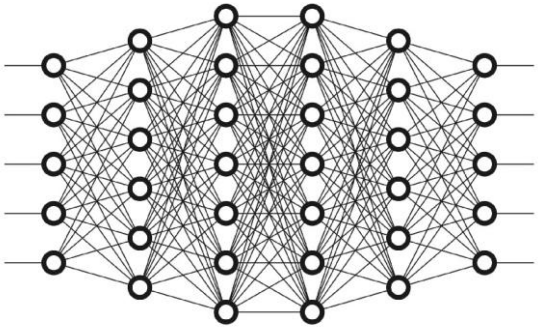
# Introduction to Question Answering

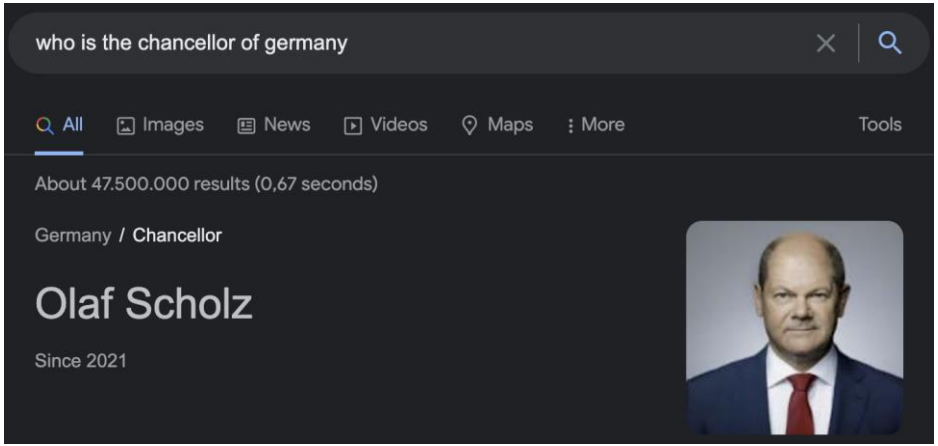
## DL4NLP Class

# Table of Contents

# 1. INTRODUCTION TO QA

# What is QA?

Question →  → Answer



# What is QA?


Google Assistant

# Why is QA important?

- Ideal testbed to evaluate the natural language understanding of AI systems
- Makes the knowledge of the world accessible
- Many other NLP tasks can be modeled as QA

## Fact Verification

**Claim:** The Rodney King riots took place in the most populous county in the USA.

**[wiki/Los_Angeles_Riots]**
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

**[wiki/Los_Angeles_County]**
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

**Verdict:** Supported

→ **Q:** Is it true that the Rodney King riots took place in the most populous county in the USA?

## Sentiment Analysis

The room was very comfortable.

↓

**Q**: Is the sentiment positive?
**Q**: What is the sentiment?

FEVER: a Large-scale Dataset for Fact Extraction and VERification (Thorne et al., NAACL 2018)

# QA Types

- Extractive QA (a.k.a. Machine Reading Question Answering)
- Multiple-Choice QA
- Open Domain QA (a.k.a. Open Retrieval)
- Visual QA
- And many others

# Extractive QA

The Rhine (Romansh: Rein, German: Rhein, French: le Rhin, Dutch: Rijn) is a European river that begins in the Swiss canton of Graubünden in the southeastern Swiss Alps, forms part of the Swiss-Austrian, Swiss-Liechtenstein border, Swiss-German and then the Franco-German border, then flows through the Rhineland and eventually empties into the North Sea in the Netherlands. The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi),[note 2][note 1] with an average discharge of about 2,900 m3/s (100,000 cu ft/s).

**What is the largest city the Rhine runs through?**

*Ground Truth Answers:* Cologne, Germany | Cologne, Germany | Cologne

*Prediction:* Cologne, Germany

(Question , Passage) → A

The answer is a contiguous span of the text

* The **passage** is sometimes called *context*

SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

# Multiple-Choice QA

Alex spilled the food she just prepared all over the floor and it made a huge mess.

**Q** What will Alex want to do next?

**A**
(a) taste the food
(b) mop up ✓
(c) run around in the mess

(Question, Passage, $Opt_1$, …, $Opt_k$) → A

Social IQa: Commonsense Reasoning about Social Interactions (Sap et al., EMNLP 2019)

# Open Domain Question Answering

Question → Answer

**Open-domain QA**
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?


Reading Wikipedia to Answer Open-Domain Questions (Chen et al., ACL 2017)

# Visual QA

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Q, Img) → A

# QA DATASETS

# Stanford QA Dataset (SQuAD)

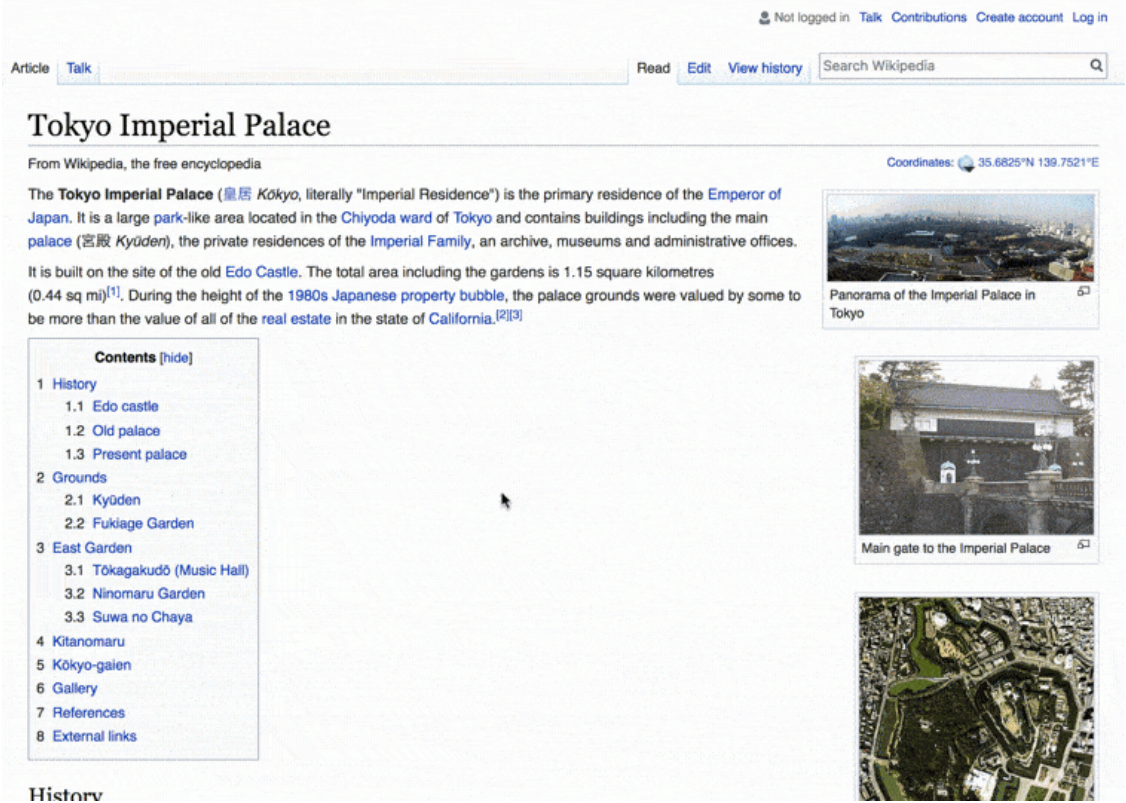
- First large QA dataset (100k QA pairs)
- Passages are from English Wikipedia (100-150 words)
- Questions are crowd-sourced
- Answers are spans in the passage
- Current QA models have “super-human” performance!
- Still one of the most popular QA datasets

| Rank | Model | EM | F1 |
|---|---|---|---|
| | Human Performance<br>*Stanford University*<br>(Rajpurkar et al. '16) | 82.304 | 91.221 |
| 1<br>Jul 24, 2021 | {ANNA} (single model)<br>*LG AI Research* | **90.622** | **95.719** |
| 2<br>Apr 10, 2020 | LUKE (single model)<br>*Studio Ousia & NAIST & RIKEN AIP*<br>https://arxiv.org/abs/2010.01057 | 90.202 | 95.379 |
| 3<br>May 21, 2019 | XLNet (single model)<br>*Google Brain & CMU* | 89.898 | 95.080 |
| 4<br>Dec 11, 2019 | XLNET-123++ (single model)<br>*MST/EOI*<br>http://tia.today | 89.856 | 94.903 |
| 4<br>Aug 11, 2019 | XLNET-123 (single model)<br>*MST/EOI* | 89.646 | 94.930 |
| 5<br>Jul 21, 2019 | SpanBERT (single model)<br>*FAIR & UW* | 88.839 | 94.635 |

SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

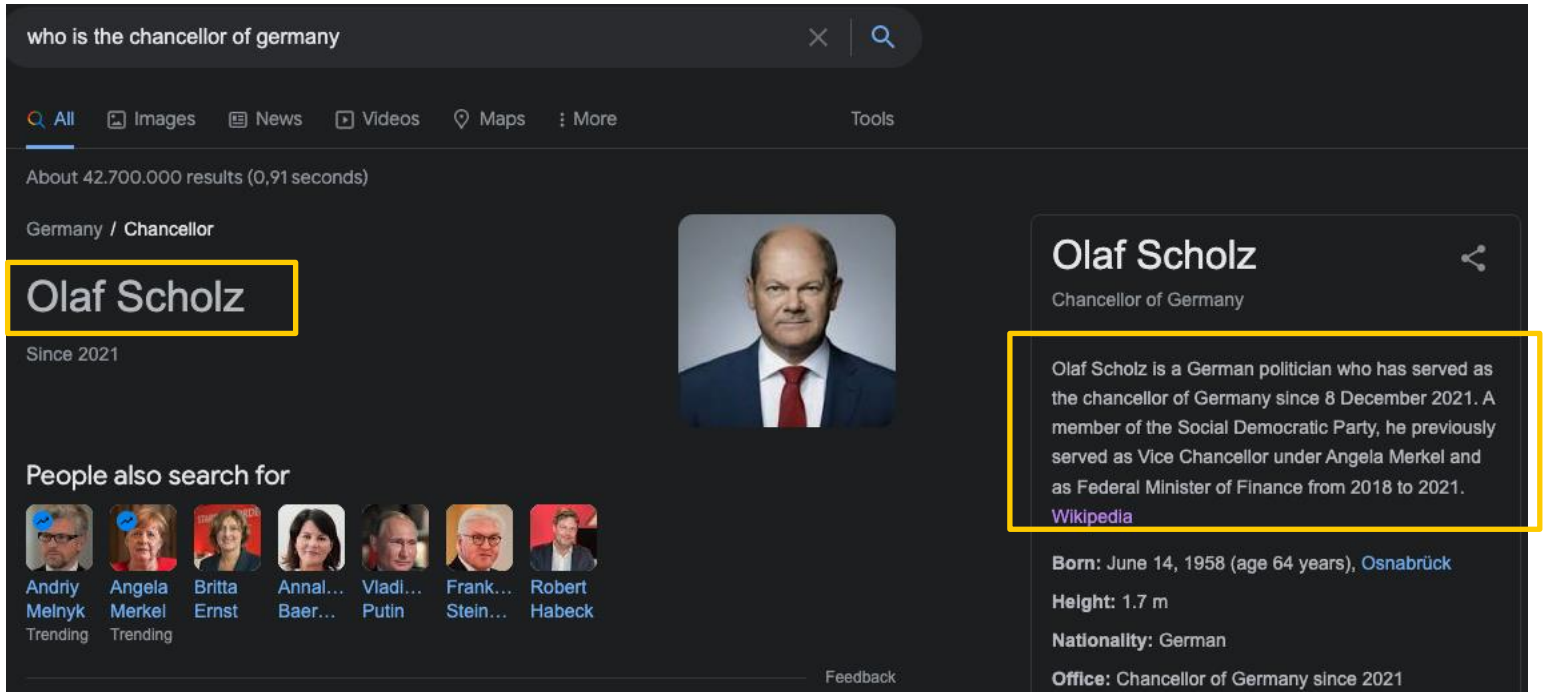
# Popular Datasets: Natural Questions

- Real questions in **Google Search**
- The context is the first paragraph of **Wikipedia**
- 2 types of answers:
  - **Long** (find a paragraph)
  - **Short** (find the exact answer)




Natural Questions: A Benchmark for Question Answering Research (Kwiatkowski et al., TACL 2019)

# NQ Is What Google Is Doing!!



# Popular Datasets: HotpotQA

*Paragraph A:* Ernest Cline

Ernest Christy Cline (born March 29, 1972) is an American novelist, spoken-word artist, and screenwriter. He is mostly famous for his novels "Ready Player One" and "Armada"; he also co-wrote the screenplay of "Ready Player One"'s upcoming film adaptation by Steven Spielberg.

*Paragraph B:* Armada (novel)

Armada is a science fiction novel by Ernest Cline, published on July 14, 2015 by Crown Publishing Group (a division of Random House). The story follows a teenager who plays an online video game about defending against an alien invasion, only to find out that the game is a simulator to prepare him and people around the world for defending an actual alien invasion.

**Q:** Which novel by the author of "Armada" will adapted as a feature film by Steven Spielberg?

**A:** Ready Player One

HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering (Yang et al., EMNLP 2018)

# Popular QA Datasets

- NewsQA:
  - **CNN** news as passages
  - Crowdsourced questions
- DROP:
  - **Numerical** reasoning
  - Crowdsourced questions
- TriviaQA:
  - **Trivia** questions
  - Passages from **Bing** search and **Wikipedia**
- SearchQA:
  - Trivia questions from **Jeopardy! TV show**
  - Passages from **Google** Search

# 2. BERT FOR QUESTION ANSWERING

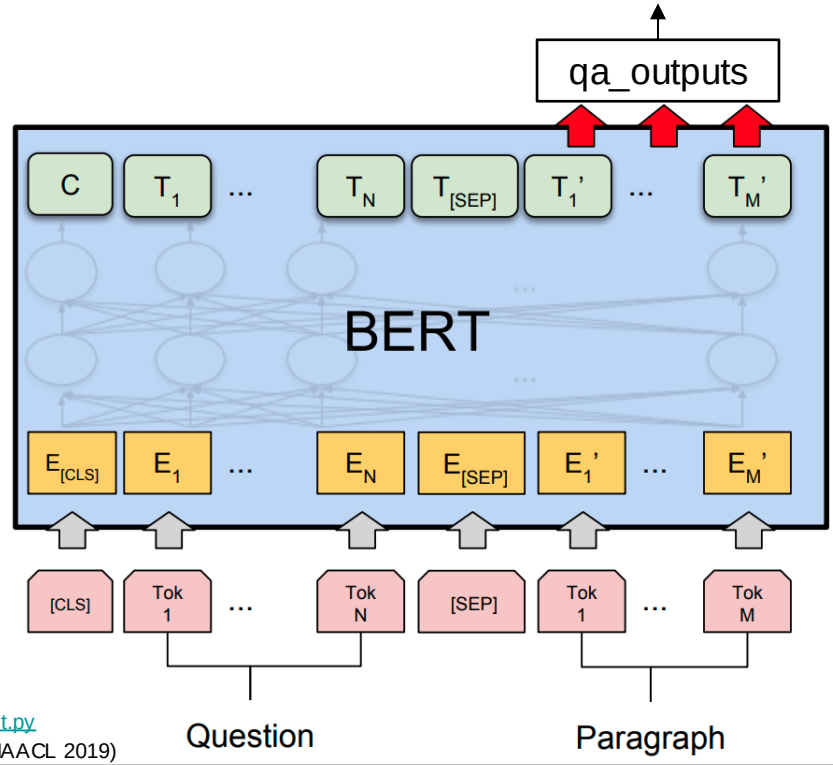
# BERT for QA

```
logits = self.qa_outputs(sequence_output)
start_logits, end_logits = logits.split(1, dim=-1)
start_logits = start_logits.squeeze(-1).contiguous()
end_logits = end_logits.squeeze(-1).contiguous()
```

```
class BertForQuestionAnswering(BertPreTrainedModel):

    _keys_to_ignore_on_load_unexpected = [r"pooler"]

    def __init__(self, config):
        super().__init__(config)
        self.num_labels = config.num_labels

        self.bert = BertModel(config, add_pooling_layer=False)
        self.qa_outputs = nn.Linear(config.hidden_size, config.num_labels)
```

2

$\hat{y}^{st} \in \mathbb{R}^{M\times 1};\ \hat{y}^{end} \in \mathbb{R}^{M\times 1}$




https://github.com/huggingface/transformers/blob/v4.19.4/src/transformers/models/bert/modeling_bert.py
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019)

# Training

Predictions for the starting span

Labels for the starting span

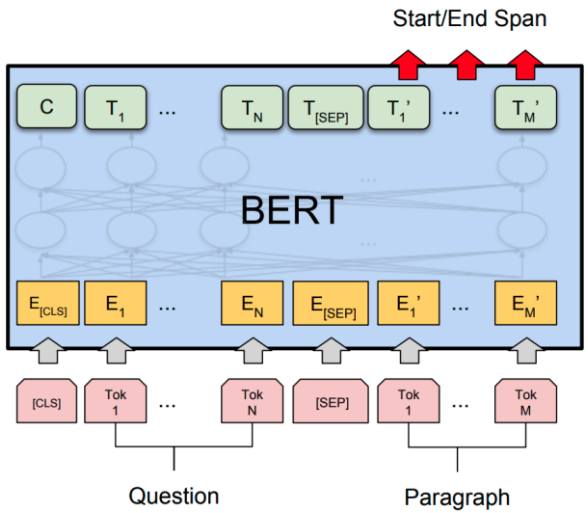
```
loss_fct = CrossEntropyLoss(ignore_index=ignored_index)
start_loss = loss_fct(start_logits, start_positions)
end_loss = loss_fct(end_logits, end_positions)
total_loss = (start_loss + end_loss) / 2
```

$$CE(p) = -ylog(p) + (1-y)\log(1-p))$$

$$CE(st) = \frac{1}{M}\sum_{i=1}^{M} CE(\hat{y}_i^{st}, y_i^{st})$$

$$CE(end) = \frac{1}{M}\sum_{i=1}^{M} CE(\hat{y}_i^{end}, y_i^{end})$$

$$\mathcal{L} = \frac{1}{2}\left(CE(st) + CE(end)\right)$$



# Getting the Top k Answers

1. Pick the top k predicted tokens as answer start
2. Pick the top k predicted tokens as answer ending
3. Sum their cartesian product
4. Pick the top k highest probabilities

$$\underset{i,j}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{y}_i{}^{st} + \hat{y}_j{}^{st}$$

# How to Evaluate the Performance?

TECHNISCHE UNIVERSITÄT DARMSTADT

## Exact Match (EM)

- Clean the prediction and label
  - Lowercase
  - Remove punctuation
  - Remove articles
  - Fix white spaces
- Prediction == label

## F1

- Harmonic mean of the token overlap between the prediction and the label
- Token = white-space tokens
- Also, cleans the prediction and label
- F1(“Hello”, “Hello World”) = 0.666



UKP

Is QA solved yet?

# HOW GOOD ARE THE MODELS?

# QA is not solved yet!

TECHNISCHE UNIVERSITÄT DARMSTADT

Training in one dataset does not generalize to others

| Fine-tuned on | Evaluated on: SQuAD | TriviaQA | NQ | QuAC | NewsQA |
|---|---|---|---|---|---|
| SQuAD | **75.6** | 46.7 | 48.7 | 20.2 | 41.1 |
| TriviaQA | 49.8 | **58.7** | 42.1 | 20.4 | 10.5 |
| NQ | 53.5 | 46.3 | **73.5** | 21.6 | 24.7 |
| QuAC | 39.4 | 33.1 | 33.8 | **33.3** | 13.8 |
| NewsQA | 52.1 | 38.4 | 41.7 | 20.4 | **60.1** |

Table 3: F1 scores of each fine-tuned model evaluated on each test set

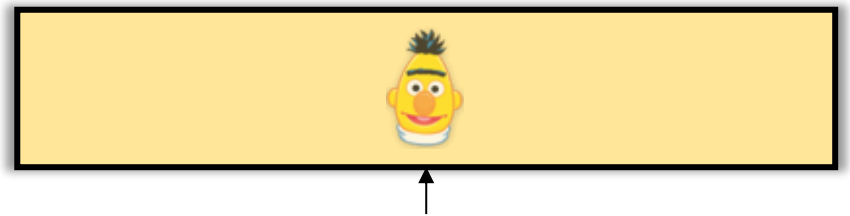
BERT Base uncased

What do Models Learn from Question Answering Datasets? (Sen & Saffari, EMNLP 2020)

# QA Generalization

## Multi-Dataset Models

- Train a model on many datasets

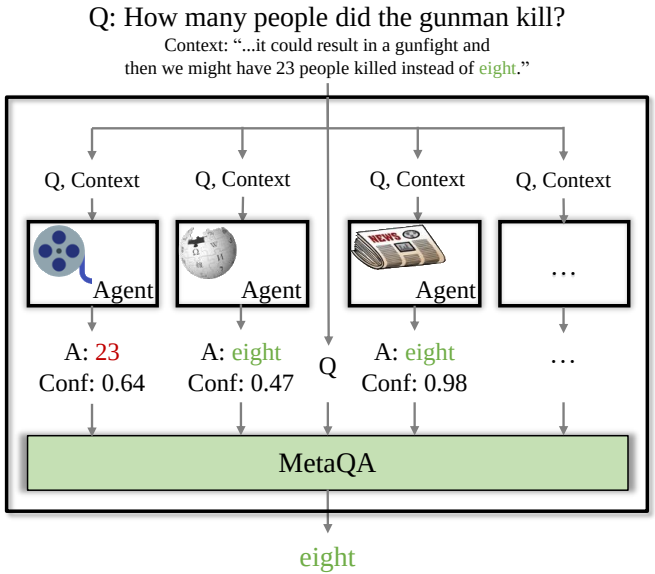


SQuAD, HotpotQA, Natural Questions, …

UNIFIEDQA: Crossing Format Boundaries with a Single QA System (Khashabi et al., Findings 2020)

## Multi-Agent Models

- Combine many models




MetaQA: Combining Expert Agents for Multi-Skill Question Answering (Puerto et al., Arxiv 2021)

# EXPLAINABILITY IN QUESTION ANSWERING

# What is Explainability?

TECHNISCHE UNIVERSITÄT DARMSTADT

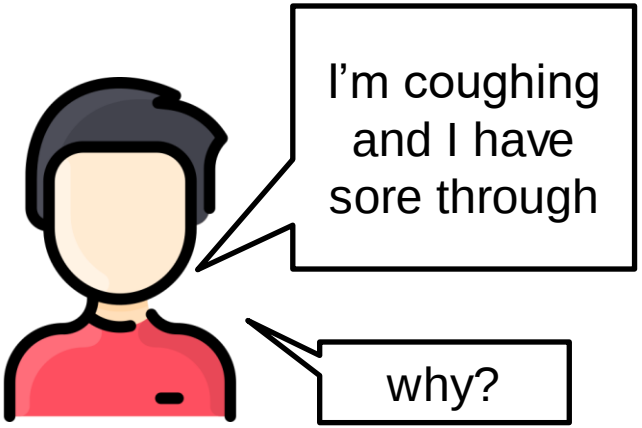
Explainability is stating “how/why” the model gives a prediction

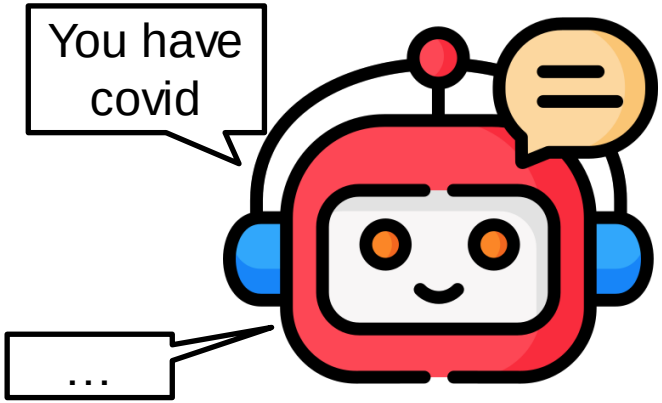
- Fundamental questions in XQA (explainable QA)
  1. Why did the QA system **choose this answer**?
  2. Why did not the QA system answer **something else**?
  3. **When** did the QA system **succeed**?
  4. **When** did the QA system **fail**?
  5. **When** does the QA system give **enough confidence** in the answer that you can trust?
  6. **How** can the QA system **correct** an error?

Shekarpour, S., & Alshargi, F. (2019). A Road-map Towards Explainable Question Answering A Solution for Information Pollution. *arXiv preprint arXiv:1907.02606*.

# Why do we need it?

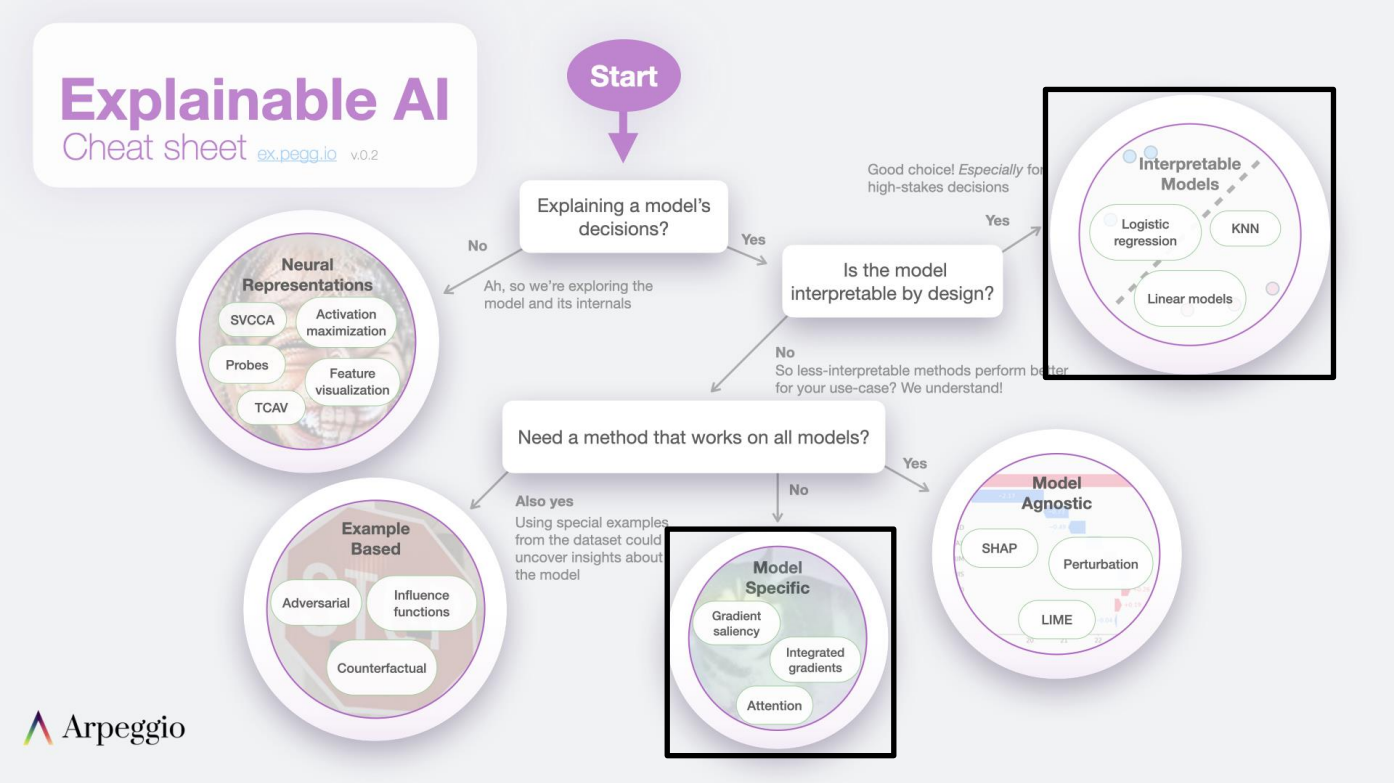
- Allows us to trust the prediction
- Can help us identify wrong predictions
- Sometimes, a prediction alone usually is not useful
  - Eg: Medical QA




Icons from flaticon.com

# Explainable AI




https://ex.pegg.io

# Interpretable Models: Supporting Facts

Supporting Fact 1

*Paragraph A:* Catawissa (tugboat)

Catawissa was a historic tugboat located at Waterford in Saratoga County, New York. She was built in 1896-1897 by Harlan and Hollingsworth of Wilmington, Delaware for the Philadelphia and Reading Railroad to tow coal barges between ports on the Eastern Seaboard. She was 158 feet in length, 19 feet in beam and 18 feet in depth. She was registered at 558 gross tons. She had a riveted steel framed and plated hull.

Supporting Fact 2

*Paragraph B:* Waterford, New York

Waterford is a town in Saratoga County, New York, United States. The population was 8,515 at the 2000 census. The name of the town is derived from its principal village, also called Waterford. The town and village are in the southeast corner of Saratoga County, and north-northwest of Troy, New York. It is located at the junction of the Erie Canal and the Hudson River.

**Q:** What was the population of the town as of 2000 where the historic tugboat Catawissa is located?

**A:** 8,515

HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering (Yang et al., EMNLP 2018)

# Interpretable Models: Supporting Facts

This type of dataset helps you create a more interpretable model

But creating these auxiliary labels is expensive

- Not all datasets have labels for supporting facts

Icons from flaticon.com

# Interpretable Models: Chain of Thought

TECHNISCHE UNIVERSITÄT DARMSTADT

Non-Explainable Model

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls How many tennis balls does he have now?

A: The answer is 11.

Model using **Chain of Thought**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls How many tennis balls does he have now?

A: **Roger started with 5 balls.**
**2 cans of 3 tennis balls each is 6 tennis balls.**
**5 + 6 = 11.**
The answer is 11.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

# Interpretable Models: Chain of Thought Problems

Not applicable to all models

You need to design the model to do this

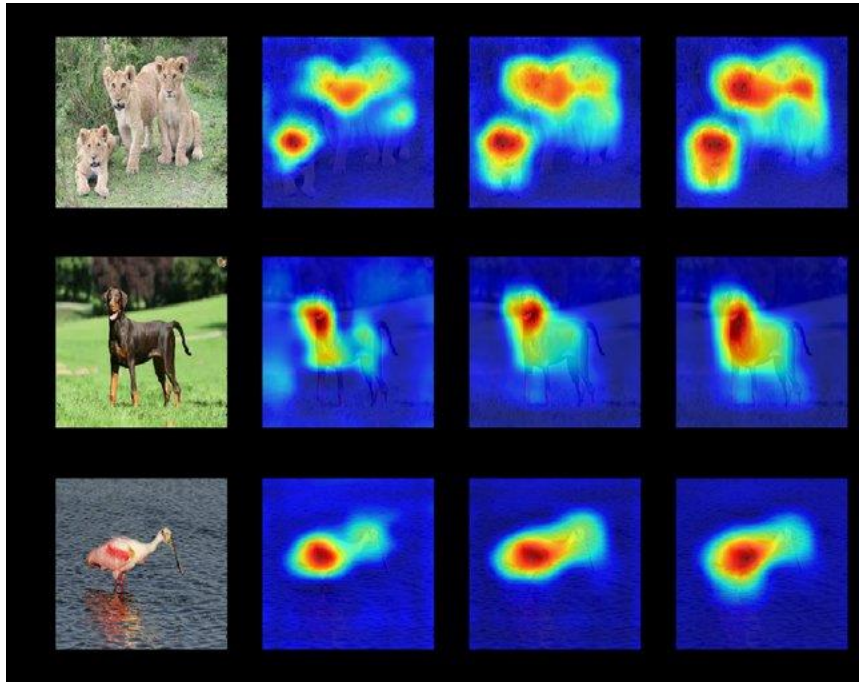
This may not always give SOTA results

Icons from flaticon.com

# Saliency Maps

- Inspired by computer vision
- Draw a map that shows the pixels that support the prediction of the class

Applicable to all neural networks



Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

# Saliency Maps in QA

QUESTION

Who stars in The Matrix ?

Visualizing the top 3 most important words.

PASSAGE

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis , starring Keanu Reeves , Laurence Fishburne , Carrie – Anne Moss , Hugo Weaving , and Joe Pantoliano . It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called " the Matrix " : created by sentient machines to subdue the human population , while their bodies ' heat and electrical activity are used as an energy source . Computer programmer " Neo " learns this truth and is drawn into a rebellion against the machines , which involves other people who have been freed from the " dream world . "

Visualizing the top 3 most important words.

https://demo.allennlp.org/reading-comprehension/bidaf-elmo

# Saliency Maps, how to compute them?

## Gradient-based Methods

- Vanilla Gradients [1]
- Integrated Gradients [2]
- SmoothGrad [3]

## Attention-based Methods

- Attention Weights
- Scaled Attention [4]

[1] Deep inside convolutional networks: Visualising image classification models and saliency maps (Simonyan et al., arXiv 2013)
[2] Axiomatic attribution for deep networks (Sundararajan et al., PMLR 2017)
[3] Smoothgrad: removing noise by adding noise (Smilkov et al., arXiv 2017)
[4] Is Attention Interpretable? (Serrano & Smith, ACL 2019)

# Gradient-based Saliency Maps

TECHNISCHE UNIVERSITÄT DARMSTADT

? What does the gradient tell us?

What weights should be **changed** to **minimize** the **loss**

? What if we use the output **prediction as label** and compute the loss?
Then, what is telling us the gradient?

What weights should be changed to minimize the loss = to maximize the selection of the prediction

Large gradient in a word → changing the word has a big effect on the prediction



UKP

# Saliency Maps in QA

QUESTION

Who stars in The Matrix ?

Visualizing the top 3 most important words.

PASSAGE

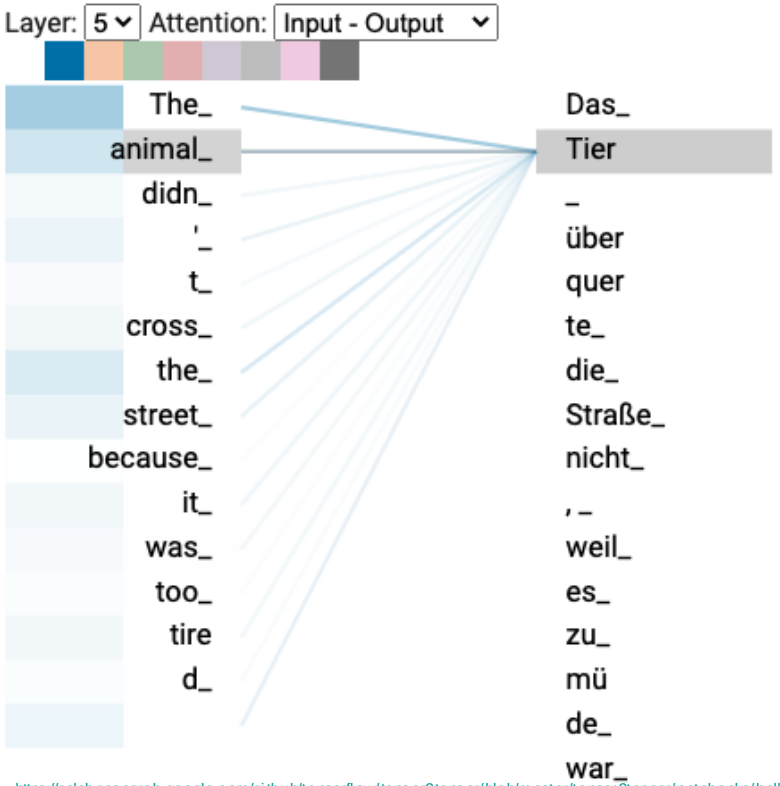
The Matrix is a 1999 science fiction action film written and directed by The Wachowskis , starring Keanu Reeves , Laurence Fishburne , Carrie – Anne Moss , Hugo Weaving , and Joe Pantoliano . It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called " the Matrix " : created by sentient machines to subdue the human population , while their bodies ' heat and electrical activity are used as an energy source . Computer programmer " Neo " learns this truth and is drawn into a rebellion against the machines , which involves other people who have been freed from the " dream world . "

Visualizing the top 3 most important words.

https://demo.allennlp.org/reading-comprehension/bidaf-elmo

# Attention-based Saliency Maps

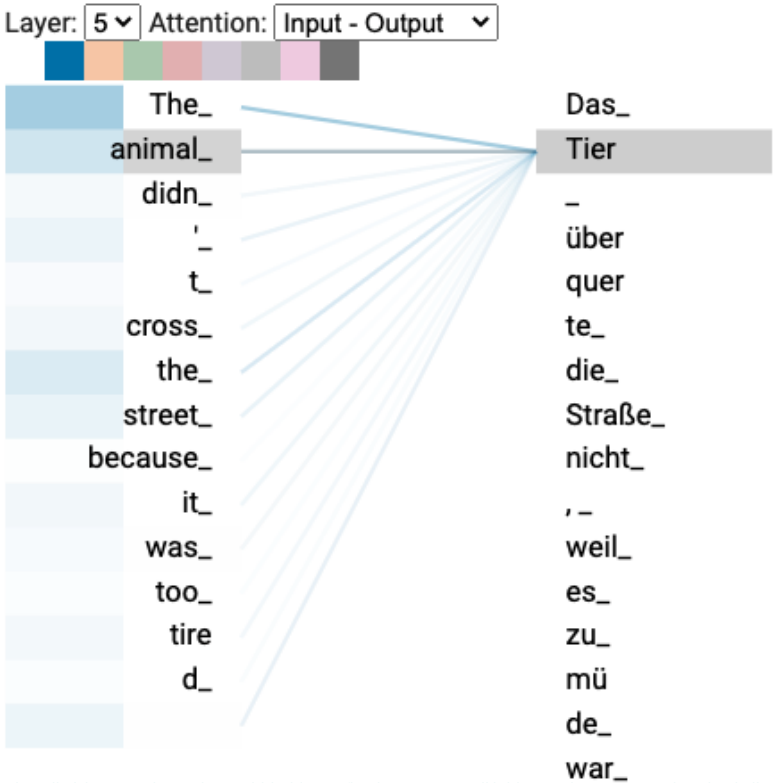
- Attention calculates a distribution over inputs
- It can naturally show the importance of the inputs




https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb#scrollTo=OJKU36QAfqOC

# Attention-based Saliency Maps

- In QA, use the [CLS] attention weights [1]
- However, attention is highly inconsistent and may not necessarily correspond to importance [1,2]
- Scaled Attention:
  - Attention Scores x Gradient




https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb#scrollTo=OJKU36QAfqOC

[1] Is Attention Interpretable? (Serrano & Smith, ACL 2019)
[2] Attention is not Explanation (Jain & Wallace, NAACL 2019)

# Behavioral Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

Validates the input-output behavior w/o knowing the model internals

List of questions and answers that evaluates the behavior of a model

Multiple types of tests

- Minimum Functionality Tests
- Invariance

## Animal vs Vehicle

Test's model's ability to understand different animals and vehicles.

**Min Func Test** test on **Taxonomy**

**Question:** What vehicle does Victoria have?

**Context:** Victoria has a snake and a SUV.

**Answer:** SUV ✓

**Prediction:** snake ✗

Icons created by Freepik - Flaticon

# Behavioral Testing: Minimum Functionality Tests

- ≈ Unit Testing
- Useful to **detect** when a model use **shortcuts** to solve a question instead of mastering the required ability

## Animal vs Vehicle

Test's model's ability to understand different animals and vehicles.

**Min Func Test** test on **Taxonomy**

**Question:** What vehicle does Victoria have?

**Context:** Victoria has a snake and a SUV.

**Answer:** SUV ✓

**Prediction:** snake ✗

# Behavioral Tests: Invariance

TECHNISCHE UNIVERSITÄT DARMSTADT

Adding perturbations that shouldn't change the output

## Question typo

Test's model's ability to handle questions typos (whether changing the spelling of words in the questions changes the model's output)

INVariance test on Robustness

**Question:** How many more males are there for every 100 females than there are → aref females → emales for every 100 males?

**Context:** In the county, the population is spread out with 24.60% under the age of 18, 6.40% from 18 to 24, 22.80% from 25 to 44, 25.40% from 45 to 64, and 20.80% who are 65 years of age or older. The median age is 42 years. For every 100 females there are 95.90 males. For every 100 females age 18 and over, there are 90.50 males.

**Answer:** 4.1 ✓

**Prediction:** 95.90 → 90.50 ✗

The typo made the model change the output → the model is not robust



UKP

Software for Question Answering Research

# UKP-SQUARE

# SQuARE Platform for QA Research

TECHNISCHE UNIVERSITÄT DARMSTADT

- Deploying models
- Comparing models
- Explainability
- Behavioral Tests
- Everything on the cloud

https://square.ukp.informatik.tu-darmstadt.de

Icons created by Freepik - Flaticon



UKP

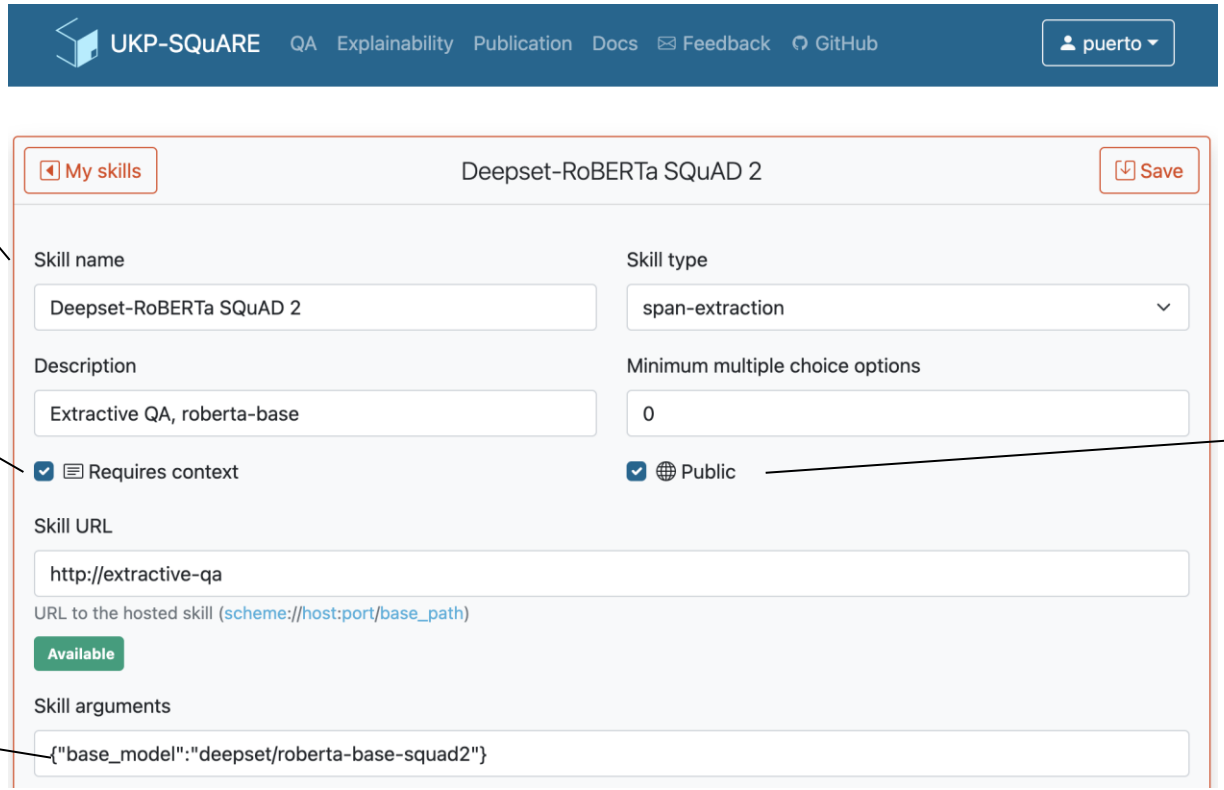
# Deploying Models on SQuARE

TECHNISCHE UNIVERSITÄT DARMSTADT

The name of your model

The input is the question and a small piece of text (context)

The ID in HF's model hub.




Publicly available for everybody



UKP

# Future Updates

TECHNISCHE UNIVERSITÄT DARMSTADT

## Explainability methods

- Launching Saliency Maps on August 1st

## Graph-based QA models

- Explore paths of entities that goes from the question to the answer
- Graph-based explainability
- Better results in some QA domains (eg: Commonsense, Open-Domain, MultiHop, etc)

## Automatic model selection based on the input question

- SQuARE as a Multi-Agent QA System

# Homework

**Goal**: Learn the basic steps to fine-tune an extractive QA model

**Format**: Jupyter Notebook tutorial on Google Colab

# Homework

- You need to choose a pretrained language model and a QA dataset
  - We will give you a list to choose from to ensure you can train it quickly
- You will:
  - Tokenize and encode the data
  - Train the model
  - Evaluate it
  - Share it on HuggingFace's Model hub
  - Deploy it on UKP-SQuARE
  - Run it on you web browser

# Homework

- Deadline: 22nd July
- Maximum score: 25 points
- Ask questions!
  - Please, always state the model and dataset you are using