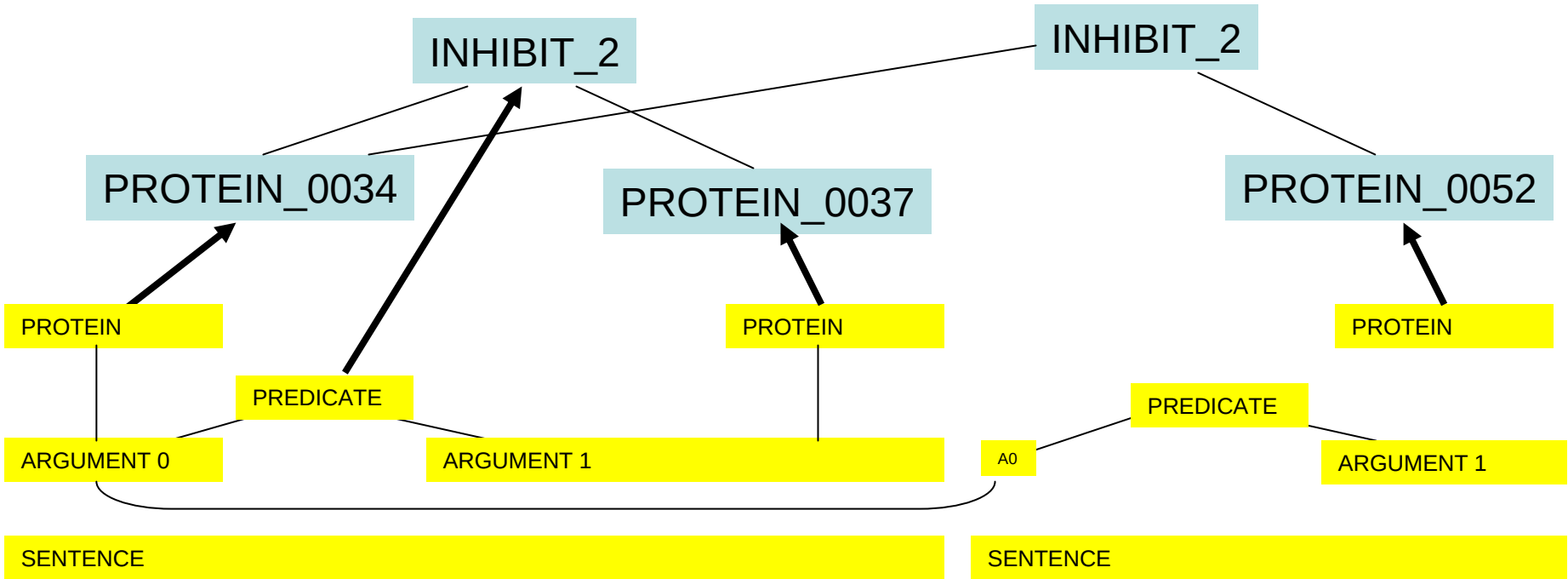
Possible Goals

- “Create a system that builds a base of knowledge about X from corpora A-Z”
  - e.g. “Extract knowledge about protein interactions and researchers publishing on them, using MEDLINE and web documents”
- What else?

“Extract Knowledge”

- Locate candidate entities and relations in the text (“mentions”)
- Associate each mention with a concept or instance in a KB



Protein A inhibits formation of Protein B. It also inhibits Protein C.

Possible Tasks

- Define Specific Learning / Understanding Tasks
  - e.g. learning named entities from context in support of named entity annotation
  - Active learning?
  - Supervision to keep the system learning over time
- Annotation / Extraction
  - “Simple” annotators (based only on text)
  - “Higher order” annotators (based on tags produced by other annotators; based on HTML structure of documents)
  - Learn annotators from training data for our domain
  - Deploy annotators based on existing 3rd-party code and resources (e.g. Brill Tagger, WordNet, etc.)
  - Assessment of results (generate and test as per URNS model)
- Reference resolution
  - Decide how to co-index annotations that refer to the same entity / relation
- Ontology and Type Systems
  - Create an ontology to represent the concepts and instances relevant to our goals
  - Formalize the system of types produced by our annotators, and map them to ontology concepts / instances

Discussion

- Scott: Separate the off-line annotation & extraction from the end-user application (e.g. question answering)
- William: relationship between queries and extracted knowledge; learning from lightly labelled data