

# Bio-YODIE, Mimir and Prospector

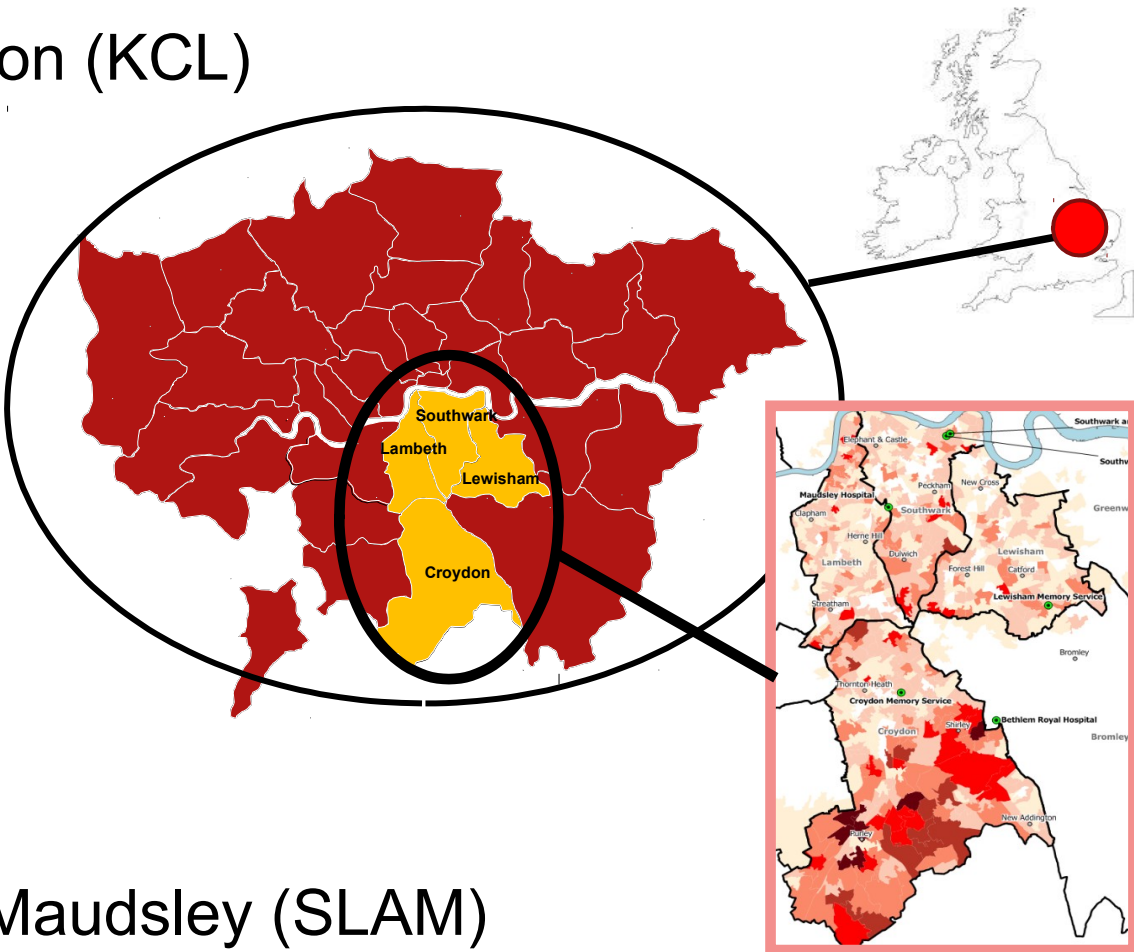
# Overview



- Bio-YODIE for semantic annotation and named entity linking
- Mimir for search
- Prospector for visualization

# SLAM Biomedical Research Centre for Mental Health (BRC-MH)

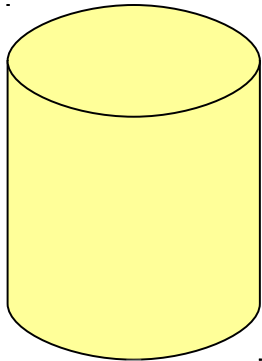
King's College London (KCL)



South London and Maudsley (SLAM)

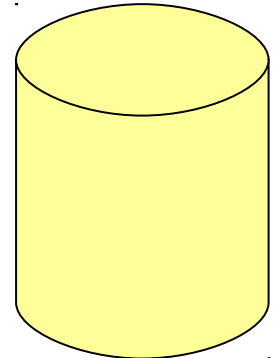
# • Context: EHR and search at SLAM

- The Patient**  
• **Journey**  
• **System (PJS)**



**Coverage:** Four London boroughs  
**Local population:** c. 1.1 million  
**Clinical area:** specialist mental health  
**Active patients:** c. 35 000  
**Total inpatients:** c. 1 000  
**Total records:** c. 250 000 people  
**Documents:** 15 million (D-CRIS: 75 m)

- CRIS**  
• **Interactive search**  
• **FAST index**  
• **SQL RDBMS**



# CRIS and Information Extraction

... finishing A levels .

.. made no eye contact ..

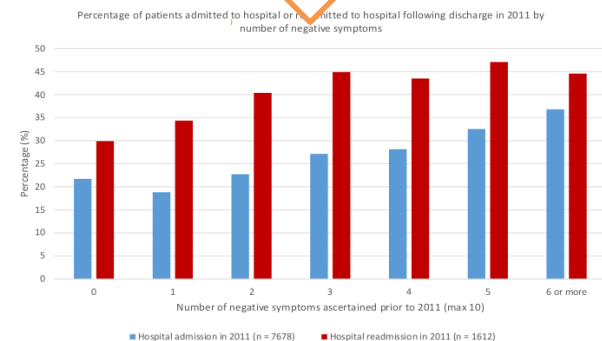
... MMSE=27 on 15 Feb ...

... meals on wheels ..

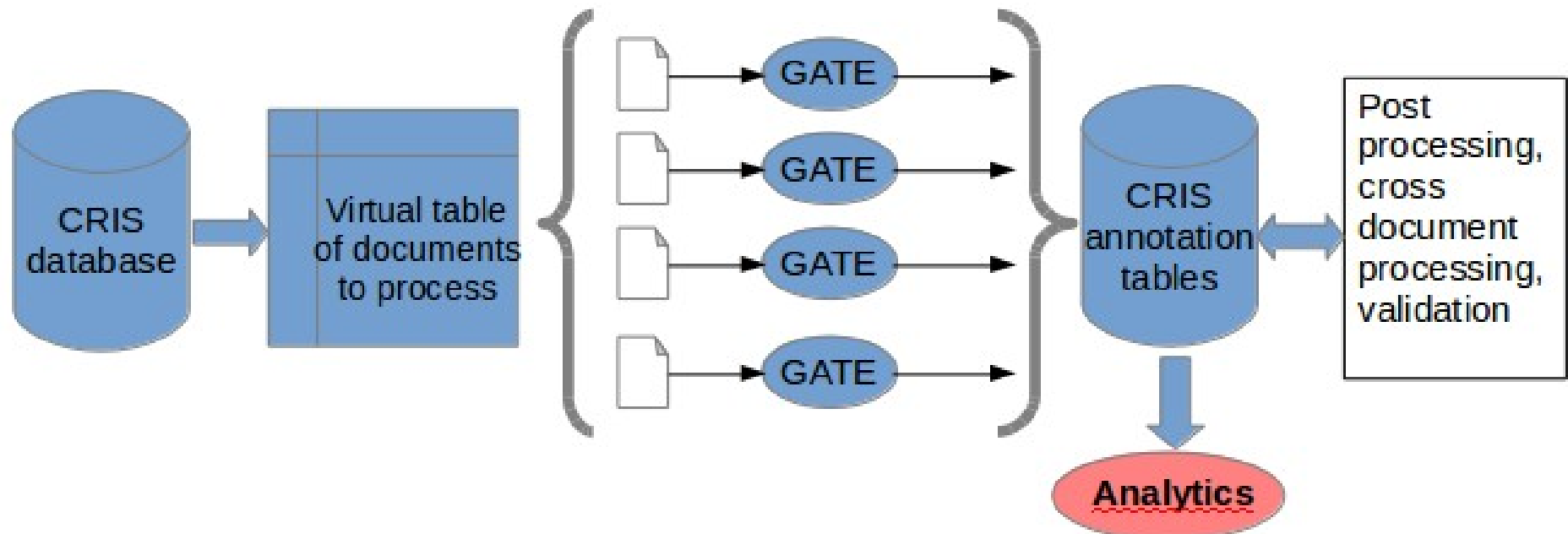
... smokes 20 a day ..



ID	Type	Date	Value
245	MMSE	15/02/16	27/30
8467	Education	12/03/16	2dary
2478	PANSS	15/04/16	
943	Smoking	10/01/16	Current
875	Social care	15/02/16	Current

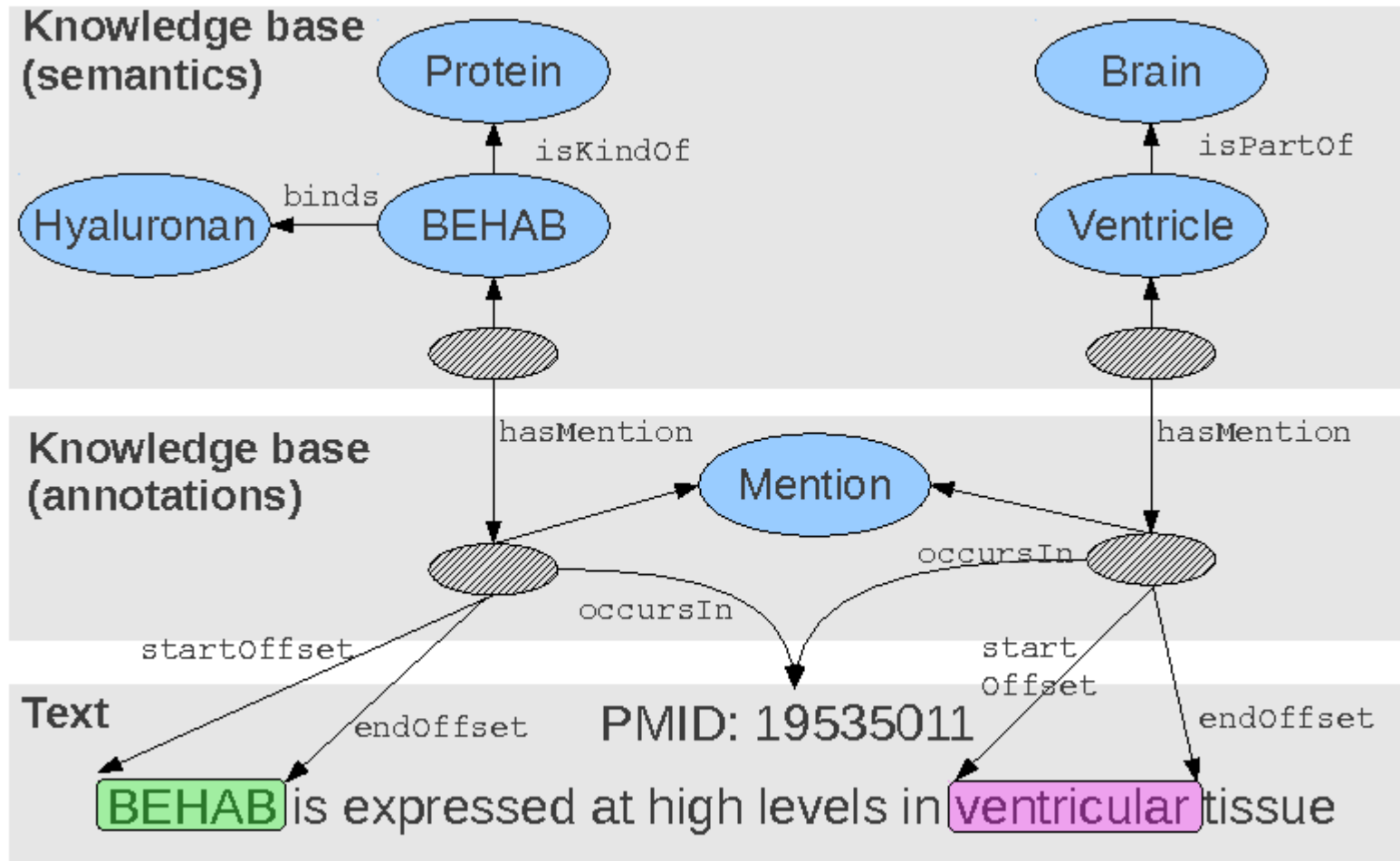


Over 60 GATE applications are routinely run over 15 million patient documents, to extract structured information



Application group	Application sub-group
Symptoms	positive, negative, disorganisation, manic, catatonic, affective, behaviour
Background	cognitive function, social care, living alone, education, smoking, HIV status, physical disorders, illicit substances
Intervention	pharmacotherapy, psychotherapy
Outcomes	trajectories, adverse events

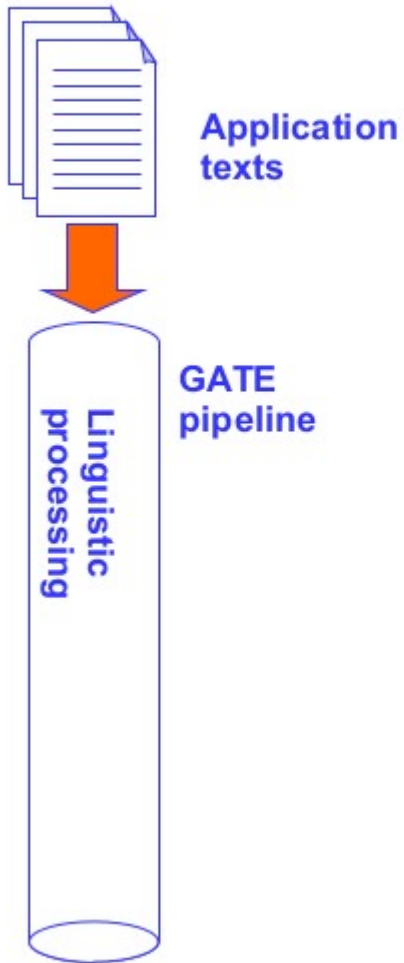
# Why semantic annotation?



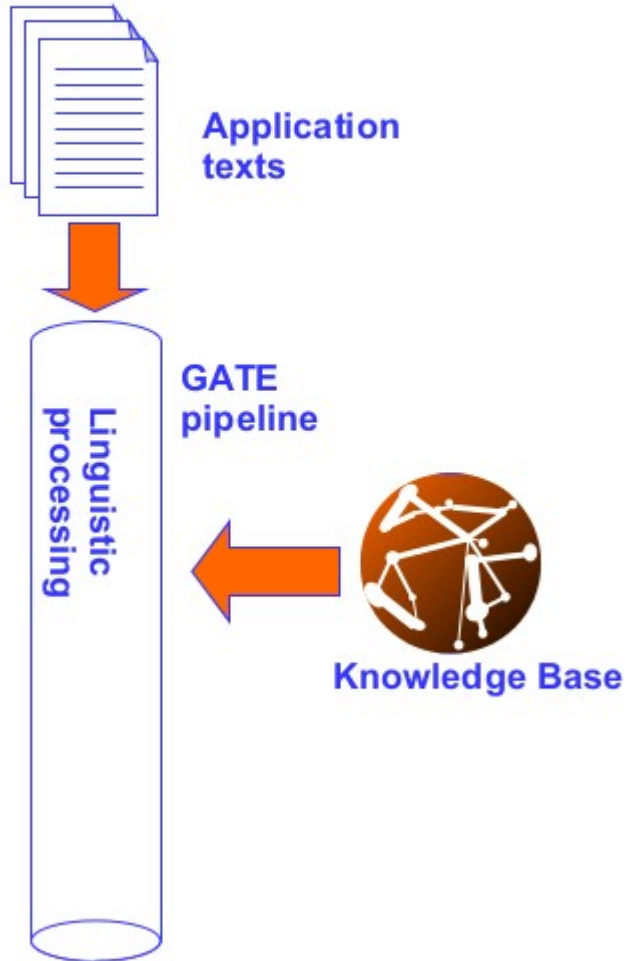
A deeper understanding of text ..



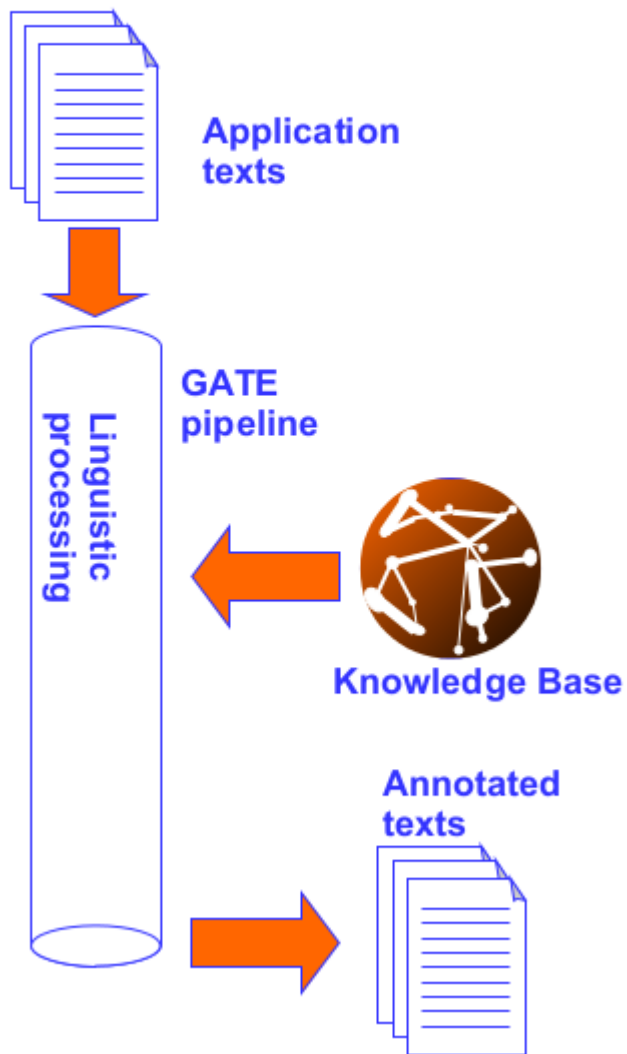
# Linking text and knowledge



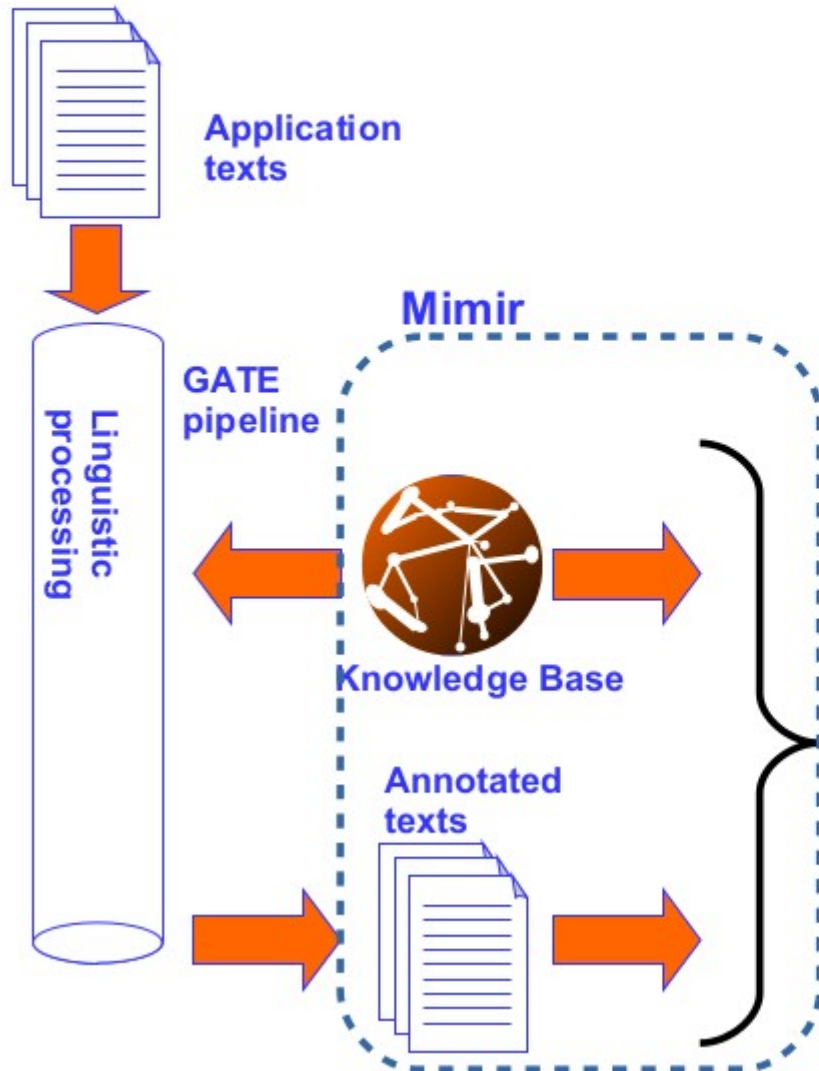
# Linking text and knowledge



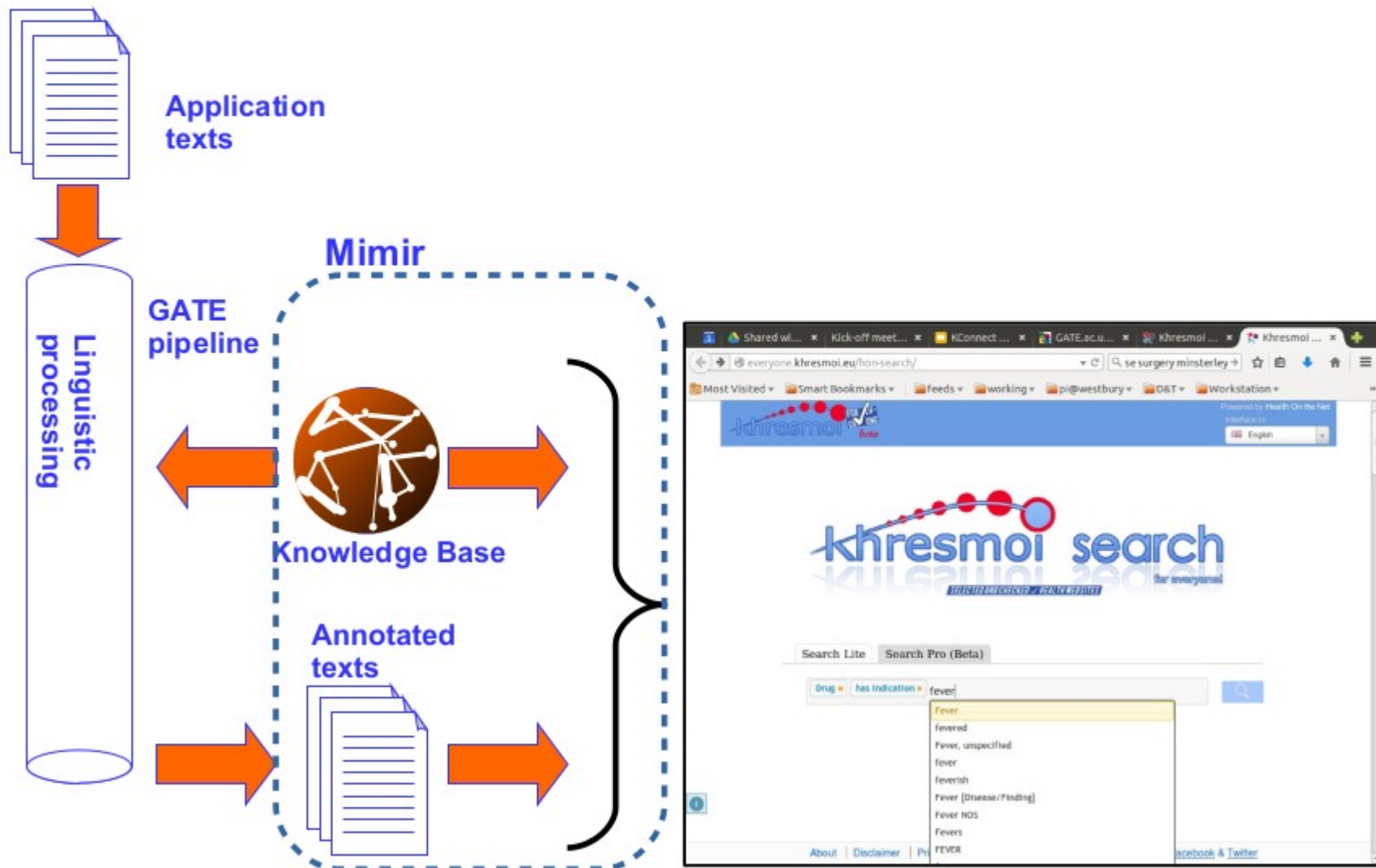
# Linking text and knowledge



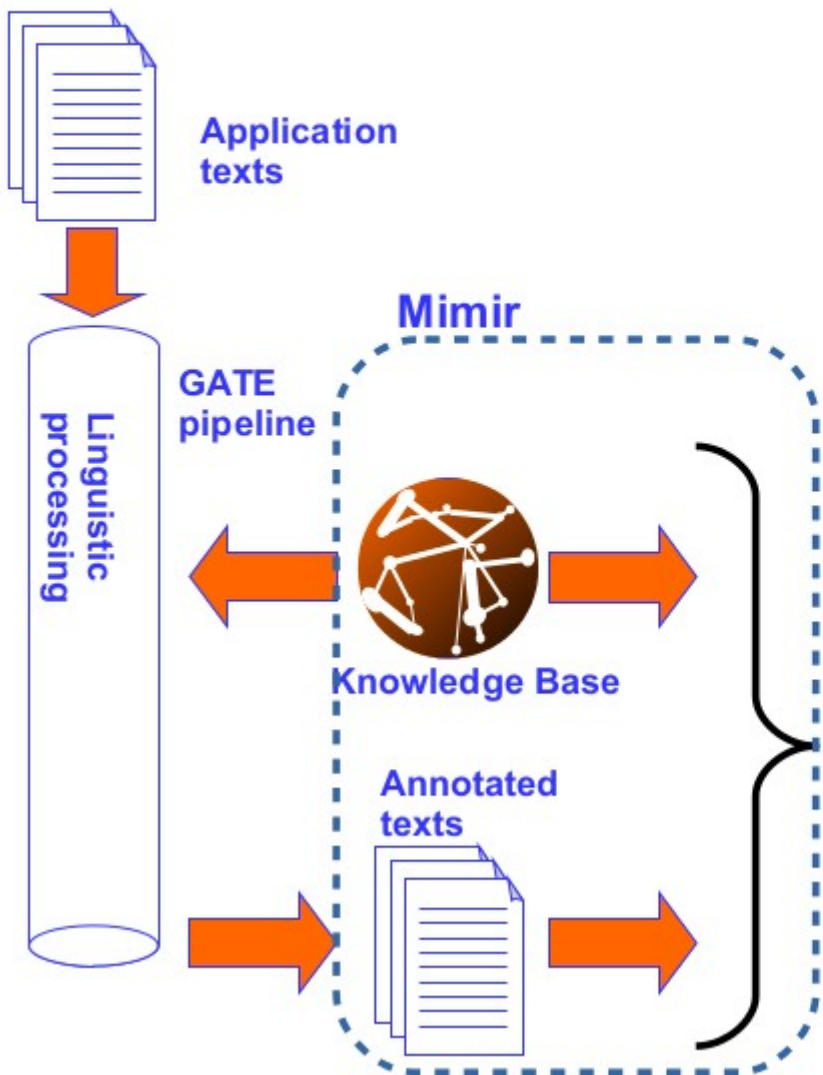
# Linking text and knowledge



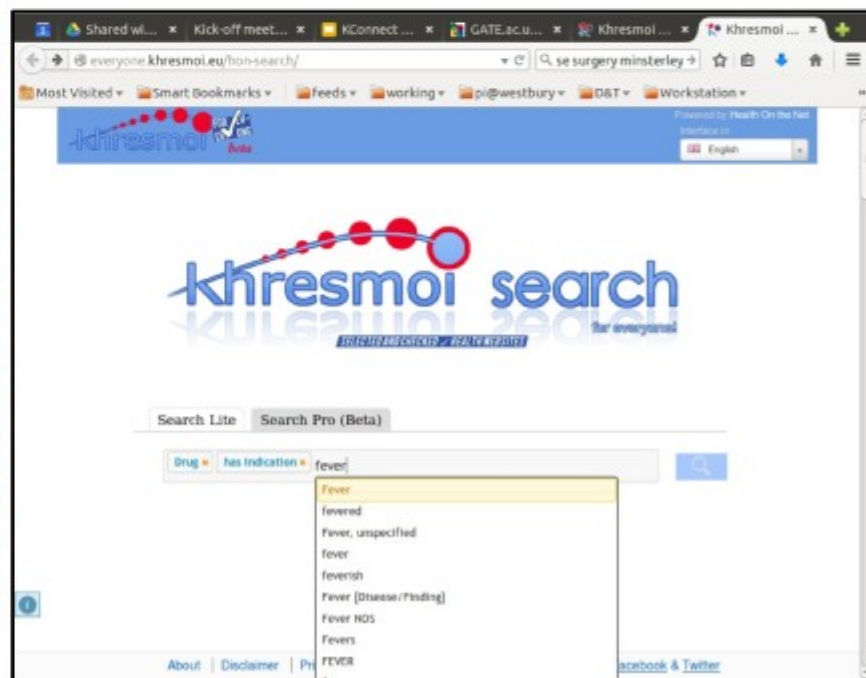
# Linking text and knowledge



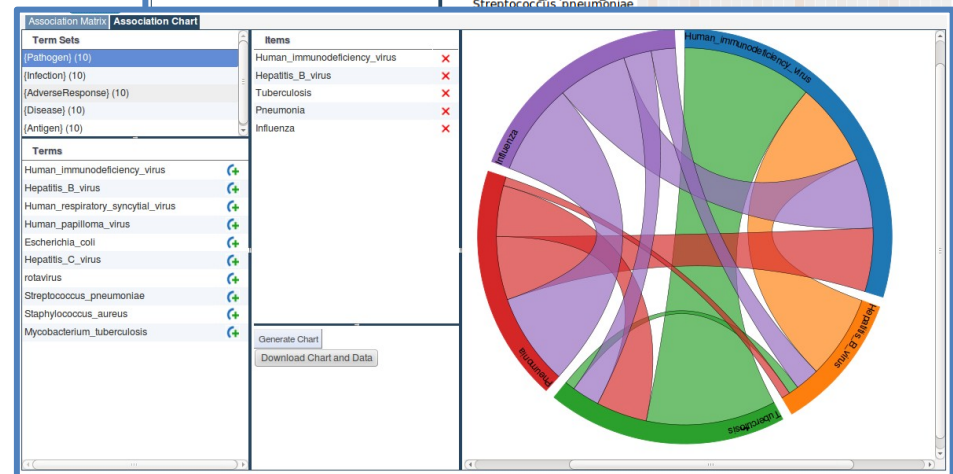
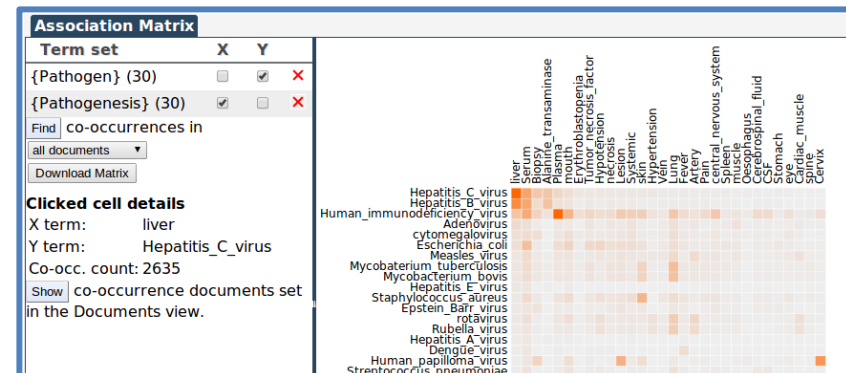
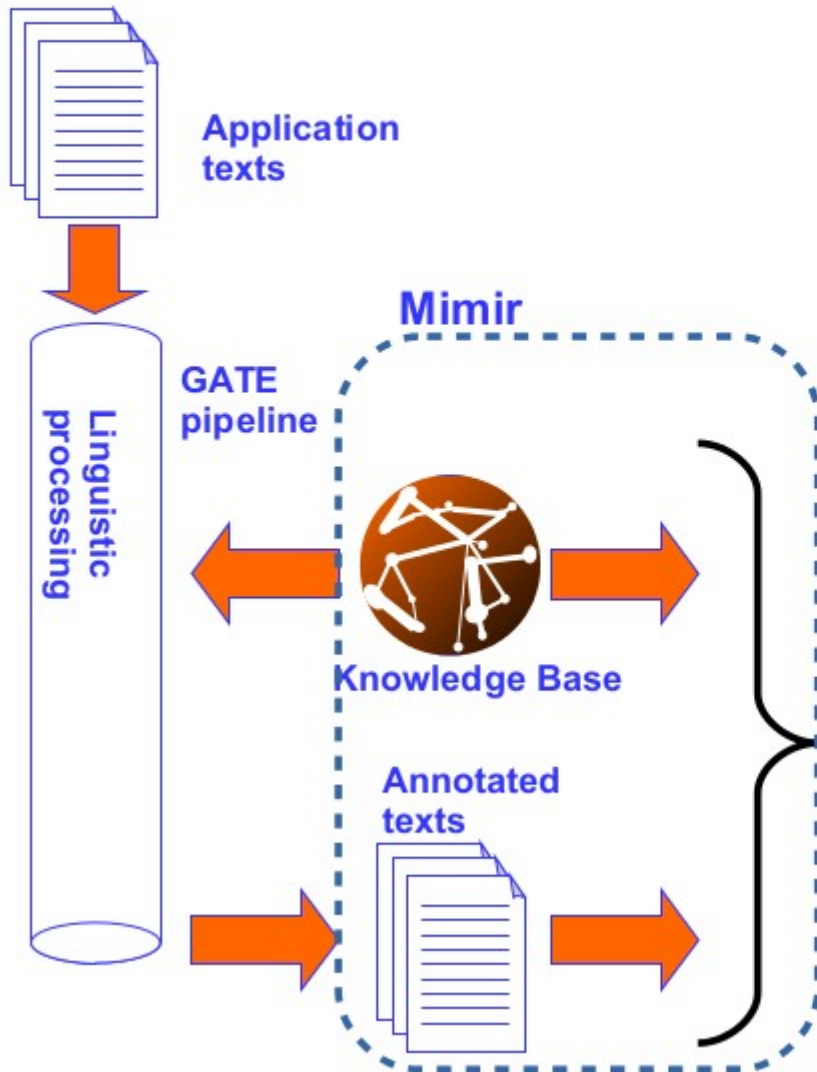
# Linking text and knowledge

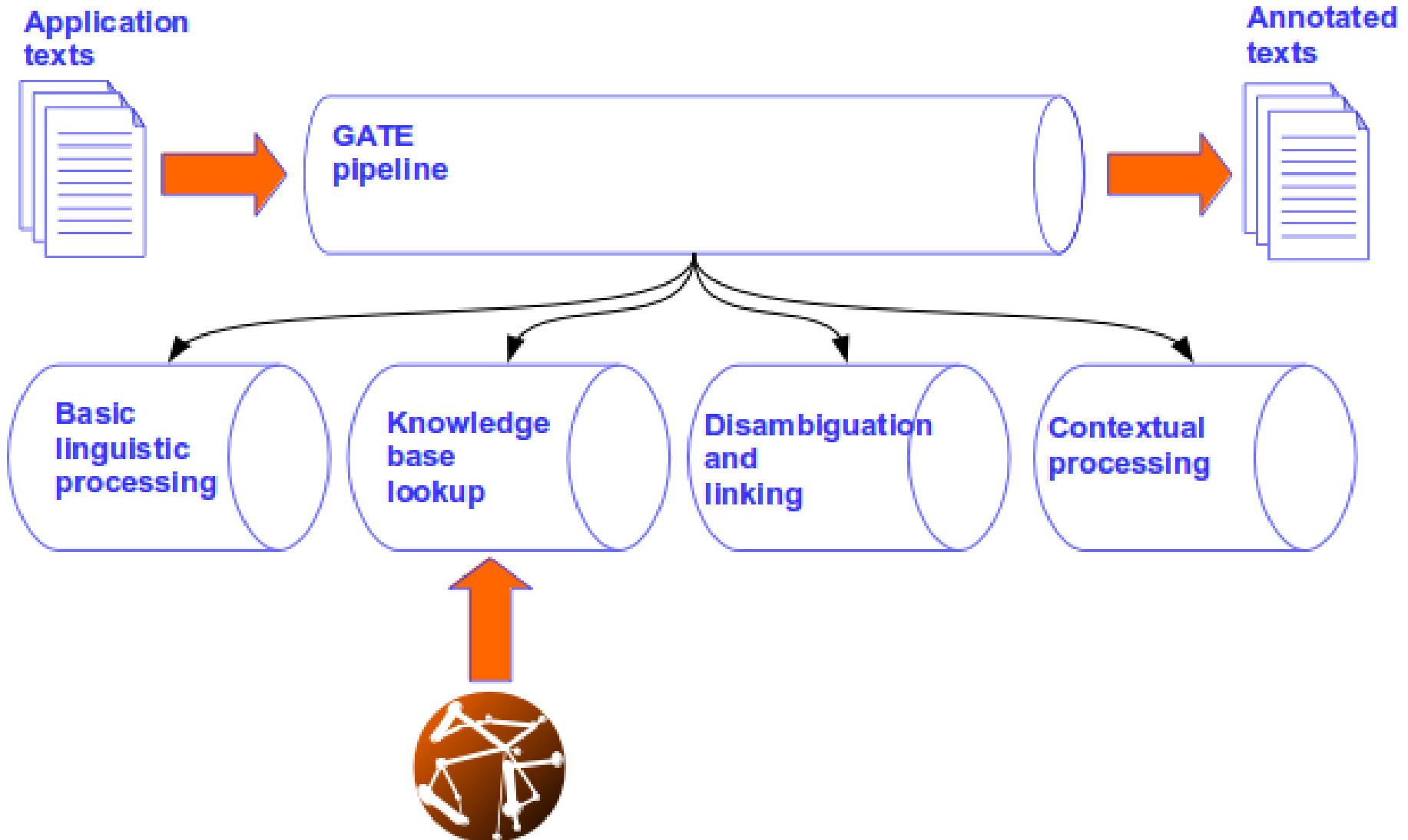


Find patient records that mention drugs used to treat disorders commonly associated with depression



# Linking text and knowledge







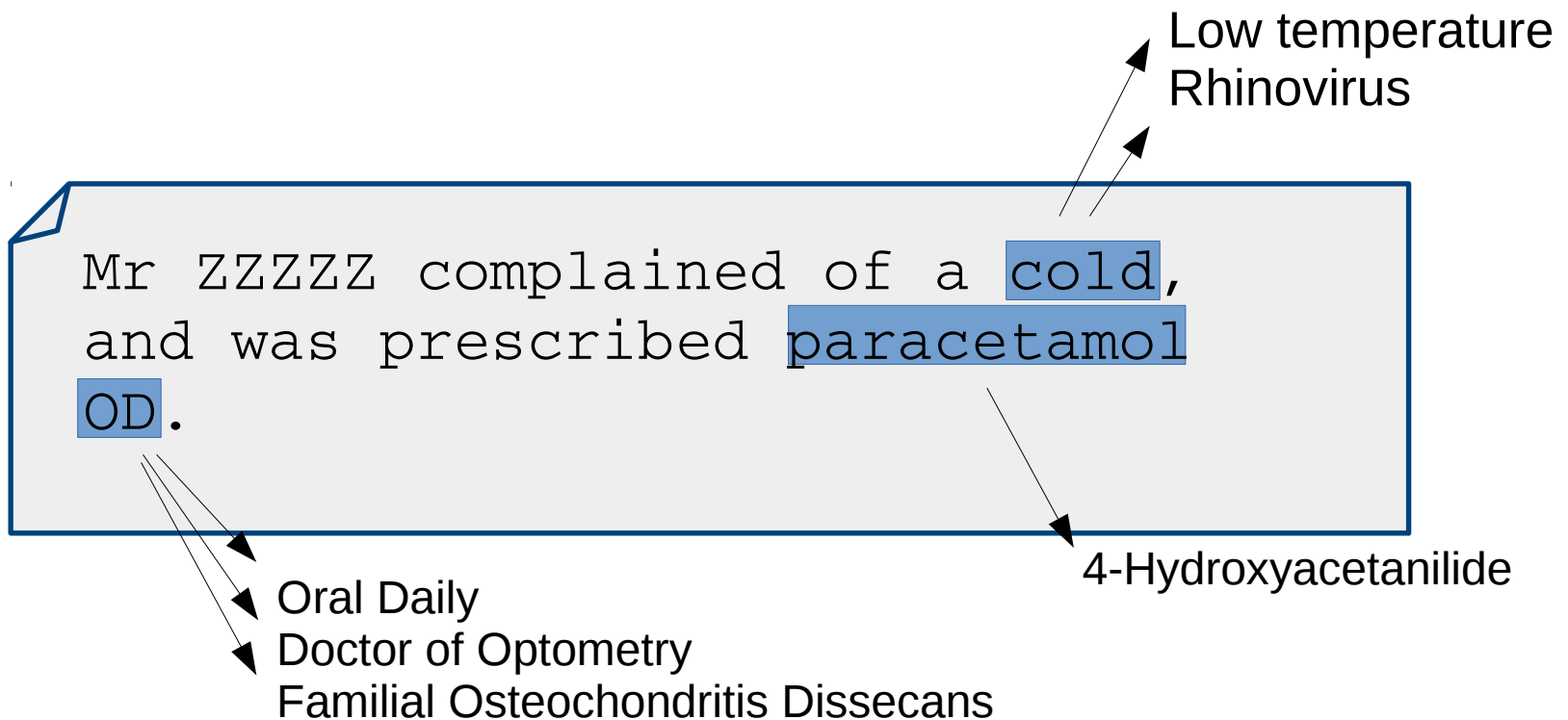
# Named Entity Linking

- A central task in semantic annotation
- Connect mentions in text to their referent in a knowledge base, so we can find out about them and coreference them
- AKA term identification and mapping, or normalization

Mr ZZZZZ complained of a cold,  
and was prescribed paracetamol  
OD.

# Named Entity Linking

- You have found mentions in text, but there are several possibilities for what something is referring to
- This is the disambiguation problem



- Dictionary based term identification based on names of UMLS entities of relevant types
- Retrieval of possible interpretations—any entity that can be called that
- Multiple disambiguation strategies gauging how well the candidates fit
- Scores weighted based on a supervised model to choose the best

- Unified Medical Language System (from the US National Library of Medicine)
- Contains over two million biomedical and health-related concepts
- Combines many thesauri (“source vocabularies”)
- Much ambiguity! E.g. “unknown” is associated with 39 concepts
- The NLP “view” (subset) is more pragmatic, excluding unhelpfully ambiguous and spurious concepts

**Search** | Tree | Recent Searches

Term  CUI  Code

Release:

Search Type:

Source:   
AIR  
ALT  
AOD  
AOT

**Search Results (6)**

- [C0009264](#) Cold Temperature
- [C0009443](#) Common Cold
- [C0010412](#) Cold Therapy
- [C0024117](#) Chronic Obstructive Airway Disease
- [C0234192](#) Cold Sensation
- [C0719425](#) Cold brand of chlorpheniramine-phenylpropanolamine

- “Cold” has six associated concepts
- “CUI” means concept unique identifier and is the unique code for the concept

- A concept has a type (in this case, “Disease or Syndrome”)
- It has definitions from different vocabularies
- It also has concept relations, i.e. things it is related to. “Common Cold” is related to “Respiration Disorders” for example

Basic View **Report View** Raw View

⊕ Concept: [C0009443] Common Cold

⊖ Semantic Types

[Disease or Syndrome](#) [T047]

⊖ Definitions

CSP/PT | catarrhal disorder of the upper respiratory tract, which may be viral or a mixed infection; marked by temperature, chilly sensations, and general indisposition.

MEDLINEPLUS/PT |

Sneezing, [sore throat](#), a stuffy nose, coughing - everyone knows the symptoms of the common cold. It is probably the course of a year, people in the United States suffer 1 billion colds.

You can get a cold by touching your eyes or nose after you touch surfaces with cold germs on them. You can usually begin 2 or 3 days after infection and last 2 to 14 days. Washing your hands and staying away from people with colds.

There is no cure for the common cold. For relief, try

- Getting plenty of rest
- Drinking fluids
- Gargling with warm salt water
- Using cough drops or throat sprays
- Taking over-the-counter pain or [cold medicines](#)

However, do not give aspirin to children. And do not give cough medicine to children under four.

NIH: National Institute of Allergy and Infectious Diseases

MSH/MH | A catarrhal disorder of the upper respiratory tract, which may be viral or a mixed infection. It generally

# Bio-YODIE: Finding Mentions in Text

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Mrs ZZZZ was seen today and appears brighter in mood and to be feeling positive. She's continuing with Sertraline and self help.

Type	Set	Start	End	Id	Features
Lookup		49	53	412	{Experiencer=Patient, Negation=Affirmed, PREF=Mood (psychological functi
Lookup		49	53	413	{Experiencer=Patient, Negation=Affirmed, PREF=Mood:-:Point in time:^Patien
Lookup		64	71	410	{Experiencer=Patient, Negation=Affirmed, PREF=Feelings, STY=Mental Proce
Lookup		64	71	409	{Experiencer=Patient, Negation=Affirmed, PREF=Emotions, STY=Mental Pro
Lookup		72	80	406	{Experiencer=Patient, Negation=Affirmed, PREF=Positive Finding, STY=Findi
Lookup		72	80	407	{Experiencer=Patient, Negation=Affirmed, PREF=Positive, STY=Qualitative C
Lookup		72	80	404	{Experiencer=Patient, Negation=Affirmed, PREF=Positive Number, STY=Conc
Lookup		72	80	405	{Experiencer=Patient, Negation=Affirmed, PREF=Positive Charge, STY=Quali
Lookup		104	114	402	{Experiencer=Patient, Negation=Affirmed, PREF=SERTRALINE, STY=Pharmac

- Date
- Lookup
- LookupList
- MISC
- NounChunk
- PERSON
- Person
- Sentence
- SpaceToken
- Split
- Token
- DEBUG\_remo
- GazetteerEN
- Shef
- debug
- deleted-prep

# Bio-YODIE disambiguation scoring strategies



<b>Strategy</b>	<b>Implementation</b>	<b>Example</b>
<b>Prior likelihood of entity mentions</b>	MeSH term frequency in psychiatry journals	For “OD”, the [overdose] entity is more common than [osteocondritis dissecans]
<b>Ranked likelihood of co-occurrence</b>	PageRank across a UMLS co-occurrence table	[overdose] is more likely to be mentioned with [depression] than is [osteocondritis dissecans]
<b>Context similarity to entity definitions</b>	UMLS definitions table	The context of “OD” is more likely to be close to the context of the [overdose] definition
<b>Count of direct and indirect relations between entities</b>	UMLS relations table	[overdose] is likely to have more links to [depression] than [osteocondritis dissecans] will have



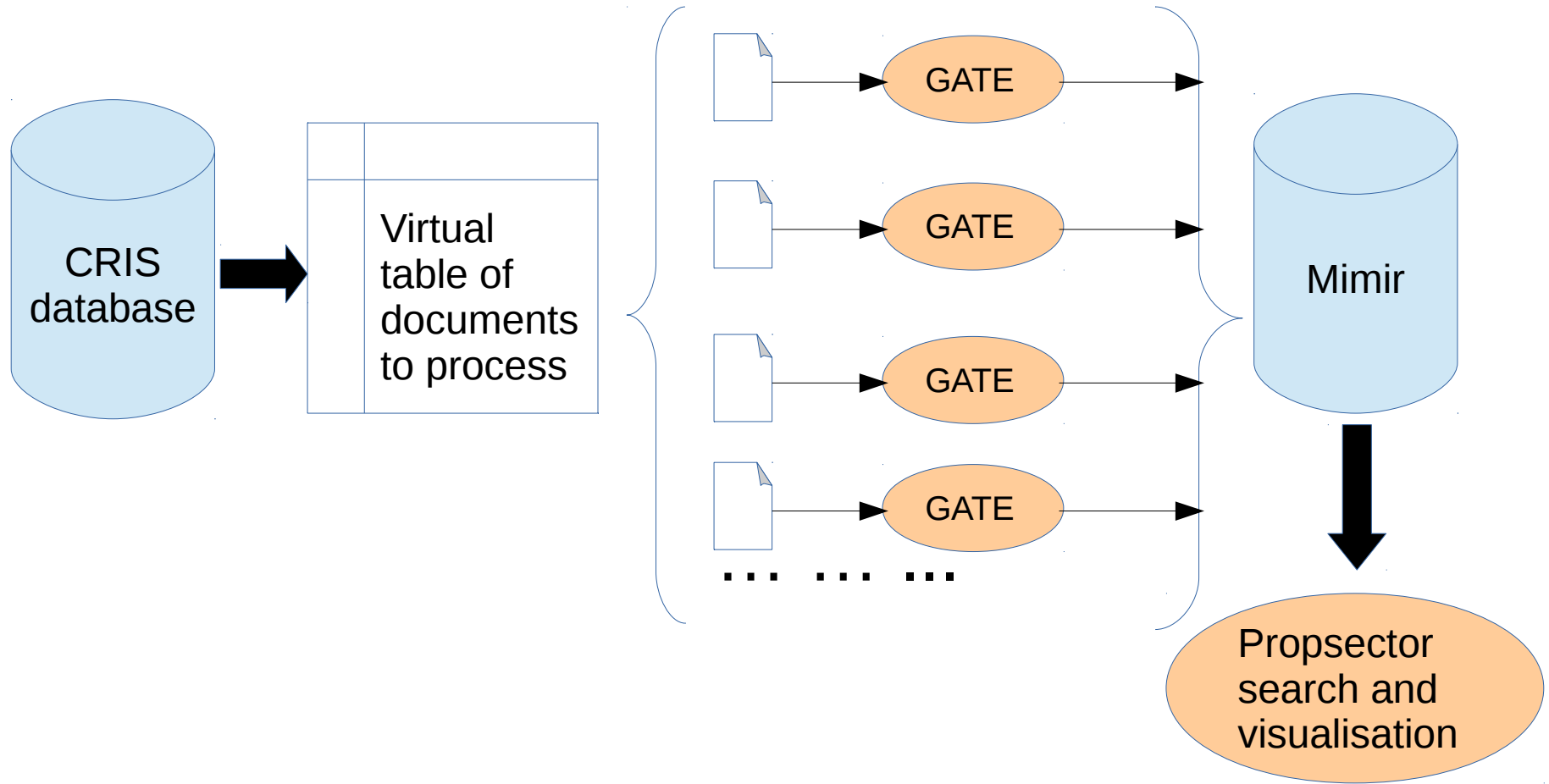
# Preliminary results

	Total correct	Correct rejections	Correct accepted candidates	Accuracy
Always accept one candidate at random	410	0	410	0.50
Accept one random candidate 63% of the time, reject all candidates otherwise	358	100	258	0.43
Bio-YODIE, reject if all scores are zero	488	82	406	0.59
Bio-YODIE, rejection threshold 0.05	435	142	292	0.53

Results obtained on 826 mention manually annotated CRIS corpus

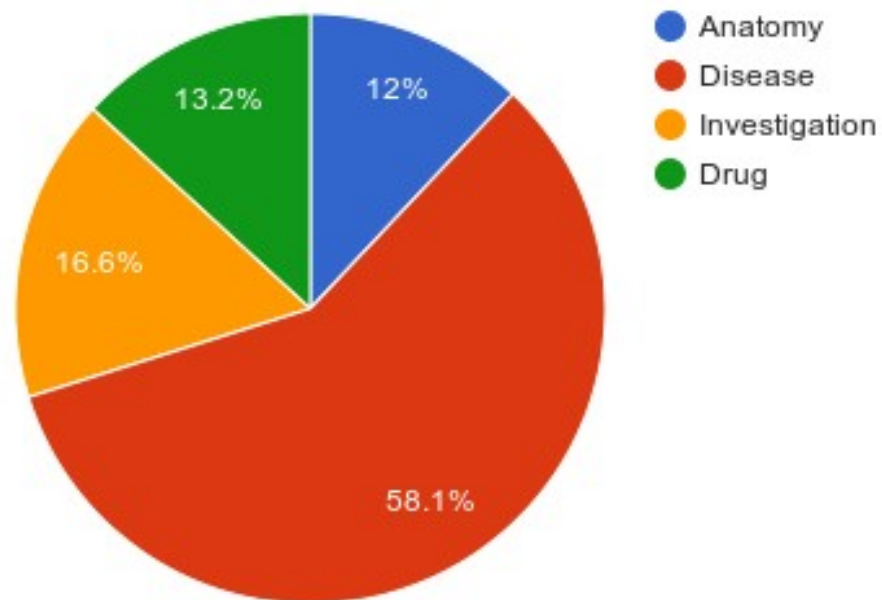
- Context—included
  - Based on the popular NegEx algorithm
  - Negation
  - Subject
  - Historicity
- Time - temporal expression extraction—still to be included
  - Supervised models (SVM)
  - Pattern matching grammars for imprecise times
- Language - now available for French

# Deployment and indexing



# Mimir index statistics

Mentions by Type



Proportion of types found on 50K CRIS test documents (patient records)

- Using GCP, we (USFD and KCL) have indexed ...

15 million CRIS documents

4 million Trip documents

- Mimir provides a semantic search interface allowing complex queries
  - E.g. find mentions of mood disorders occurring in the same patient record as heart conditions
- Prospector is Mimir's visualization layer
  - Visualize relationship between heart conditions and mood disorders
- Integrates D3 to allow visualization of complex semantic queries

# Prospector visualisations

Documents **Terms**

Select top 30 terms with type: {Pathogen} from retrieved documents ranked: <All>

Term	Count
Human_imn	762
Hepatitis_B_	632
Hepatitis_C_	515
Human_pap	260
Human_resp	93
Hepatitis_A_	57
cytomegalo	57
Haemophilu	57
Mycobacter	44
Mycobaterit	44
Hepatitis_E	35

Download Term Set  
Save Term Set with the name: {Pathogenesis} (30)

**Saved term sets**

- {Pathogen} (30)
- {Pathogenesis} (30)

# Prospector visualisations



Documents **Terms**

Select top 30 terms with type: {Pathogen} from retrieved documents ranked: <All>

Term	Count
Human_imn	762
Hepatitis_B_	632
Hepatitis_C_	515
Human_pap	260
Human_resf	93
Hepatitis_A_	57
cytomegalo	57
Haemophilu	57
Mycobacter	44
Mycobaterit	44
Hepatitis_E	35

Download Term Set  
Save Term Set with the name: {Pathogenesis} (30)

**Association Matrix**

Term set	X	Y
{Pathogen} (30)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
{Pathogenesis} (30)	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Find co-occurrences in  
all documents  
Download Matrix

**Clicked cell details**  
X term: liver  
Y term: Hepatitis\_C\_virus  
Co-occ. count: 2635  
Show co-occurrence documents set in the Documents view.

# Prospector visualisations

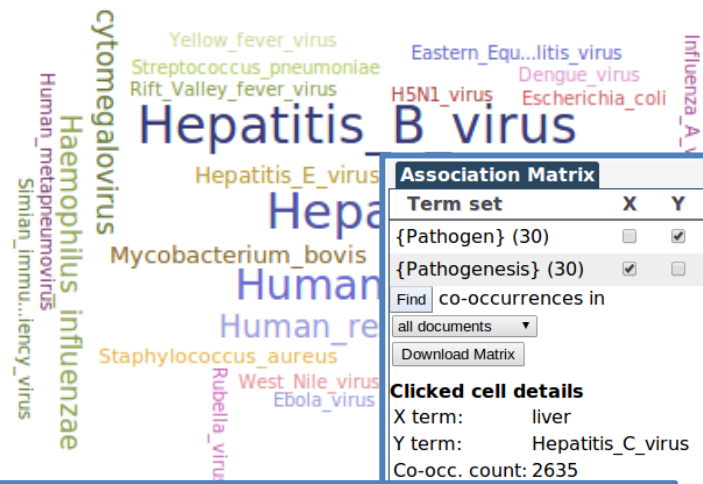


Documents **Terms**

Select top 30 terms with type: {Pathogen} from retrieved documents ranked: <All>

Term	Count
Human_imn	762
Hepatitis_B	632
Hepatitis_C	515
Human_pap	260
Human_resp	93
Hepatitis_A	57
cytomegalo	57
Haemophilu	57
Mycobacter	44

Download Term Set  
Save Term Set with the name: {Pathogenesis} (30)



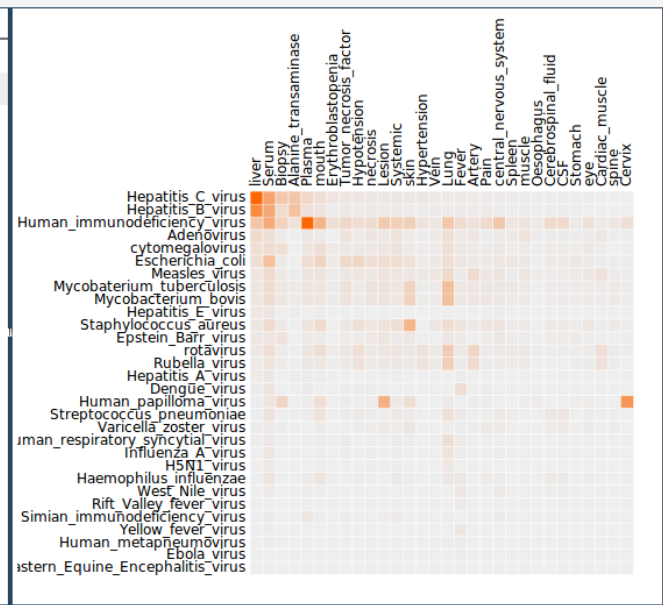
**Association Matrix**

Term set	X	Y
{Pathogen} (30)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
{Pathogenesis} (30)	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Find co-occurrences in all documents

Download Matrix

**Clicked cell details**  
X term: liver  
Y term: Hepatitis\_C\_virus  
Co-occ. count: 2635



**Association Chart**

Term Sets	Items
{Pathogen} (10)	Human immunodeficiency_virus
{Infection} (10)	Hepatitis_B_virus
{AdverseResponse} (10)	Tuberculosis
{Disease} (10)	Pneumonia
{Antigen} (10)	Influenza

Terms: Human immunodeficiency\_virus, Hepatitis\_B\_virus, Human\_respiratory\_syncytial\_virus, Human\_papilloma\_virus, Escherichia\_coli, Hepatitis\_C\_virus, rotavirus, Streptococcus\_pneumoniae, Staphylococcus\_aureus, Mycobacterium\_tuberculosis

Generate Chart  
Download Chart and Data

Circular chord diagram showing relationships between terms. The diagram is divided into segments for Influenza, Tuberculosis, and Hepatitis B virus, with chords connecting them to represent associations.



