

# The Great NLP and AI Swindle: Why State-of-the-art AI Technologies Still Fail(?)

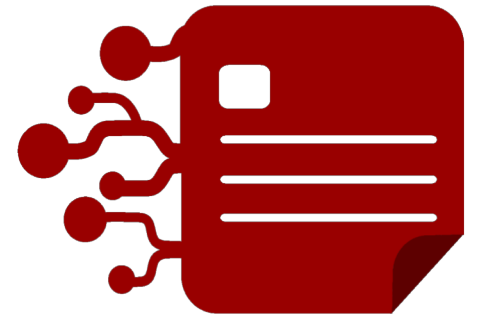
Damir Cavar

February 2024

Language Processing Brown Bag

# NLP-Lab

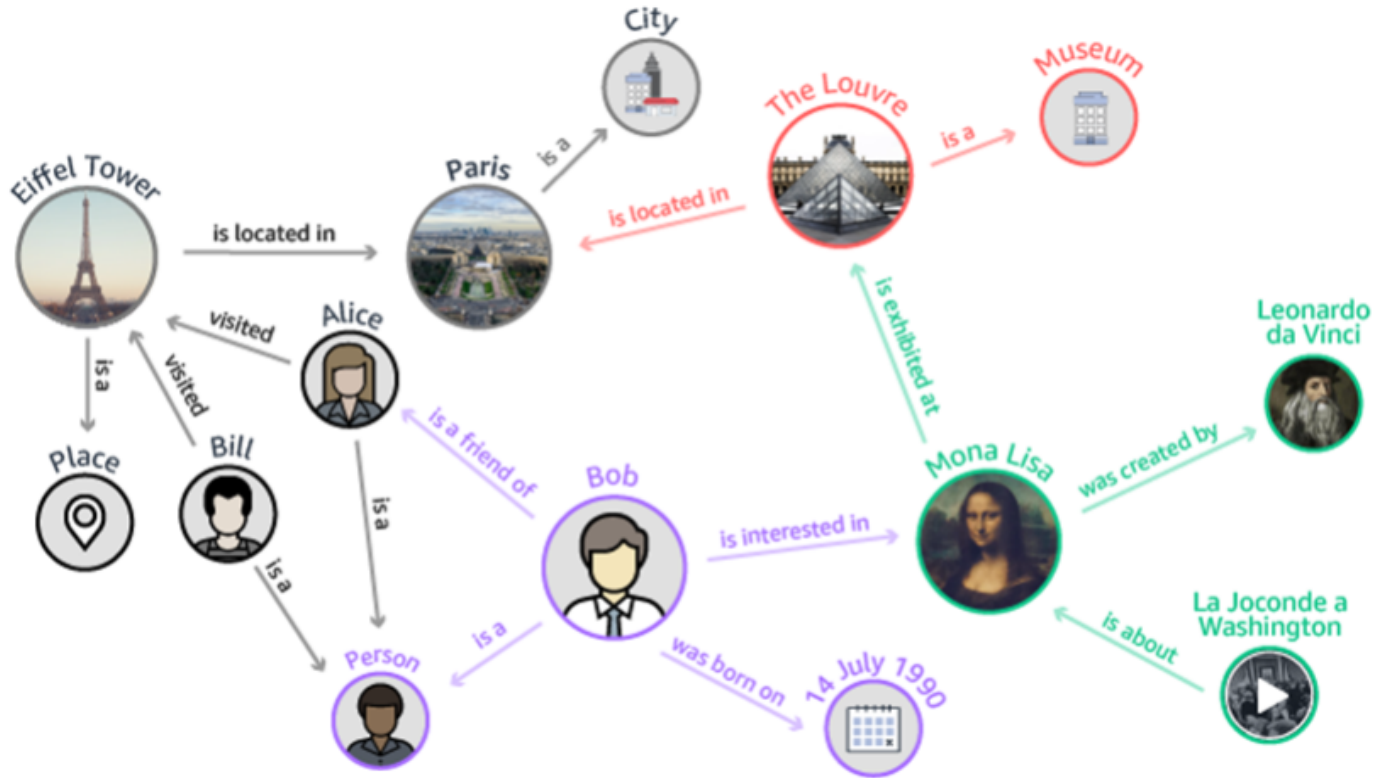
- Team:
  - Zoran Tiganj, Billy Dickson, Ludovic Mompelat, Andrew Davis, Chi Zhang, Rahul Gupta, Calvin Josenhans, Yuchen Yang, Luis Abrego, John Macintosh Phillips, Shane Sparks, Khai Willard, Nicholas Kilo, Van Holthenrichs, Soyoung Kim, Amit Singh, ...
- <http://nlp-lab.org/>
  - Regular NLP meetings: Wednesday 5 PM
  - Quantum NLP meetings: Friday 4 PM



# Research Goals

- Build the models, infrastructure, technologies to
  - Computational semantics and pragmatics
  - Knowledge Representations
  - Event Semantics, Temporal Logic, Common Sense Reasoning
- Using:
  - Neuro-symbolic models:
    - Graphs and Graph Embeddings
    - Neural Architectures

# Knowledge Graphs

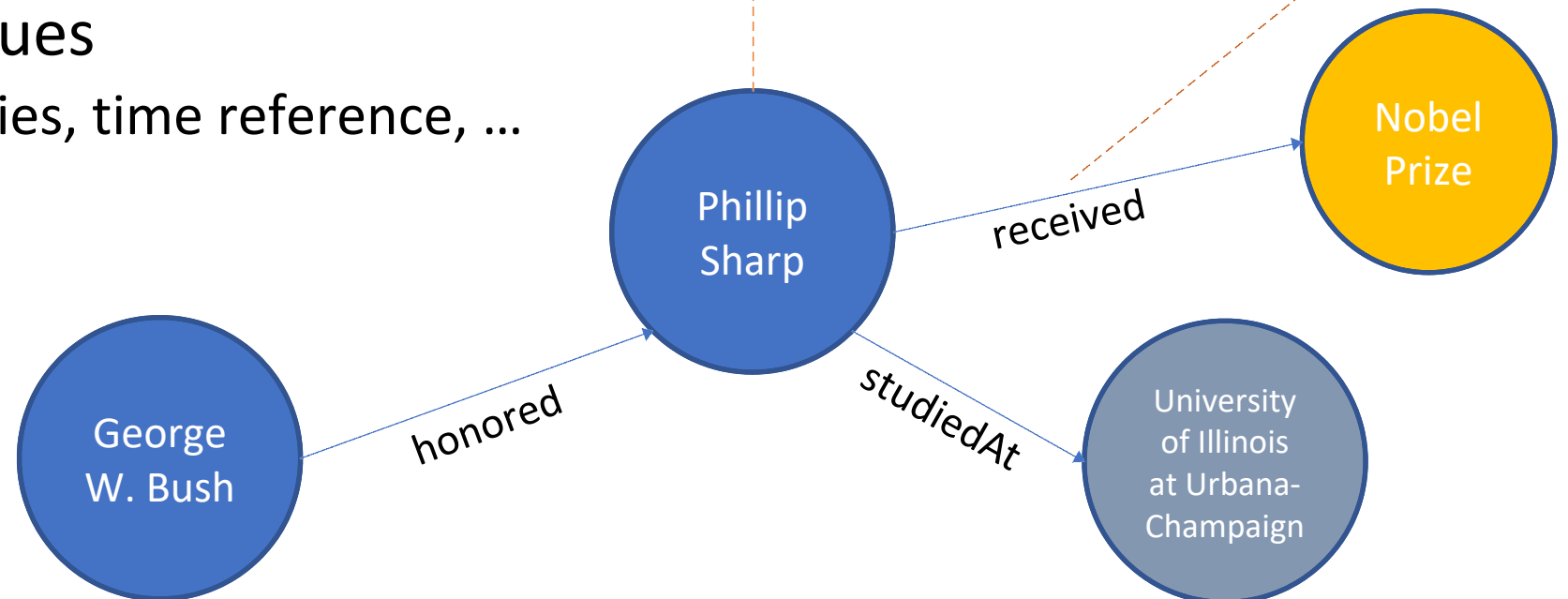


# Knowledge Graphs

- Concepts and Relations
  - Mostly unconstrained
  - Domain specific or free
- Attributes and Values
  - encoding properties, time reference, ...

Attribute	Value
Birth place	Falmouth, KY
DOB	06/06/1944
gender	male
...	

Attribute	Value
year	1993
...	



# Knowledge Representation

- Static: Concept and Relation Properties
  - Even when dynamically growing or changing
- Problem to encode events or procedures
  - *Mary gave John a book.*
    - Event as a state change / transformation:
      - Mary owns a book, John does not → John owns a book, Mary does not
  - Peter was fetching his daughter from school.
    - Intermediate states:
      - Peter is at home, daughter at school → Peter is at school, daughter at school → Peter is at home, daughter at home

# Temporal Relations



- Sequencing of events or sub-events
  - *Wash the veggies, chop them, fry them.*
    - Presentation and Temporal event sequence: 1 2 3
  - *Before you fry the veggies, wash and chop them.*
    - Presentation sequence: 3 1 2
    - Temporal event sequence: 1 2 3
- Duration of events
  - Clear reference: “for 30 minutes”
  - Common sense

# Temporal Relations

- Duration of events
- Unfolding over time
  - Events relate to time
  - States are points in time
- Temporal sequencing relates to
  - Causal reasoning



# Temporal Scope

- Simple temporal relations
  - Past tense: *Tim Cook bought Google.*
    - Assumptions: factive, true event
  - Future tense: *Tim Cook will buy Google.*
    - Assumptions: non-factive, hypothetical
- Complex relations: temporal scope
  - *Reuters reported that*  *Tim Cook bought Google*
  - *Reuters will report that*  *Tim Cook bought Google*

# Pragmatics

- Implicatures:

- John to Peter: *I bought the blue car.*

- John and Peter talked about cars earlier.

- There should be a set with at least one more car that John could have bought but did not, and

- None of the cars in the set is blue.

- Clues: Definiteness of NP via **the**, and specificity of NP

- Presuppositions:

- *John fed his cat this morning.*

- Assumptions:

- John owns/has a cat/pet.

- John owned cat food this morning.

- Clues: Possessive pronoun as modifier of Direct Object.

# Predicates

- Veridicality

- Factive predicates: *know, regret, realize, notice, ...*
  - *I regret that ...* (X did something to Y)
  - Complements are assumed to be true
- Non-factive predicates: *believe, think, claim, ...*
  - *I believe that ...* (X did something to Y)
  - Complements cannot be assumed to be true
- Counter-factive predicates: *pretend, ...*
  - *John pretends that he is ill.*
  - Complement cannot be true: *John is not ill*

- Question:

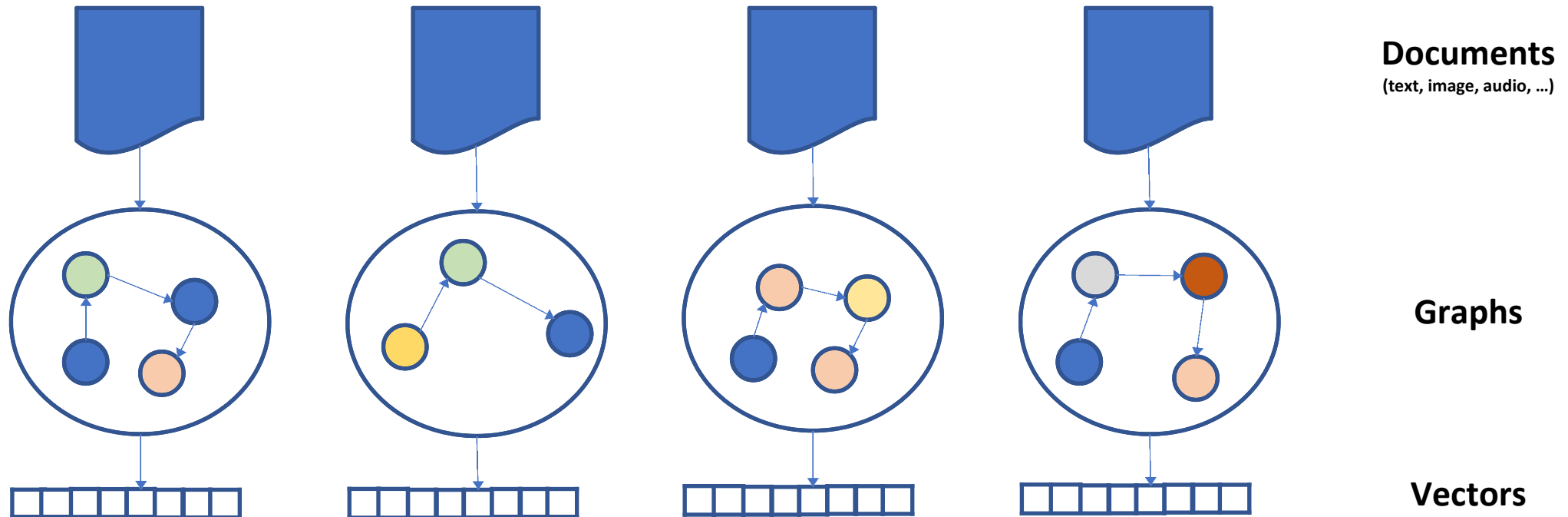
- Cross-linguistic similarity = universal properties related to factivity

# Solution

- Multi-modal information input to knowledge representation
  - Language input (speech and text)
  - Information in images
  - Haptic information
  - Secondary information: sound it makes, properties when shaking, tossing, etc.
- Graphs generated using:
  - General or common sense knowledge
  - Domain specific knowledge
  - Semantic restrictions over graphs: ontologies, taxonomies

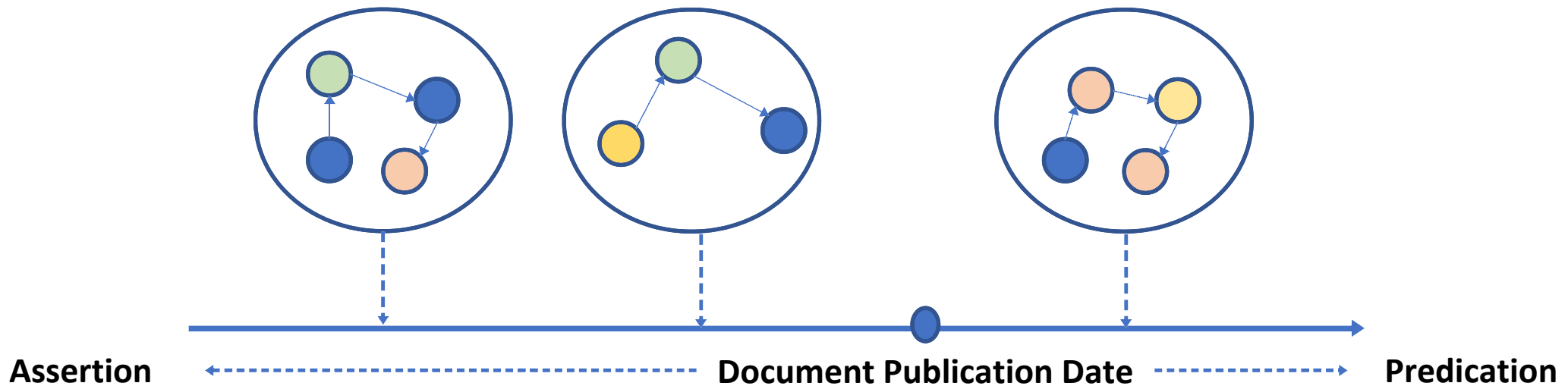
# Document Graphs

- Concept/Knowledge Graph Document Representation

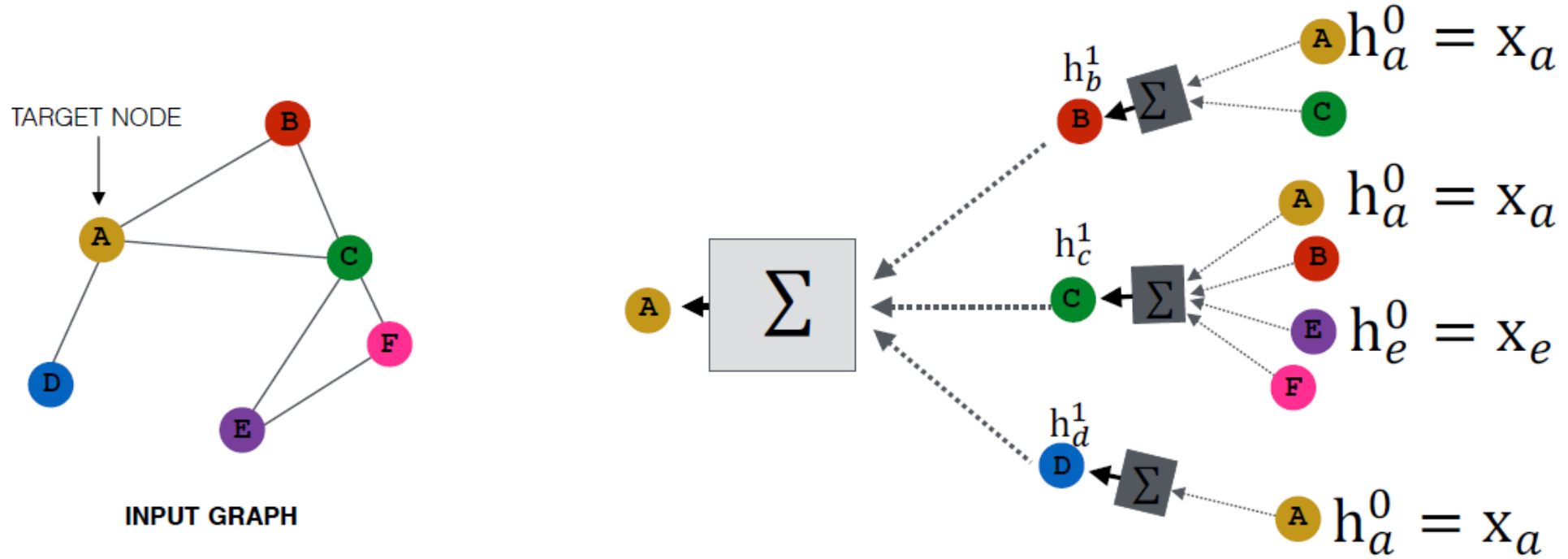


# Event Graphs

- Arrangement of sub-events along time axis
  - Approximation of duration
  - Identification of geo-location
  - Linked entities and relations



# Graph Neural Network Models



$$h_v^{(l+1)} = \sigma\left(W_l \sum_{u \in N(v)} \frac{h_u^{(l)}}{|N(v)|} + B_l h_v^{(l)}\right)$$

# Language 2 Graph Approach

- NLP technologies to:
  - Identify concepts in text/speech
  - Identify anaphoric relations or coreference
  - Parse sentences and determine grammatical roles (subject, object, adjunct)
  - Process temporal properties of propositions
  - Link concepts to world knowledge
  - Etc.



# What's the Swindle about?

- NLP and AI algorithms/systems judged SOTA failing with:
  - Discontinuities and Long Distance Dependencies
  - Ellipsis constructions
  - Maybe:
    - Empty subject (pro-drop) or empty object constructions, missing functional words (e.g., missing copula)
- Mostly: NLPs and AIs systematically fail with
  - NL constructions that involve implied lexical elements (invisible words)
  - Complex constructions that triggered the concept of transformation or movement in syntax

# Discontinuities and Dislocations

- Words belonging together semantically appear discontinuously:
  - *John [ bought a book ]*
  - *[ A book ] John did not [ buy \_\_\_\_ ]*
  
  - *John said that Peter claimed that Mary [ bought [ a book ] ]*
  - *[ What ] did John say that Peter claimed that Mary [ bought \_\_\_\_ ]*

# Limitations of Dislocation

- Not all such dislocations are possible:

\* [ *Who* ] *did John say that Peter claimed that \_\_\_ bought a book?*

- Grammar theories provide explanations for:
  - The possible constructions
  - Ungrammaticality

# Data

- Different ellipsis types:
  - Gapping and Pseudo-gapping
  - Sluicing
  - VP-ellipsis
  - Forward or backward conjunct reduction
  - ...
- Rich body of literature over the last century
- Best overview: Van Craenenbroeck and Temmerman (2018)  
[The Oxford Handbook of Ellipsis](#) (see [IUCAT](#))

# Examples

- Forward Conjunction Reduction
  - My sister lives in Utrecht and (my sister/she) works in Amsterdam.
- Gapping
  - Paul and John were watching the news, and Mary \_\_\_ a movie.
  - Will Jimmy greet Jill first, or \_\_\_ Jill \_\_\_ Jimmy \_\_\_ ?
  - John always tried to finish reading a book in the evening, and Mary \_\_\_ a newspaper \_\_\_ .
- Note: mismatching word forms: *was watching*

# Examples

- Sluicing
  - She will high-five someone, but I don't know who \_\_\_\_ .
- Stripping
  - Why did Sam call, and \_\_\_\_ Bill \_\_\_\_ too?
- VP-ellipsis
  - Charles scratched his arm and Devin did \_\_\_\_ too.

# Examples

- Answers and responses
  - Who wants to marry whom?
    - Susan \_\_\_ Larry. → Susan wants to marry Larry.
- Complex morphological mismatches: e.g. Croatian
  - *Ana je zagrlila Marka, a Petar i Marko \_\_\_ Anu.*
  - *Ana je zagrlila Marka, a Petar i Marko **su zagrlili** Anu.*

# Reduced Relative Clauses as Ellipsis?

- Discussion with Zeping Liu:
  - Reduced Relative Clause as Garden-path
    - The horse raced past the barn fell.
    - The horse ~~that was~~ raced past the barn fell.
- Can SOTA NLP-pipelines process the reduced relatives?
  - Yes, but we get useless trees from the parser.



# Discontinuities

- Croatian: Split Islands, maintaining underlying linear sequencing
  - A B C -> A ... B ... C
  - Ivan se penje [ na [ veliko stablo ] ]
  - [ Na kakvo ] se Ivan penje [ stablo ] ?
  - [ Na kakvo ] se Ivan [ stablo ] penje?
  - C ... A B is ungrammatical in Croatian/Polish, etc.
  - \* [ Stablo ] se Ivan penje [ na veliko ]
- Polish: as Croatian
  - [ Na jakie ] się Marek [ drzewo ] wspina?

# Discontinuities

- German: NP Split generating reverse order
  - A B C -> C ... B ... A

Ich habe bislang noch [ keine grünen Gummibärchen ] gegessen.

[ Gummibärchen ] habe ich bislang [ grüne ] noch [ keine ] gegessen.

- A ... B ... C is ungrammatical in German:

\* [ Grüne ] habe ich noch [ keine ] bislang [ Gummibärchen ] gegessen.

# Parsing Discontinuities

- Crucial evaluation:
  - Dislocated and discontinuous elements in syntax have to be associated with their syntactic function and canonically unmarked position
- Corpus: Evaluation using ca. 100 examples
  - German: (Discontinuous Noun Phrases, Wh-movement)
  - Croatian (Split islands)
  - Polish (Split islands)
  - English (including Preposition stranding, topicalization)

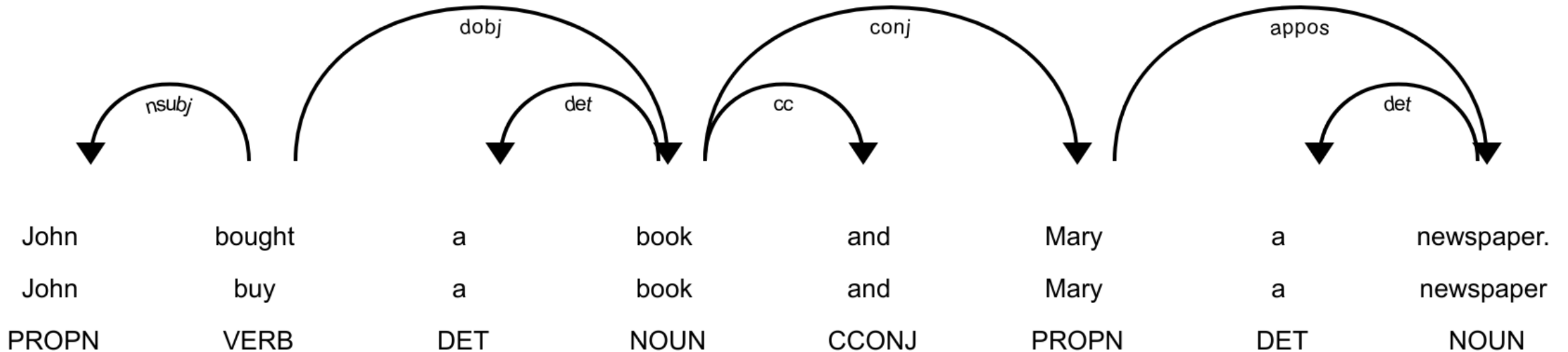
# Our Conclusion

- All current parsers are useless for downstream tasks processing:
  - Relations between concepts in clauses and sentences:
    - Concept → predicate → Concept relations/triples
    - Semantic analysis of events, temporal relations, and durations (temporal logic, event logic) using Graph-based approaches impossible using SOTA NLP
      - Graph-based models of events and temporal relations as Description Logic models with Graph transformations
  - Confirmed by an informal questionnaire study online
- Reasons: many → data, theory, and models

# Problem with Ellipsis and Discontinuities

- Current State of the Art (SOTA) Natural Language Processing-pipelines and parsers perform poorly (or not at all)
- Tested SOTA parsers:
  - Stanford CoreNLP
  - Stanford Stanza (V 1.6) (Dependency and Constituent Parser)
  - Berkley Neural Parser (benepar)
  - SpaCy 3.6
  - XLE (Web-XLE, LFG Parser for English)
- All parsers fail with Ellipsis and specific Discontinuities → not useful for downstream NLP tasks (e.g., relation extraction)

# Dependency Parsing Errors

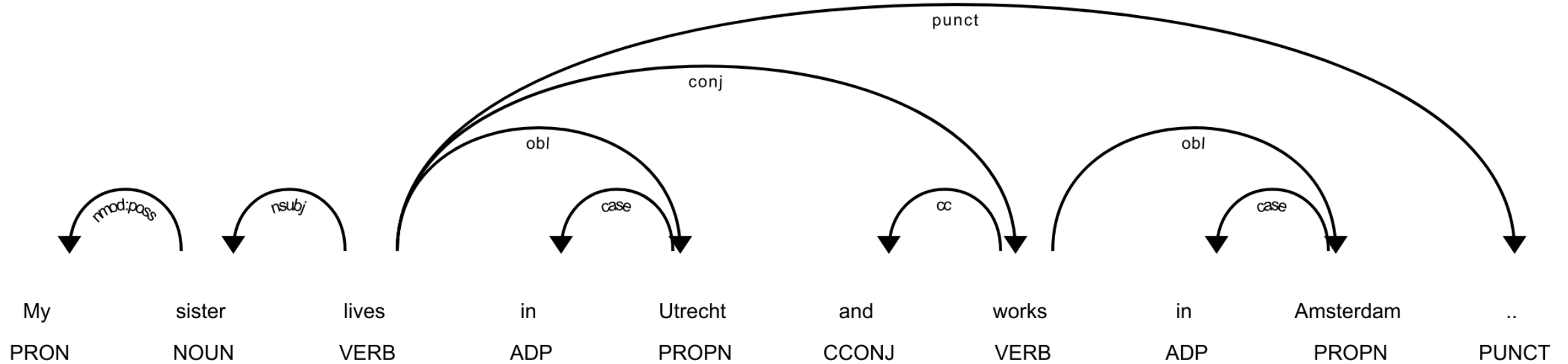


SpaCy 3.6

Resulting assumption:

John bought: (a book and Mary) (local coordination of two noun phrases); "a newspaper" is assumed to be a modifier or specifier of "Mary"

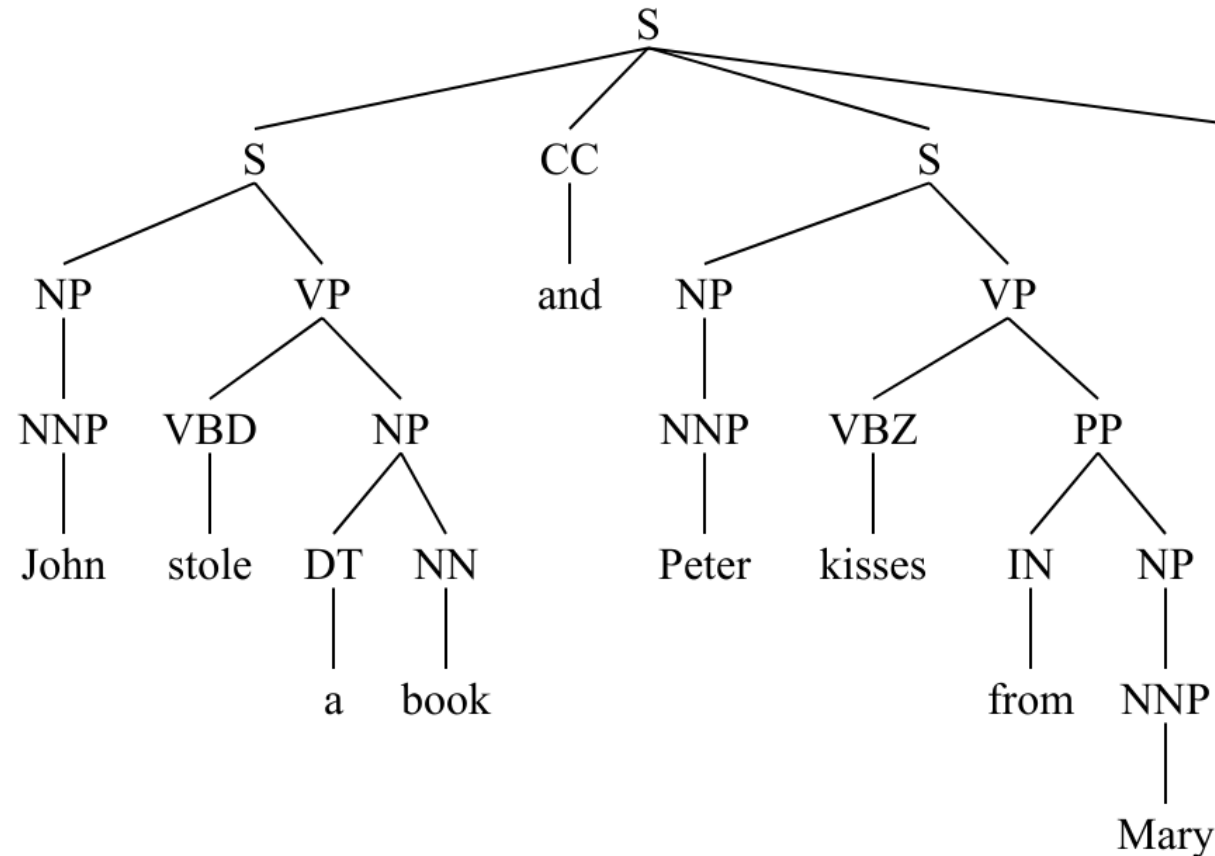
# Dependency Parsing Errors



Stanza V 1.6:

Correct coordination link (*conj*) for the two predicate heads, but missing subject in second conjunct.

# Constituent Parsing Errors

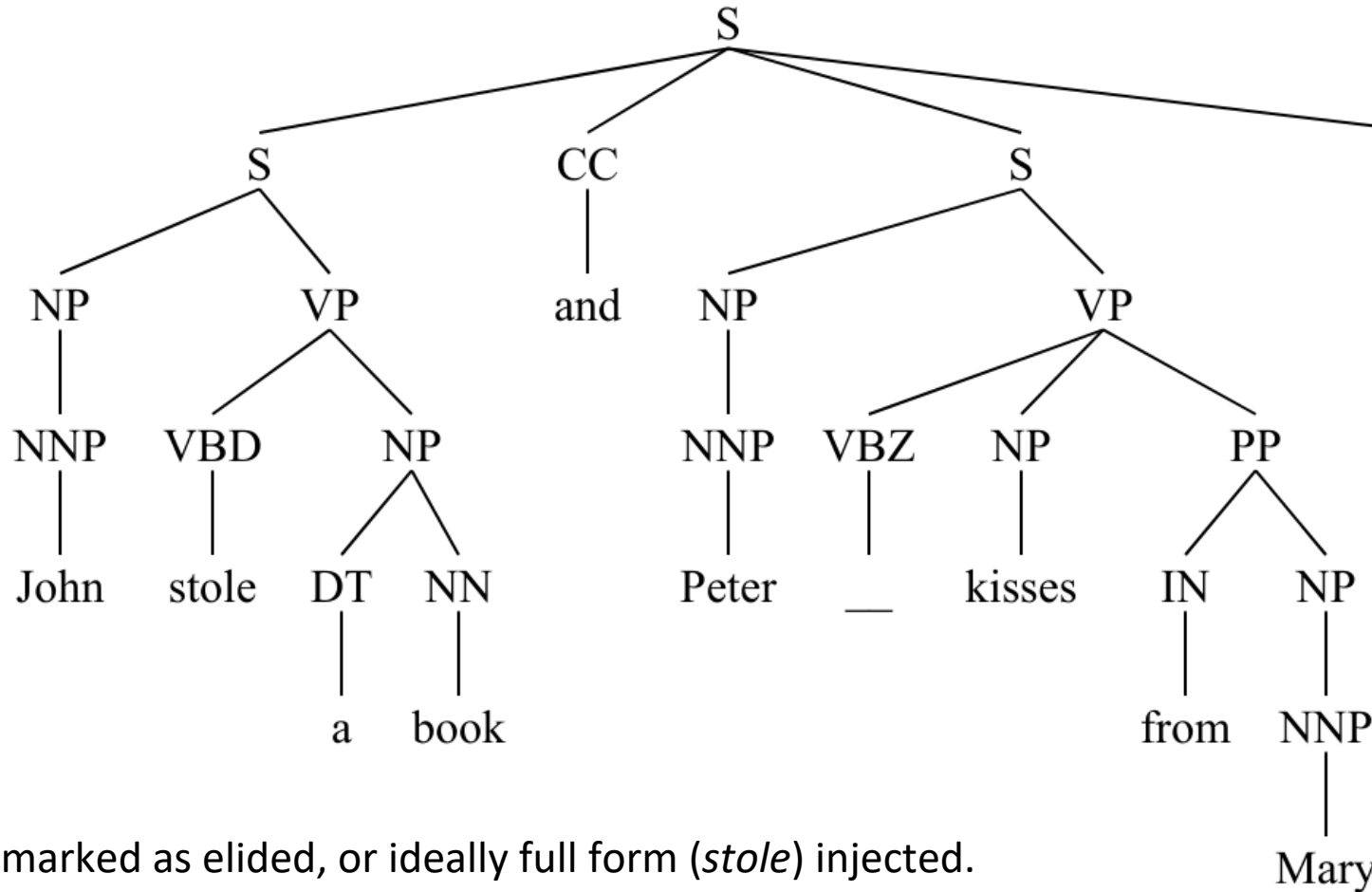


Berkley Neural Parser

Head Noun of the object (*kisses*) is assumed to be the predicate head of the second conjunct.



# Constituent Parsing Desired Output

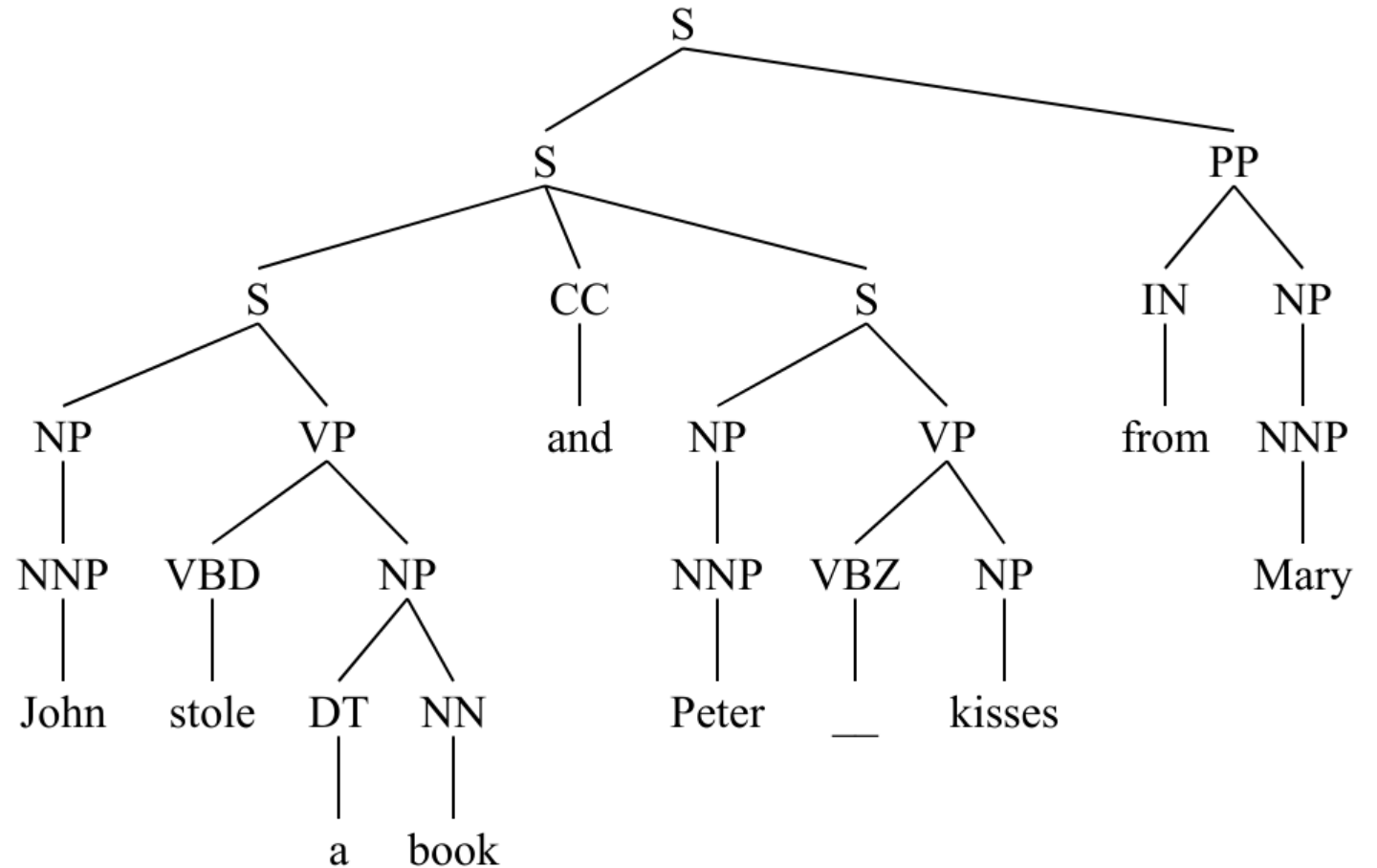


Predicate head marked as elided, or ideally full form (*stole*) injected.

# Constituent Parsing Desired Output

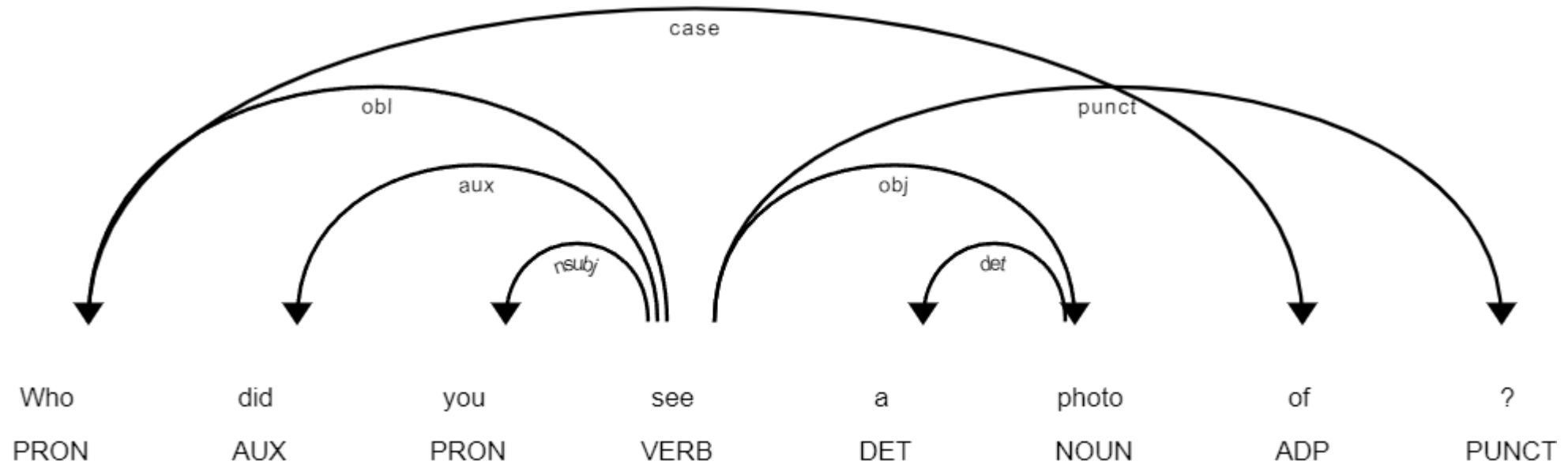
Alternative tree with Preposition Phrase (PP) indicating scope over both conjunct clauses.

(In previous tree the scope is only over second conjunct clause.)



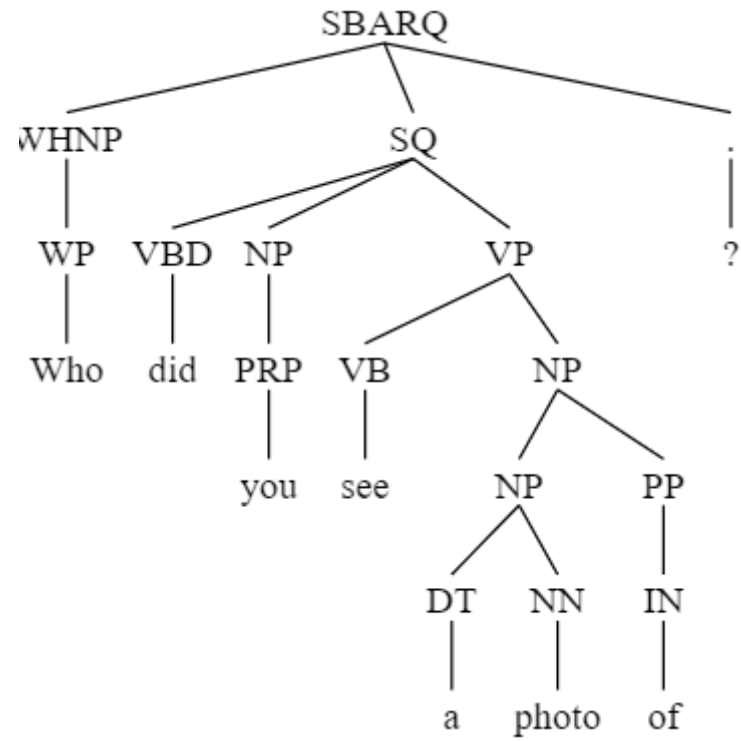
# Discontinuities

Preposition stranding in English using Stanza Dependency Parser:



# Discontinuities

Preposition stranding in English using Berkley Constituent Parser:

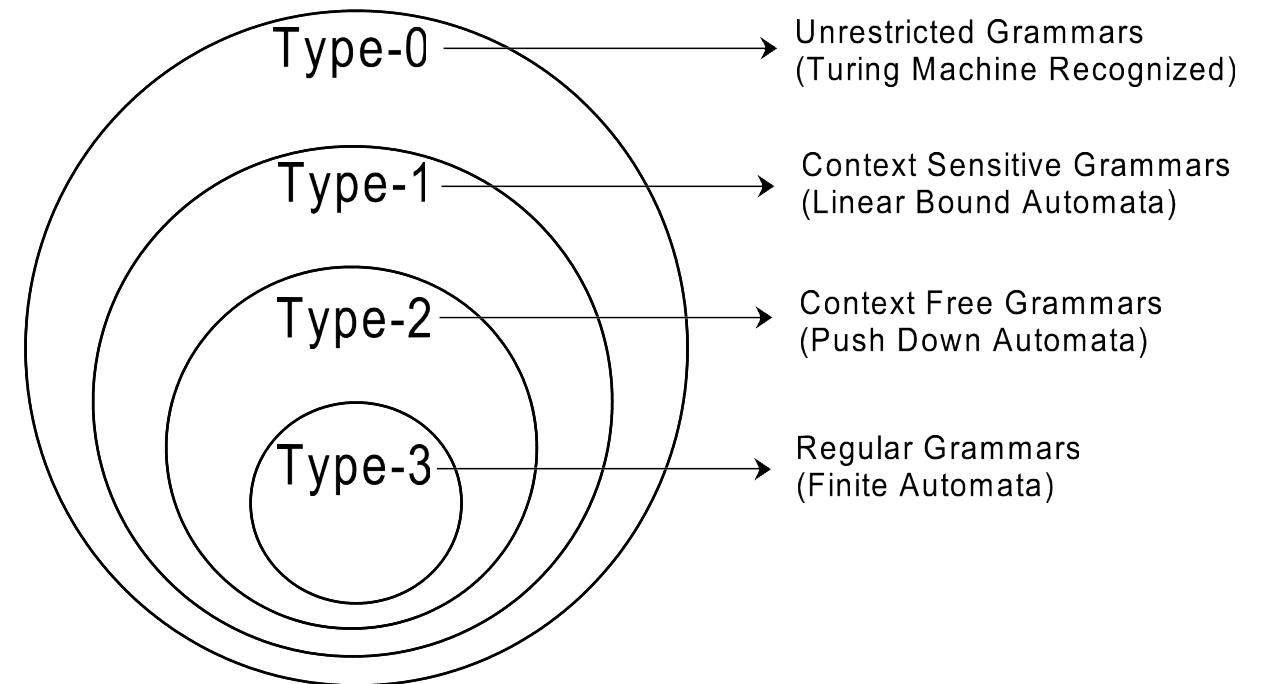


# Limitations of Models

- Complexity of Language and Expressivity of Grammar Formalisms
- Limitations of Neural Models

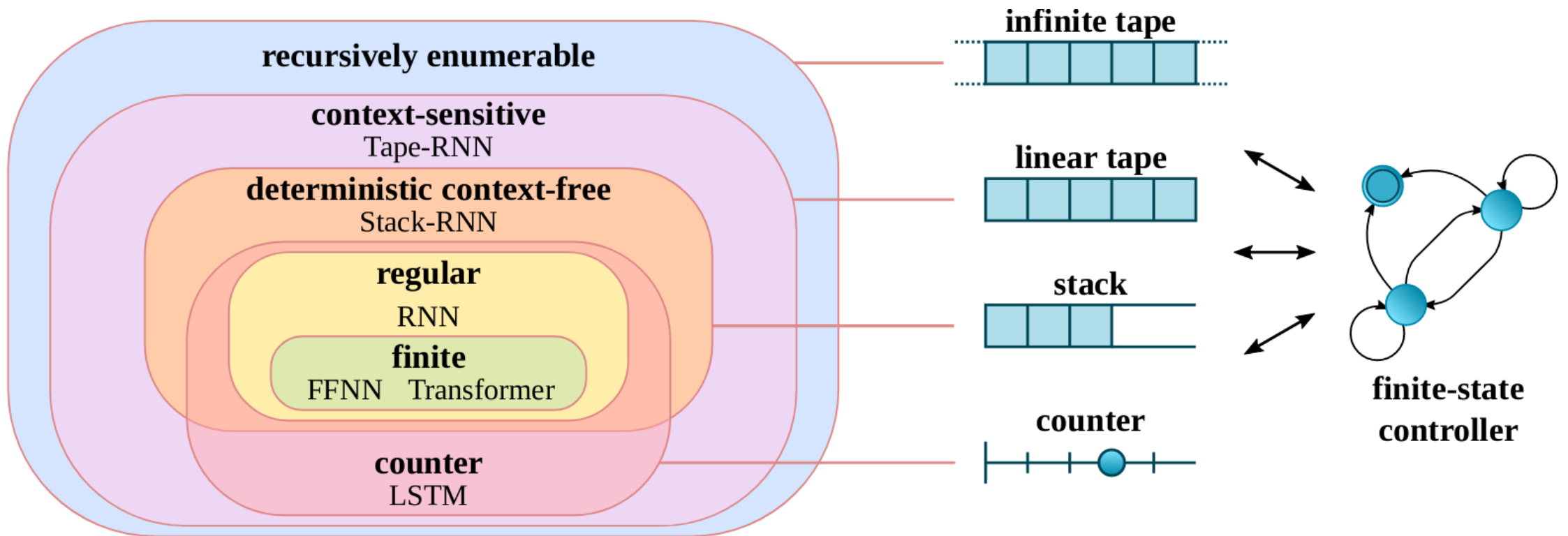
# Complexity of Language

- Chomsky Hierarchy
  - Phonology, Morphology in Type-3
  - Syntax in Type-2, partially in Type-1
  - Semantics ?



# Complexity of Language

- Neural Network Architectures and their corresponding Formal Languages Classes (Delétang et al., 2023)



# Neural Models

- Reliable generalization in Machine Learning and AI
  - How do neural networks generalize?
  - Insights from the theory of computation can predict the limits of neural network generalization in practice.
  - Grouping tasks according to the Chomsky hierarchy allows us to forecast whether certain architectures will be able to generalize to out-of distribution inputs



# Limitations of Neural Models

- Results in Delétang et al. (2023)
  - RNNs and Transformers fail to generalize on non-regular tasks
  - LSTMs can solve regular and counter-language tasks
  - Only networks augmented with structured memory (such as a stack or memory tape) can successfully generalize on context-free and context-sensitive tasks.
    - Stack-RNNs
    - Tape-RNNs

# Research Task

- Identify
  - a grammar formalism and/or
  - a network architecture
- that can predict empty elements and underlying continuities of phrasal segments.

# Hoosier Discontinuity Corpus

- Various types of discontinuities:
  - Wh-constructions
  - Topicalizations
  - Split Noun Phrases
  - Split Islands
- Other variables:
  - Within on clause
  - Across clause boundaries
- Languages:
  - German, English, Croatian, Polish, ...

# Indiana Ellipsis Corpus

- All ellipsis types
- 17 languages, including: English, German, Russian, Polish, Croatian, Ukrainian, Mandarin Chinese, Japanese, Korean, Navajo, Hindi, Gujarati, Arabic, Swedish, Spanish...
- Real examples:
  - SEC 10k Reports
  - Penn Treebank (Wall Street Journal)
  - Research papers and corpora

# Tasks

- Identify Ellipsis:
  - Random presentation of sentences
    - 800 with ellipsis, 800 without ellipsis
  - Identify the position of the ellipsis in the sentence only for examples with ellipsis
    - One elided position (potentially multiple words elided)
    - Scattered ellipsis
  - Identify the elided words

# Expectations

- Classification challenge:
  - LLMs should outperform statistical/ML baseline, and BERT-style transformers  
(→ remember, LLMs can replace NLP, is a common idea)
  - Guessing the position of elided elements: LLMs should outperform all other approaches
  - Guessing the elided elements:  
should be challenging for even LLMs in highly inflecting languages (e.g., Spanish, all Slavic, Hindi, Arabic)

# Baseline Classifier Task 1

- LR → supervised training
  - Training: 1,600 (50% ellipsis constructions)
- Accuracy: 72%
- Can be improved with a few more features, incl. unsupervised feature generation.

# Advanced Classifier Task 1

- Using BERT / transformer classifier
  - 10-fold rotation over 600 positive and 600 negative examples
  - Average accuracy: 0.94



# LLM Classifiers Task 1

- 0-shot classification: “Does this sentence contain ellipsis?”
- LLMs:
  - GPT 3.5
  - GPT 4
  - Llama2
  - Zephyr

# LLM Classifier Task 1

- GPT 3.5
  - Precision: 0.33, Recall: 0.44, F1-Score: 0.38, Accuracy: 0.35
- GPT 4
  - Precision: 0.55, Recall: 0.67, F1-Score: 0.60, Accuracy: 0.60
- Llama2
  - Precision: 0.40, Recall: 0.67, F1-Score: 0.50, Accuracy: 0.40
- Zephyr
  - Precision: 0.25, Recall: 0.11, F1-Score: 0.15, Accuracy: 0.42

# Preliminary Results

- Supervised ML/NLP methods outperform all LLMs on 0-shot
- GPT with default temperature (0.7)
  - Randomizes 20% of the output decisions, i.e. for 20% of repeated tasks with the same data the classifier will be switched.
- GPT with temperature set to 0
  - No random decisions → deterministic, but:
  - Drop of accuracy by 10% over sample data