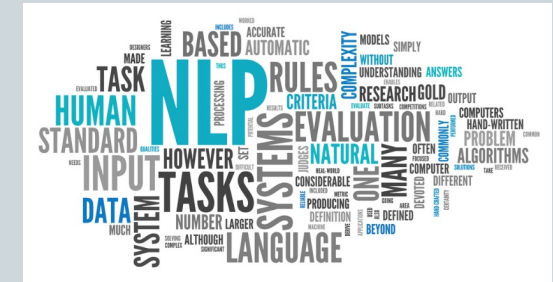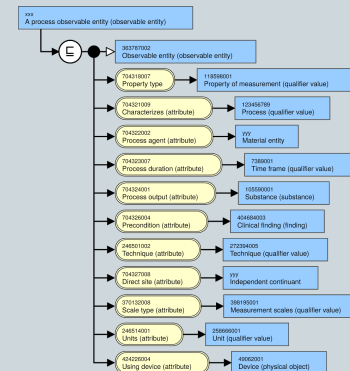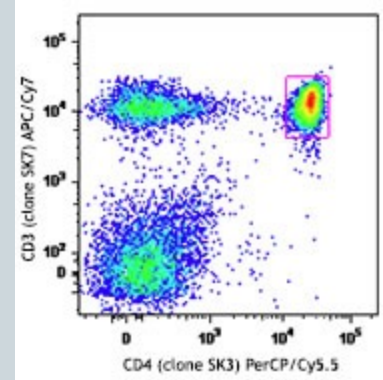# Natural Language Processing

Introduction

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

Introduction – my journey

Motivation – reusing the health record

- 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 e.g. for
    - Observational studies
    - Trial recruitment
    - Case identification
    - Retrospective cohorts
    - Individual risk prediction

Electronic health records (EHRs) – structured and unstructured information

HAEMATOLOGY
HAEMOGLOBIN (g/L)   151
HCT   0.422
RED CELL COUNT   4.83
MCV   87.4
MCH   31.3
MCHC (g/L)   * 358
RDW
PLATELET COUNT
MPV
WHITE CELL COUNT
    Neutrophils
    Lymphocytes
    Monocytes
    Eosinophils
    Basophils
ESR

BIOCHEMISTRY
SODIUM
POTASSIUM
CHLORIDE
BICARBONATE
UREA
CREATININE
estimated GFR

Clinical Findings

Concept ID	Preferred term
16932000	Nausea and vomiting
68566005	Urinary tract infectious disease
38341003	Hypertensive disorder
49436004	Atrial fibrillation
49218002	Hip pain
301011002	Escherichia coli urinary tract infect
40835002	Coffee ground vomiting
167667006	Fecal occult blood: negative

Procedures

Concept ID	Preferred term
52734007	Total replacement of hip
117010004	Urine culture
76009000	Esophagogastroduodenoscopy
91251008	Physical therapy procedure

Reason: CHECK ETT TUBE PLACEMENT, ?PNA, CHF
[**Signature 1**]
UNDERLYING MEDICAL CONDITION:
 85 y/o male s/p acute mi and catherization now in ccu with
cardiogenic shock
REASON FOR THIS EXAMINATION:
CHECK ETT TUBE PLACEMENT
?PNA
CHF
[**Signature 1**]
 FINAL REPORT
CLINICAL INDICATION: Assess endotracheal tube placement in
patient with congestive heart failure.

Comparison is made to previous study of one day earlier.
An endotracheal tube is present, in satisfactory position.
A Swan-Ganz catheter terminates in the proximal left
pulmonary artery and has been withdrawn in the interval.
An intraaortic balloon pump terminates about 3.3 cm below
the superior aspect of the aortic knob, and a nasogastric
tube terminates in the region of the gastroduodenal
junction.

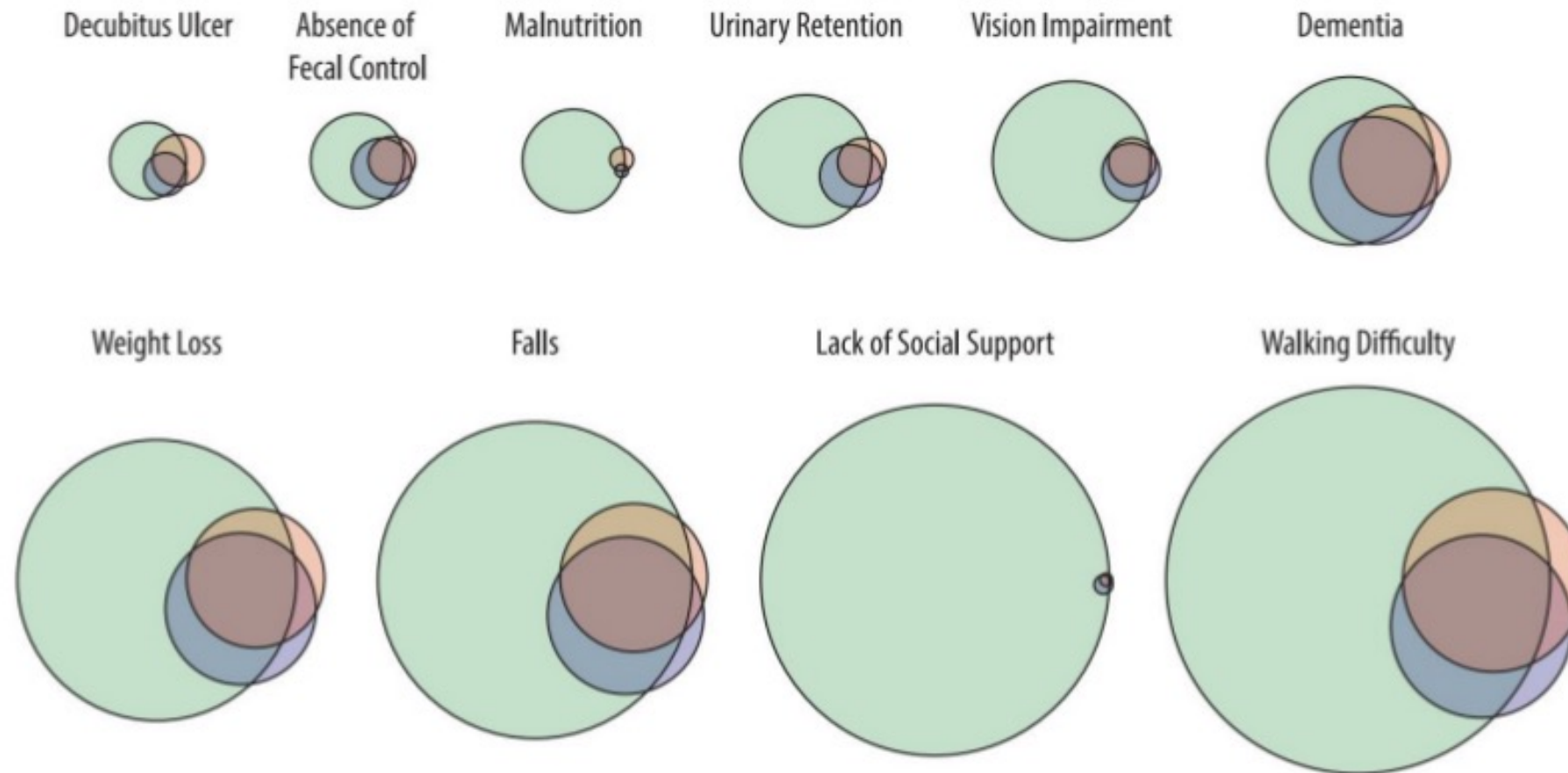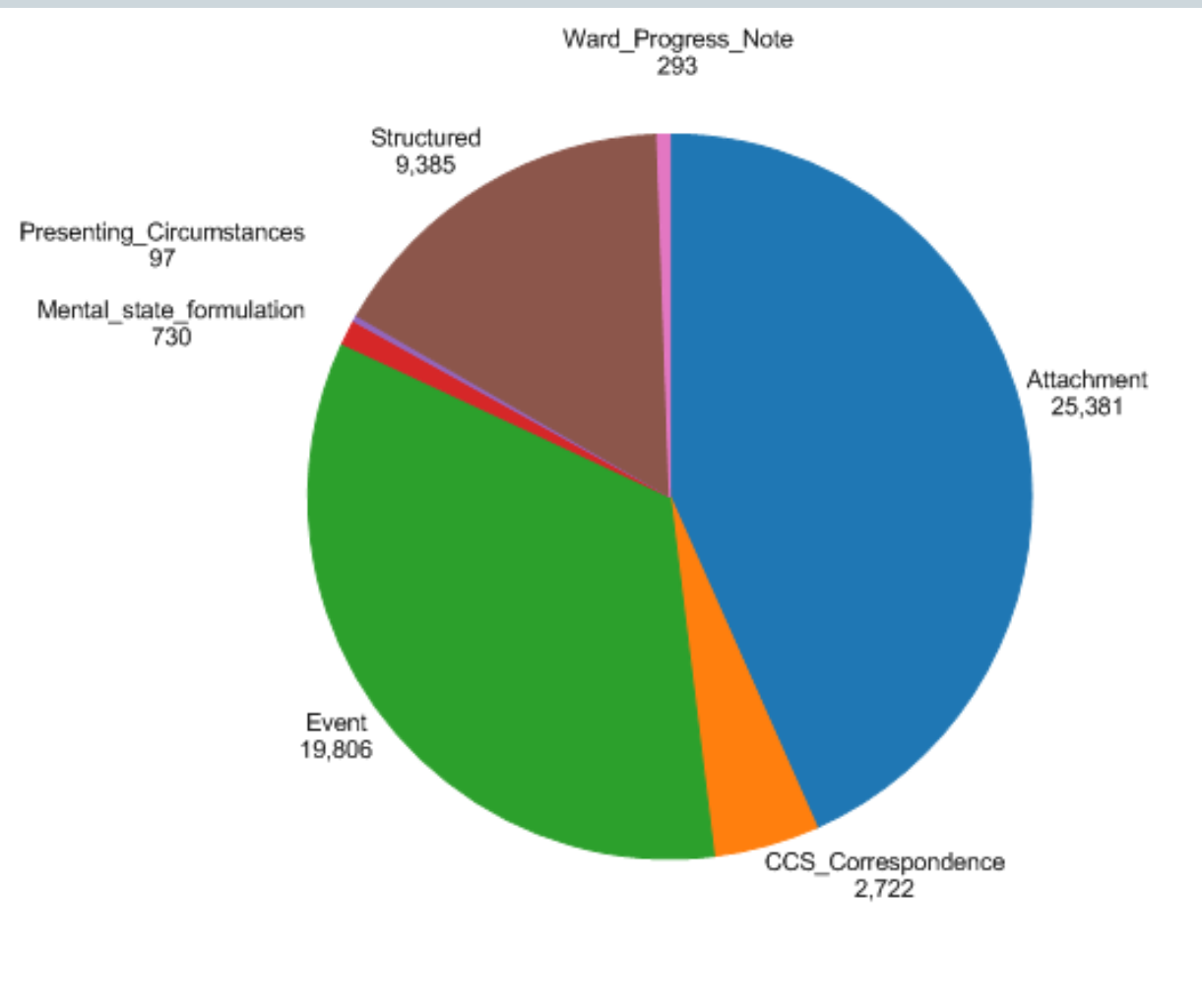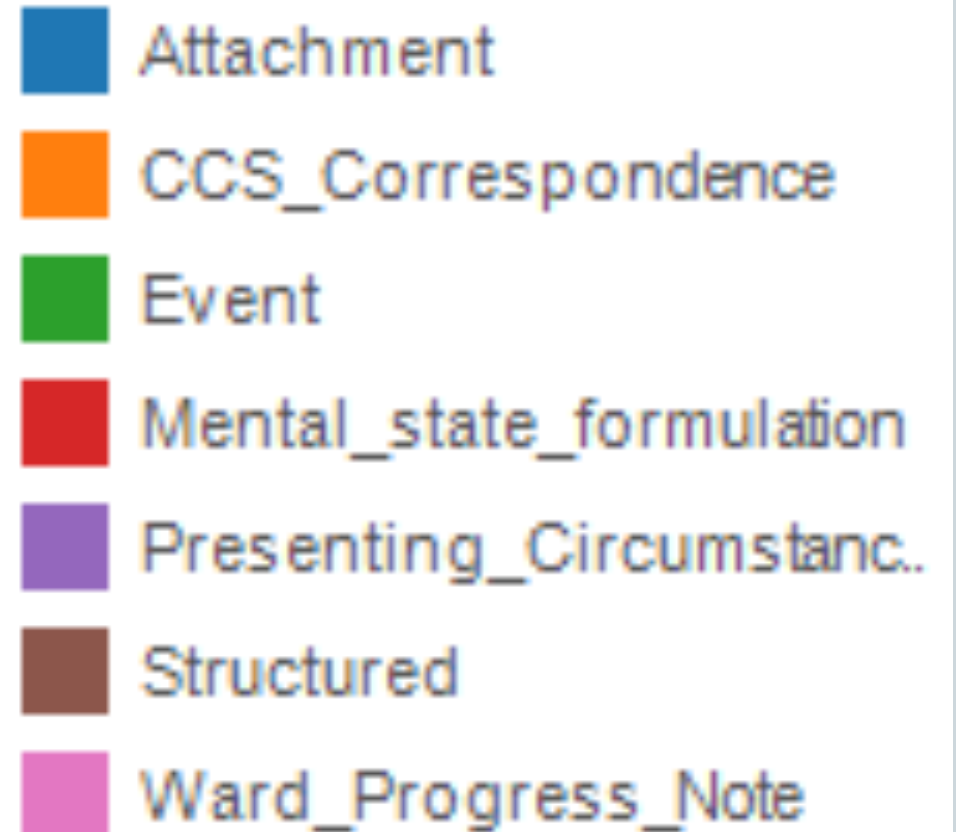# How much unstructured, textual information is in an EHR?

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)

An example – Mini Mental State Examinations

Ward_Progress_Note
293

Structured
9,385

Presenting_Circumstances
97

Mental_state_formulation
730

Attachment
25,381

Event
19,806

CCS_Correspondence
2,722

source
- Attachment
- CCS_Correspondence
- Event
- Mental_state_formulation
- Presenting_Circumstanc..
- Structured
- Ward_Progress_Note

Natural Language Processing

Write a short definition of natural language processing

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It involves analyzing and processing natural language data, such as text or speech, to perform tasks like translation, sentiment analysis, speech recognition, and information extraction. NLP combines linguistics, computer science, and machine learning to bridge the gap between human communication and computer understanding.

Natural Language Processing

Write a short definition of natural language processing

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It involves analyzing and processing natural language data, such as text or speech, to perform tasks like translation, sentiment analysis, speech recognition, and information extraction. NLP combines linguistics, computer science, and machine learning to bridge the gap between [...] and computer understanding.

plus
- Search
- Question answering
- Summarisation
- Document classification
- Dialogue

Information extraction

the process of deriving disambiguated quantifiable data from natural language texts in service of some pre-specified precise information need
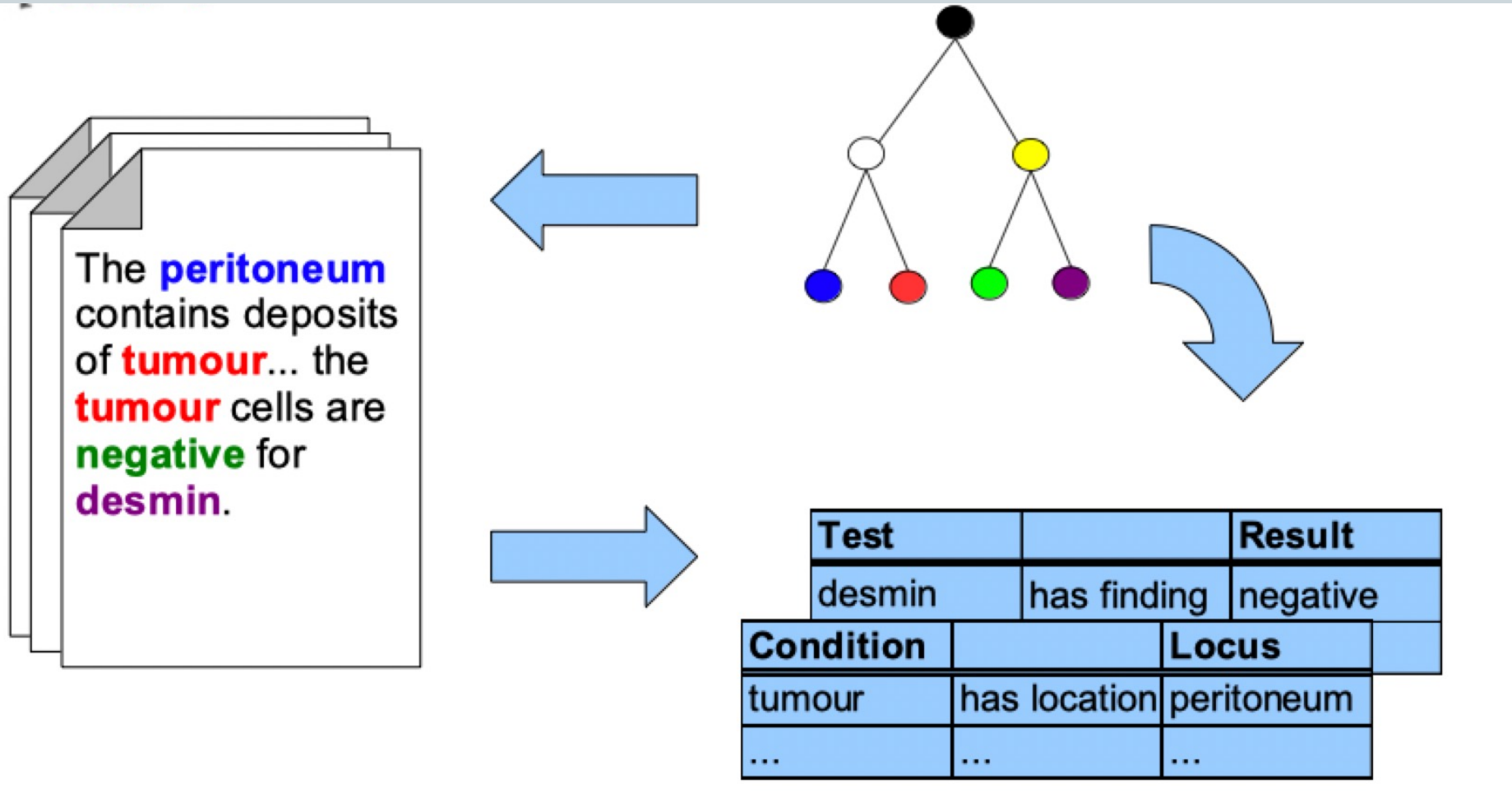
(Cunningham, in Encyclopedia of Language and Linguistics, 2nd Edition, pages 665–677, 2005.).

Information extraction
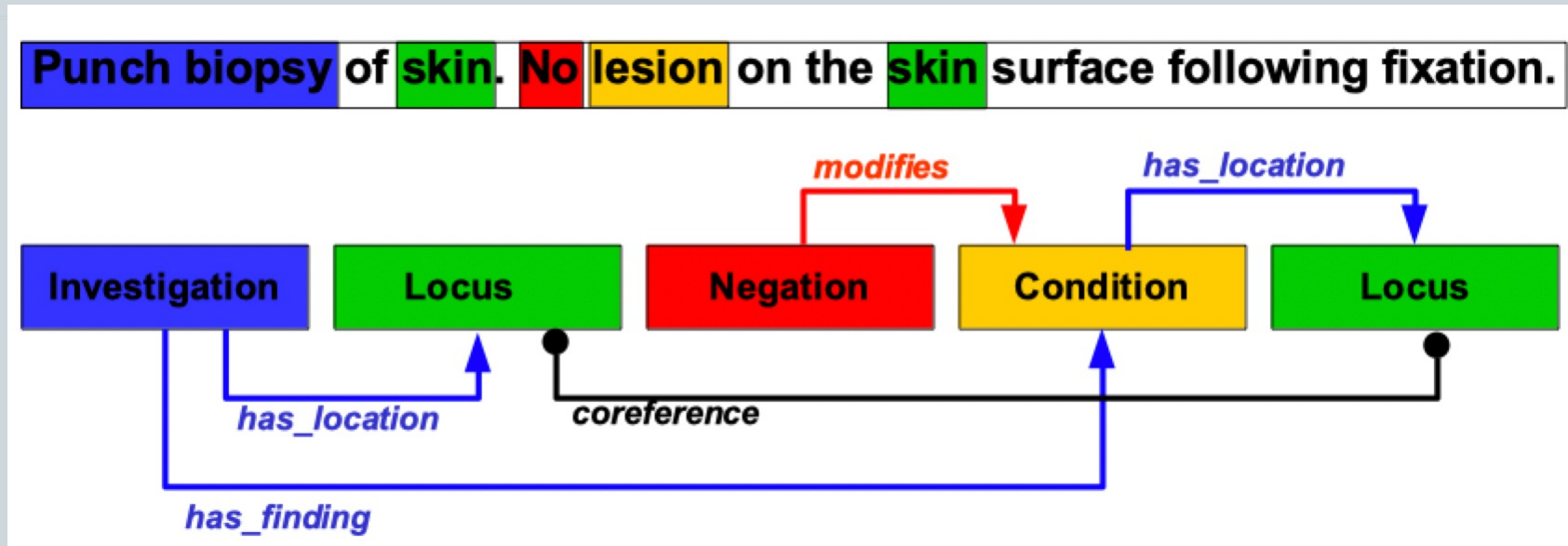
- We might extract:
  - Entities and their co-referents
  - Negation, certainty
  - Relations
  - Events
  - Temporal expressions and relations

Entity linking

This half-day session

- Introduce some of the basic ideas behind natural language processing
- Give intuitive explanations for these ideas
- Make NLP a little bit less of a black box within hidden workings
- Introduce some health use cases

This half-day session

How do we represent or *model* language in the computer?

How do we train or *supervise* the computer to solve problems involving language?

What can we use NLP of the health record for?

Content

- Presentations explaining the key ideas

- Examples and demonstrations to illustrate these ideas

- Some practicals use the Python programming language. Don't worry – you don't need to know Python to follow them

- A demo of a working system

- A presentation of some real-world use cases

Course material

All material can be found on GitHub

Let's take a look:

http://bit.ly/4pKxAJa

Thank you

angus.roberts@kcl.ac.uk

https://www.kcl.ac.uk/people/angus-roberts