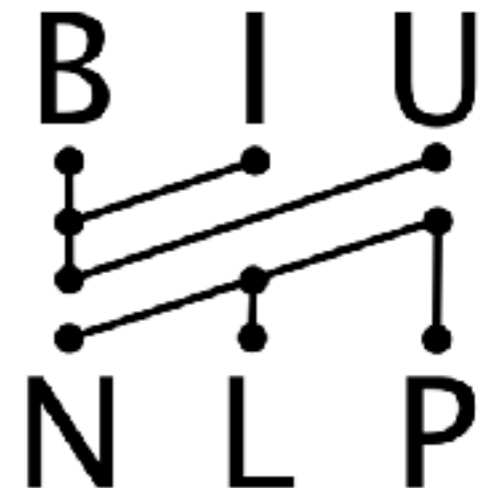


# Near human performance in question answering?

**Yoav Goldberg**  
Bar Ilan University



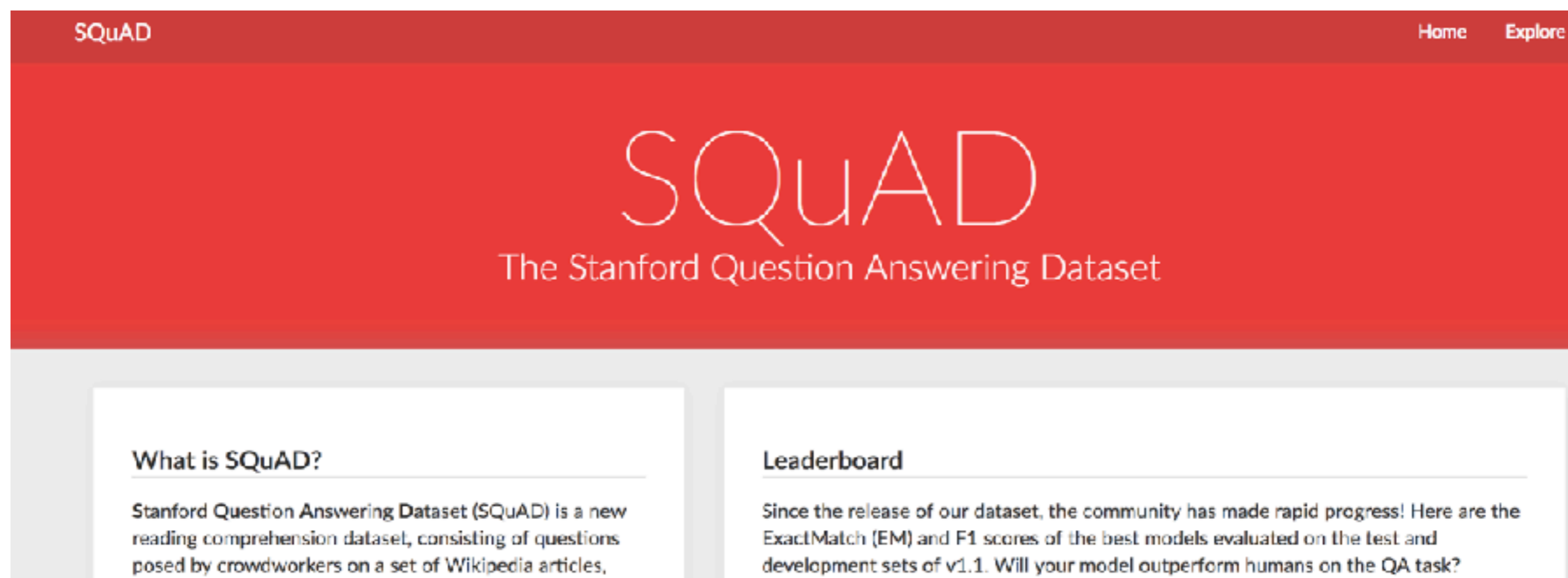
# SQuAD: 100,000+ Questions for Machine Comprehension of Text

**Pranav Rajpurkar** and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**

`{pranavs, zjian, klopyrev, pli}@cs.stanford.edu`

Computer Science Department

Stanford University



The image shows a screenshot of the SQuAD website homepage. The top navigation bar is red with the text 'SQuAD' on the left and 'Home' and 'Explore' on the right. The main content area has a red background with the word 'SQuAD' in large white letters, and below it, 'The Stanford Question Answering Dataset' in smaller white text. Below this, there are two white boxes with grey borders. The left box is titled 'What is SQuAD?' and contains the text: 'Stanford Question Answering Dataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles,'. The right box is titled 'Leaderboard' and contains the text: 'Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1. Will your model outperform humans on the QA task?'

SQuAD

Home

Explore

# SQuAD

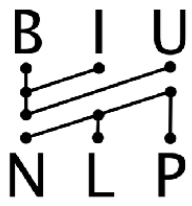
The Stanford Question Answering Dataset

## What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles,

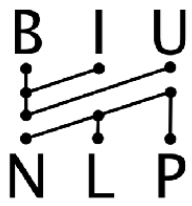
## Leaderboard

Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1. Will your model outperform humans on the QA task?



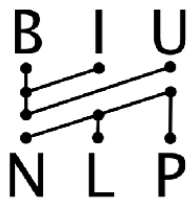
# Near Human Performance on the SQUAD Dataset

- "Neural systems achieve near-human performance on Question Answering".
- Not really, let's see why.



# Restricted QA Setup

- Restricted to questions that can be answered by **span selection**.
- Need to find the answer in a **given paragraph**.
- The answer is **guaranteed to be** in the paragraph.
- Annotators see the paragraph when asking the question, resulting in **high lexical similarity** between question and answer.



# Near Human Performance?

- Human performance: 91.2 F1.  
Current best system: 84.0 F1.
- Humans on MTurk. Who are instructed to answer 5 answers in 2 minutes. That's 16 cent per question.
- Humans are wrong mostly in span boundaries.
- Also, when doing max-vote between several humans, human perf goes up substantially.



# How hard is the dataset?

- Do the questions require complex reasoning, or can they be "cheated" using superficial clues?

| Reasoning                           | Description  | Example  | Percentage |
|-------------------------------------|--|--|------------|
| Lexical variation (synonymy)        | Major correspondences between the question and the answer sentence are synonyms.   | Q: What is the Rankine cycle sometimes <b>called</b> ?<br>Sentence: The Rankine cycle is sometimes <b>referred</b> to as a <u>practical Carnot cycle</u> .   | 33.3%      |
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve.   | Q: Which <b>governing bodies</b> have veto power?<br>Sen.: <u>The European Parliament and the Council of the European Union</u> have powers of amendment and veto during the legislative process.  | 9.1%       |
| Syntactic variation                 | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar is <b>currently on the faculty</b> ?<br>Sen.: <b>Current faculty include</b> the anthropologist Marshall Sahlins, ..., Shakespeare scholar <u>David Bevington</u> .  | 64.1%      |
| Multiple sentence reasoning         | There is anaphora, or higher-level fusion of multiple sentences is required.   | Q: What collection does <b>the V&amp;A Theatre &amp; Performance galleries</b> hold?<br>Sen.: <b>The V&amp;A Theatre &amp; Performance galleries</b> opened in March 2009. ... <b>They</b> hold the UK’s biggest national collection of <u>material about live performance</u> . | 13.6%      |
| Ambiguous                           | We don’t agree with the crowdworkers’ answer, or the question does not have a unique answer.   | Q: What is the main goal of criminal punishment?<br>Sen.: <b>Achieving crime control via <u>incapacitation and deterrence</u></b> is a major goal of criminal punishment.  | 6.1%       |

**Table 3:** We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

| Reasoning                    | Description  | Example  | Percentage |
|------------------------------|--|--|------------|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes <b>called</b> ?<br>Sentence: The Rankine cycle is sometimes <b>referred</b> to as a <u>practical Carnot cycle</u> . | 33.3%      |

The Rankie cycle is sometimes referred to as \_\_\_\_\_

words from questions      some verb      **only NP in sent**

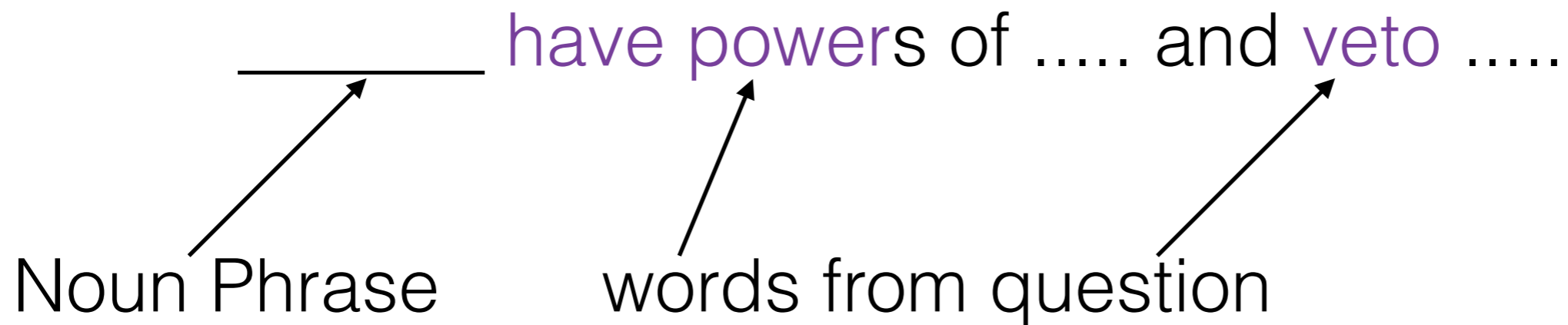


Lexical variation  
(world knowledge)

Major correspondences between the question and the answer sentence require world knowledge to resolve.

Q: Which **governing bodies** have veto power?  
Sen.: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.

9.1%

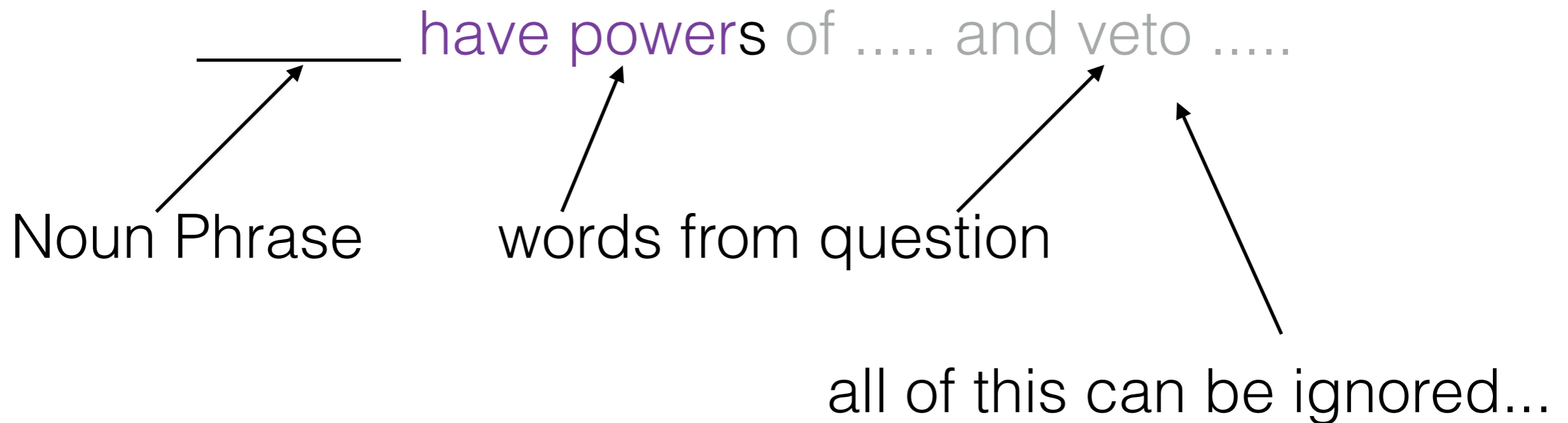


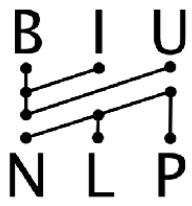
Lexical variation  
(world knowledge)

Major correspondences between the question and the answer sentence require world knowledge to resolve.

Q: Which **governing bodies** have veto power?  
Sen.: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.

9.1%



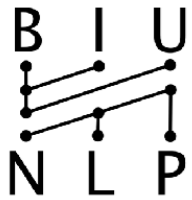


---

|                     |  |   |       |
|---------------------|--|---|-------|
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar is <b>currently on the faculty</b> ?<br>Sen.: <b>Current faculty include</b> the anthropologist Marshall Sahlins, ..., Shakespeare scholar <u>David Bevington</u> . | 64.1% |
|---------------------|--|---|-------|

What Shakespeare scholar is ....?

... Shakespeare scholar \_\_\_\_\_



hold

collection

Multiple sentence reasoning

There is anaphora, or higher-level fusion of multiple sentences is required.

Q: What collection does **the V&A Theatre & Performance galleries** hold?

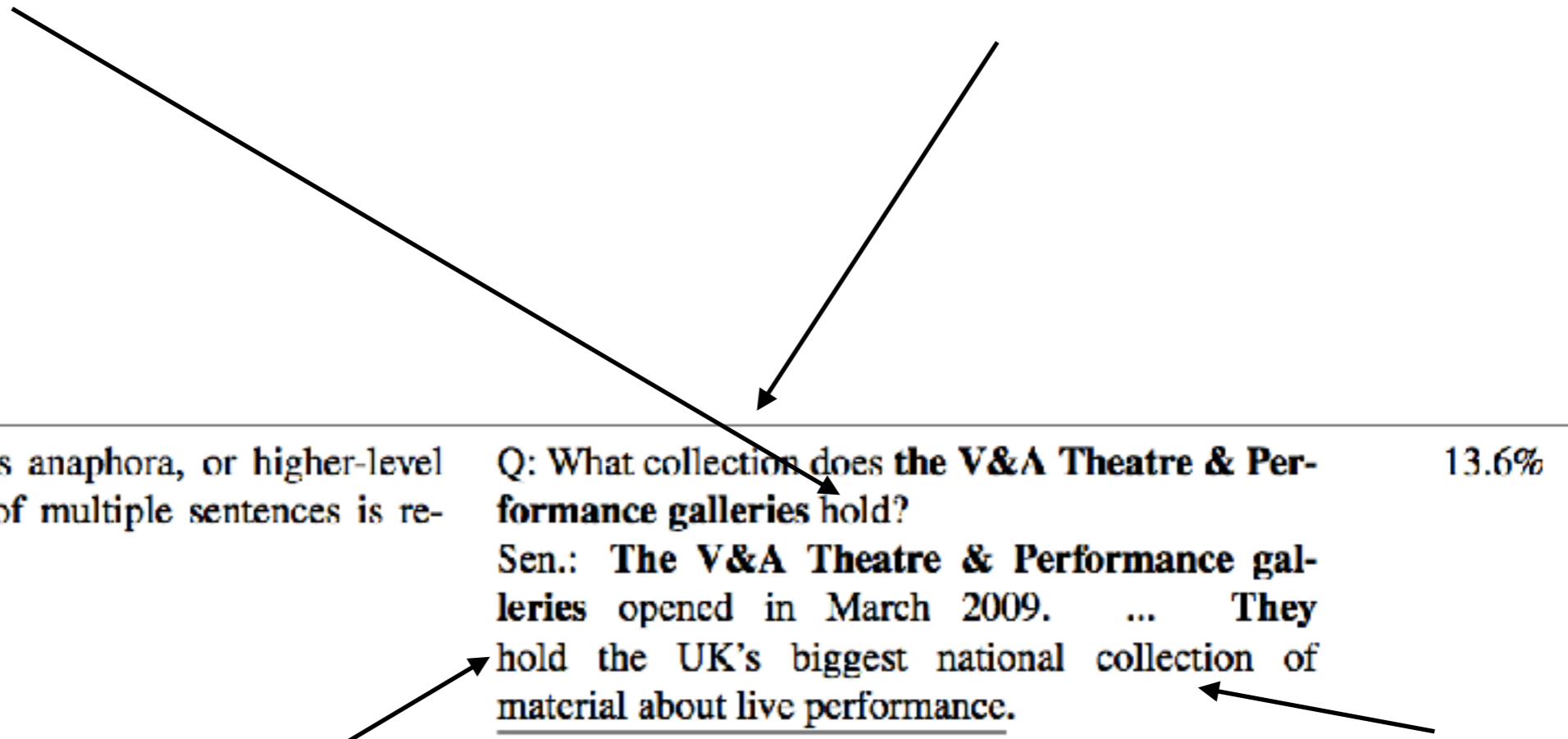
13.6%

Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.

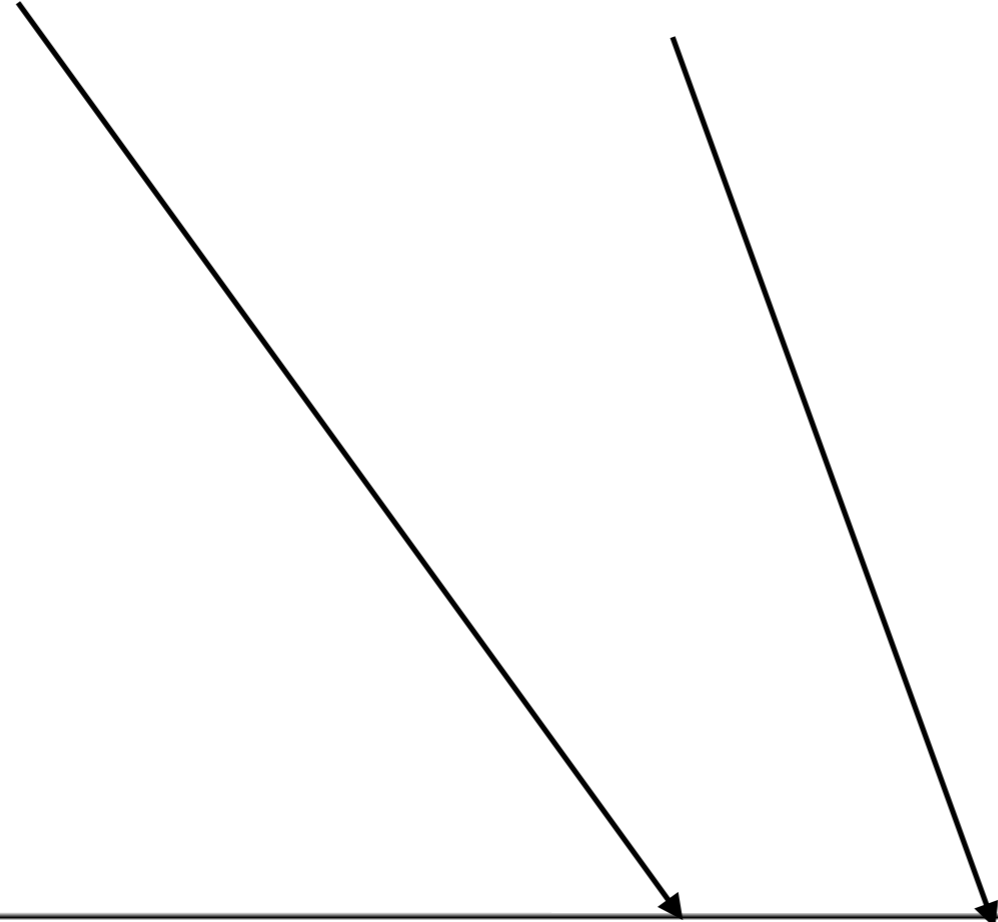
hold

Noun Phrase

collection

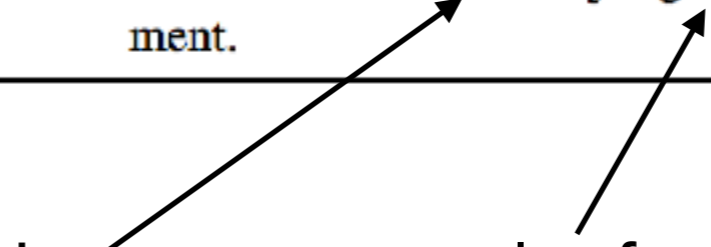


what is goal of criminal punishment



|                  |   |   |             |
|------------------|---|---|-------------|
| <p>Ambiguous</p> | <p>We don't agree with the crowdworkers' answer, or the question does not have a unique answer.</p> | <p>Q: What is the main goal of criminal punishment?<br/>Sen.: <b>Achieving crime control via <u>incapacitation and deterrence</u></b> is a major goal of criminal punishment.</p> | <p>6.1%</p> |
|------------------|---|---|-------------|

\_\_\_\_\_ is goal of criminal punishment





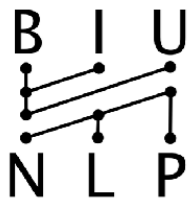
all can be "cheated away" using  
some smart template matching.



all can be "cheated away" using  
some smart template matching.

and the template-matching systems  
can be easily fooled by tailored examples  
that don't fool humans

(Percy Liang, personal communication)



all can be "cheated away" using  
some smart template matching.

and the template-matching systems  
can be easily fooled by tailored examples  
that don't fool humans

**Update:** paper by Jia and Liang demonstrates this:

**Adversarial Examples for Evaluating Reading Comprehension Systems**

**Robin Jia**

Computer Science Department  
Stanford University

robinjia@cs.stanford.edu

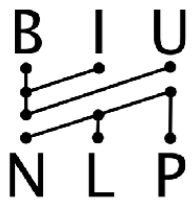
**Percy Liang**

Computer Science Department  
Stanford University

pliang@cs.stanford.edu

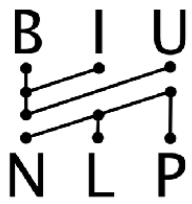
<https://arxiv.org/abs/1707.07328>





# To Summarize

- DL methods gets near human performance on SQUAD but:
  - Still 84 F1 vs. 91.2 F1.
  - Restricted QA Setting (span selection, within paragraph, answer always present, high lexical overlap).
  - Compared to under-incentivized humans.
    - (91.2 is a low estimate of human performance)
  - Questions can be answered with "cheating".
    - (84.0 is a high estimate of DL performance)



# Take away

- Neural systems / RNNs / ConvNets do very clever pattern matching. Not "intelligence", not "reasoning".
- Not everything can be pattern-matched.
- Pattern matchers can be easily fooled.