

# Natural language processing for healthcare

## Motivation

Angus Roberts, Senior Lecturer in Health Informatics  
Institute of Psychiatry, Psychology and Neuroscience  
King's College London

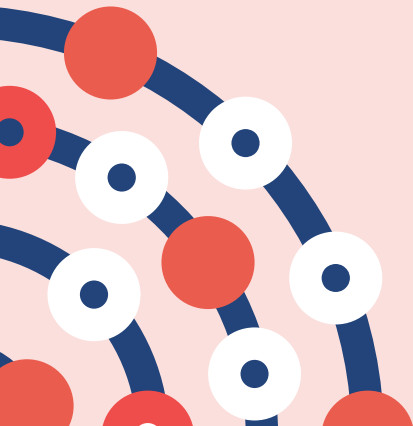


# Contents

- Background and motivation
- Free text in the Electronic Health Record (EHR)
- Tackling text with Natural Language Processing (NLP)
- Information extraction
- Tools for the job



# Background and motivation



# Motivation

- Thomas Willis in *Dr. Willis's Practice of Physick Being All the Medical Works of That Renowned and Famous Physician. 1684.*
  - weigh all the symptoms, and to put them, with exact Diaries of the Diseases, into writing; then diligently to meditate on these, and to compare some with others; and then [begin] to adopt general Notions from particular Events
- Reuse of the medical record
- Computerisation of the record allows us to magnify the efforts of Willis by many degrees

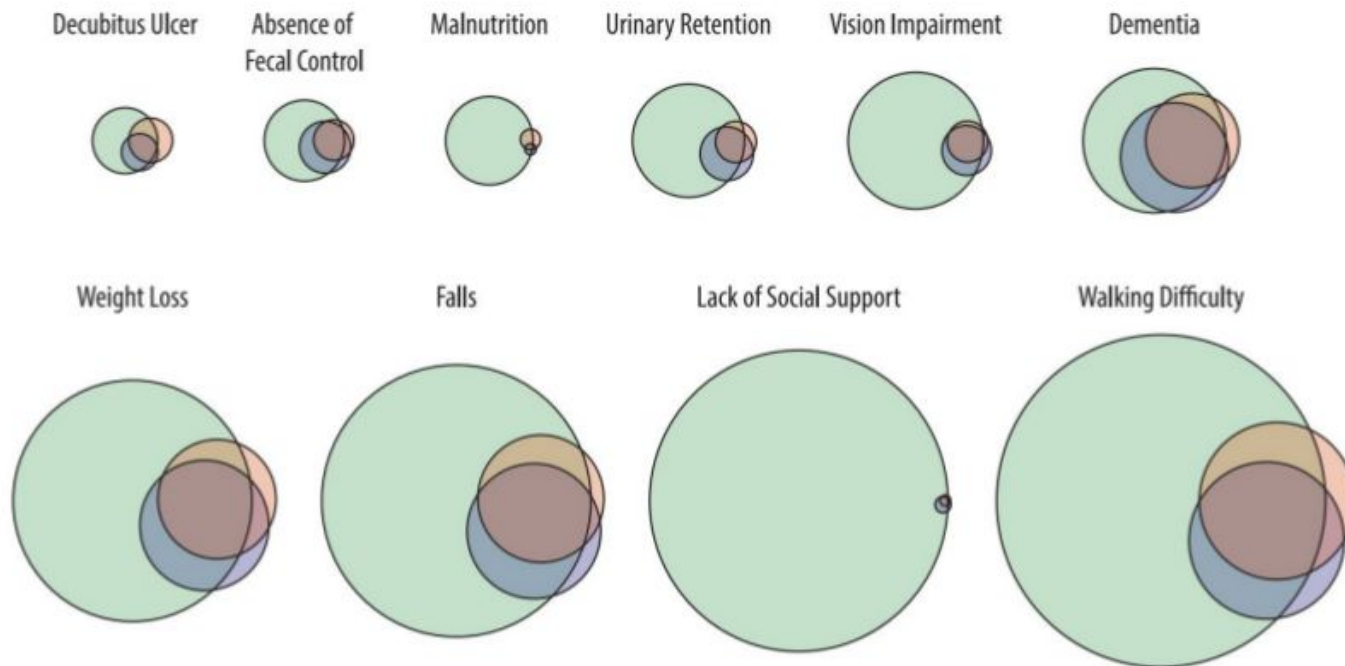
# EHRs vs traditional studies

- Clinical cohort studies...  
*expensive and time consuming, especially for rarer disorders, to recruit and follow enough people*
- RCTs / experimental studies...  
*Often cannot do these for ethical reasons*
- Relapse of mental disorder...  
*Lack of patient capacity to consent*
- Allows us to include...  
*Representative samples, patients who may not be able to take part in clinical studies – e.g. suicidal patients, very sick and marginalised*



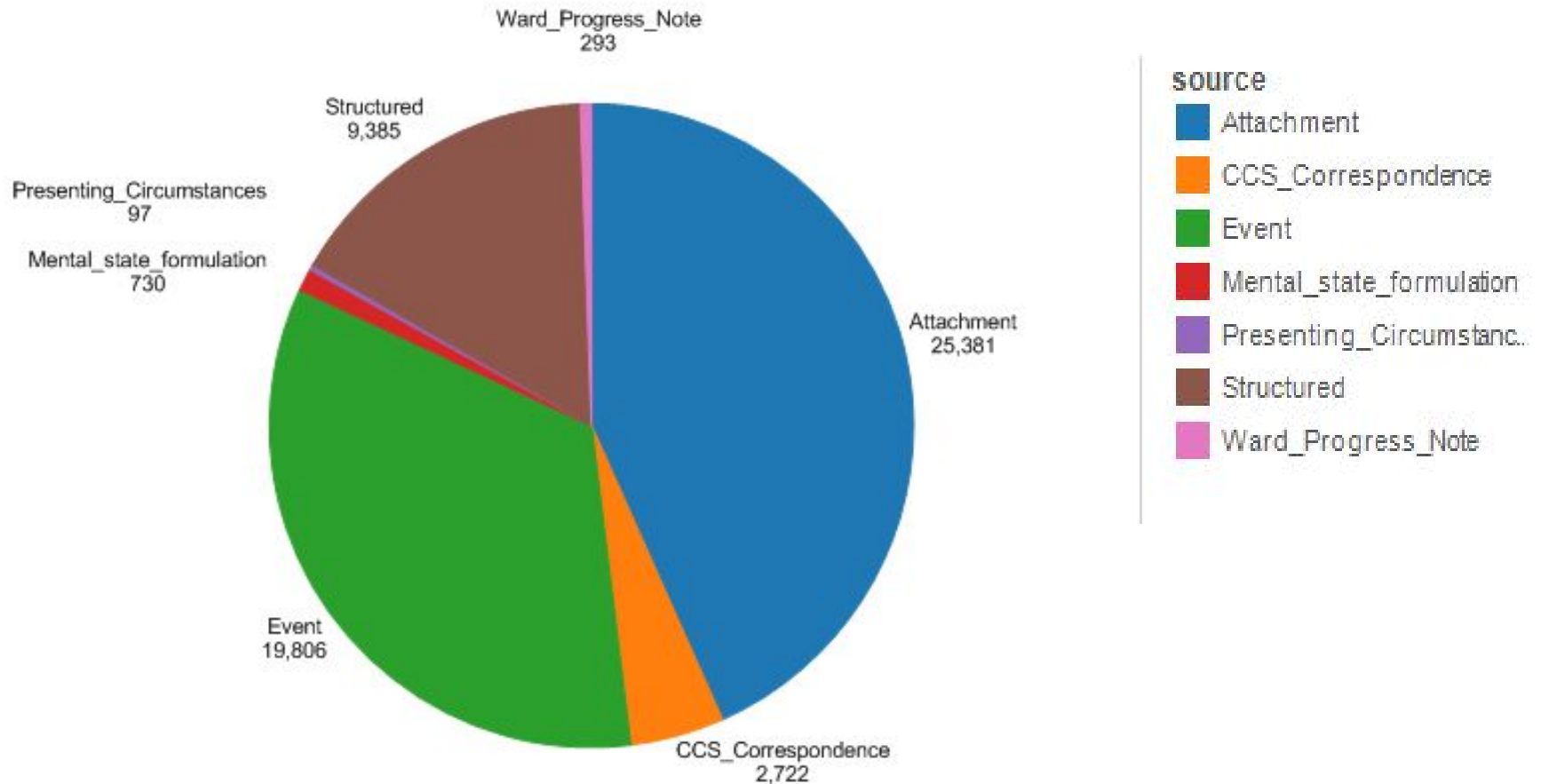
# Free text in the electronic health record (EHR)





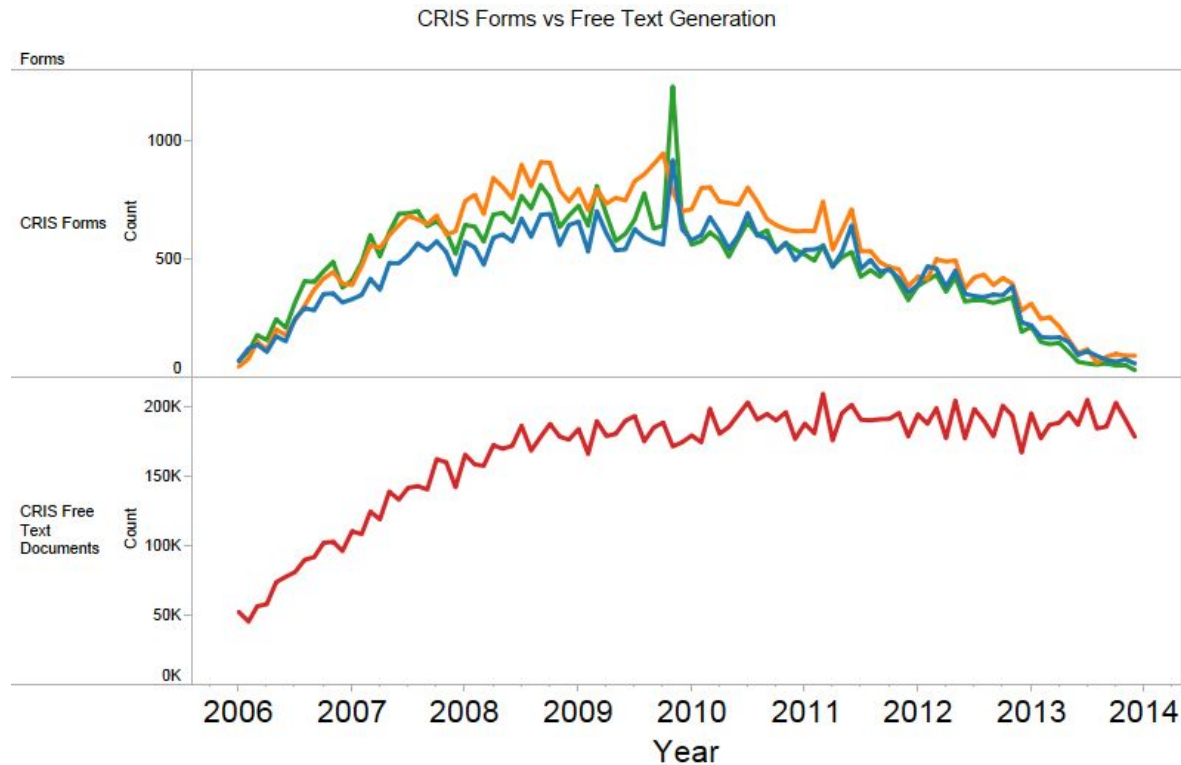
**Figure 2.** Green: Unstructured free text EHR data; Other colours: structured data. “The value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification”. (Kharrazi et al., 2018)

# MMSE: text vs structured data





# Structured data capture is 'unpopular'



The trend of count of CN\_Doc\_ID for date Month broken down by Forms. Color shows details about event\_category. The data is filtered on date, which includes the last 8 years relative to 31/12/2013. The filter associated with this field ranges from 01/01/2006 to 31/12/2013.

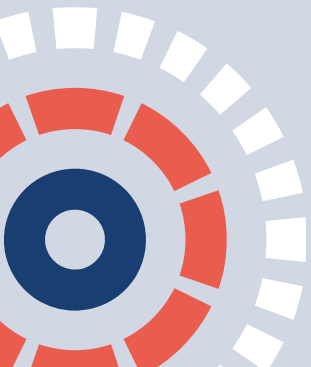
event\_category  
History  
Mental State Formulation  
Presenting Circumstances  
Text\_Document

# The problem with language

<b>Ambiguity</b>	<ul style="list-style-type: none"><li>• To some degree</li><li>• Finishing her degree</li></ul>
<b>Grammatical subject</b>	<ul style="list-style-type: none"><li>• She smokes</li><li>• Her mother smokes</li></ul>
<b>Hedging</b>	Probably a possible tumour
<b>Negation</b>	He no longer smokes
<b>Synonymy, abbreviations, acronyms</b>	MMSE, Mini Mental State Exam, Folstein
<b>Extracting patterns</b>	<ul style="list-style-type: none"><li>• 17<sup>th</sup> percentile</li><li>• 5 weeks and 3 days</li></ul>
<b>Relationships and events</b>	His MMSE was 24/30 a week before the appointment



# Natural language processing (NLP)



# NLP origins

- 50s and 60s
  - Machine Translation
    - Russian – English
    - Hand coded rules and large dictionaries
- Late 60s to late 70s
  - Failure to deliver and cut in funding
- Mid 80s to mid 90s
  - Competitive challenges, e.g. Message Understanding Conferences
    - Organised by Naval Command, Control and Ocean Surveillance Center
    - Originally, extraction of information from military messages
    - Moved on to news reports on terrorist attacks, and trade union disputes...
    - Largely rule based systems

# Natural Language Processing

- **a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications**

*(Liddy, in Encyclopedia of Library and Information Science, 2nd edition, 2003. page 137)*

# Natural Language Processing

- a theoretically motivated range of computational techniques for **analyzing** and **representing** naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications

*(Liddy, in Encyclopedia of Library and Information Science, 2nd edition, 2003. page 137)*

# Natural Language Processing

- a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more **levels of linguistic analysis** for the purpose of achieving human-like language processing for a **range of tasks or applications**

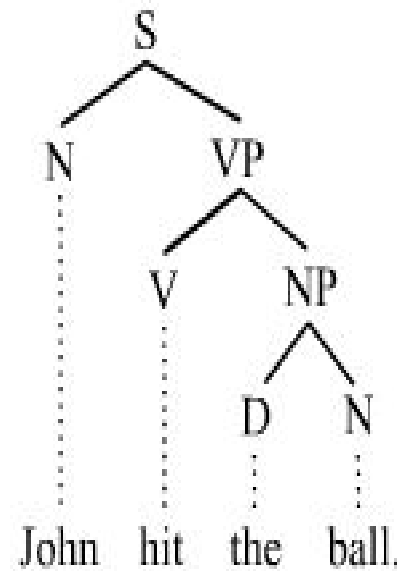
*(Liddy, in Encyclopedia of Library and Information Science, 2nd edition, 2003. page 137)*

# Levels of linguistic analysis

## Lexicon - the words

- A foul and pestilent congregation of vapours.
- 1.30pm: Cx: 3cm. Contractions q2-3 min. FHR: reassuring.

## Syntax - the grammatical structure





# Levels of linguistic analysis

## Semantics - meaning

- Expression of atrial natriuretic factor gene in ventricular tissue
- BEHAB expression in ventricular tissue

## Pragmatics - context

- I saw this 12 year old girl in clinic today with her mother. She is morbidly obese.

# Levels of linguistic analysis

This lady attended outpatients today. In 1984 she had a right simple mastectomy of a carcinoma of the breast and was commenced on Tamoxifen. There was no sign of tumour recurrence on follow up.

Her new symptoms are of lymphoedema in the right arm which has developed over the last six weeks. She has also complained of pain in the right hip. I note her recent FBC was normal.

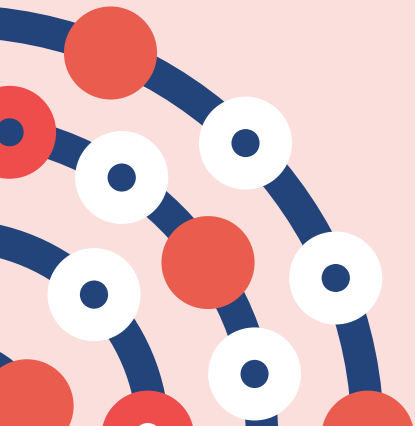
I have taken the precaution of doing an X-ray of the pelvis and given her a tubigrip bandage to use for the lymphoedema in her arm. We plan to see her again in two weeks time with the result of the X-ray.

# Many applications...

- Search - information retrieval
- Information extraction
- Question answering
- Document summarisation
- Dialogue
- Machine translation
- Document classification
- Social media tracking
- ....



# Information extraction



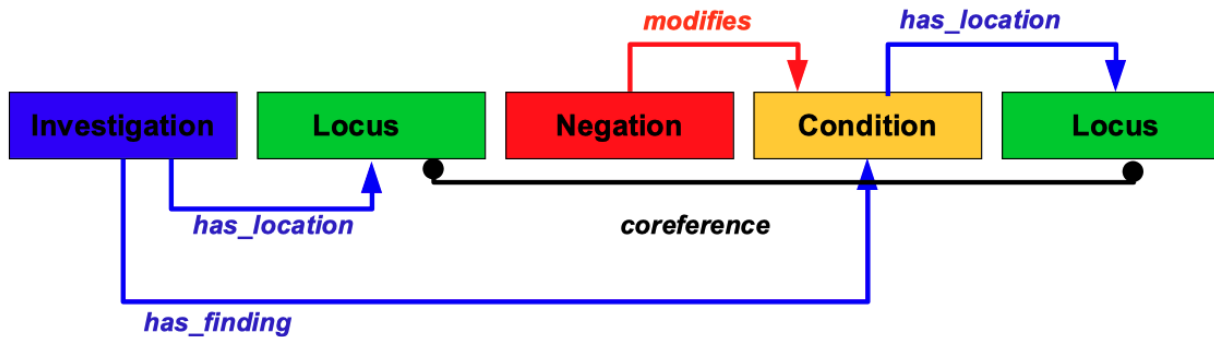
# Information extraction

- **the process of deriving disambiguated quantifiable data from natural language texts in service of some pre-specified precise information need**

*(Cunningham, in Encyclopedia of Language and Linguistics, 2nd Edition, pages 665–677, 2005.).*

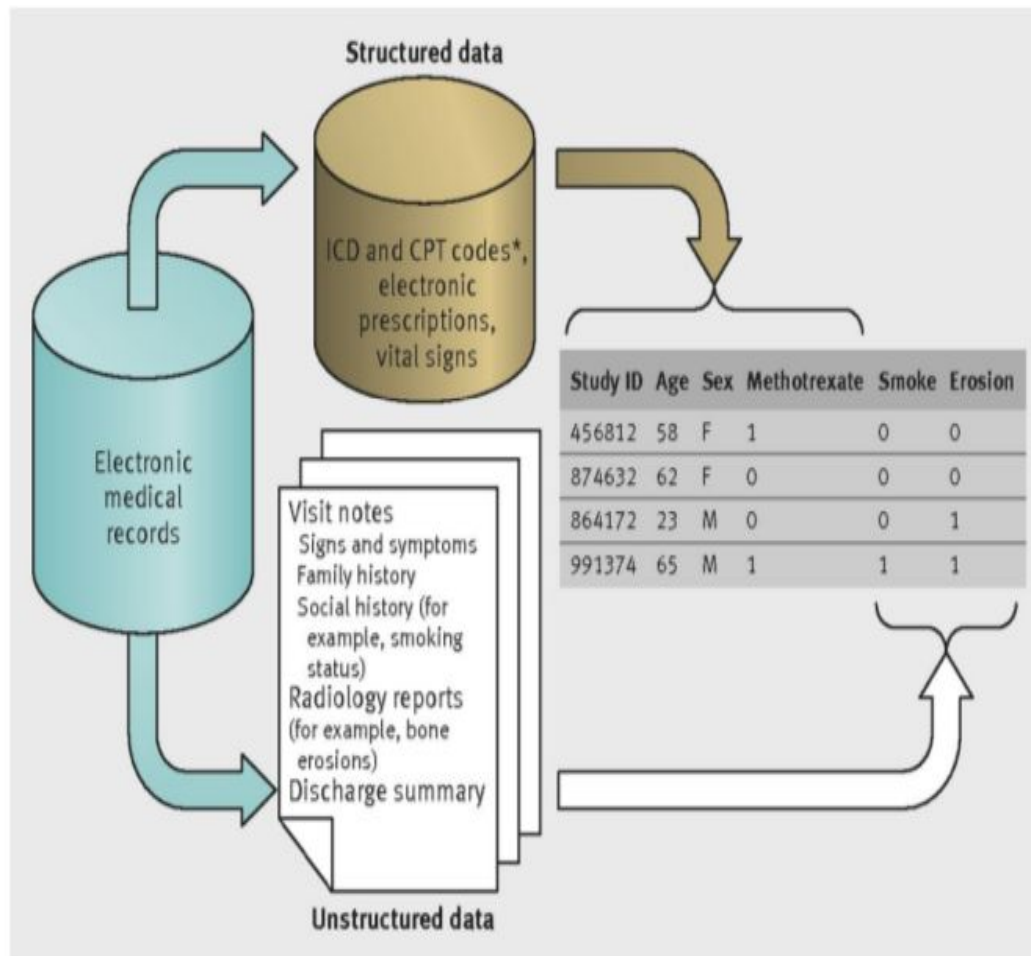
# Information extraction

Punch biopsy of skin. No lesion on the skin surface following fixation.



- We might extract:
  - Entities and their co-referents
  - Negation, certainty, time
  - Relations
  - Events

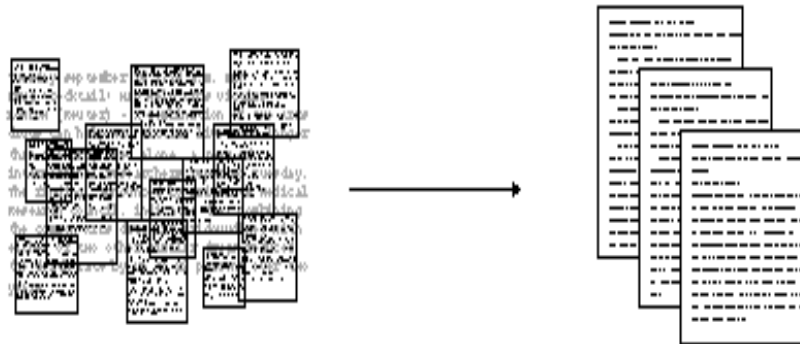
# Information extraction



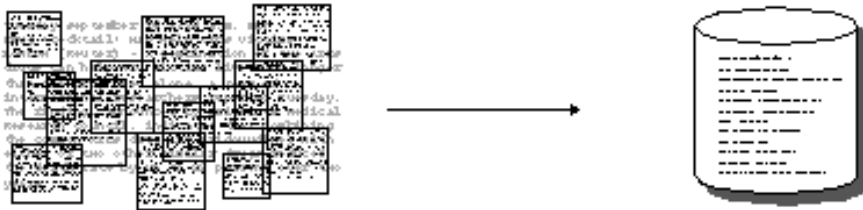
Liao et al. *Development of phenotype algorithms using electronic medical records and incorporating natural language processing.*  
*BMJ* 2015;350:h1885

# Information extraction vs information retrieval

IR pulls **documents** from large text collections in response to specific keywords or queries. You analyse the **documents**.



IE pulls **facts** and **structured information** from the content of large text collections. You analyse the **facts**.







# Tools for the job



# GATE

- A widely used open source NLP toolkit, 20+ years old, 35 000+ downloads per year
- Graphical user interface
- Plug and play approach
- No programming skills required
- Large user community
- 100s of modules for all kinds of language processing tasks
- Scales to large distributed systems and data
- Limited support for state of the art NLP



# Python

- Currently the most popular way of doing NLP
- Requires programming skills
- Most popular Python packages:
  - RE – python regular expressions
  - NLTK – general NLP and pre-processing
  - spaCy – “best of breed” statistical models
  - SciPy – lots of machine learning
  - Hugging Face – language models
  - TensorFlow – neural nets



**HUGGING FACE**





Thank you.  
Any questions?

[angus.roberts@kcl.ac.uk](mailto:angus.roberts@kcl.ac.uk)

