

# 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



File Options Tools Help

Annotations pipeline Case\_006.htm\_00...

Annotation Sets Annotations List Co-reference Editor Text

1:30pm  
Cx: 3cm. contractions q2-3min. FHR: reassuring. reactive.

4:00pm  
BP: 140/90.  
PV: 6cm, 60%; -1; soft consistency, anterior position; cephalic; intact membranes; no vaginal bleeding.  
Contractions: 3/10min; regular; moderate  
On urinalysis: Protein > 300mg  
BP before 20 weeks gestation: 120/80

5:15pm  
Plan: monitor Vital Signs by protocol for elevated BP

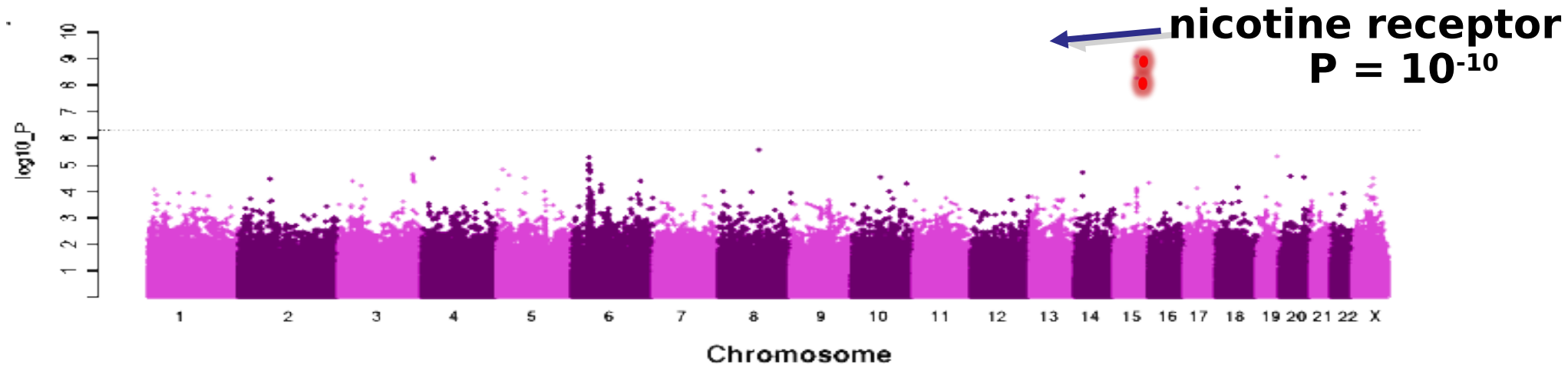
18 Annotations (0 selected) Select:

- CesareanSectionInPriorDelivery
- DiastolicBloodPressure
- DiastolicBloodPressureBefore20W
- Dinoprostone
- EstimatedFetalWeight
- FHREvaluation
- GBSNeonatalSepsisAfterAPrevious
- Gravidity
- HighRiskForAnaphylaxis
- MagnesiumSulfate
- MembranesStatus
- MyastheniaGravis
- PatientAge
- PelvicAdequacy
- PenicillinAllergy
- PreviousCesareanSectionType
- SystolicBloodPressure
- SystolicBloodPressureBefore20We
- TimeStamp
- UrineProtein

MimeType text/  
currentGravidity 3  
day 20  
gate.SourceURL file:/  
month 8  
shift 12



# WHO Genome Wide Association Studies



- $1 \times 10^5$  to  $1 \times 10^6$  SNP markers, 3 to  $7 \times 10^3$  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 ( <u>don't filter keywords</u> )
<u>1138</u>	<p><u>18783506</u>: 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 <b>cigarette</b>; these sensations were associated highly with current <b>smoking</b> (OR = 8.2, P = 0.0001).</p> <p><u>18618000</u>: In the 2,827 long-term smokers examined, common susceptibility and protective haplotypes at the CHRNA5-A3-B4 locus were associated with <b>nicotine dependence</b> severity.</p> <p><u>18759969</u>: Correlated SNPs in the <b>cholinergic</b> nicotinic receptor gene cluster CHRNA5-CHRNA3-CHRNA4, in a case-control study of cocaine <b>dependence</b> composed of 504 European-American and 583 African-Americans.</p> <p><u>18450646</u>: Environmental tobacco smoke and <b>nicotine</b> up-regulated the levels of alpha5 and alpha7 expression in a time-dependent fashion</p> <p><u>18559515</u>: Increased levels of <b>cholinergic</b> receptor nicotinic alpha 5 is associated with squamous cell lung carcinoma</p> <p><u>18957677</u>: The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for <b>nicotine dependence</b> and for lung cancer.(</p> <p><u>18227835</u>: A common haplotype in the CHRNA5/CHRNA3 gene cluster on chromosome 15 contains alleles, which predispose to <b>nicotine dependence</b>.</p> <p><u>19029397</u>: Two distinct variant groups in the CHRNA5-CHRNA3-CHRNA4 gene cluster are strongly associated with heavy <b>smoking</b>. The snp rs16969968 alters the coding sequence of these genes.</p>



# 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

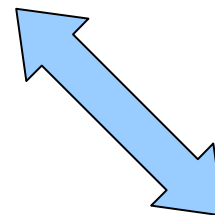
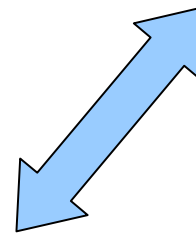
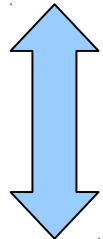


- A framework for language processing
- Open Source – a large community of users and developers
- Mature: over ten years old, currently at version 7.1
- Funded by a mix of EU, UK RC and commercial funding
- The most widely used toolkit of its kind, with 1000s of users at 100s of sites
  - BBC World Cup 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

GATE



# The GATE Team

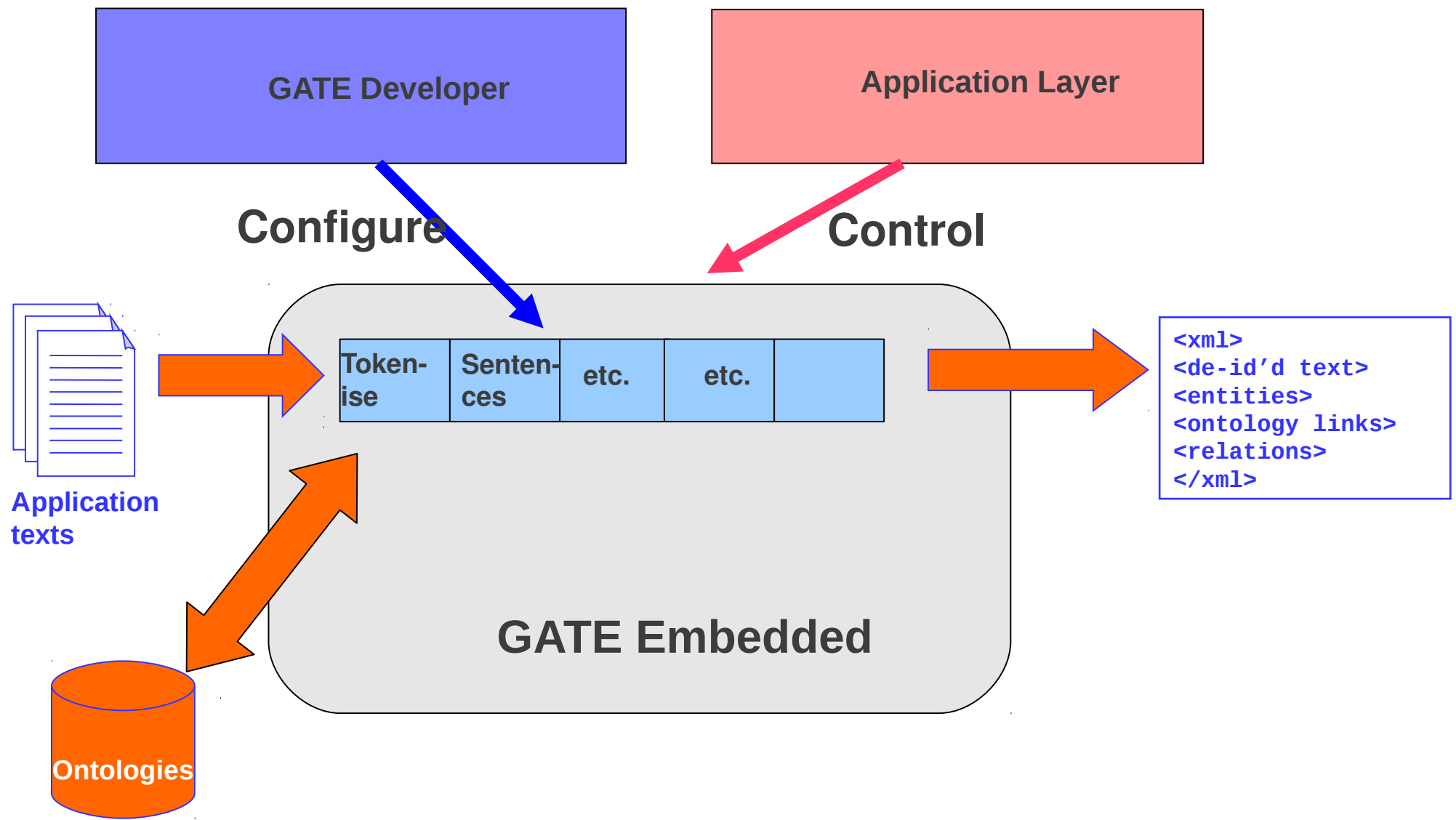


- 
- University of Sheffield
  - Department of Computer Science
    - Natural Language Processing Group
    - 25 academics and researchers, 14 postgrads
    - Links to Information School, speech and hearing
  - GATE Team: 14 researchers

# 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

general architecture

abac defg hijik



0101010101010101

mono rustu x

for text engineering

Thank you.  
Any questions?