Diagnosis Certainty and Progression: A Natural Language Processing Approach to Enable Characterisation of the Evolution of Diagnoses in Clinical Notes

Alfredo Madrid[1,3,5], Caitlin Bullin[1], Lifeng Han[3], Judith Andrew[1,3], Warren Del-Pinto[1,3], Ghada Alfattni[1,3], Oswaldo S. Pabón[4], Ernestina M. Ruiz[4], Luis Rodríguez[5], Ana María Aldana[2], Robyn Hamilton[2], Karim Webb[2], Meghna Jani[2], Goran Nenadic[3], William G. Dixon[1,2]

[1]Centre for Epidemiology Versus Arthritis, University of Manchester, Manchester, UK; [2]Department of Rheumatology, Salford Royal Hospital, Salford, UK; [3]School of Computer Science, The University of Manchester, Manchester, UK; [4]Universidad Politécnica de Madrid, Madrid, Spain; [5]Rheumatology Department, and Health Research Institute (IdISSC), Hospital Clínico San Carlos, Madrid, Spain

**HealTAC 2022: the 5th Healthcare Text Analytics Conference**

Introduction

The accurate identification of **diagnoses** in **free clinical narratives** is decisive for characterizing the patients in a medical cohort. Thefore, the **knowledge extraction and information retrieval** tasks must be addressed carefully.

Clinical notes might present **multiple qualifiers** that could change the meaning of a statement: negation, speculation, temporal information, family history and so on.

Sentence	Qualifier	Status
Rheumatoid Arthritis can be ruled out	**Negation**	Discarded
Rheumatoid Arthritis under investigation	**Speculation**	Not confirmed
Family history of Rheumatoid Arthritis	Family member	Not confirmed
Under biologic treatment for Rheumatoid Arthritis	Confirmed	Confirmed

Table 1. Clinical notes qualifiers

It is not unusual for caregivers to **preserve uncertainty using broad and ambiguous terms** when they have not full evidence of the disease status of a patient.
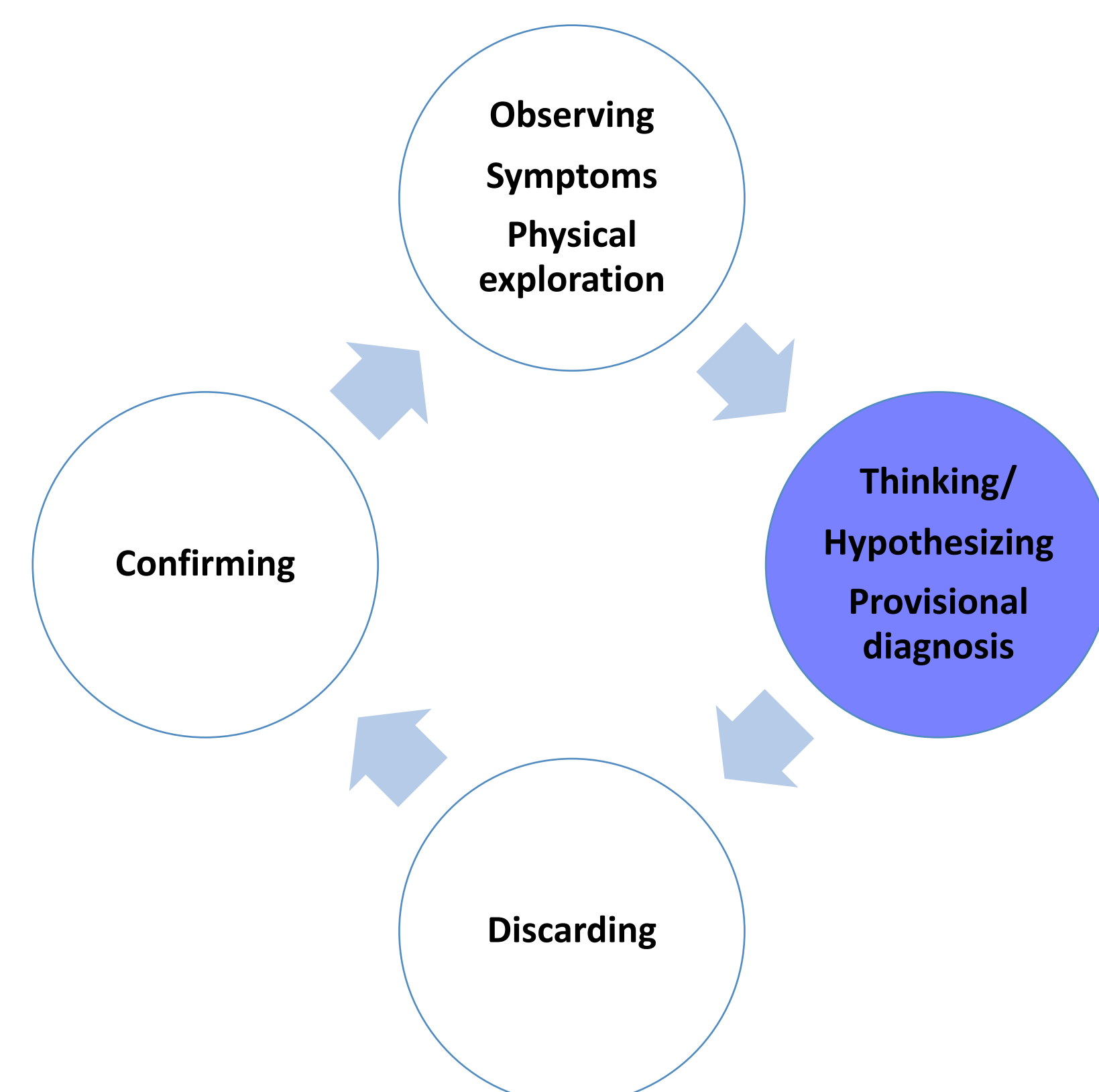
The **percentage** of speculative sentences may range between the **11% and the 22.7%**[1]

It is estimated that **over 40%** of the data in an EHR is stored as free text.


**Image 1.** Clinical notes with speculation qualifier

Objectives

1. To appropriately identify the **uncertainty and negation qualifiers** in Rheumatic and Musculoskeletal narratives building a **classification model** that handles four different categories: **Definitive, probable, possible, not**
2. To demonstrate the extent to which that influences possible population health research about **incidence/prevalence** of disease and its evolution
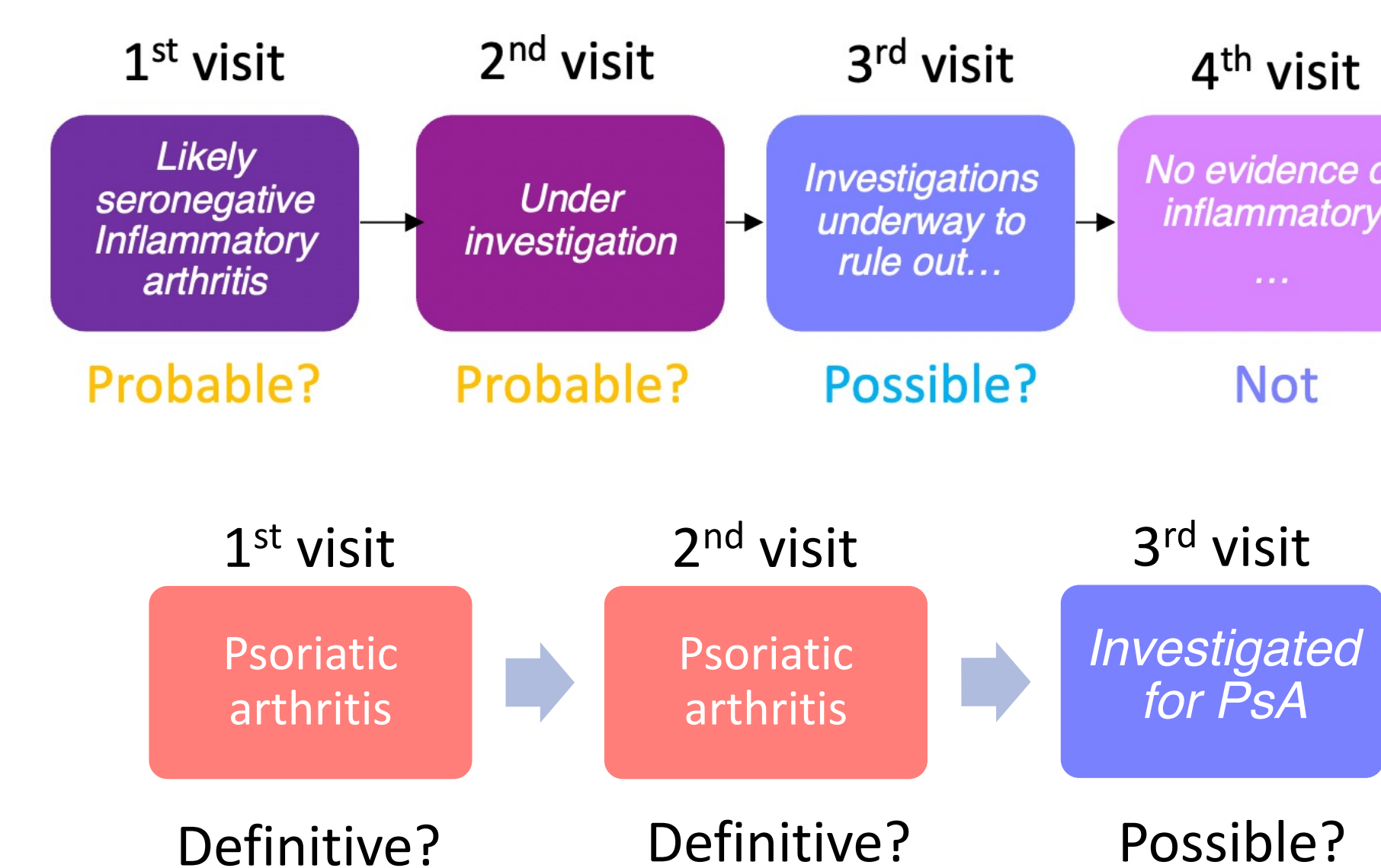
Levels of speculation

More annotation classes lead to **lower agreement results**. Less granular models **lose expressive power** [2]
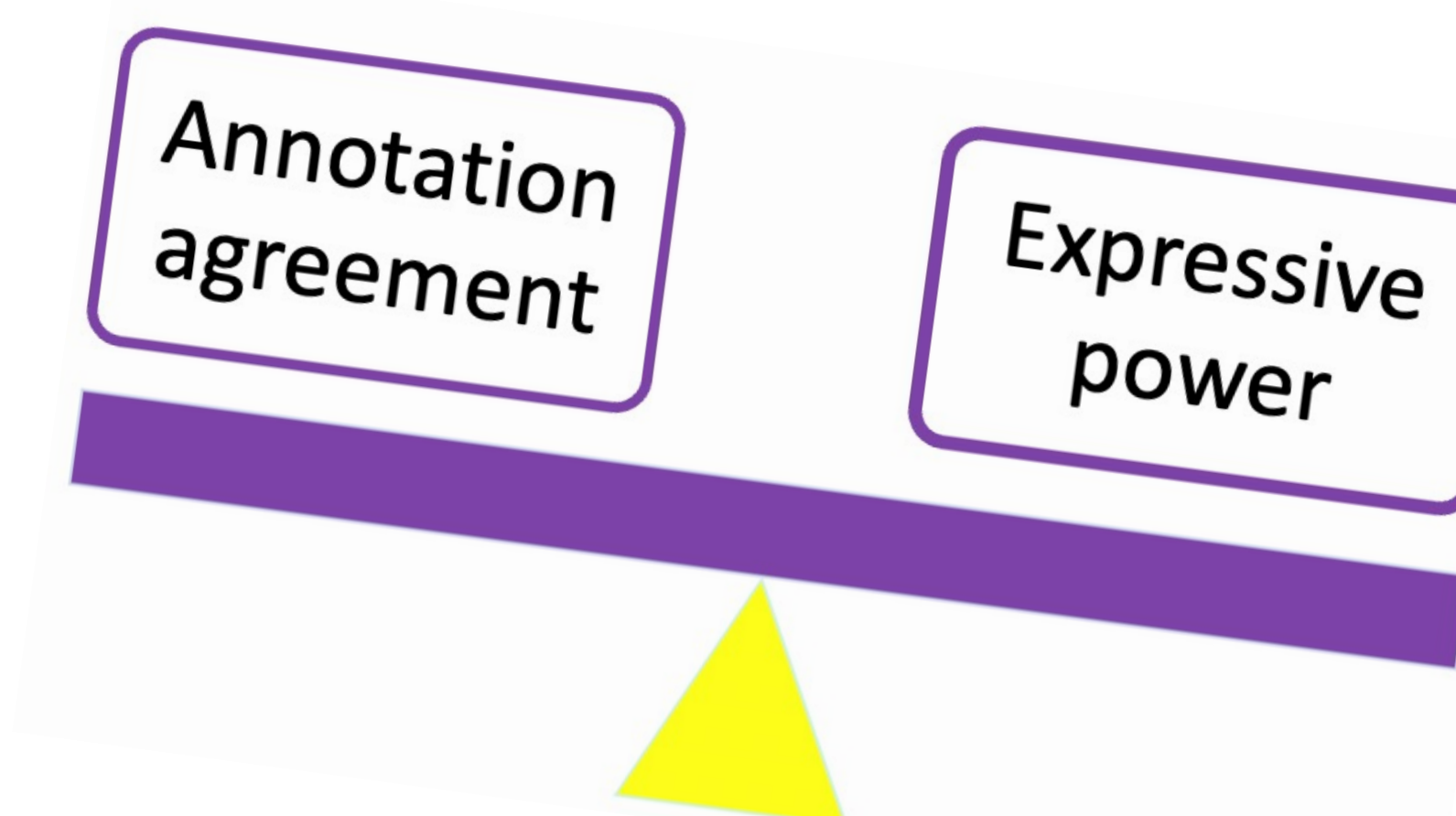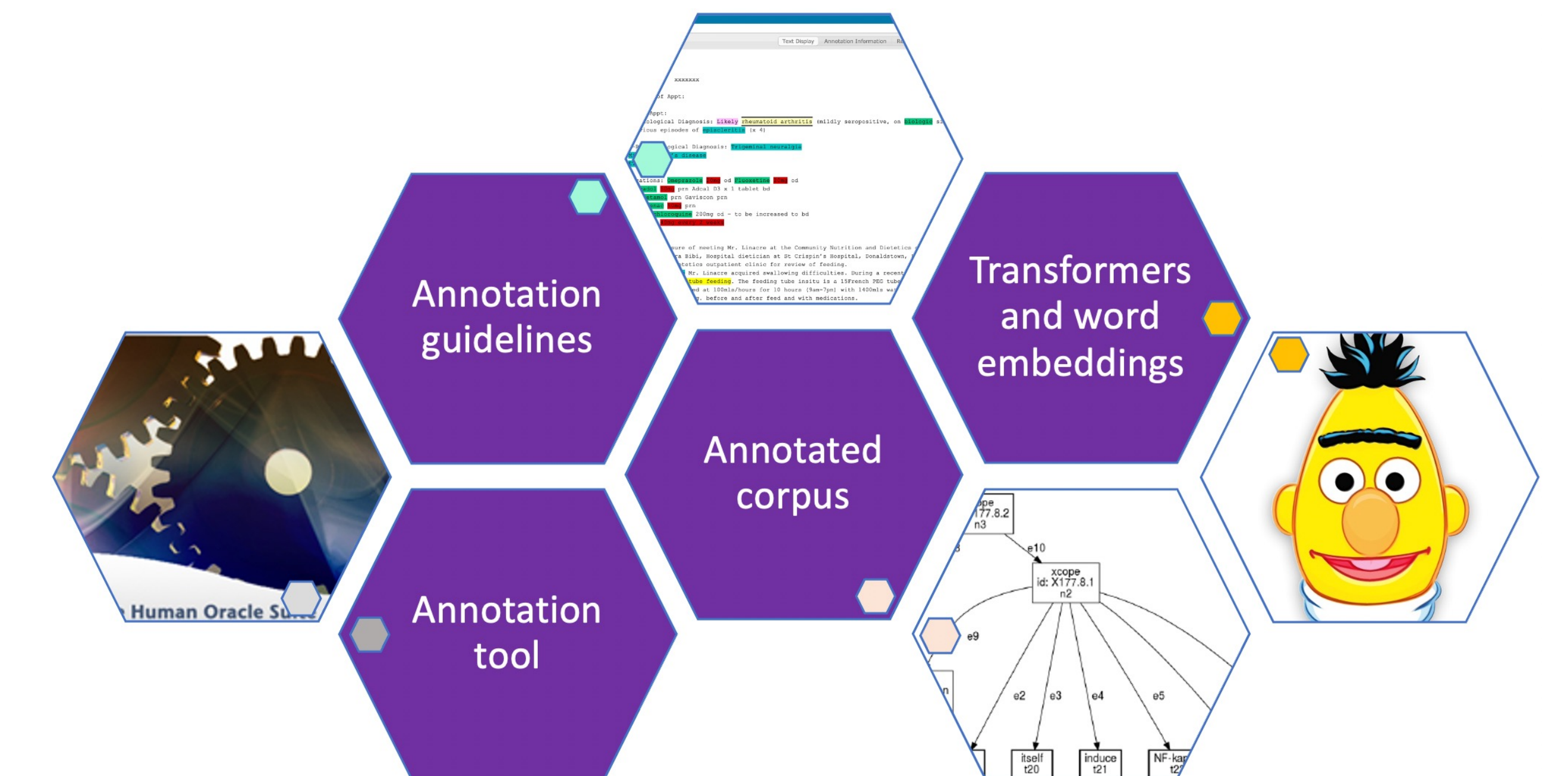

A **binary classifie**r is good for identification but don't allow us to characterize the speculation evolution over clinical notes.
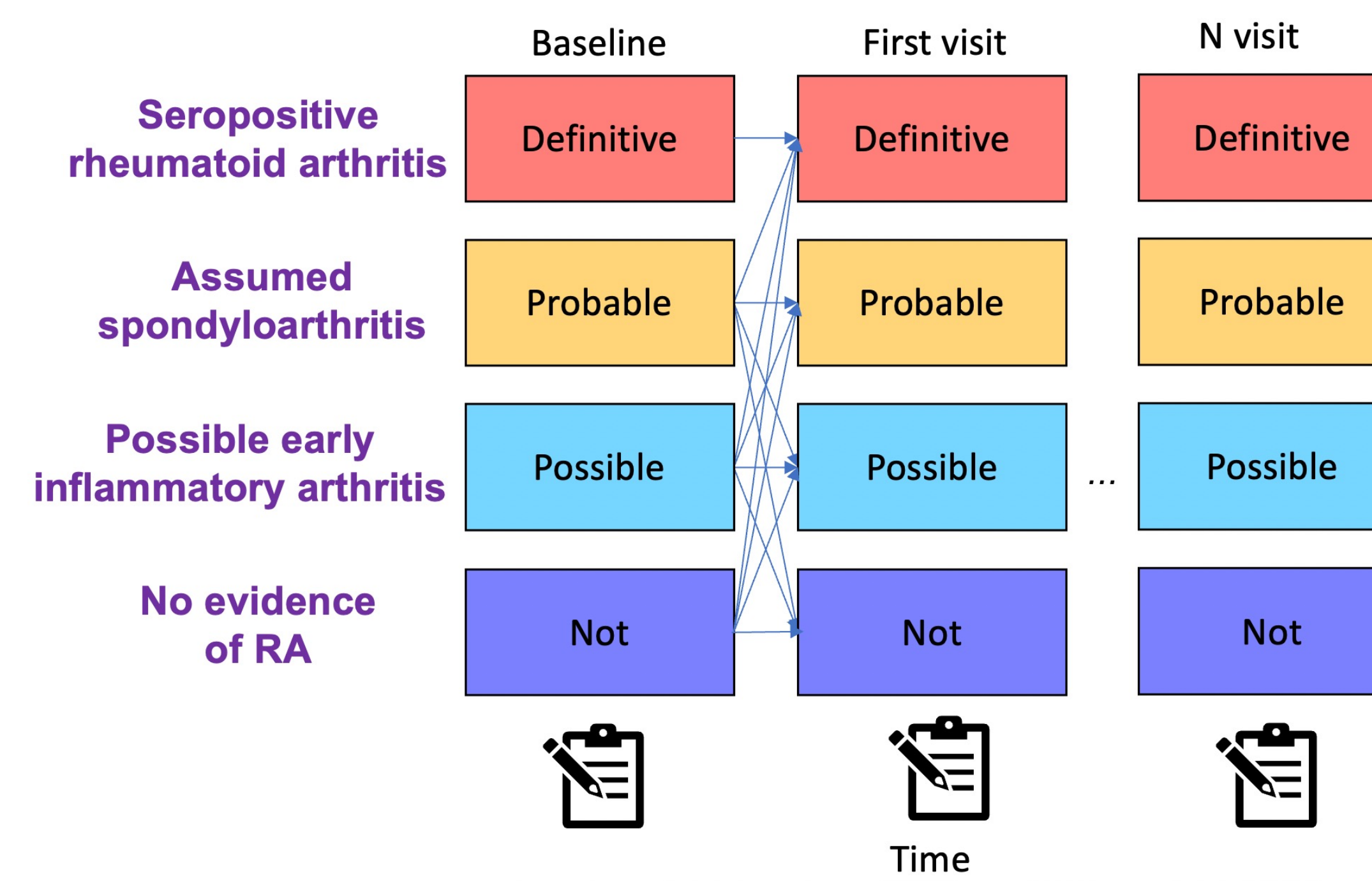
Examples

NLP and text mining

Roadmap

Annotation guidelines, and definition of the four different categories: Definitive, Probable, Possible, Not.
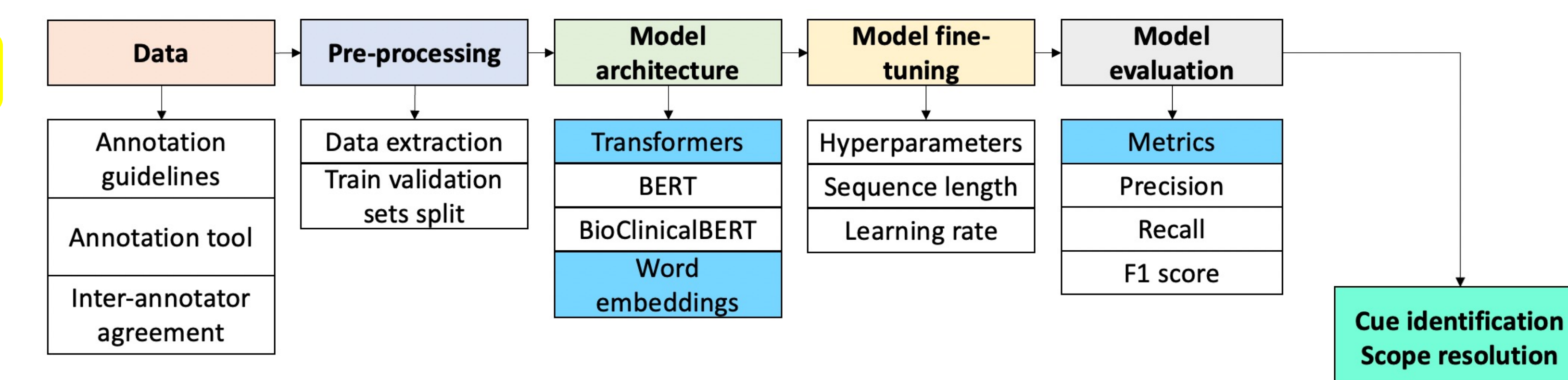**Manual annotation** of clinical notes by a physician: *eHost, spreadsheet*
Inter-annotator agreement measure with a subset of clinical notes
Enriching dataset with *Probable* and *Possible* cases
Word embeddings + classification model: Snomed2Vec
Transformers models: BioClinicalBert, MedBert
Cue identification and scope resolution: Precision, recall, F1-Score

References

[1]Cruz Díaz, N. P., & Maña López, M. J. (2019). Negation and Speculation Detection (Vol. 13). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.13
[2]Velupillai, S. (2012). *Shades of certainty: annotation and classification of swedish medical records* (Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University).