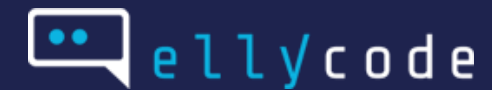




global AI developer days Torino 2022



28 OTTOBRE
OGR TECH - TORINO



Football Transfer Market News Generation

Federico Barbiero **deltatre**

28 OTTOBRE
OGR TECH - TORINO



Football Transfer Market News Generation

Can we generate automatically the football transfer market news text?

Manchester United agree £85m deal to sign Antony from Ajax

Brazil international Antony will arrive in Manchester on Monday; Ajax rejected Man Utd's £76.3m (€90m) bid for player; winger worked with Erik ten Hag during United manager's time with Eredivisie champions; United have until Thursday's transfer deadline to complete the deal



Fabrizio Romano @FabrizioRomano · 7min

"It's very difficult to answer that [Chelsea rumours]. So that will have to wait until after the World Cup", Leandro Trossard said when asked about Chelsea. 🇳🇱🇧🇪 #CFC

Important: Trossard deal with Brighton expires in 2023, but #BHAFC have an option to extend contract until 2024.

Erik ten Hag delaying contract decision on Marcus Rashford and David de Gea

- Ten Hag: 'In this moment we only think about performing'
- Ronaldo and Luke Shaw also out of contract in June

Manchester United complete Lisandro Martinez signing from Ajax in £57m deal

27 July 2022 | Man Utd

Juventus agree deal to make Di Maria second free signing with Pogba

- Di Maria poised to sign one-year deal after leaving PSG
- Pogba due in Turin on Saturday to complete his return

Paris Saint-Germain are closing in on Fabián Ruiz deal. Final details now discussed with Napoli as personal terms were agreed weeks ago. 🇪🇸🇮🇹 #PSG

It has always been matter of time and it's finally being completed. Five year deal for Fabián at PSG.

Juventus, in chiusura per Di Maria: i dettagli dell'ingaggio

MILAN-DE KETELAERE, TRATTATIVA BLOCCATA. ROZIELINSKI. INTER IN PRIMA FASE

Mbappé-Psg, dalla Francia: è rottura, potrebbe lasciare a gennaio. Ma Campos smentisce

Where will Cristiano Ronaldo leave Manchester United for? Here are the 7 options available to CR7



MailOnline Sport @MailSport

Romelu Lukaku arrives back in Italy and is greeted by Inter Milan fans chanting his name ahead of Chelsea switch

Gleison Bremer: Juventus sign Brazilian defender from Torino

Inter refuse latest PSG offer for Skriniar: the latest

Luis Diaz: Liverpool sign Porto winger on five-and-a-half-year deal for initial fee of £37m

Porto forward Luis Diaz has joined Liverpool for an initial £37m, with the fee potentially rising to £49m; Liverpool also interested in Fulham's Fabio Carvalho; Bournemouth are keen on taking Liverpool right-back Neco Williams on loan until the summer

Erling Haaland: Man City confirm signing of Borussia Dortmund striker in £51m deal

Manchester City activated the Norway striker's release clause and the player will join on July 1 for a total of £85.5m including agents fees and other add-ons; Liverpool manager Jurgen Klopp has said the Haaland deal will "set new levels" in the transfer market

Transfer Market News Generation

Inputs

Player Name

Cristiano Ronaldo

Club Name

Manchester United

Transfer Market Term

For Sale

League

Premier League

TeamB Name

Paris Saint Germain

Model Input String

<|PERSON|> Cristiano Ronaldo <|PERSON|> <|CLUB|>Manchester United<|CLUB|> <|TRANSFER_MARKET|> For Sale
<|TRANSFER_MARKET|> <|COMPETITION|> Premier League <|COMPETITION|> <|CLUB|> Paris Saint Germain <|CLUB|>

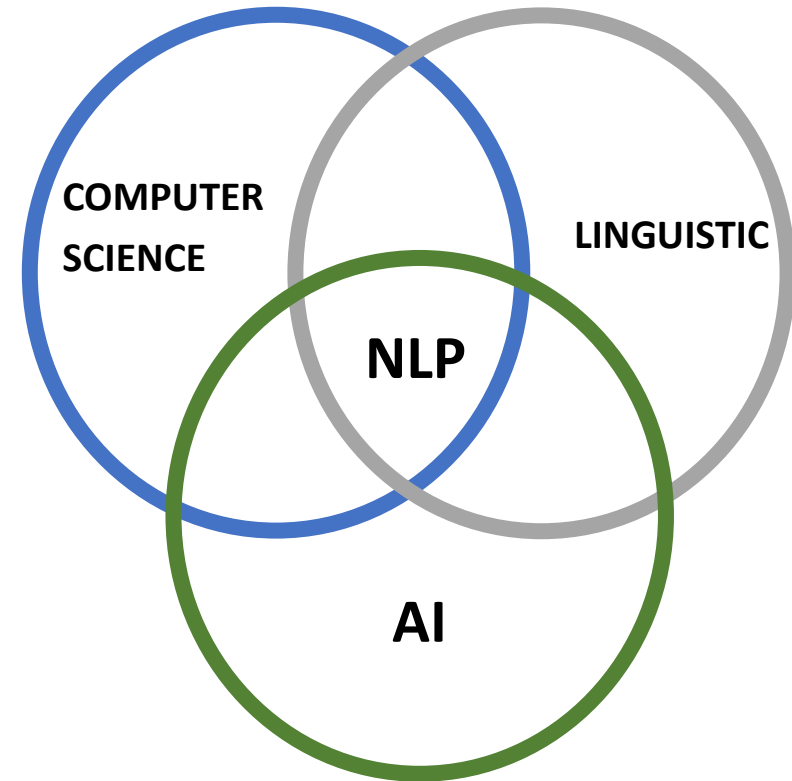
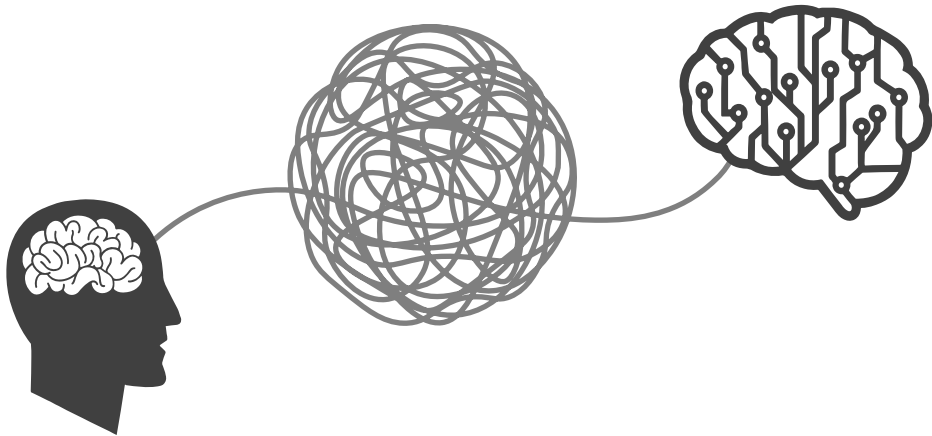
Output

Cristiano Ronaldo is set to leave Manchester United as hes not for sale in the Premier League. transfers Paris SaintGermain board already had direct talks with Cristiano before discussing about future options



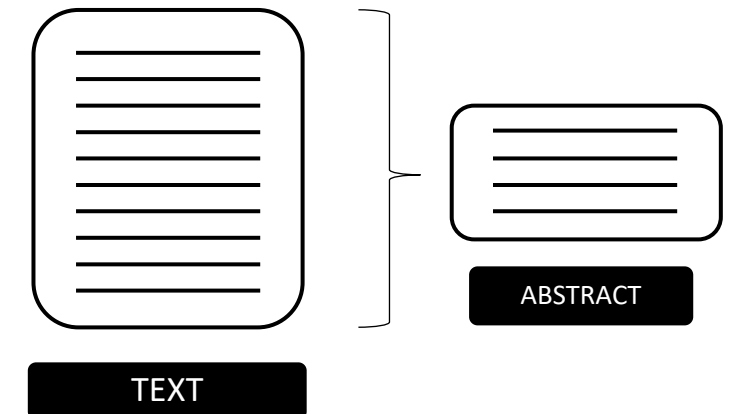
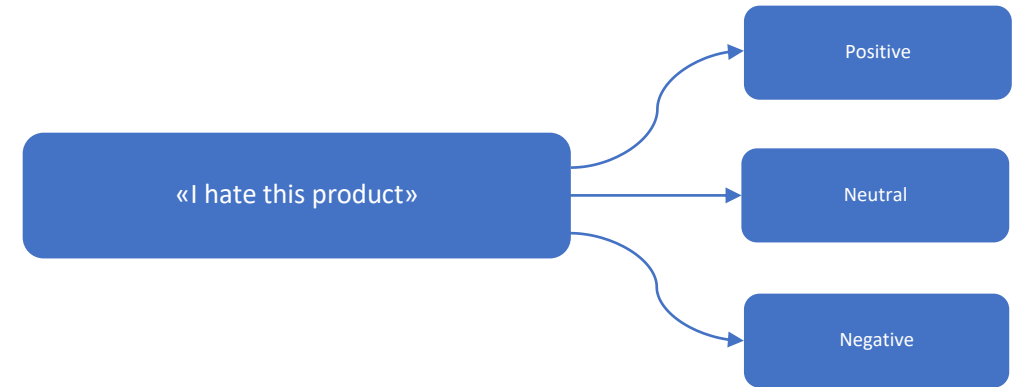
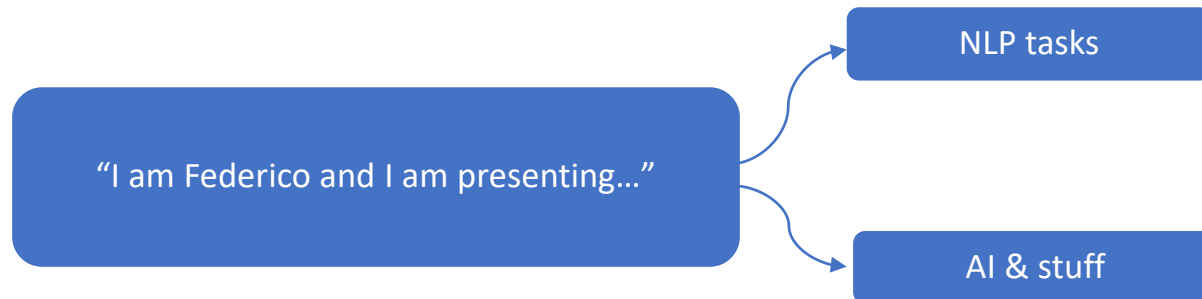
Natural Language Processing (NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human through language.



NLP common tasks

- Text Classification
- Named Entity Recognition (NER)
- Automatic Text Summarization
- Translation
- Text Generation
- Many others (and hybrids, heard about DALL-E?)...



Natural Language Generation (NLG)

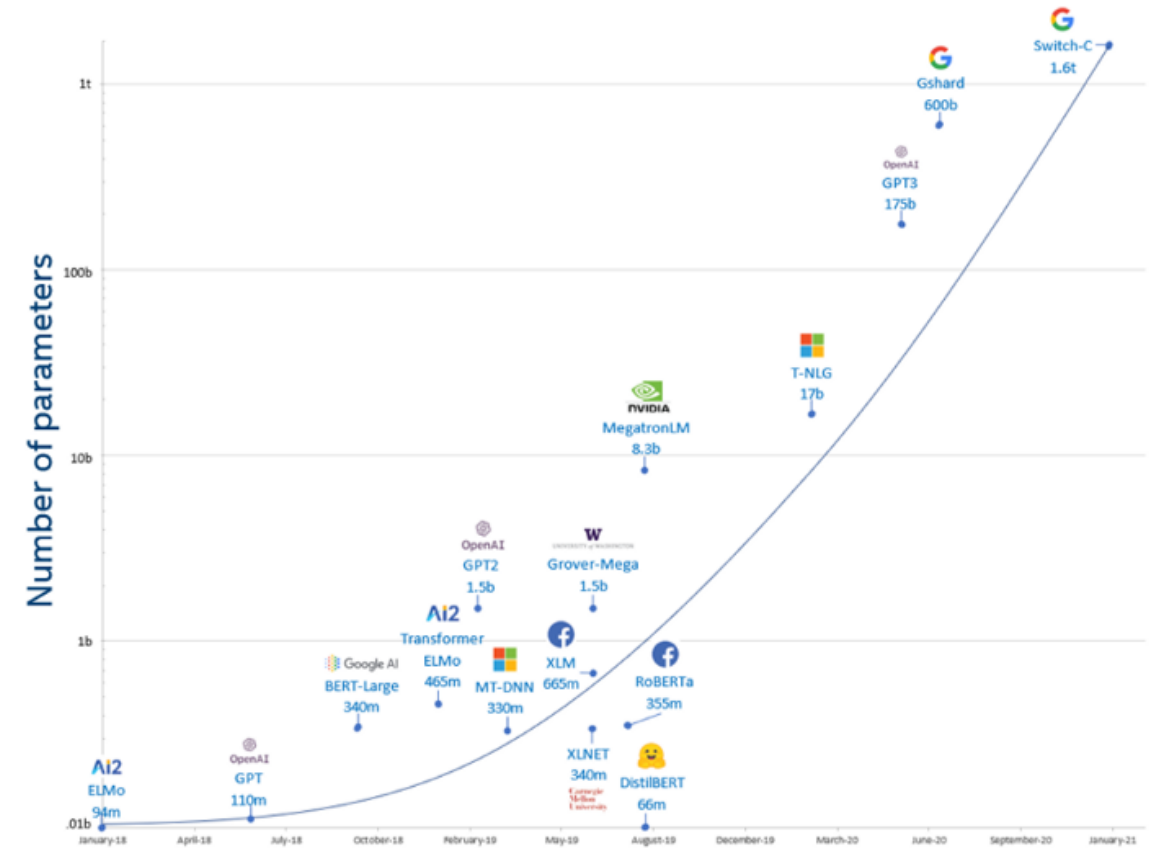
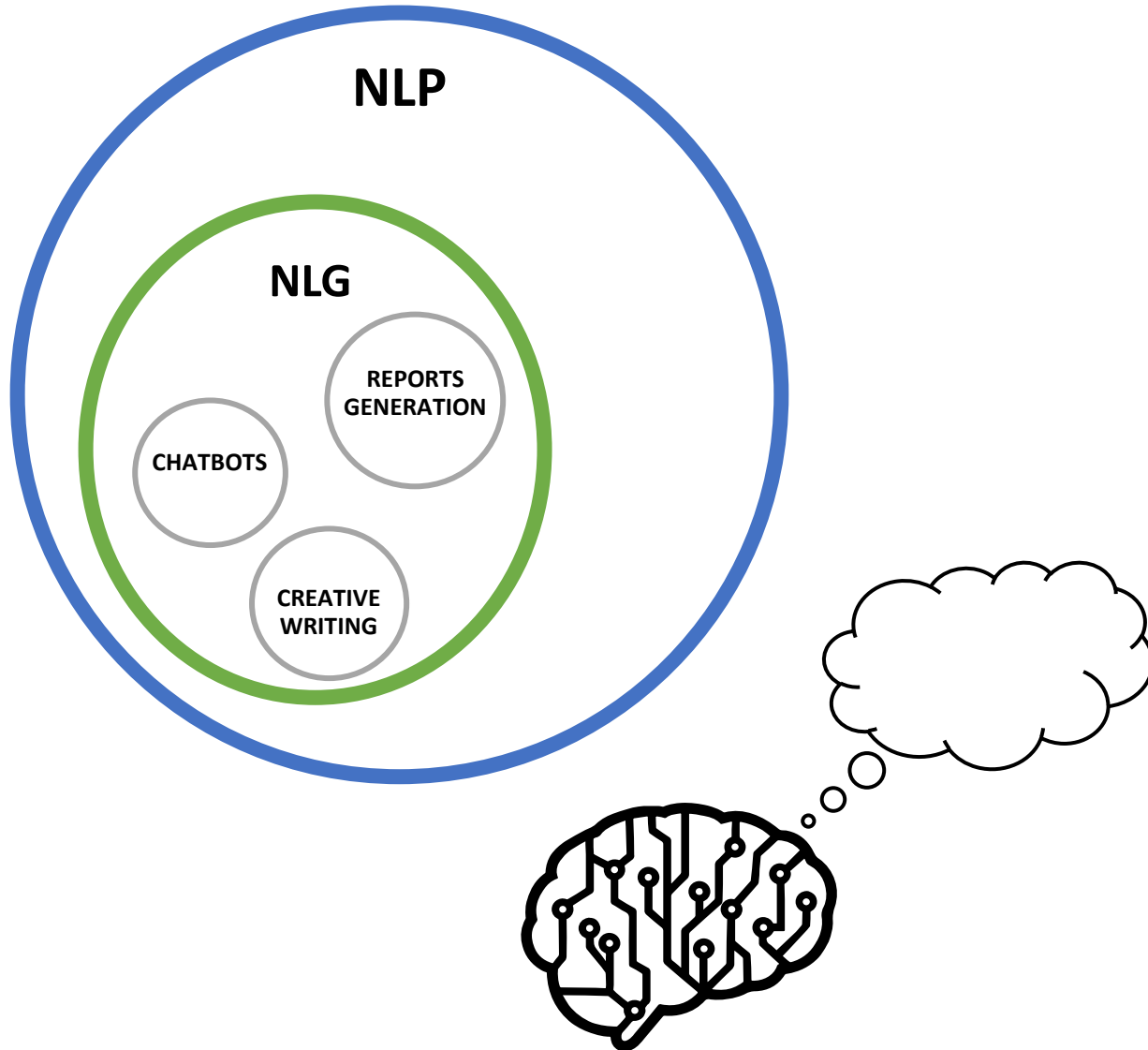
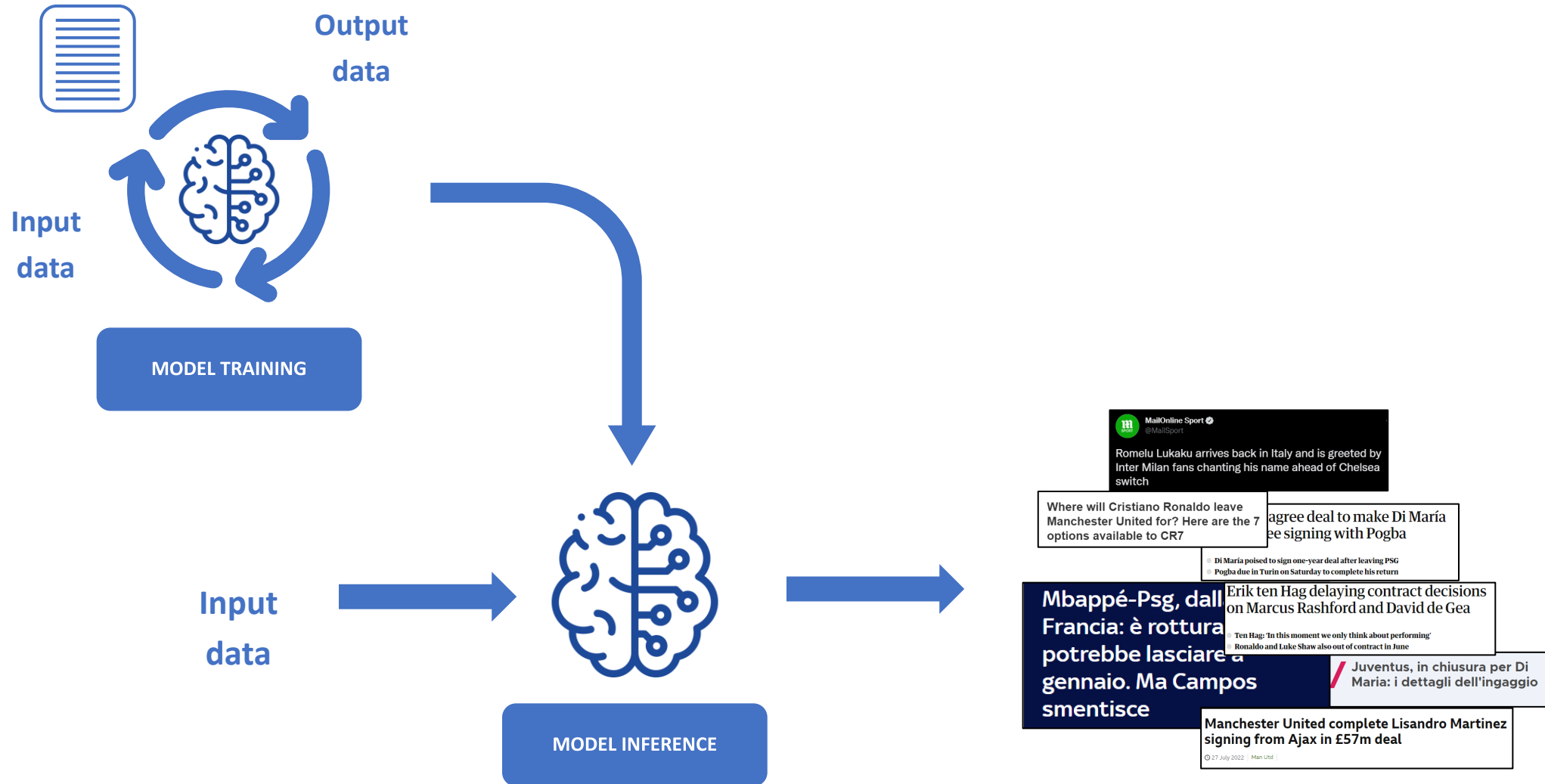


Figure 1: Exponential growth of number of parameters in DL models

The standard AI workflow



What do we need to start?

Most important things to start an AI project:

- **You need data!**
- **You need a lot of data!**
- **You need a lot of good data!**

Apart from that, **you need data**. And you need to **label your data**.

“Where can we find data to train a model like to generate football news?”

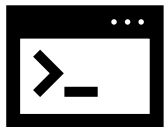
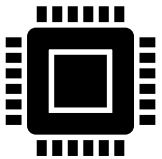


Our DATA

I thought that probably the best source of data I could find were **tweets**!

Twitter has its own API to download tweets, you just need a **developer account**.

To interact with Twitter, I used Python and its dedicated library Tweepy.



```
config = configparser.ConfigParser()
config.read('config.ini')

api_key=config['twitter']['api_key']
api_key_secret=config['twitter']['api_key_secret']

access_token=config['twitter']['access_token']
access_token_secret=config['twitter']['access_token_secret']

# authentication

auth = tweepy.OAuthHandler(api_key, api_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

# get tweets from account
user = 'username'

tweets = tweepy.Cursor(api.user_timeline, screen_name=user, count=200, tweet_mode='extended').items(limit)
columns = ['Time', 'Text', 'Entities', 'In_reply_to_status_id', 'Language']
data = []

for tweet in tweets:
    text = str(tweet.full_text.encode('ascii', errors='ignore'))
    data.append([tweet.created_at, text, tweet.entities, tweet.in_reply_to_status_id, tweet.lang])

df = pd.DataFrame(data, columns=columns)
df.to_csv(f'{output_name}.csv', sep=',')

print(df)
```

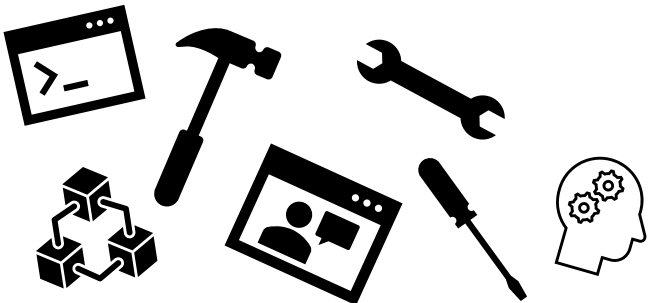
Cool we have DATA, but:

They are a mess!

We need to prepare them in order to make as clean as possible.

In AI/ML this phase is called **preprocessing**.

And this is what we want generate, what do we show as input to our model?



Napoli are set to sign Kim Min Jae as new centre back from Fenerbahçe by triggering €19.5m release clause as Koulibaly replacement. 🇸🇰🇵🇰🇰🇷 #Napoli

South Korean centre back was close to join Rennes but Napoli hijacked the deal. Medical today. @SkySport

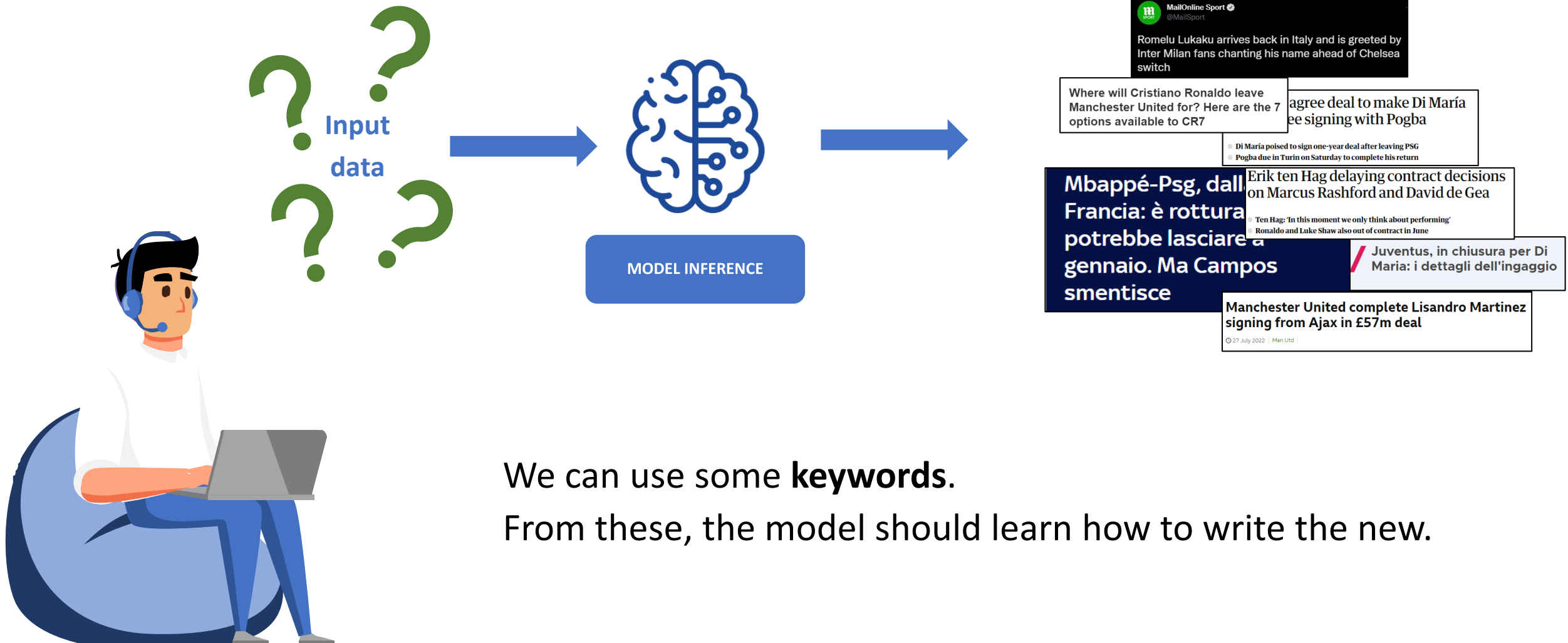


b'Napoli are set to sign Kim Min Jae as new centre back from Fenerbahe by triggering 19.5m release clause as Koulibaly replacement. #Napoli\n\nSouth Korean centre back was close to join Rennes but Napoli hijacked the deal. Medical



Napoli are set to sign Kim Min Jae as new centre back from Fenerbahe by triggering 19.5m release clause as Koulibaly replacement. Napoli South Korean centre back was close to join Rennes but Napoli hijacked the deal. Medical today.

Input DATA



We can use some **keywords**.

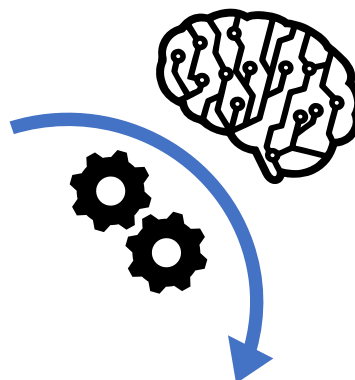
From these, the model should learn how to write the new.

NER

Antonio Conte admits Tottenham 'cannot work miracles' after Newcastle defeat as Jamie Redknapp questions delayed contract talks

Person Transfer Market
Team Result

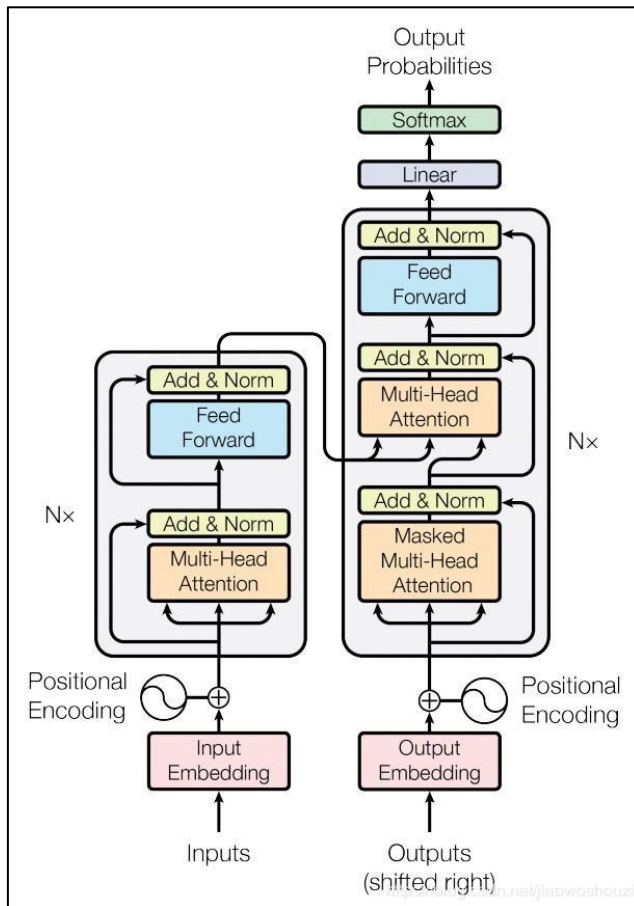
Cristiano Ronaldo will meet with Erik ten Hag to discuss about his future. Man Utd insist hes not for sale while Cristiano wants to go - Mendes, still pushing MUFC No changes on Frenkie de Jong, as of now - no intention to accept a salary cut. |



<|PERSON|> cristiano ronaldo <|PERSON|> <|PERSON|> erik ten hag <|PERSON|>
<|CLUB|> man <|CLUB|> <|TEAM_NICKNAME|> utd <|TEAM_NICKNAME|>
<|TRANSFER_MARKET|> for sale <|TRANSFER_MARKET|> <|PERSON|> cristiano
<|PERSON|> <|PERSON|> mendes <|PERSON|> <|CLUB|> mufc <|CLUB|>
<|PERSON|> frenkie de jong <|PERSON|> <|MONEY|> salary cut <|MONEY|>

CLUB
PERSON
COMPETITION
NATION/NATIONAL_TEAM
SCORE
RESULT
DATE
ORG
ROUND
GPE
STADIUM
POSITIONING
TEAM_NICKNAME
PERSON_NICKNAME
ROLE
SHOT_TYPE
MATCH_PHASE
SET_PIECE
AWARDS
CUSTOMS_&_TRADITIONS
BETTING
TRANSFER_MARKET
TACTICAL
DISCIPLINARY
INJURY
ESPORT
FAC
LOC
NORP
EVENT
TIME
PERCENT
MONEY
ORDINAL
CARDINAL

Training BERT



```
def compute_metrics(p: EvalPrediction) -> Dict:
    preds_list, out_label_list = align_predictions(p.predictions, p.label_ids)
    return {
        "precision": precision_score(out_label_list, preds_list),
        "recall": recall_score(out_label_list, preds_list),
        "f1": f1_score(out_label_list, preds_list),
    }

# Initialize our Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
    compute_metrics=compute_metrics,
)

# Training
if training_args.do_train:
    trainer.train(
        model_path=model_args.model_name_or_path if os.path.isdir(model_args.model_name_or_path) else None
    )
    trainer.save_model()
    # For convenience, we also re-save the tokenizer to the same directory,
    # so that you can share your model easily on huggingface.co/models
    if trainer.is_world_master():
        tokenizer.save_pretrained(training_args.output_dir)
```

```
tokens = tokenizer.tokenize(tokenizer.decode(tokenizer.encode(text)))
inputs = tokenizer.encode(text, return_tensors="pt")
outputs = model(inputs)[0]
predictions = torch.argmax(outputs, dim=2).detach().numpy()[0, :]
score_matrix = outputs.detach().numpy()
meaningful_ent = []

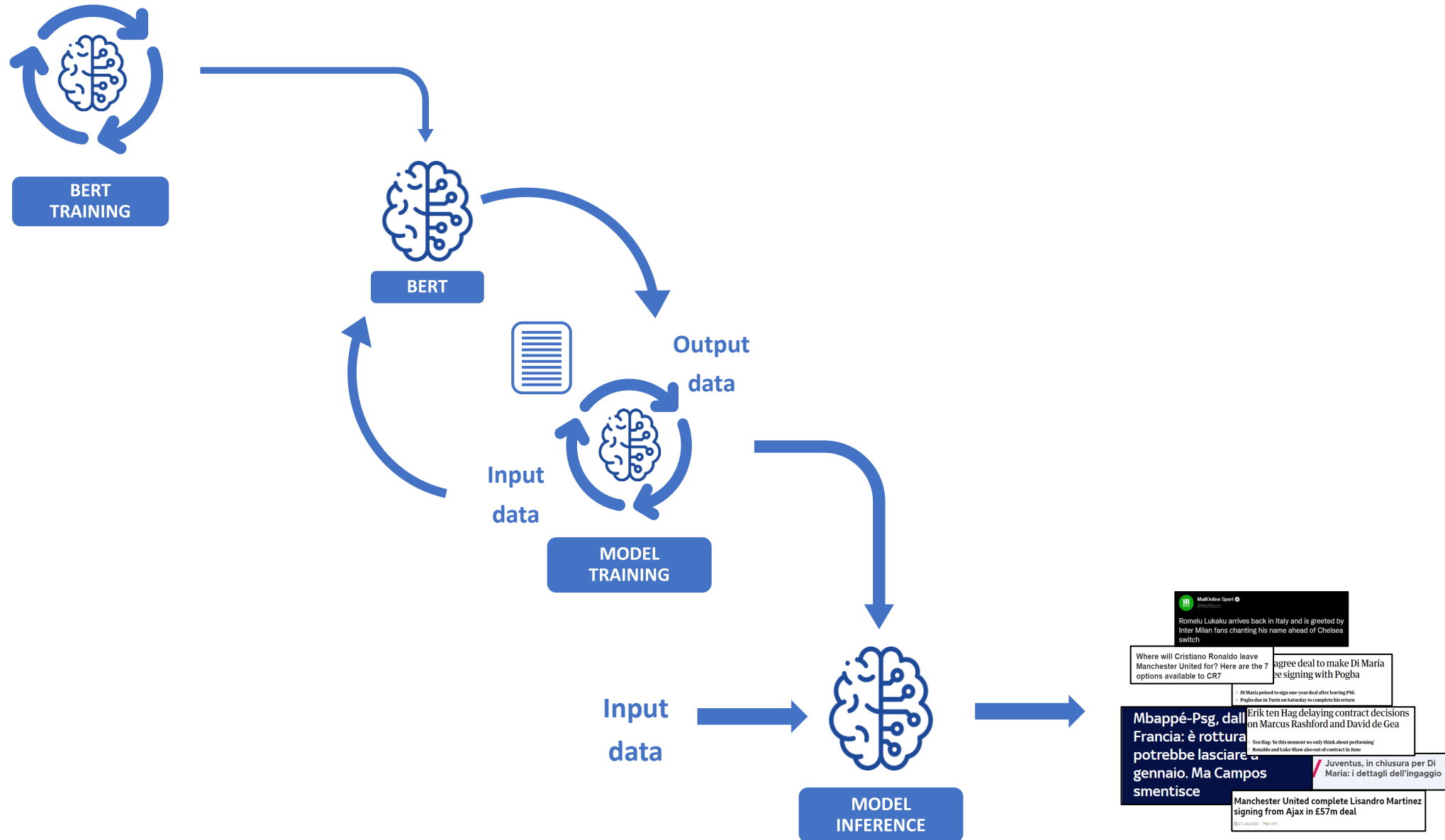
for i in range(0, len(predictions)):
    meaningful_ent.append(entity(tokens[i], label_list[predictions[i]], soft_max(score_matrix[0, i, :])))

entities_to_show=[m for m in merge_same_labels(remove_sobtokens(meaningful_ent)) if m.label!='0'and len(m.token)

return [{'text': ent.token , 'label': ent.label, 'score': ent.confidence} for ent in entities_to_show]
```

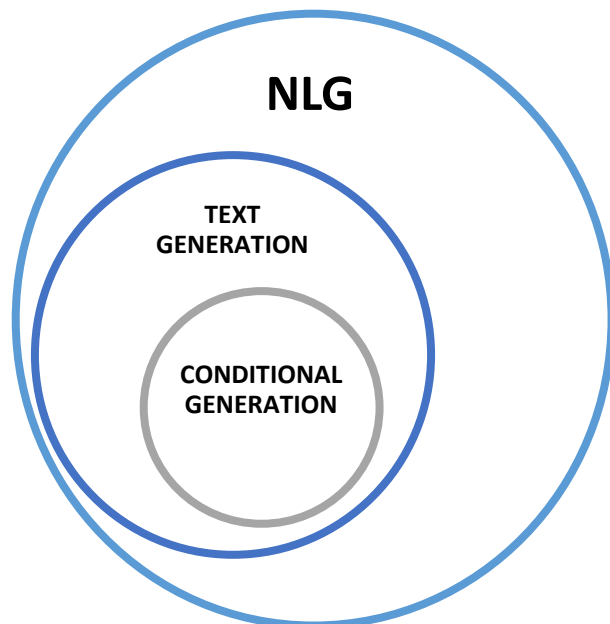
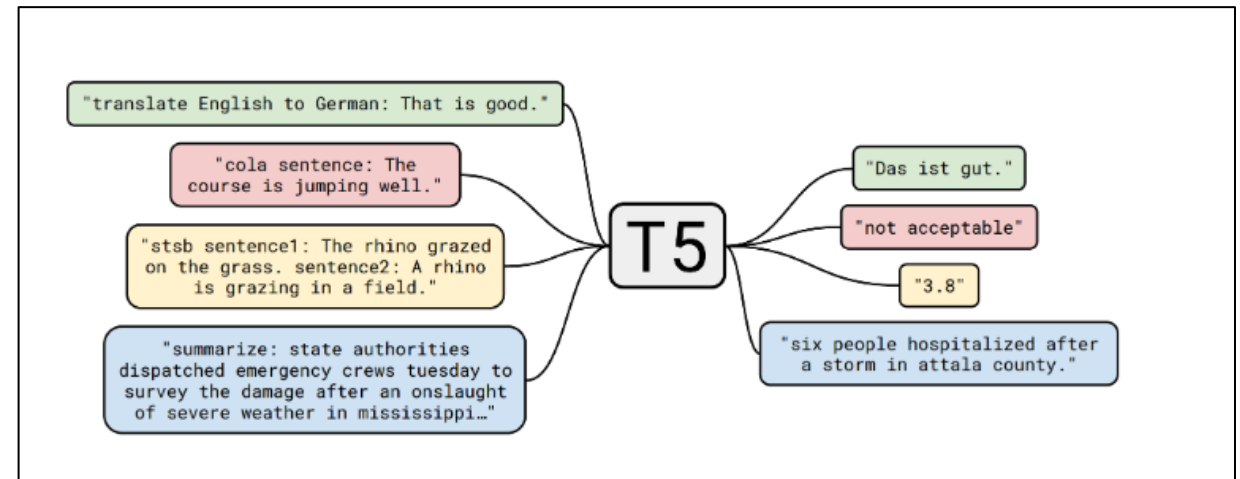


Our AI workflow



And now what model to generate text?

T5
(Text-to-Text
Transfer
Transformer)



ToTTo

Table Title: Cristhian Stuani
Section Title: International goals
Table Description: As of 25 March 2019 (Uruguay score listed first, score column indicates score after each Stuani goal)

No.	Date	Venue	Opponent	Score	Result	Competition
1.	10 September 2013	Estadio Centenario, Montevideo, Uruguay	Colombia	2-0	2-0	2014 FIFA World Cup qualification
2.	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	2-0	5-0	2014 FIFA World Cup qualification
3.	31 May 2014	Estadio Centenario, Montevideo, Uruguay	Northern Ireland	1-0	1-0	Friendly
4.	5 June 2014		Slovenia	2-0	2-0	

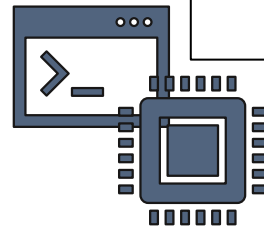
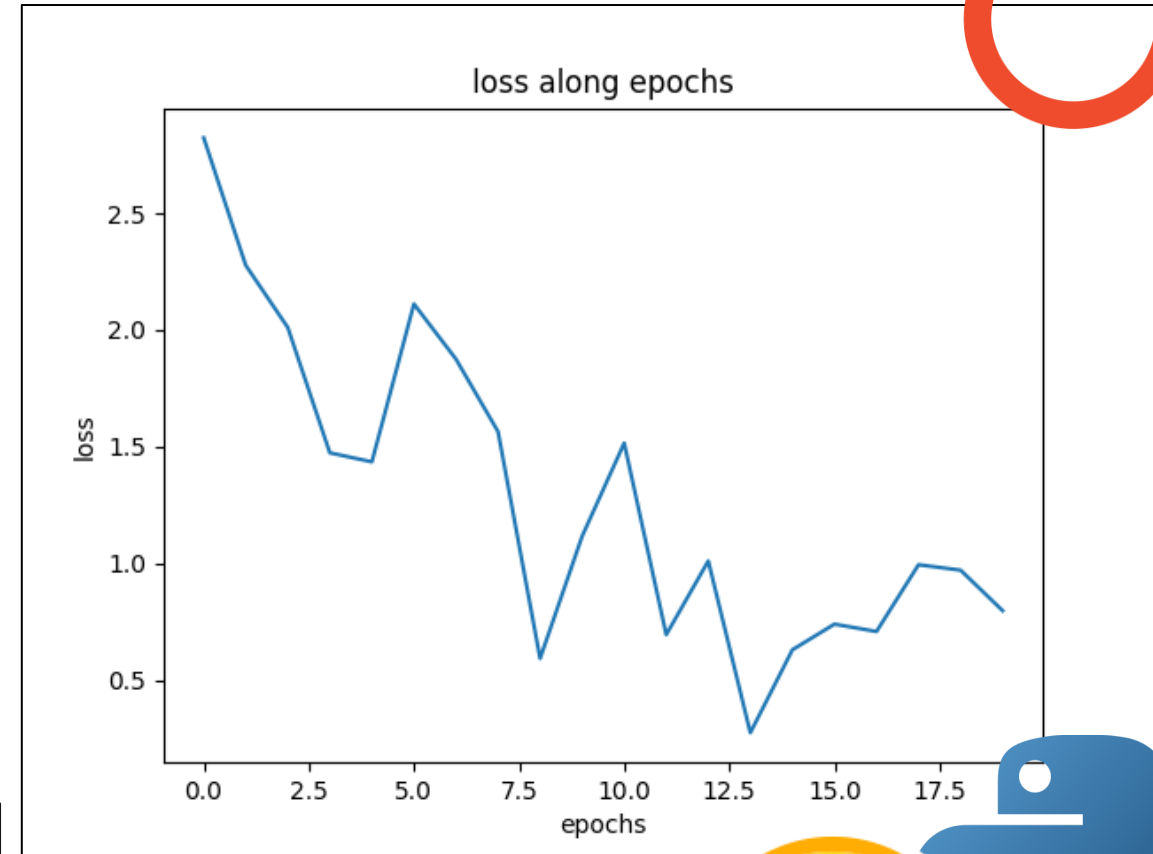
Original Text: On 13 November 2013, he netted the Charruas' second in their 5 – 0 win in Jordan for the playoffs first leg, finishing Nicolas Lodeiro's cross at close range.
Text after Deletion: On 13 November 2013, he netted the second in their 5 – 0 win in Jordan.
Text after Decontextualization: On 13 November 2013, Cristhian Stuani netted the second in 5 – 0 win in Jordan.
Final Text: On 13 November 2013 Cristhian Stuani netted the second in a 5 – 0 win in Jordan.

Table 1: Example in the ToTTo dataset. The goal of the task is given the table and set of highlighted cells, to produce the final text. Our data annotation process revolves around annotators iteratively revising the original text to produce the final text.

Train T5

- T5-large
- 20 epochs
- Batch size: 8
- Max length: 256
- Hardware: NVIDIA Titan RTX
- Loss: CrossEntropyLoss (default)
- Lr: 0.001

"I never use loss functions because I never lose"



AND NOW DEMO!

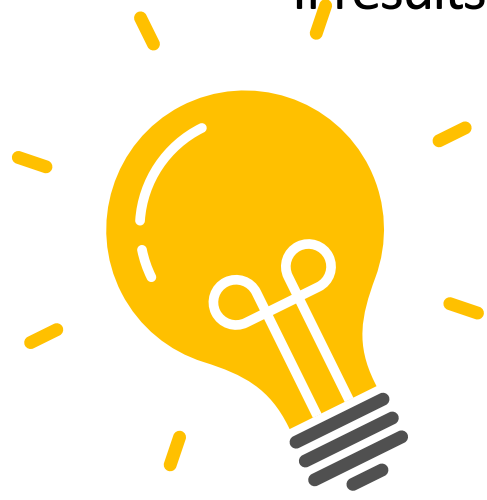
me: showing results' demo again

everyone else hearing at this presentation:



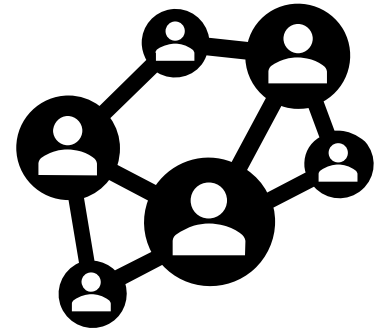
TODO list:

- Identify a way to evaluate results apart from loss and human evaluation.
- Find a way to better represent input data structure and elements relations.
- Train bigger models (T5-3b, T5-11b) using more GPUs in parallel.
- Try other models (like T5X)
- If results are robust and reliable create an inference structure/API.



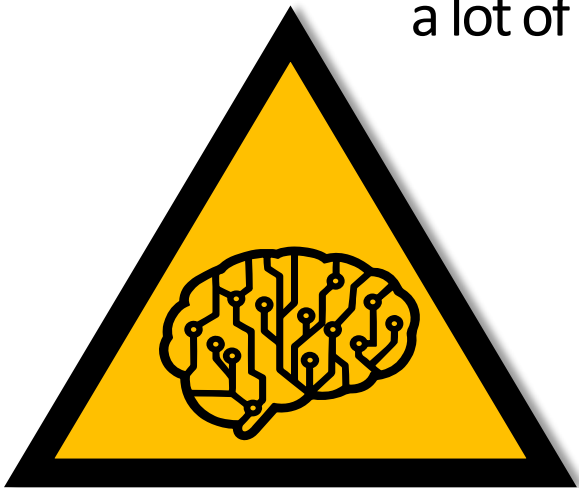
Houston, as usual, we have problems

- System invents, differently from GPT-3, but we still do not have coherence for sure.
- I need labelling resources
- Find a better data source maybe



Danger

- These systems are powerful but dangerous too.
- In wrong hands they be used to create bots/fake news or to help cyber criminals.
- OpenAI at first released GPT-3 just to developers.
- They are trained on a large corpus of data coming from different sources: they absorbed a lot of bias.



**See you next time with session titled
“Sessions generation with AI”**

Thank You!

ευχαριστώ Salamat Po متشكراً شكراً Grazie

благодаря ありがとうございます Kiitos Teşekkürler 谢谢

ໂພນດຸນດຣັບ Obrigado شكریه Terima Kasih Dziękuję

Hvala Köszönöm Tak Dank u wel ДЯКУЮ Tack

Mulțumesc спасибо Danke Cám ơn Gracias

多謝晒 Ďakujem תודה நன்றி Děkuji 감사합니다

Slides/Demo repository



<https://github.com/deltatrelabs/deltatre-global-ai-dev-days-2022-demo>





Useful links

- <https://arxiv.org/pdf/1910.10683.pdf>
- <https://arxiv.org/pdf/2004.14373.pdf>
- https://huggingface.co/docs/transformers/model_doc/t5
- <https://developer.twitter.com/en/portal/dashboard>
- <https://www.tweepy.org/>
- <https://ai.googleblog.com/2021/01/totto-controlled-table-to-text.html>
- <https://paperswithcode.com/sota/data-to-text-generation-on-totto>
- <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>
- <https://github.com/google-research/bert>
- <https://huggingface.co/bert-base-uncased?text=The+goal+of+life+is+%5BMASK%5D>
- <https://github.com/google-research/t5x>