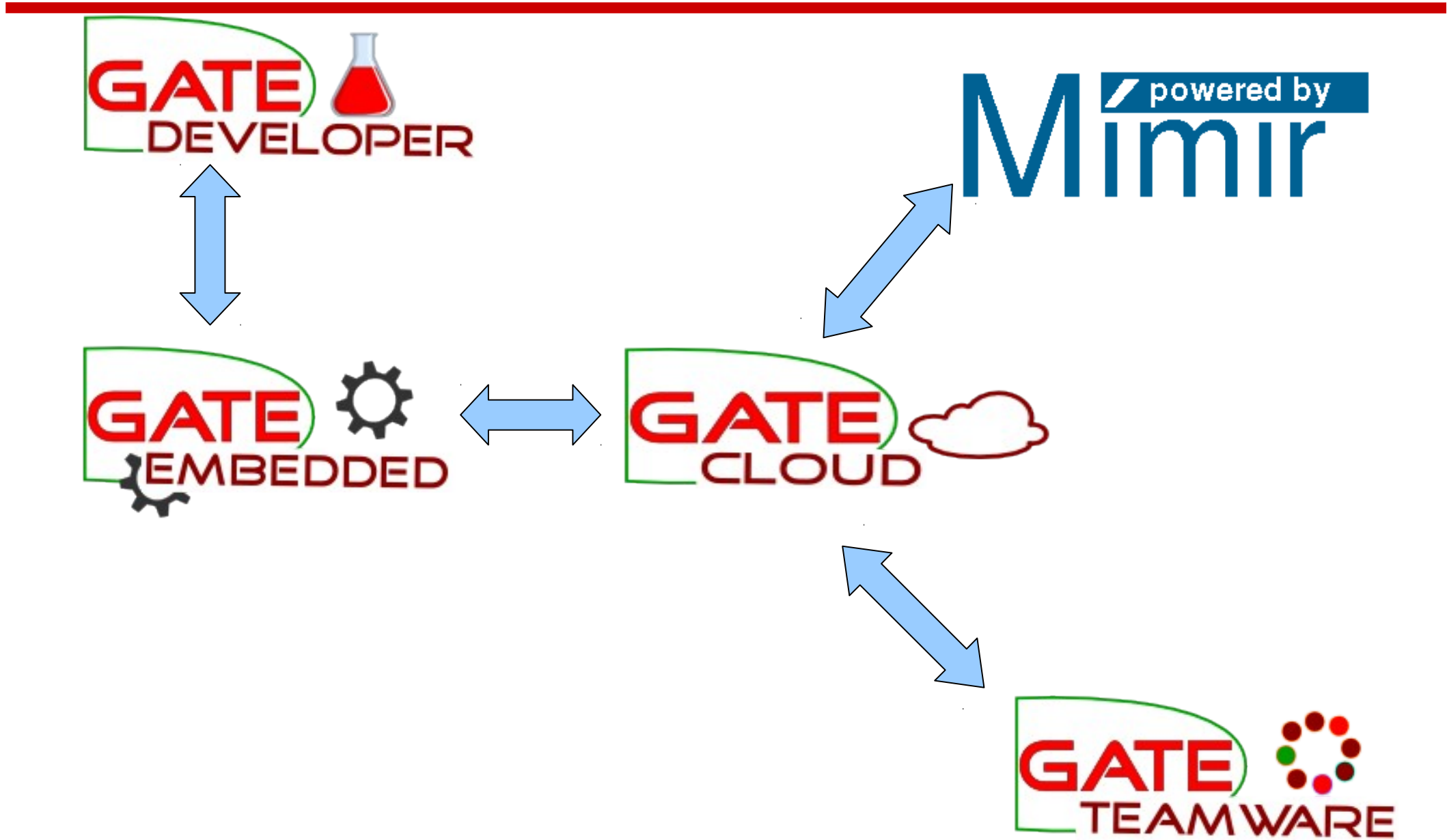
SNP ID	Old method	New method
rs8034191	1	1
rs1051730	2	
rs4324798	4	
rs401681	74	6
rs2736100	77	8
rs3117582	124	10

Extra 3000 samples needed to find these with old method
Potential saving: 900 000 euros

GATE: a framework for Human Language Technology

- A framework for language processing
- Open Source – a large community of users and developers
- Mature: over ten years old, currently at version 7.1
- Funded by a mix of EU, UK RC and commercial funding
- The most widely used toolkit of its kind, with 1000s of users at 100s of sites
  - BBC World Cup and Olympics sites; The Press Association; The National Archives; Elsevier; IBM and Oracle integration; various pharma; many other multi-nationals and SMEs
- Biggest single installation supports 10 000 concurrent users
- An architecture: simplifying the construction of NLP software.

The GATE family

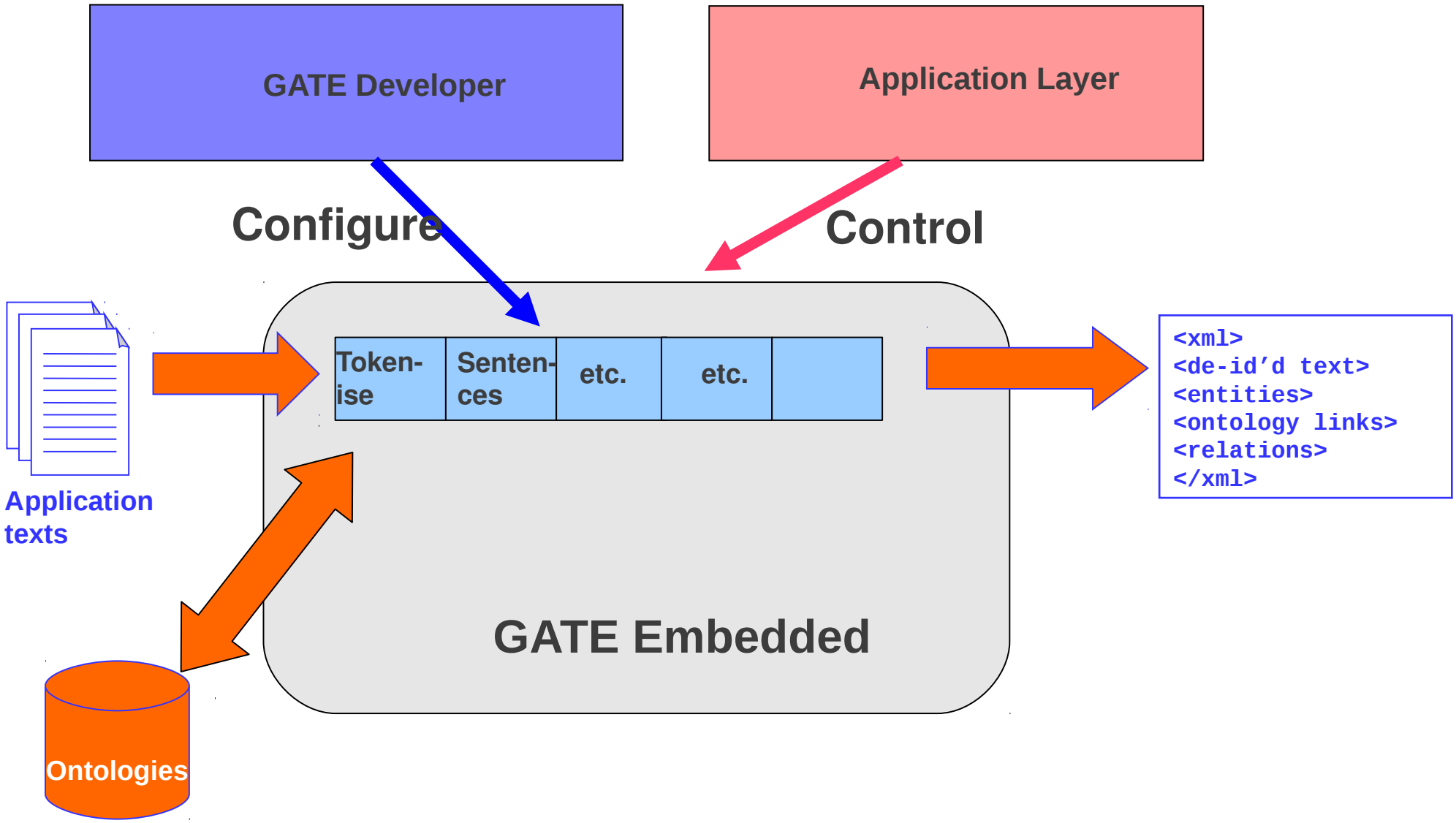


The GATE Team

- University of Sheffield
- Department of Computer Science
  - Natural Language Processing Group
  - 25 academics and researchers, 14 postgrads
  - Links to Information School, speech and hearing
- GATE Team: 14 researchers

Who we are

GATE Developer and Embedded



GATE and medical records

- CaTIES
- HiTEX
- GATE systems often highly ranked in I2B2 challenges
- Commercial use
- University of Sheffield
  - CLEF – a Clinical e-Science Framework
  - German radiology reports
  - Obstetrics system
  - BRC at South London and Maudsley

Types of NLP systems

- Various types of NLP system - GATE is agnostic and can act as an architecture for any
- Deep or shallow analysis
  - Simple lexico-syntactic features
  - Chunking
  - Parsing
- Knowledge Engineering or Machine Learning approaches
  - Supervised
  - Unsupervised
  - Active learning

Applications

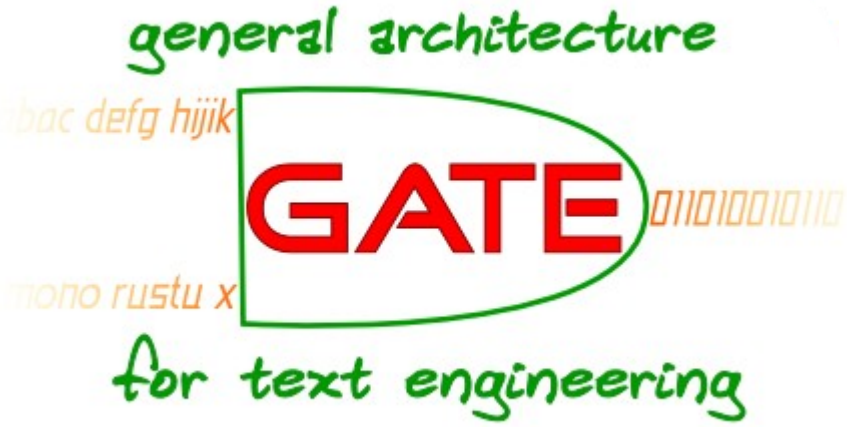
- Information extraction
- Classification
-
-
- Building resources for machine translation
- Sentiment analysis

Outline

Format of the training

Beyond GATE

- NLP
  - Dan Jurafsky and Christopher Manning
- Software
  - Apache UIMA
  - NLTK (Python)
- Biomedical NLP
  - JAMIA
  - JBI
  - BMC Bioinformatics



Thank you.
Any questions?