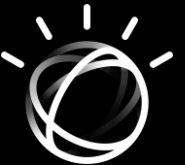


Engineered AI Still Matters for Question Answering

J. William Murdock, Lin Pan, Chung-Wei Hang, Mary Swift,
Zhiguo Wang, Chris Nolan, Prathyusha Peddi, Nisarga Markandaiah,
Eunyoung Ha, Kazi Hasan, Yang Yu, Wei Zhang

IBM Watson



Stanford Question Answering Dataset (SQuAD)

- Reading comprehension data
 - Passages
 - Questions about the passages
 - Answers in the passages
- Wikipedia passages
- Crowd workers saw the passages, wrote questions, and selected answers
- Very popular for statistical reading comprehension research

... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. ...

Where do water droplets collide with ice crystals to form precipitation?

Why should anyone care about SQuAD?

- Answering reading comprehension questions is an interesting AI challenge
- **Not** a particularly useful capability by itself
 - Users do not want to provide a passage + a question and ask for an answer from that passage
- Important subtask of factoid question answering
- Combine a system built for SQuAD with a passage search capability

Hypothesis

A system that excels at SQuAD will also excel at factoid question answering

Engineered
Multi-Strategy

SQuAD

Statistical
Single-Strategy

Factoid

Factoid-1527

- Factoid question answering data
 - Answers are typically entities or numbers
- Fairly small (1,527 questions total)
- Questions written without being tied to a specific piece of text
- We use Wikipedia and Wiktionary as sources
- IBM confidential

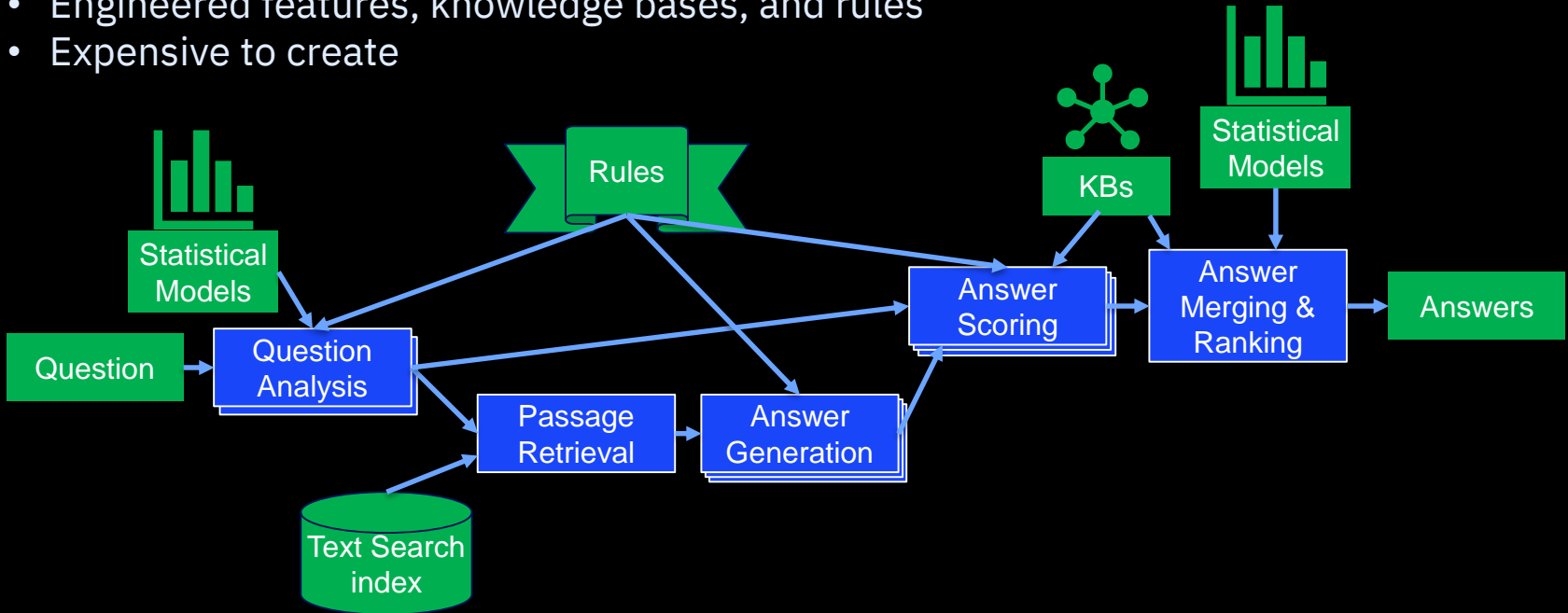
In what year did William Bligh arrive in Tahiti?

1788

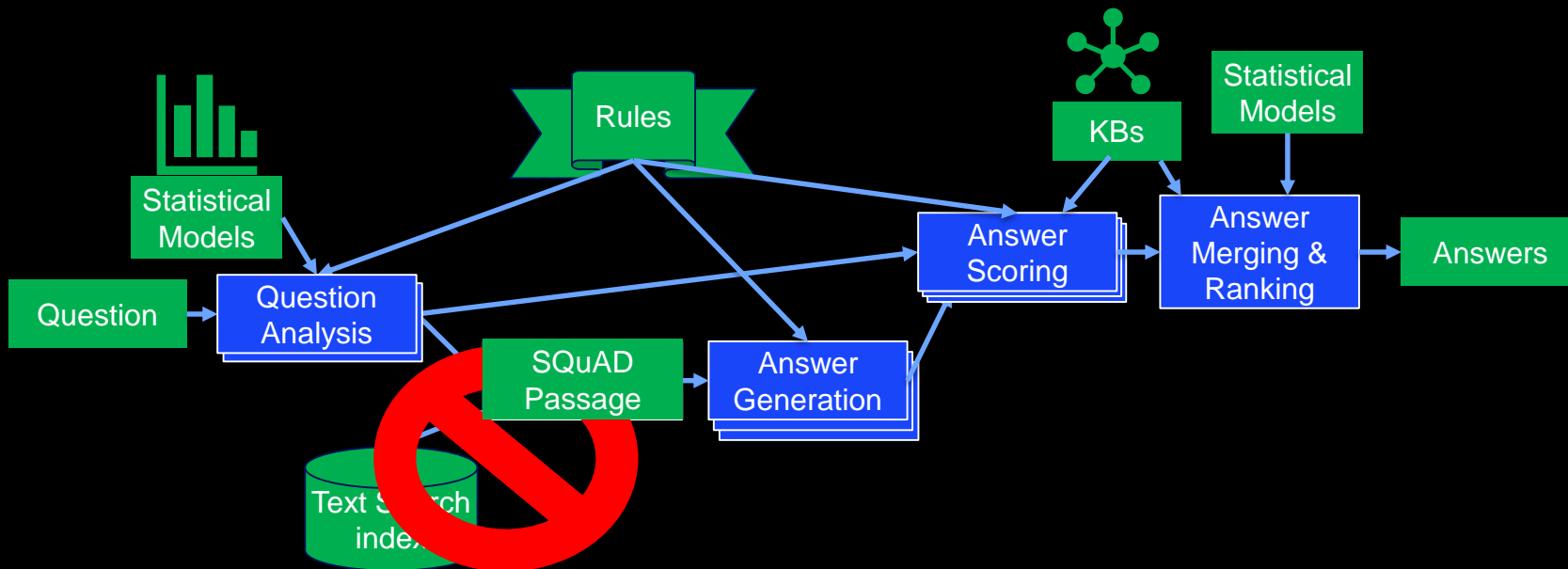
(not a real example)

DDQA: Multi-Strategy Factoid Question Answering

- DDQA = Discovery DeepQA
- Simpler version of IBM Watson 1.0, designed for cloud
- Engineered features, knowledge bases, and rules
- Expensive to create

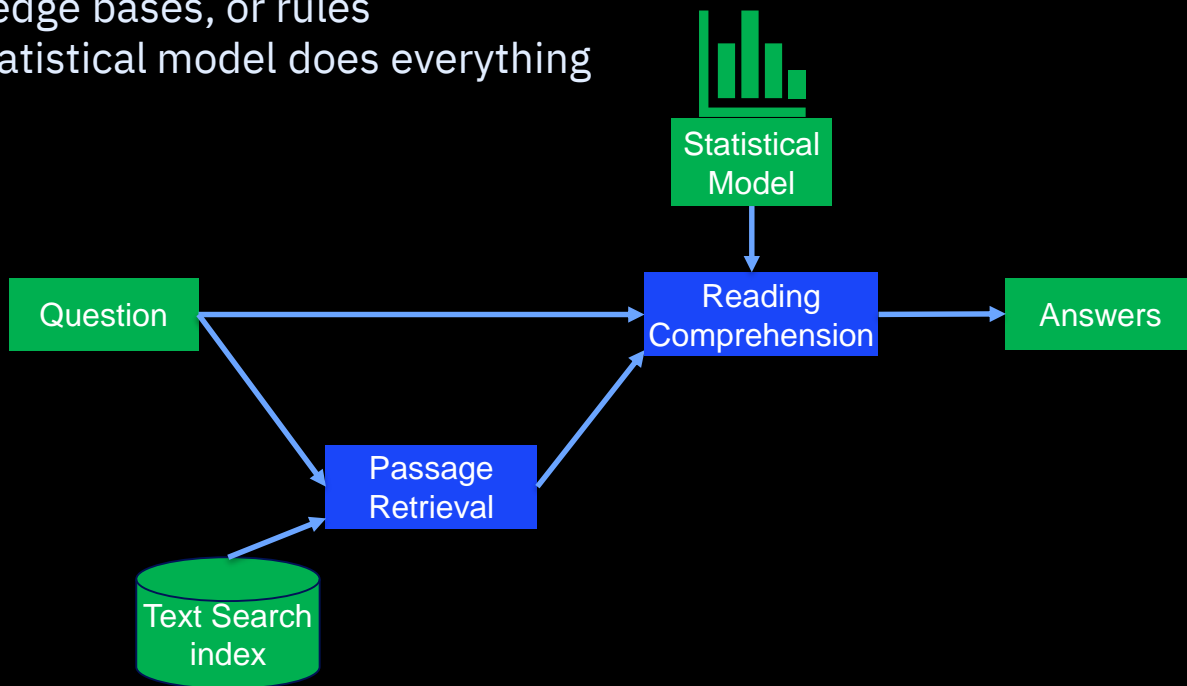


DDQA for SQuAD

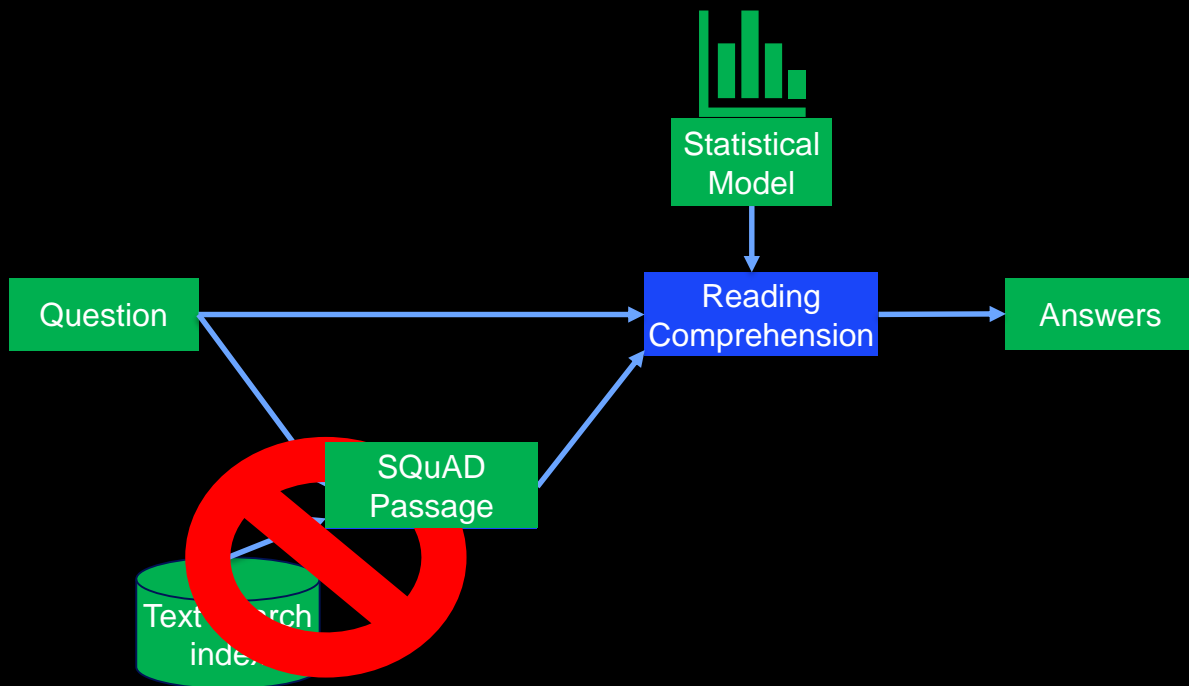


Single-Strategy Statistical Factoid Question Answering

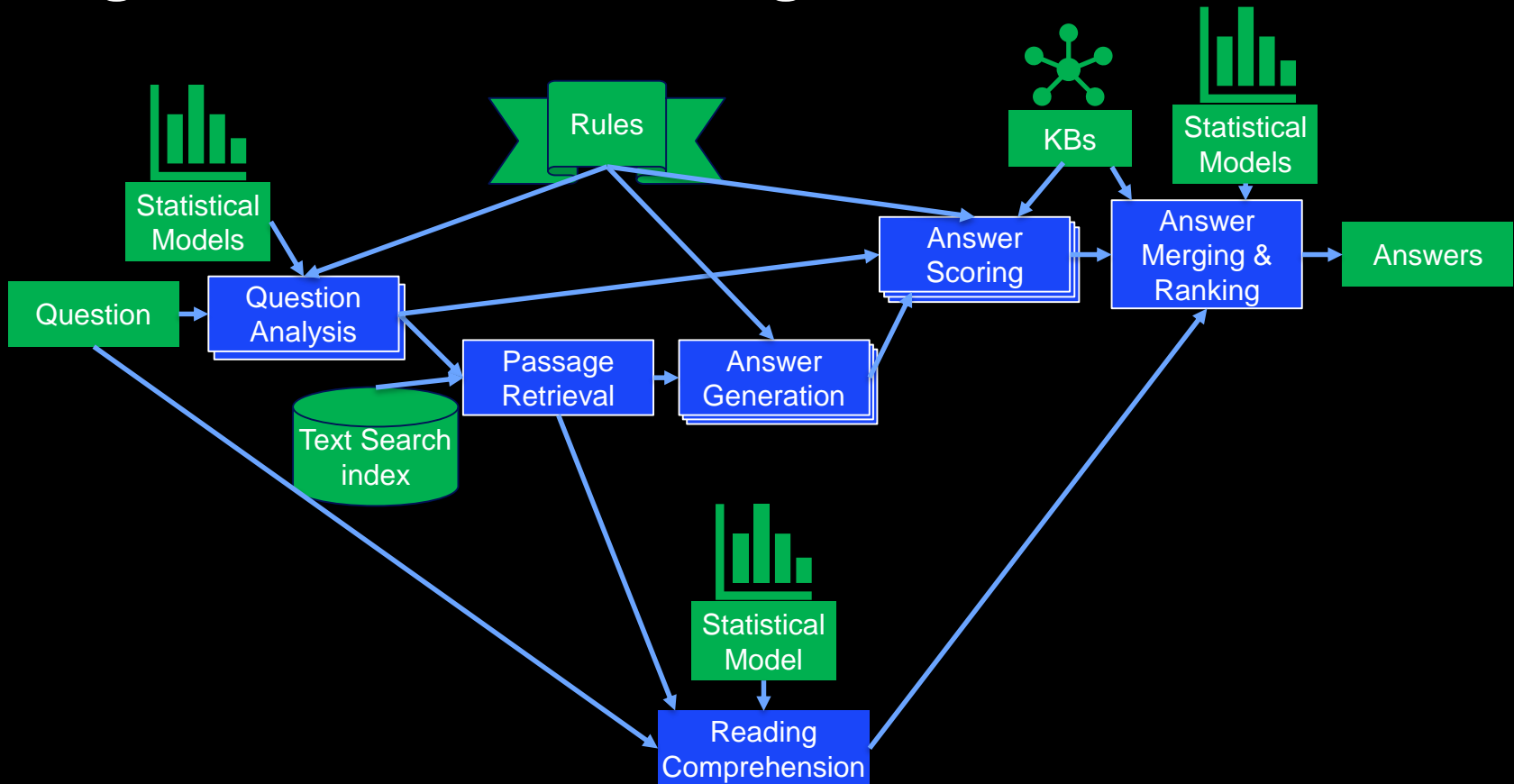
- No manually engineered features, knowledge bases, or rules
- One statistical model does everything



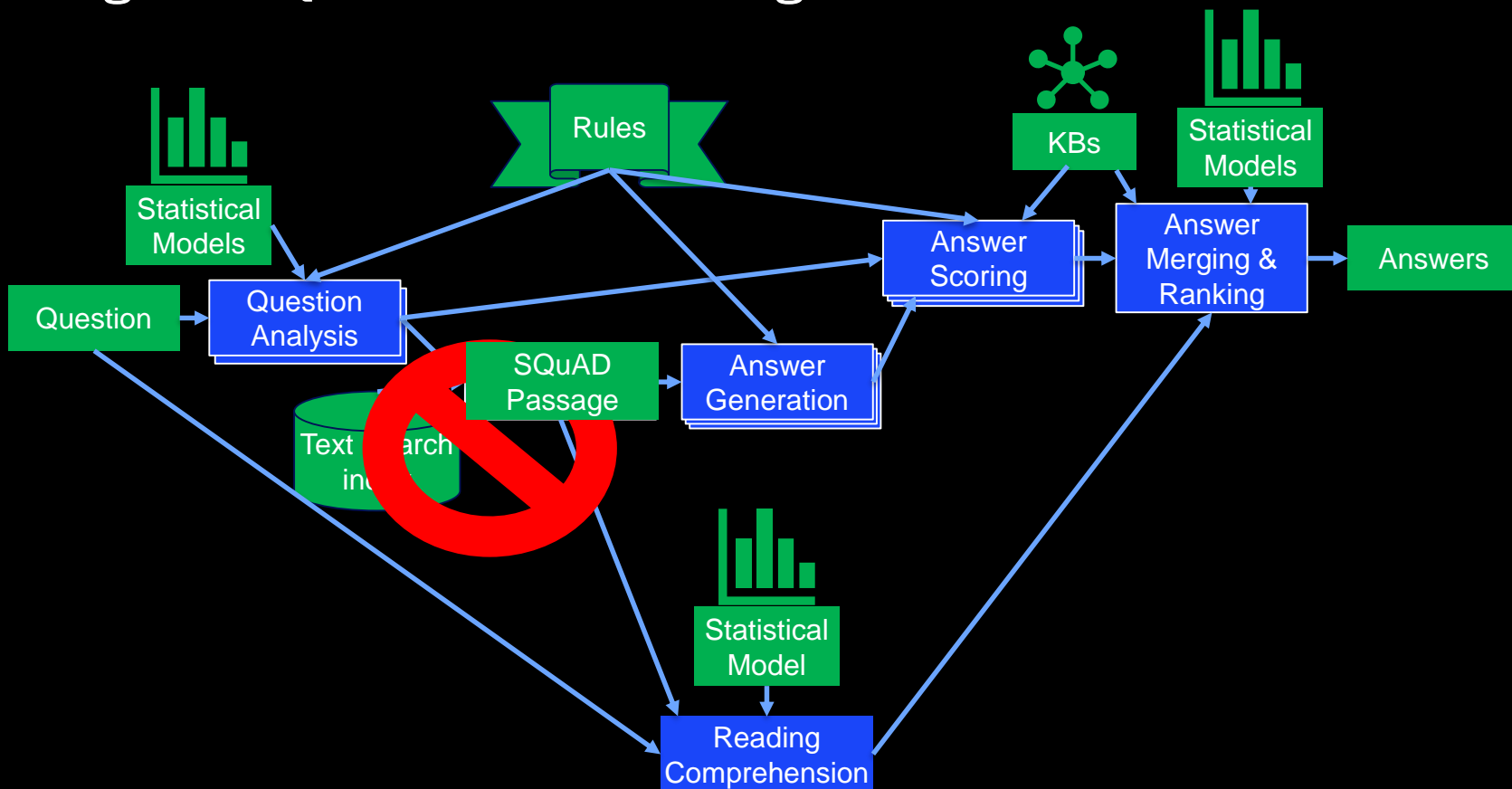
Single-Strategy Statistical Question Answering for SQuAD



Integrated Question Answering

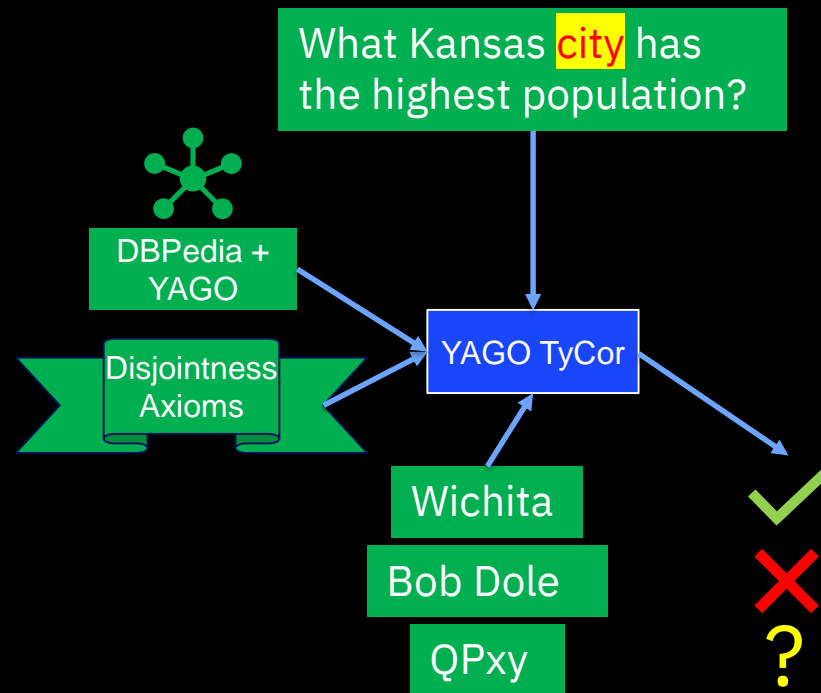


Integrated Question Answering



YAGO TyCor: A DDQA example component

- DBpedia*: knowledge-base of structured information from Wikipedia
- YAGO**: semi-automatically constructed taxonomy from WordNet, Wikipedia, etc.
- YAGO Disjointness Axioms***: axioms built for IBM Watson 1.0 specifying disjoint types, e.g., *person, location*



* F. M. Suchanek, et al. YAGO: A core of semantic knowledge-unifying WordNet and Wikipedia. *WWW 2007*.

** C. Bizer, et al. Dbpedia: A crystallization point for the web of data. *Journal of Web Semantics*, 2009.

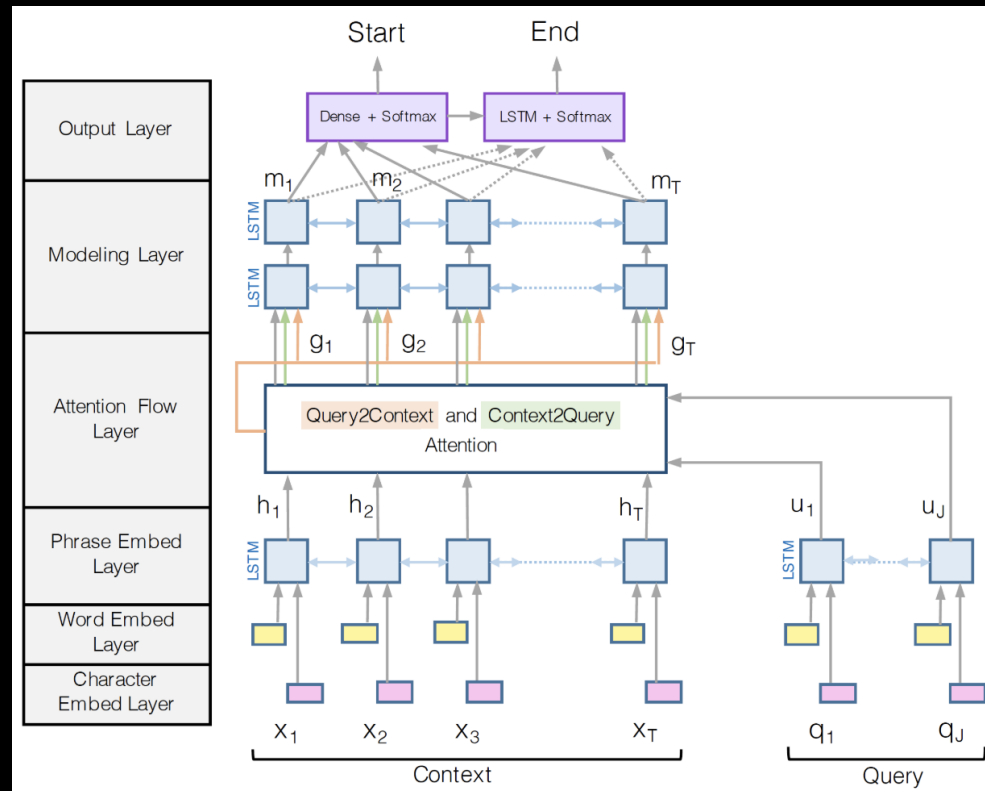
*** J. Murdock, et al. Typing candidate answers using TyCor. *Journal of IBM R&D*, 2012.

YAGO TyCor is very expensive!

- Dbpedia and YAGO are free for us because they already exist
- WordNet and Wikipedia already existed
- WordNet and Wikipedia both took huge investments of effort
- If we wanted to do something similar for copper mining it would cost a lot
- We created the disjointness axioms
- And the logic to reason about entities and types
- *And this is just one of dozens of components!*

Bidirectional Attention Flow

- Off-the-shelf Deep Neural Net
- Some engineering on structure
- No manually engineered features



IBM would like statistics alone to be best

- Recall: YAGO TyCor is very expensive and one of many
- Recall: Bidirectional Attention Flow is relatively cheap
- We would like to hear that the cheap system is also the best

Metrics

– Exact Match

- % of questions where the top-ranked answer exactly matches the answer-key

– Mean Reciprocal Rank

- Average across all questions of the reciprocal rank of the highest ranked correct answer
- First answer correct gets 1, second gets $\frac{1}{2}$, third gets $\frac{1}{3}$, etc.

– Other metrics: See paper

SQuAD Results

	Exact Match	Mean Rec. Rank
Statistical	66%	71%
DDQA + Statistical	67%	73%

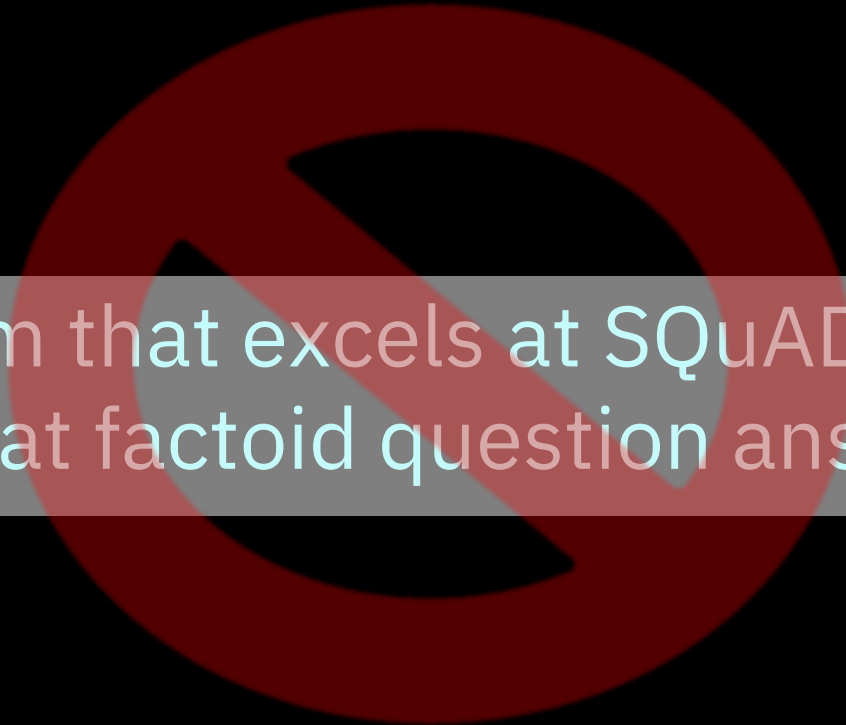
- The statistical system alone provides nearly all of the power.
- Adding DDQA provides very little benefit despite all of its great cost.

Factoid-1527 Results

	Exact Match	Mean Rec. Rank
Statistical	15%	21%
DDQA + Statistical	47%	56%

- The statistical system alone provides very little power.
- Adding DDQA provides enormous benefit.

Hypothesis: Not Confirmed



A system that excels at SQuAD will also excel at factoid question answering

Needs more evidence

Is SQuAD a toy problem?

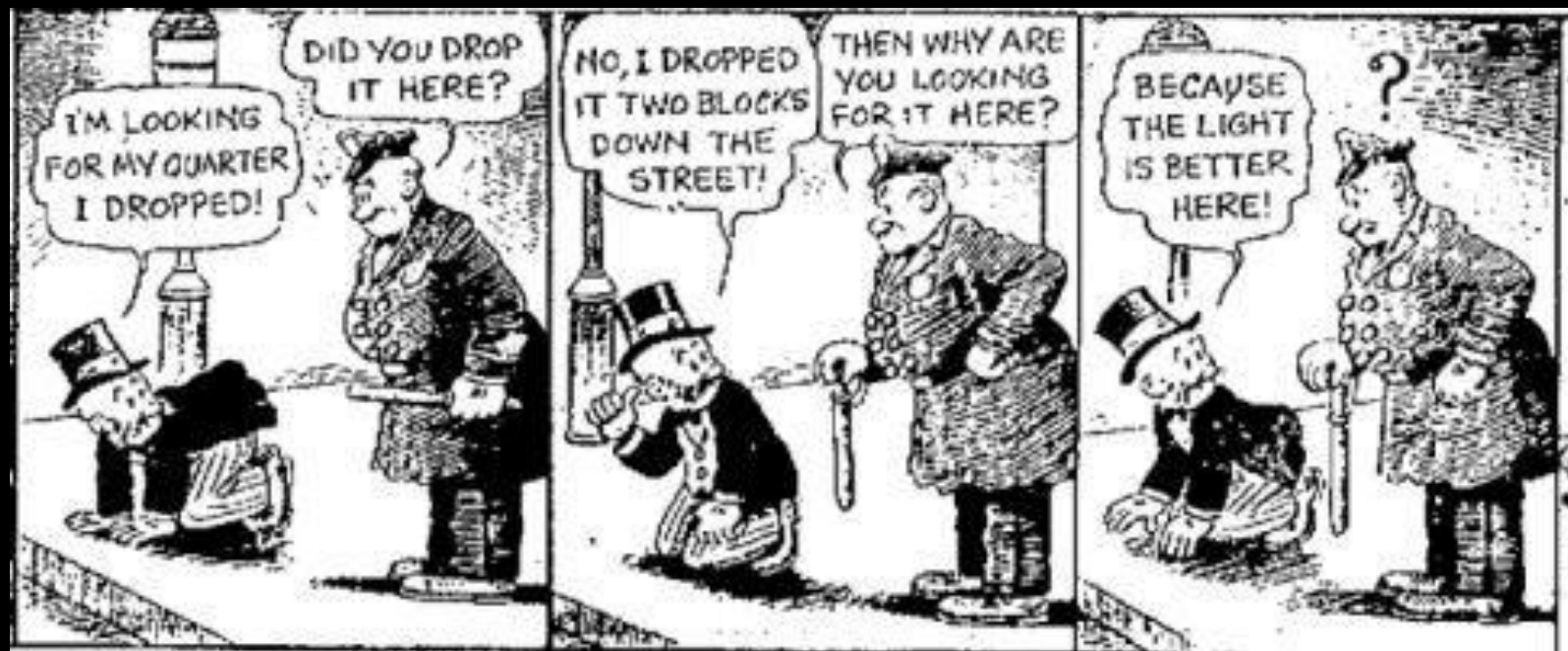
- Recall: Not a particularly useful capability by itself
- Is a statistical system learning reading comprehension? Or just statistical trends in how people write questions for given passages?
- (Not talking about the metaphysical question)
- Are the systems learning to identify answers given (*passage, question*) or given (*passage, question-written-for-that-passage*)?

... Various species of poison dart frogs secrete lipophilic alkaloid toxins through their flesh ...

What are dart frogs are known to secrete?

Amphibians secrete a wide diversity of chemicals from skin glands as defense against predators, parasites, and pathogens. Most defensive chemicals are produced endogenously through biosynthesis, but poison frogs sequester lipophilic alkaloids from dietary arthropods. *

* A. M. Jeckel, et al. The relationship between poison frog chemical defenses and age, body size, and sex. *Frontiers in Zoology*, 2015



Is factoid a toy problem?

- Users do not really want a system that only answers factoid questions
- *Most* information needs require more than just an entity or quantity to address
- *Sometimes* users do ask factoid questions
- When they do, it is nice to provide a (correct) factoid answer
- The complete factoid question answering problem seems *closer* to a real-world use case than SQuAD reading comprehension.
- This will only be *proven* when we show that it adds value as part of a comprehensive information finding system.

- Google and WolframAlpha work very well for some kinds of factoid questions
- To our knowledge, nobody has a big commercial success doing **narrow-domain** factoid question answering using **customer supplied content**.

Google Search Results:

What is the capital of New Jersey?

About 206,000,000 results (1.01 seconds)

New Jersey / Capital

Trenton

People also search for New Jersey Newark

Trenton, New Jersey - <https://en.wikipedia.org/wiki/Trenton>
Trenton became New Jersey's within Trenton Township on No Wilbur, Trenton, New Jersey · T

New Jersey - Wikipedi https://en.wikipedia.org/wiki/New_Jersey
New Jersey is a state in the M Camden, Paterson, Newark, Tr ... George Washington in Morr Revolution".

Capital: Trenton Popul Largest city: Newark Largest

New Jersey State Capital | Trenton - State Symbols USA <https://statesymbolsusa.org/place/new-jersey/capitals-cities-towns/trenton>
Located on the Delaware river, the city of Trenton is the state capital of New Jersey and the seat of Mercer County.

WolframAlpha Search Results:

What is the capital of New Jersey?

Input Interpretation: New Jersey capital

Result: Trenton, New Jersey, United States

city population	84964 people (country rank: +41 st) (2017 estimate)
urban area population	258472 people (Trenton, NJ urban area) (country rank: 120 th) (2000 estimate)
metro area population	367063 people (Trenton-Camden metro area) (country rank: 140 th) (2011 estimate)

Location: Mercer County, New Jersey

World map Show coordinates

DISCOVER WHAT'S POSSIBLE with Wolfram|Alpha
Take the Tour

COMPUTE & VISUALIZE JUST ABOUT ANYTHING

Boltzmann Bosons
Fermions

Engineered AI still matters

- Mounting evidence from data like SQuAD: single-strategy deep neural networks are the state-of-the-art for answering questions
- Might be an artifact of the limitations of SQuAD and similar data sets
- Need more experiments with more data
- We believe that there is still a significant role for multi-strategy systems that make extensive use of engineered knowledge and rules.