

# 문제+데이터+AI=?

바이오지능 연구실 신입세미나

Ceyda Cinarel  
재이다 츠나렐  
21 February 2022



서울대학교

# About Me

Ceyda Çınarel  
재이다 츠나렐



 [@cceyda](#)

 [@ceyda\\_cinarel](#)

 [cceyda.github.io](#)

 Contributing to Open Source



Mathematics  
Computer Engineering



서울대학교

Computer Science & Engineering  
-> BioIntelligence Lab



2015 ↓



AI Researcher (NLP) 1년



Senior Researcher & Engineer (FashionAI) 2년 +

Attention  
Embeddings  
Color  
Augmentation  
GAN  
Style Transfer  
Transformers  
NLP ❤️ CV  
Adversarial  
Multimodal  
QA  
NER

# 문제 + 데이터 + AI

In Practice



\*made

# ML in practice이란?

Real world 데이터를 사용해 real world 문제를 푸는것

- More concerns; Monitoring, Resource Use, Privacy
- Stakes are higher
  
- Real world / real data : dataset에 없는 data  
(not even on test set)
  
- From scratch: 0 부터 시작해서
- Applied ML: ML 적용해서 쓰는것
- MLOps: Like DevOps but for ML



\*made

발표자  
조국현  
의선



문제 + 데이터 + AI

# Model~Data~Problem

## 관계를 예시로 생각

- Data  $\Leftrightarrow$  Model
  - *Zero-shot, One-shot Learning*
  - *Un/semi/supervised Learning*
- Model  $\Leftrightarrow$  Problem
  - *Multi-task learning, fine-tuning, pre-training*
- Data  $\Leftrightarrow$  Problem
  - *Reinforcement Learning* 자체



\*made

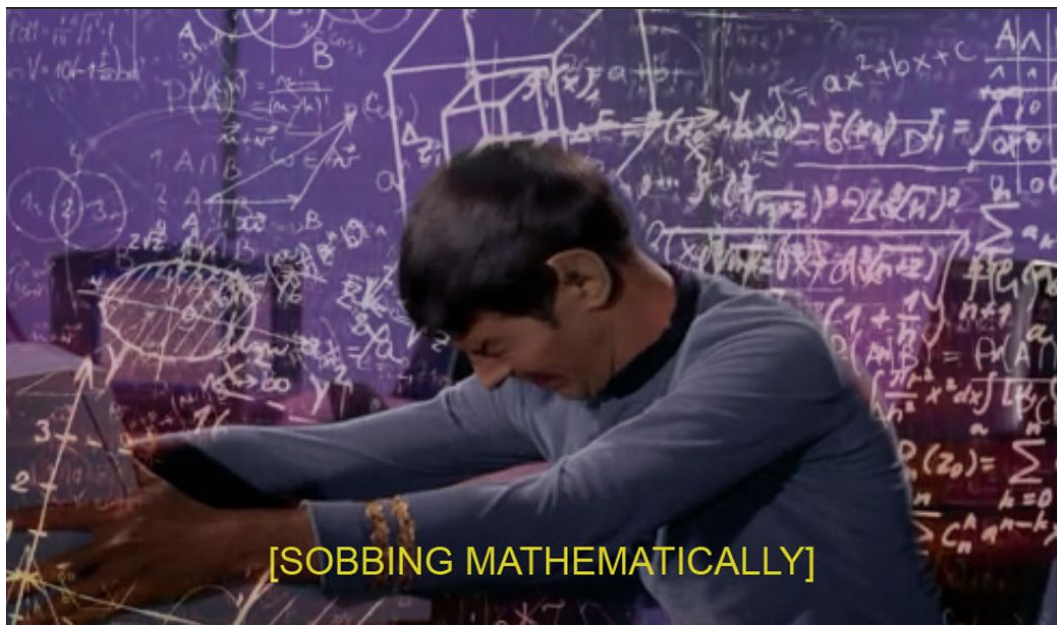


문제 + 데이터 + □

연구 관점

## AI 연구자의 영역?

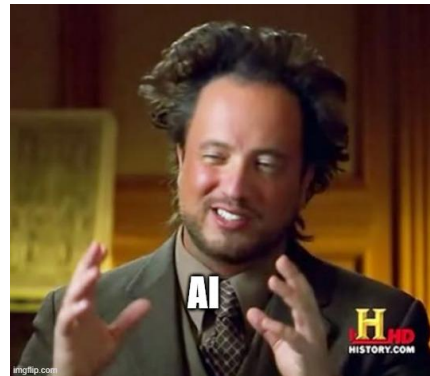
- 새로운 model architecture 만들기?
- 수학적인 증명/이해하기?



더...더 복잡하죠

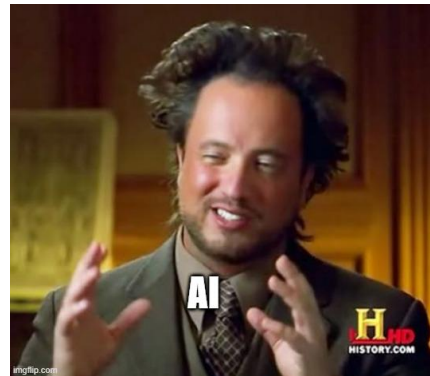
- Prediction (예측, 확률)
- Approximation (근사)
- Modelling (분포)

$$E_{\theta}(x)$$



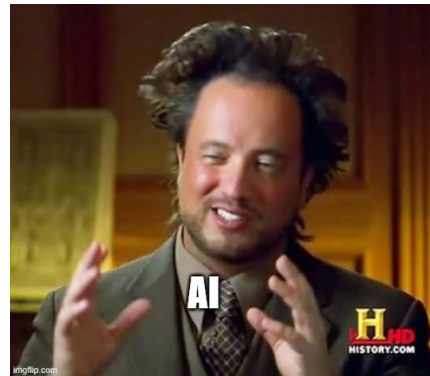
# 더...더 복잡하죠

- Prediction (예측, 확률)
- Approximation (근사)  $E_{\theta}(x)$
- Modelling (분포)
- 가능성:
  - increase 효율(expert 대신하기, 대량, 근사를 통해...)
  - 인간이 할수 없던 일 하기? ( 뛰어난 Pattern recognition )
  - 재미ㅋㅋㅋ (GAN)



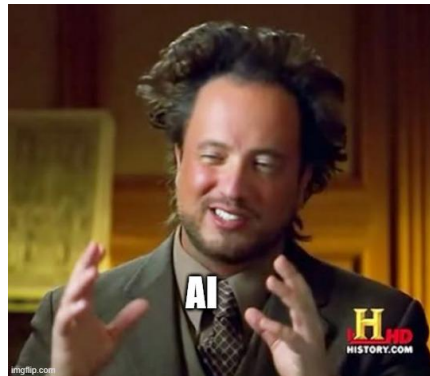
## 더...더 복잡하죠

- Prediction (예측, 확률)
- Approximation (근사)  $E_{\theta}(x)$
- Modelling (분포)
- 가능성:
  - increase 효율(expert 대신하기, 대량, 근사를 통해...)
  - 인간이 할수 없던 일 하기? ( 뛰어난 Pattern recognition )
  - 재미ㅋㅋㅋ (GAN)
- 예시:
  - 어떤 제품 매출 예측, stock price 예측
  - 제일 가까운 path 찾기
  - 로또 예측



## 더...더 복잡하죠

- Prediction (예측, 확률)
- Approximation (근사)  $E_{\theta}(x)$
- Modelling (분포)
- 가능성:
  - increase 효율(expert 대신하기, 대량, 근사를 통해...)
  - 인간이 할수 없던 일 하기? ( 뛰어난 Pattern recognition )
  - 재미ㅋㅋㅋ (GAN)
- 예시:
  - 어떤 제품 매출 예측, stock price 예측 ✓(~)
  - 제일 가까운 path 찾기 ✓
  - 로또 예측 X



# DeepMind's AI helps untangle the mathematics of knots



BLOG POST  
RESEARCH

30 NOV 2020

**AlphaFold: a solution to a 50-year-old grand challenge in biology**



BLOG POST

16 FEB 2022

**Accelerating fusion science through learned plasma control**

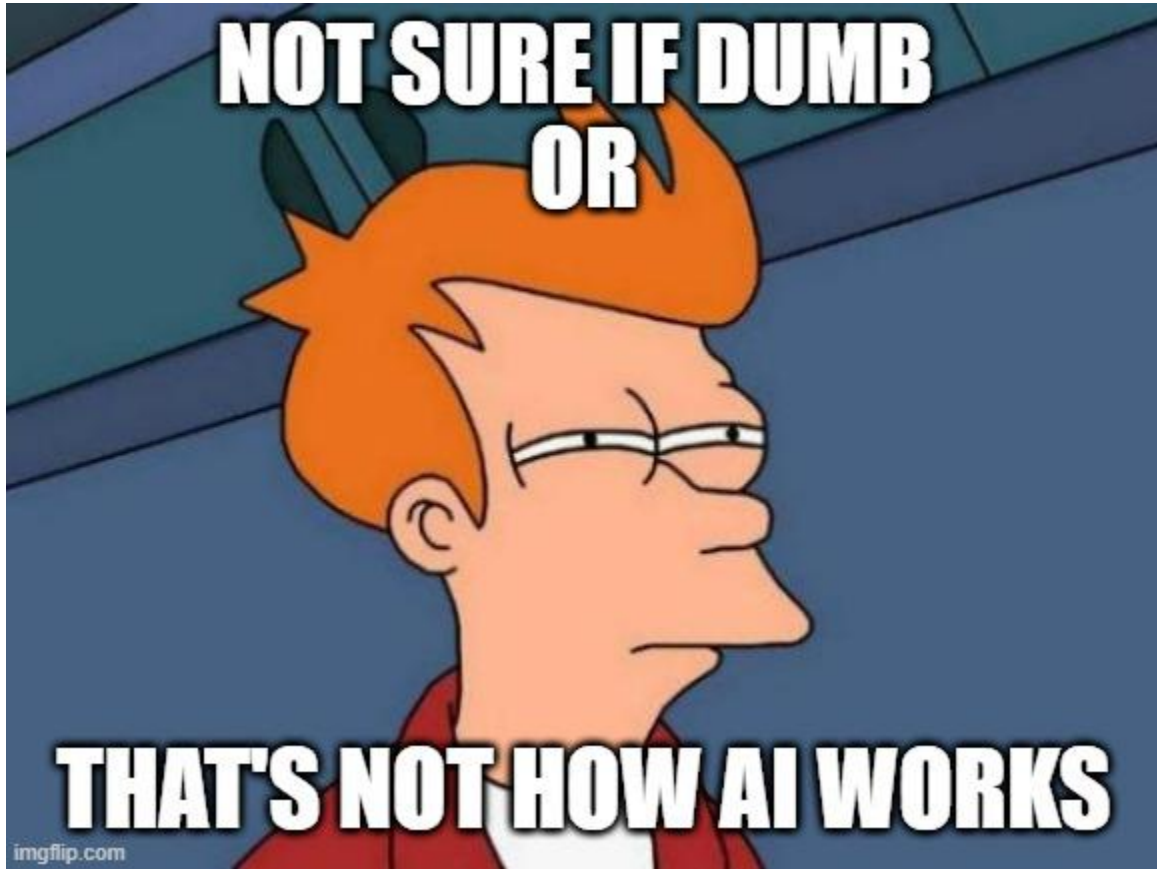






\*made





\*made

# AI Hype 주의!!!



AI 아닌것



\*made

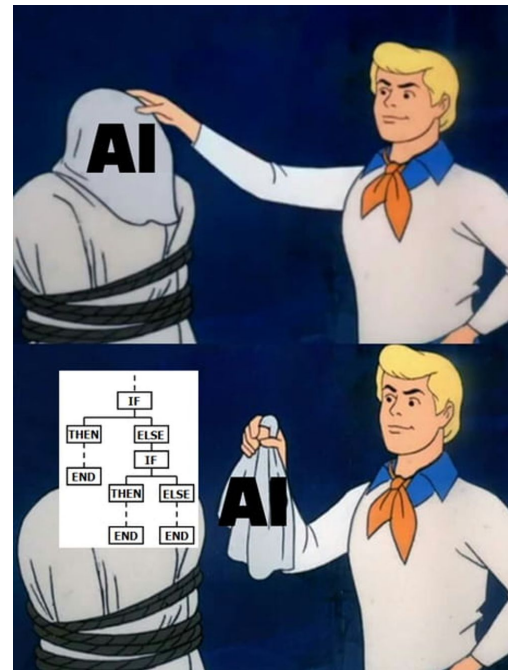
\*made

# AI Snake Oil 🐍

AI로 할수 있는것

AI로 할수 없는것

AI 아닌것







출처:직접 찍음  
장소: 양재천



가짜

AI(인공지능)

확산방지 및 예방을 위하여  
출입을 금하여 주시기 바랍니다.

서초구청



□ + 데이터 + AI

# 문제를 어떻게 정의하는가에 나름이다 (Depends on how you frame the problem)

Problem Definition & 어떤 Task을 풀건지

- Classification (많음. QA, segmentation, translation 이든)
- Regression
- Embedding
- ...
- “ Input - [Model] - Output “ 단위 abstraction
  - 예: Token/patch, pixel 단위
  - 예: Next token classification , 중간 token classification

Loss을 어떻게 생각할지의 영향

# 굳이 그 문제를 풀어야 하는 이유?

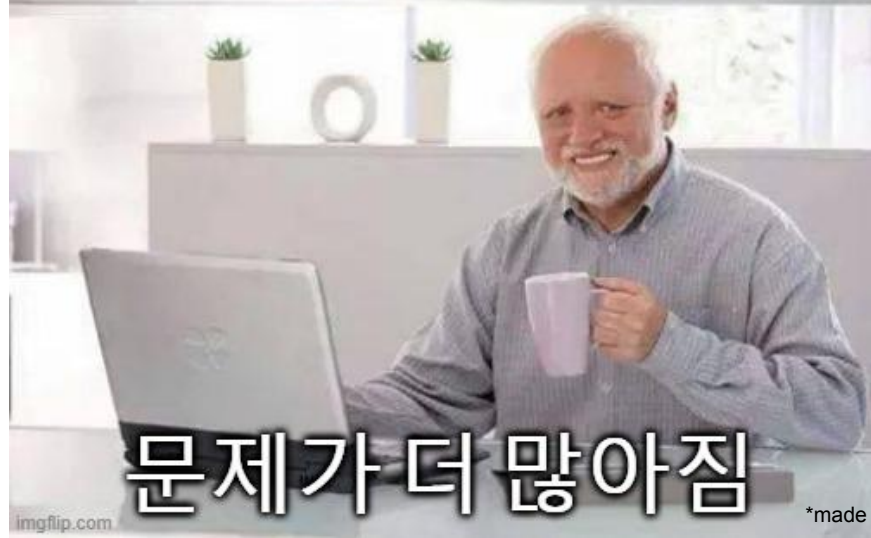
- 그 문제를 회피할수 있나  
(Can you avoid solving the problem)
  - 예시:
    - 문제:sentence split해라
    - 이유:downstream task에 필요
    - No; use model with longer context etc.
- 다른 문제로 같이 풀수 있나 (일석이조)
- 다른 문제를 풀면서 할수 있나
  - Multi-task learning, fine-tuning, pre-training






## 문제 유래 AI

- 사람을 대신하는 문제 풀자
- 사람이 하기 어려운 문제를 풀자
- 사람이 하기 가능하지 않은 문제를 풀자
- ( Novel ) 독특한 문제를 풀자
- 재미 있는 문제 풀자



문제 +  + AI

Data Scientist 관점

# $P$ given data $\{x_i\}_{i=1}^N$

주어진 뭔가 벌써 있다고 가정 ... Nothing is given!

- Scraped Data
- 만들어진 Data
- RL (가상 환경에 만들어진 데이터)
- Generated data (GAN 등으로)
- Label필요?



# Know Your Data

은행에 Know your Customer(KYC) 제도 있다면

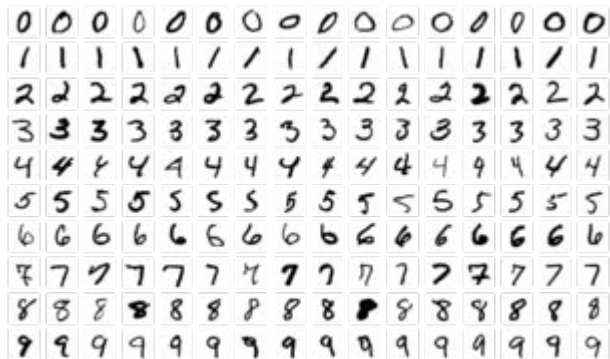
AI에 Know Your Data!

- Data의 제한적인 면을 알아라
- 수집법을 알아라
- 특성
  - Image: resolution, format
  - Audio: sampling rate, background noise
  - Text: Unicode encoding
- 기관 / License
- 역사



\*made

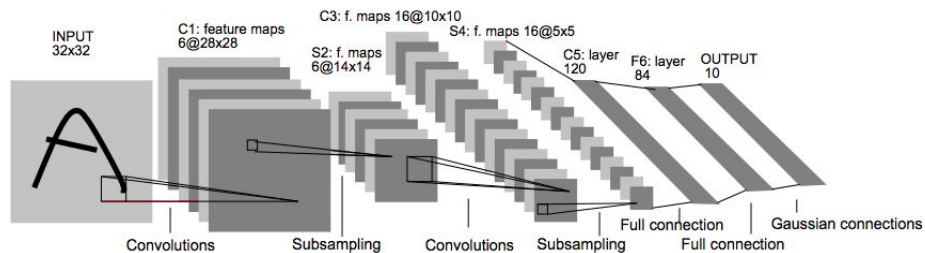
저희가 잘 아는 평범한 그 MNIST이...



Kyunghyun Cho @kchonyc 9 Dec 2021

@keunwoochoi just pointed out in his lecture that @ylecun's LeNet-5 figure has an oddity (which i've failed to see after seeing it M's of times...)

if it's for "digit recognition", why is the input "A"?????



Yann LeCun

@ylecun

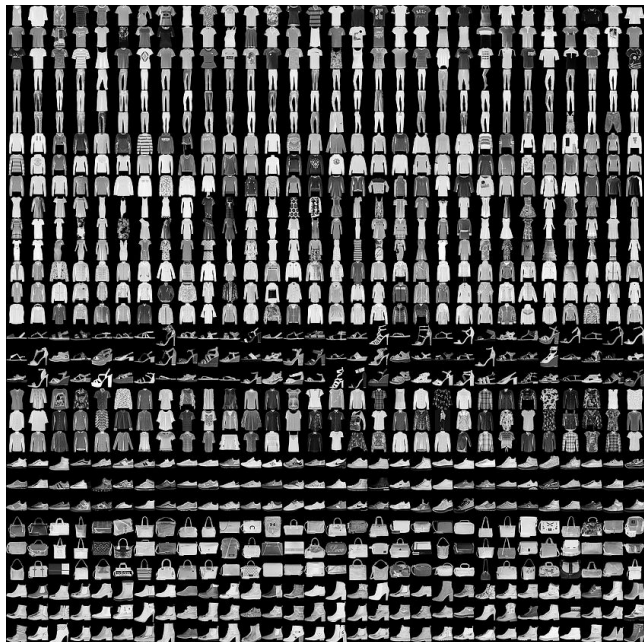
Replying to @kchonyc @keunwoochoi

The real system was trained on alphanumeric characters, which made it work better on just digits.

12:30pm · 9 Dec 2021 · Twitter for Android

2 Replies 2 Retweets 67 Likes





## Use more than 1 dataset!

기본:

- Cleaning
  - data 중복 제거, mislabeled
- Data Augmentation
- Inbalance (분포)
  - Bias의 원인을 수도

# MIT pulls down Racist and Misogynistic **Tiny Image Dataset** used for training AI after 14 Years

By **Ajit Abhyankar** - July 3, 2020

## ImageNet

ImageNet will remove 600,000 images of people stored on its database after **an art project exposed racial bias in the program's artificial intelligence system**. Created in 2009 by researchers at Princeton and Stanford, the online image database has been widely used by machine learning projects.

Sep 23, 2019



\*made



# Data 유래 AI

Scraping...

주의 점:

저작권 문제

License

Data Poisoning



Aha 🎵 All the things I could do 🎵 if I had a little data  
Abba - Money, Money, Money



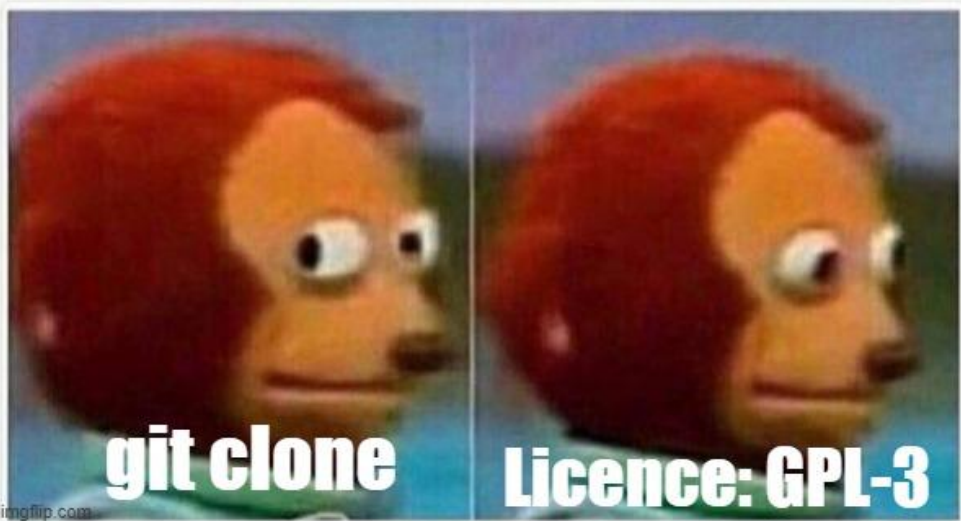
It's big data's world ↓

\*made





streamlit/streamlit is licensed under the  
**Apache License 2.0**



\*made

### Permissions

- ✓ Commercial use
- ✓ Modification
- ✓ Distribution
- ✓ Patent use
- ✓ Private use

### Limitations

- ✗ Trademark use
- ✗ Liability
- ✗ Warranty

### Conditions

- ⓘ License and copyright notice
- ⓘ State changes

출처: github

Can	
▶ Commercial Use	
▶ Modify	
▶ Distribute	
▶ Place Warranty	
▶ Use Patent Claims	

Cannot	
▶ Sublicense	
▶ Hold Liabile	

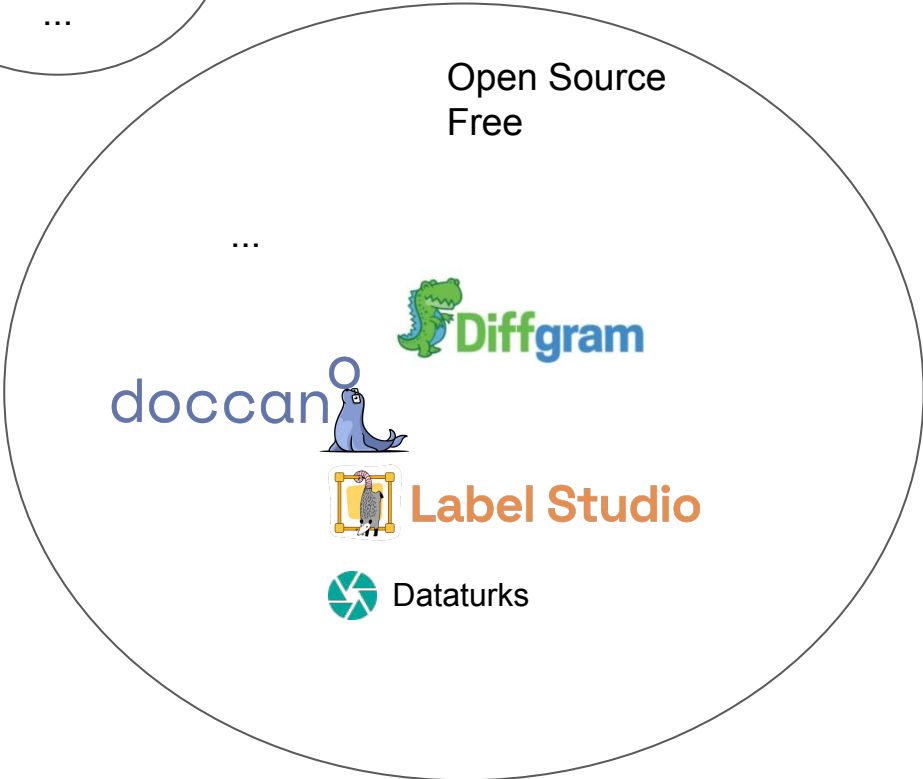
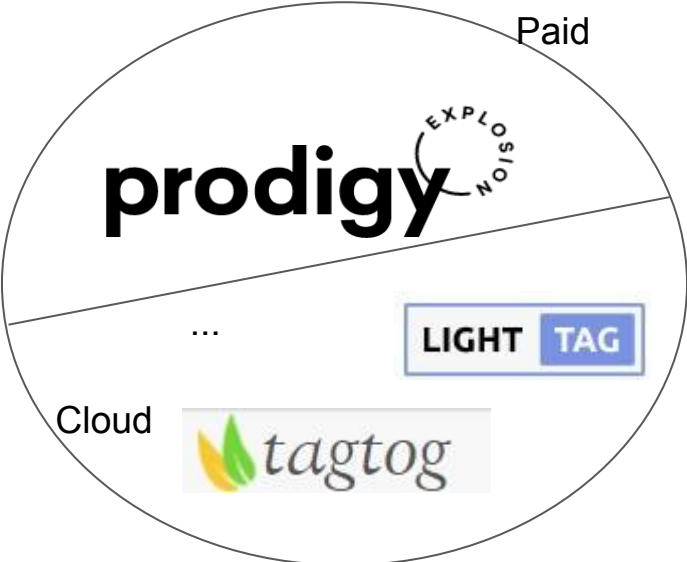
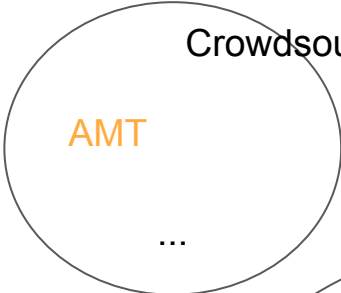
Must	
▶ Include Original	
▶ State Changes	
▶ Disclose Source	
▶ Include License	
▶ Include Copyright	
▶ Include Install Instructions	

출처: <https://tldrlegal.com/>

# Dataset 모으기

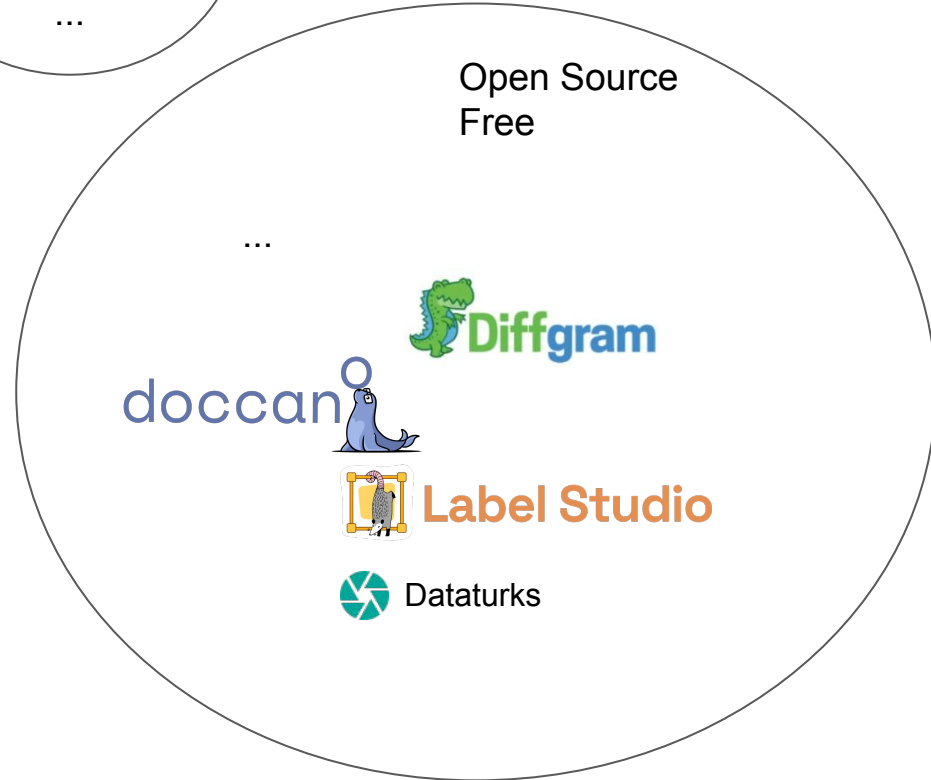
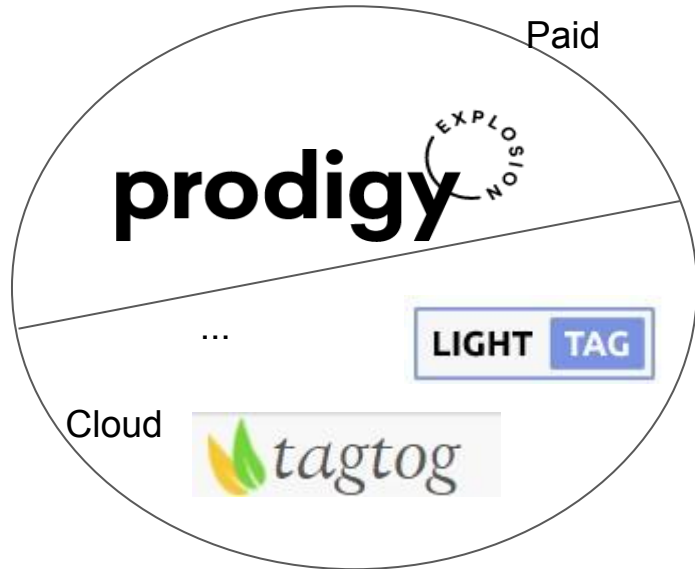
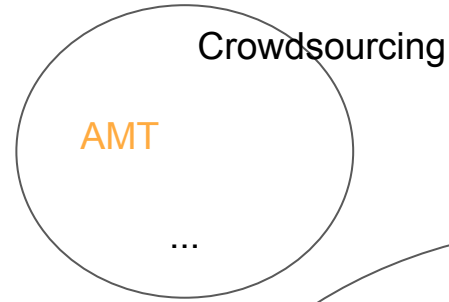
- 과제로 **dataset** 모으기 요청 오기도한다
- 학회에 **dataset** 소개 페이퍼 제출할 때 유의점
  - Annotators' guideline 같이 제출 (NeurIPS)
  - Ethical concerns
- 하고 싶은 **task**을 학습할수 있는 데이터가 어떻게 생길까
  - Not always straightforward (뻘하지 않다)
  - 생각 없이 하다 다 쓰레기 될수 있음!
  - 성능에 영향 줄수 있음
    - 예:Recommendation System ( Contextual & Collaborative )

# Data Annotation



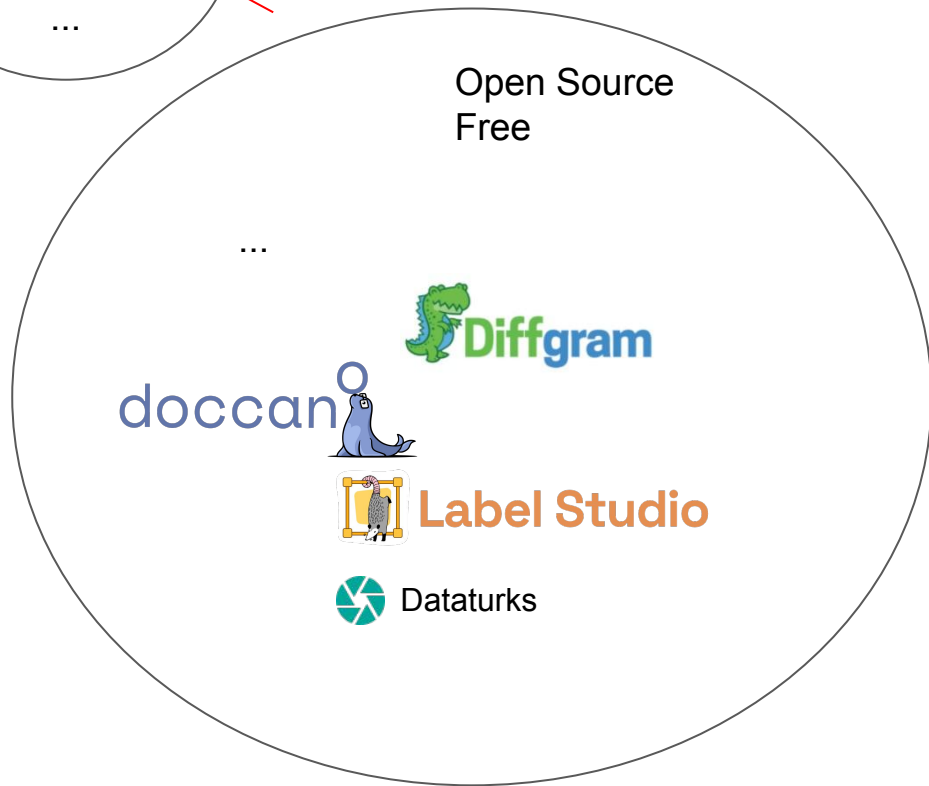
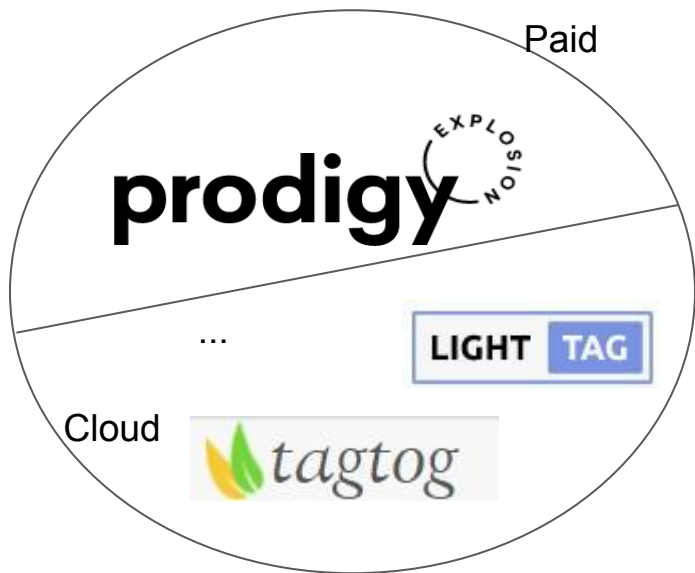
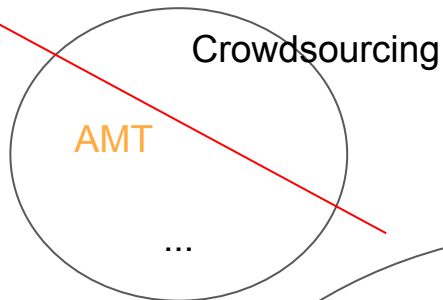
# Data Annotation

- Who are your annotators?
  - Experts
  - Annotator guide/training



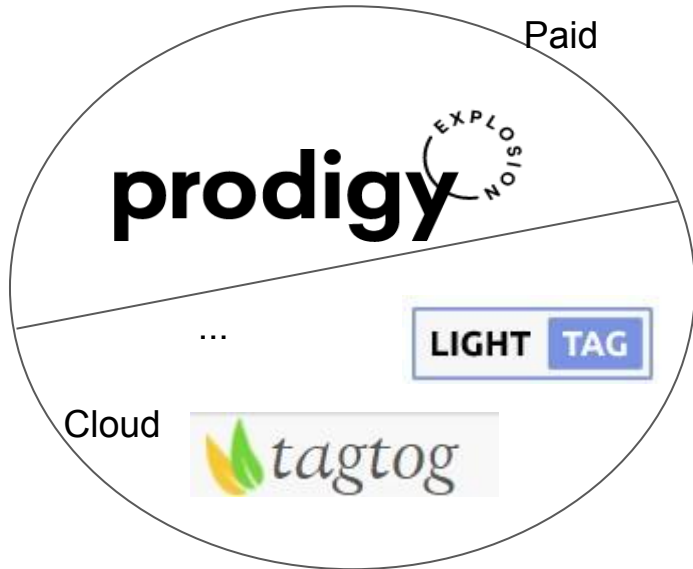
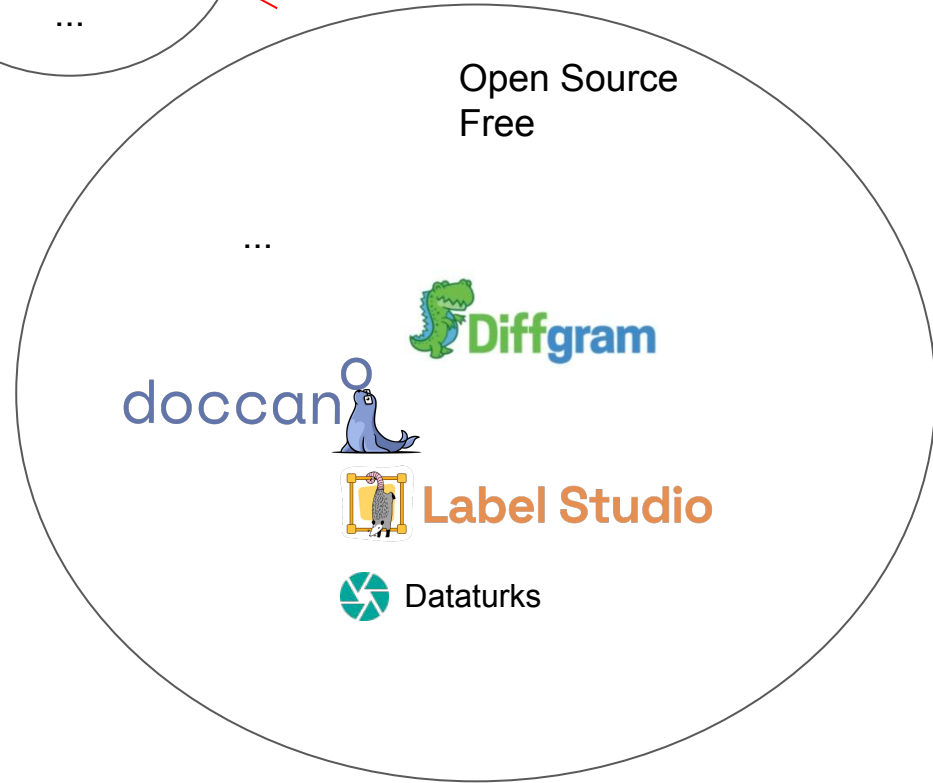
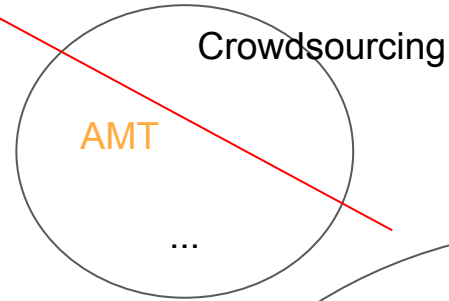
# Data Annotation

- Who are your annotators?
- Data privacy issue?
  - Self-hosted options
  - Pipeline access



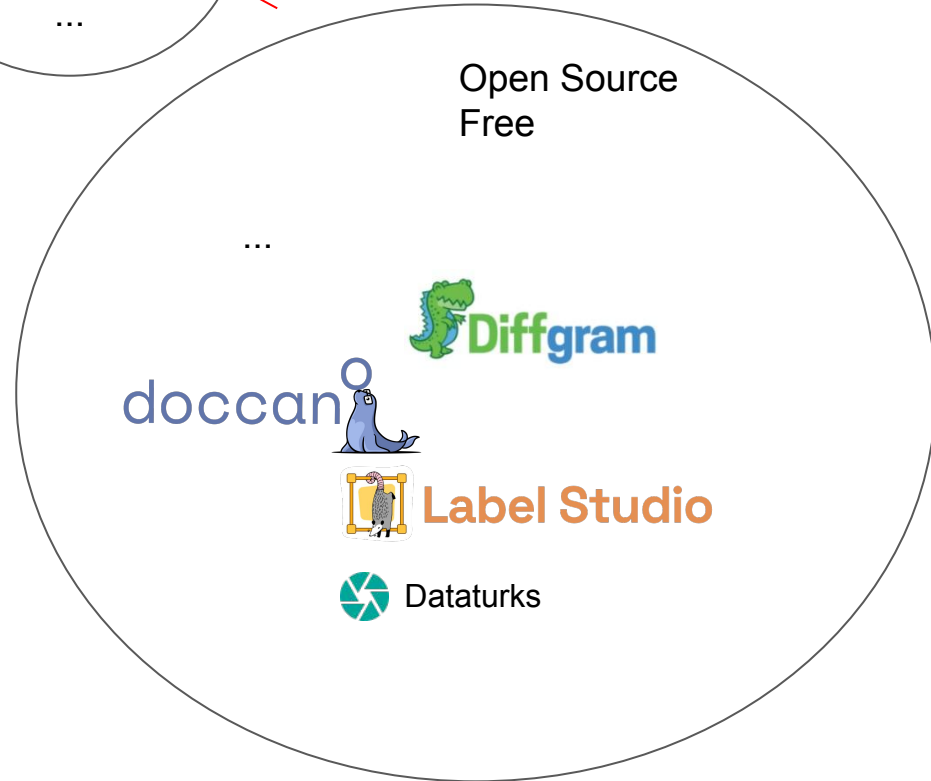
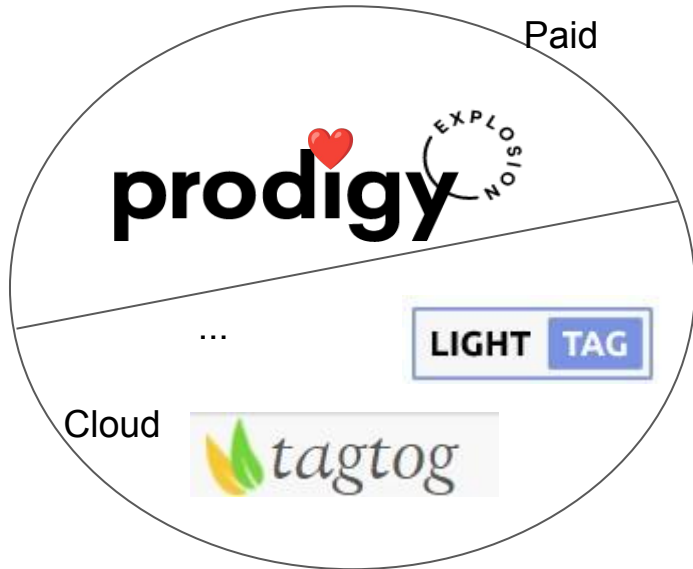
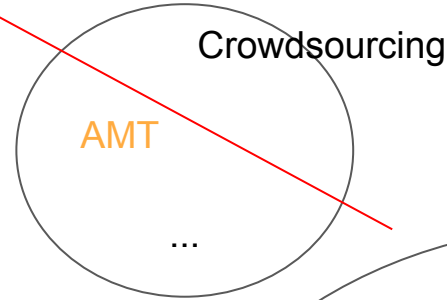
# Data Annotation

- Who are your annotators?
- Data privacy?
- UI comfort / Ease of use?



# Data Annotation

- Who are your annotators?
- Data privacy?
- UI comfort / Ease of use?



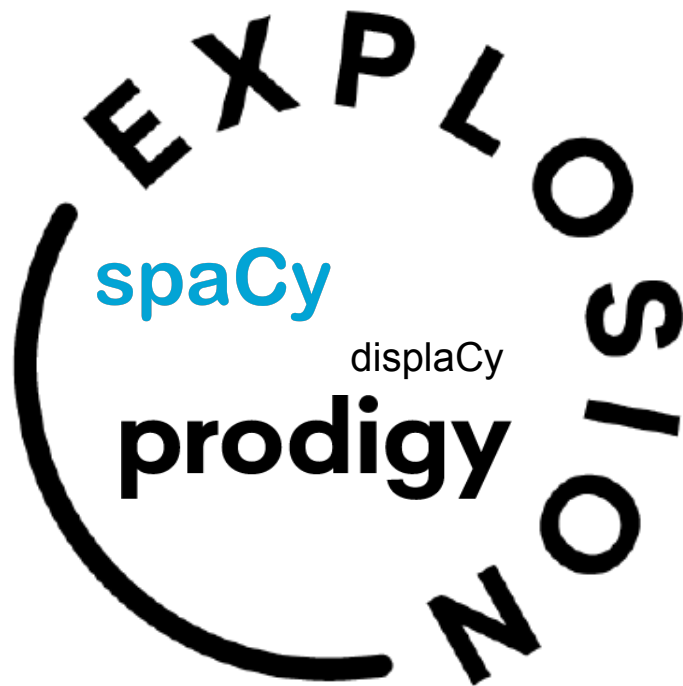
# Why

## Prodigy

- Multiple Annotation Types (CV+NLP+Audio)
- Python API pluggable custom annotation scripts
  - Streams / Callbacks
- Pretty UI
- Support

Format compatibility across tools makes life easier on the researcher 😊~

- + Warm fuzzy feeling you get for supporting spaCy development





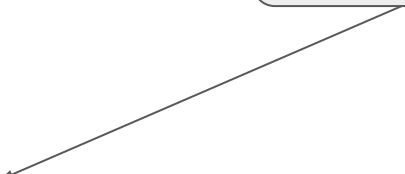
Data Ingress



Annotation



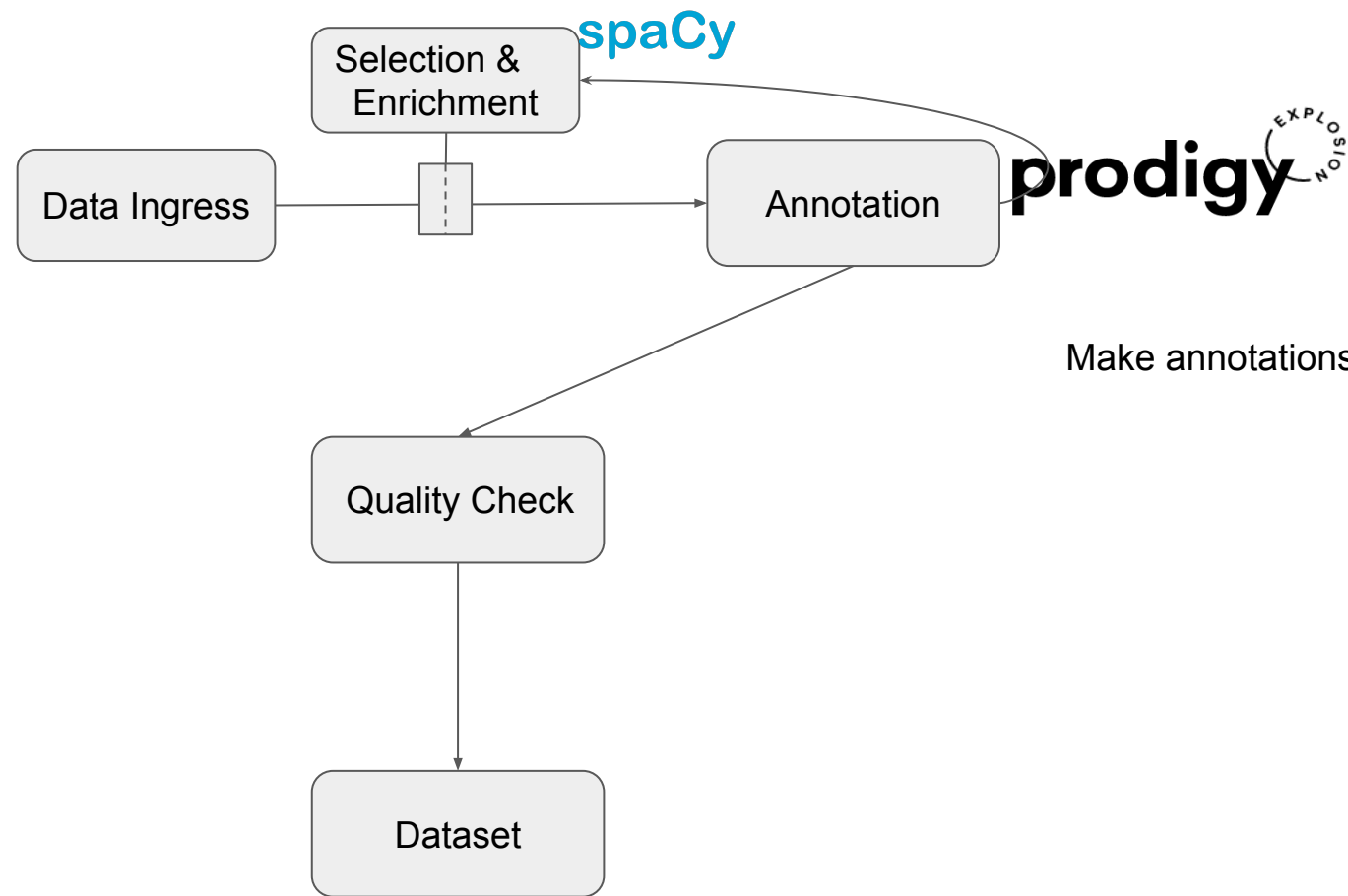
- Provide Guidelines for Annotators



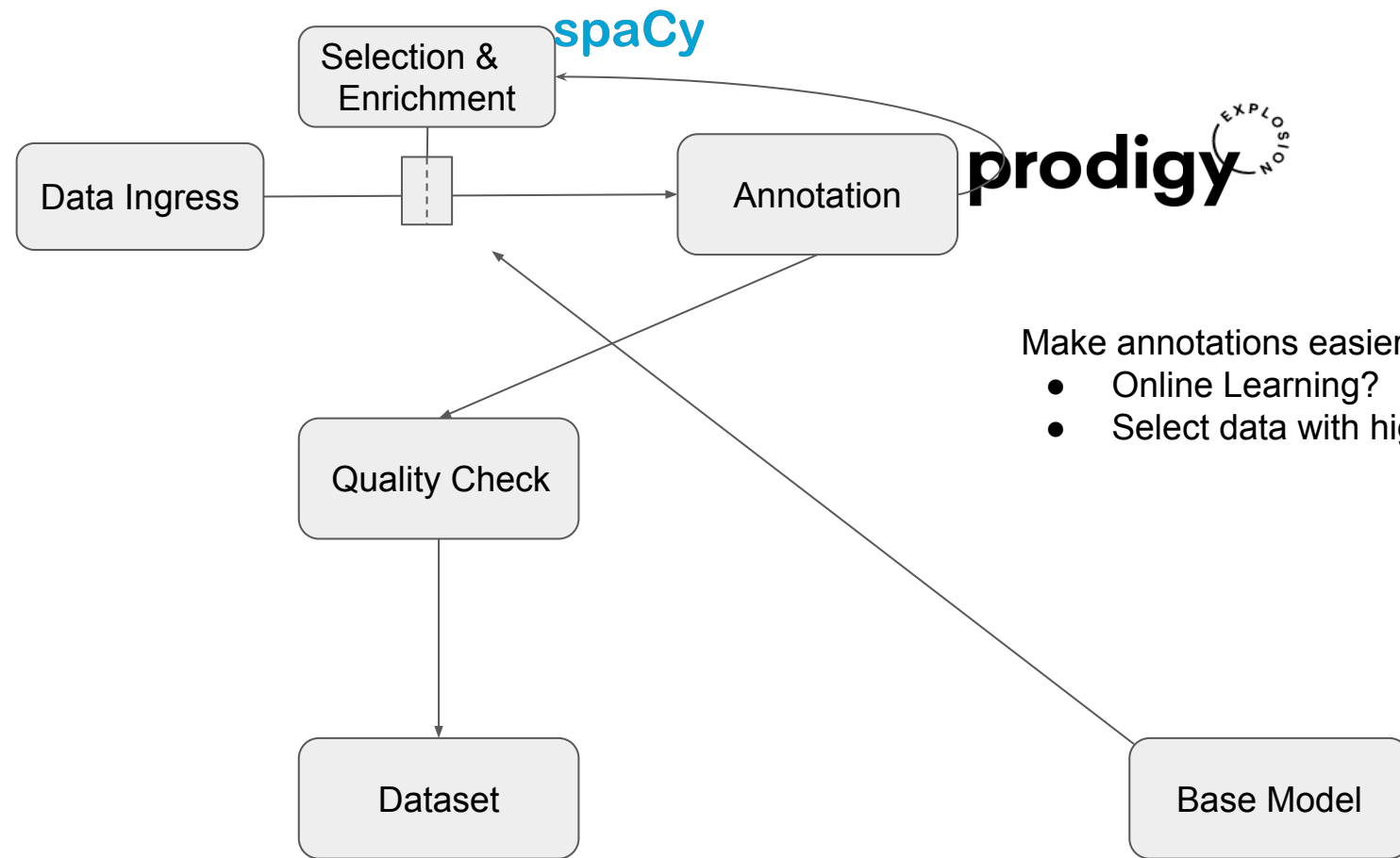
Quality Check



Dataset

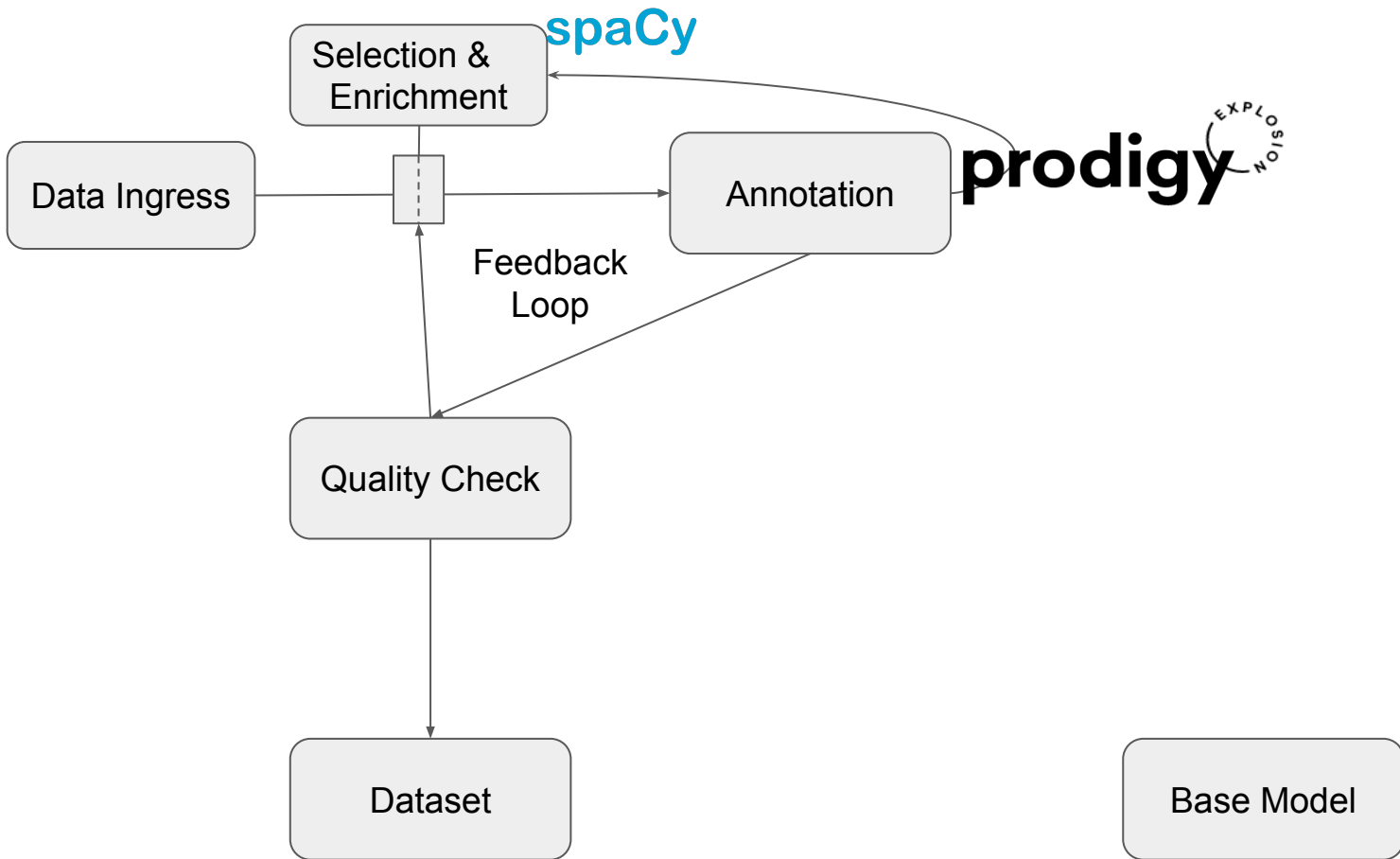


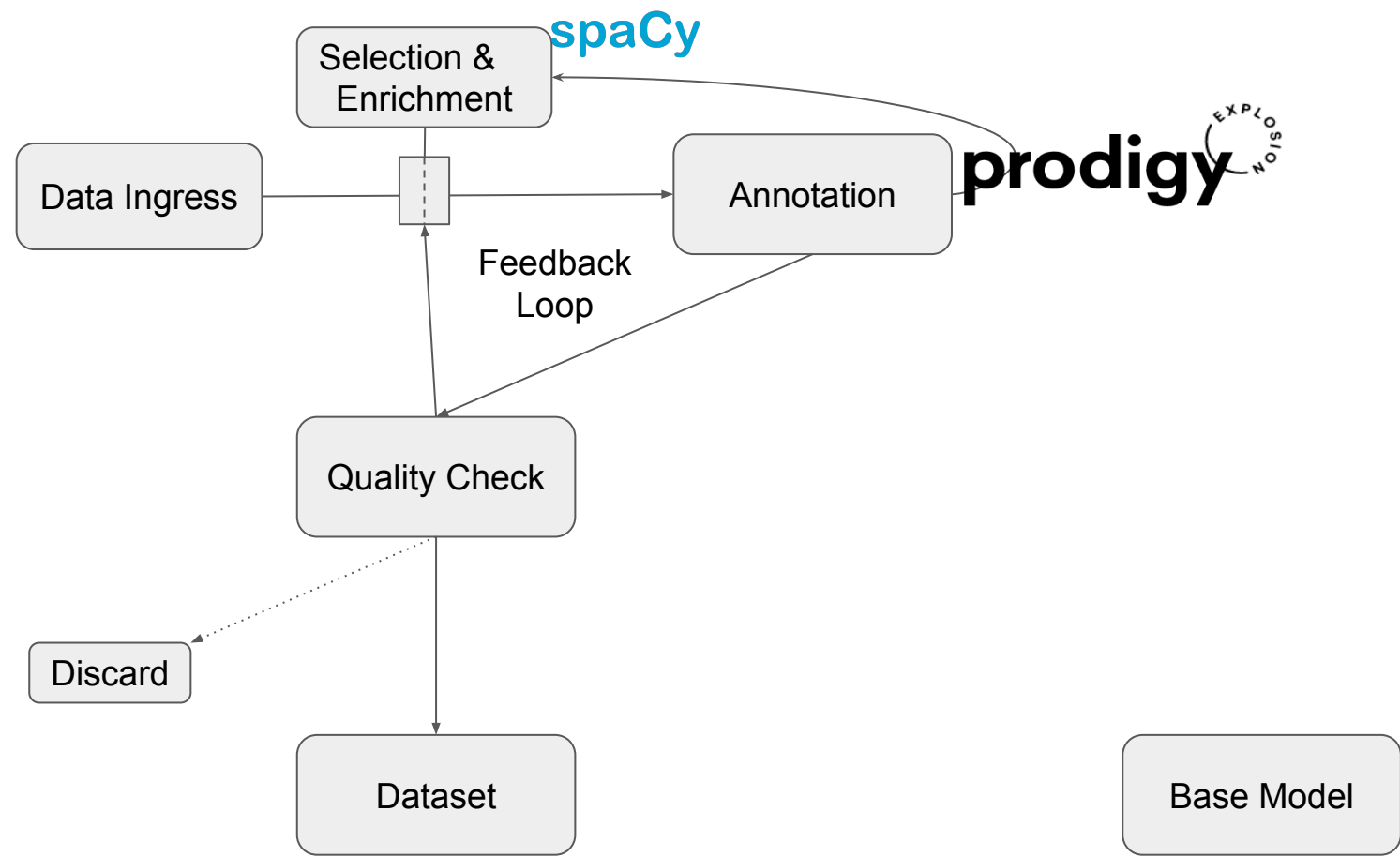
Make annotations easier to generate

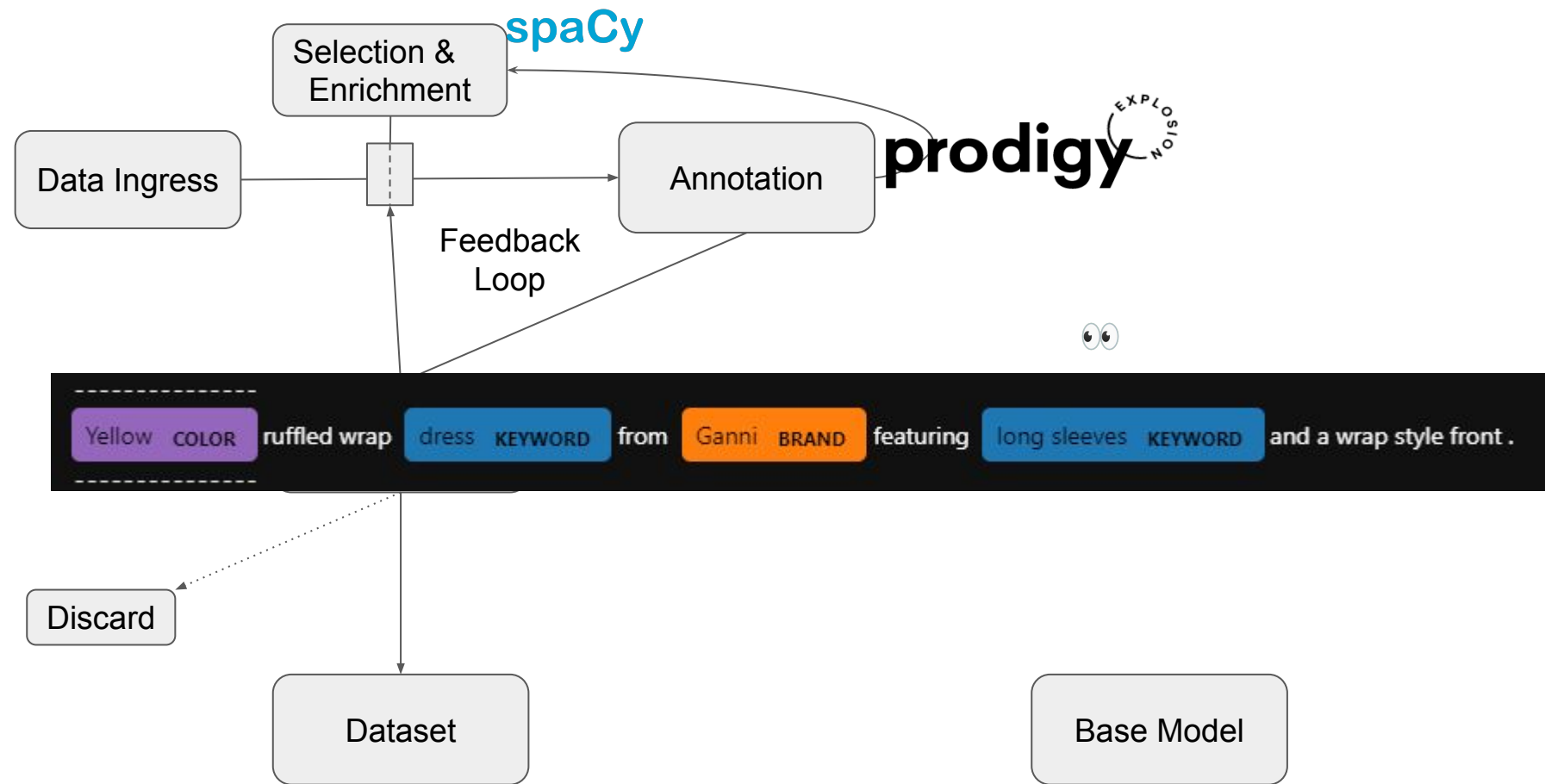


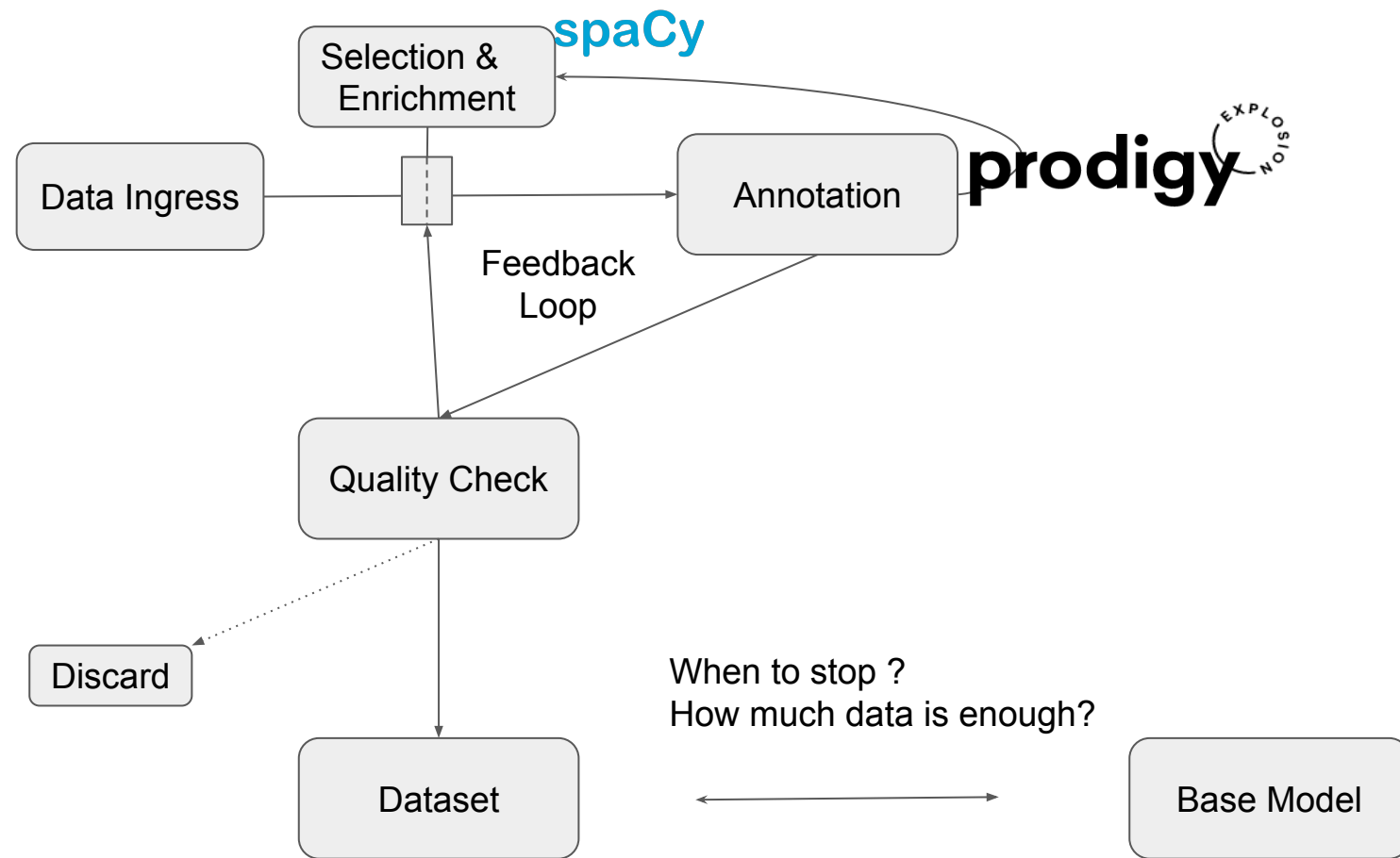
Make annotations easier to generate

- Online Learning?
- Select data with high information gain

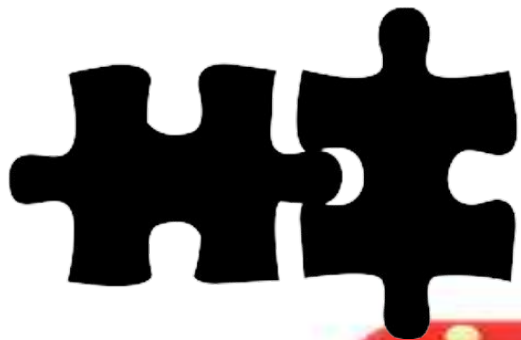










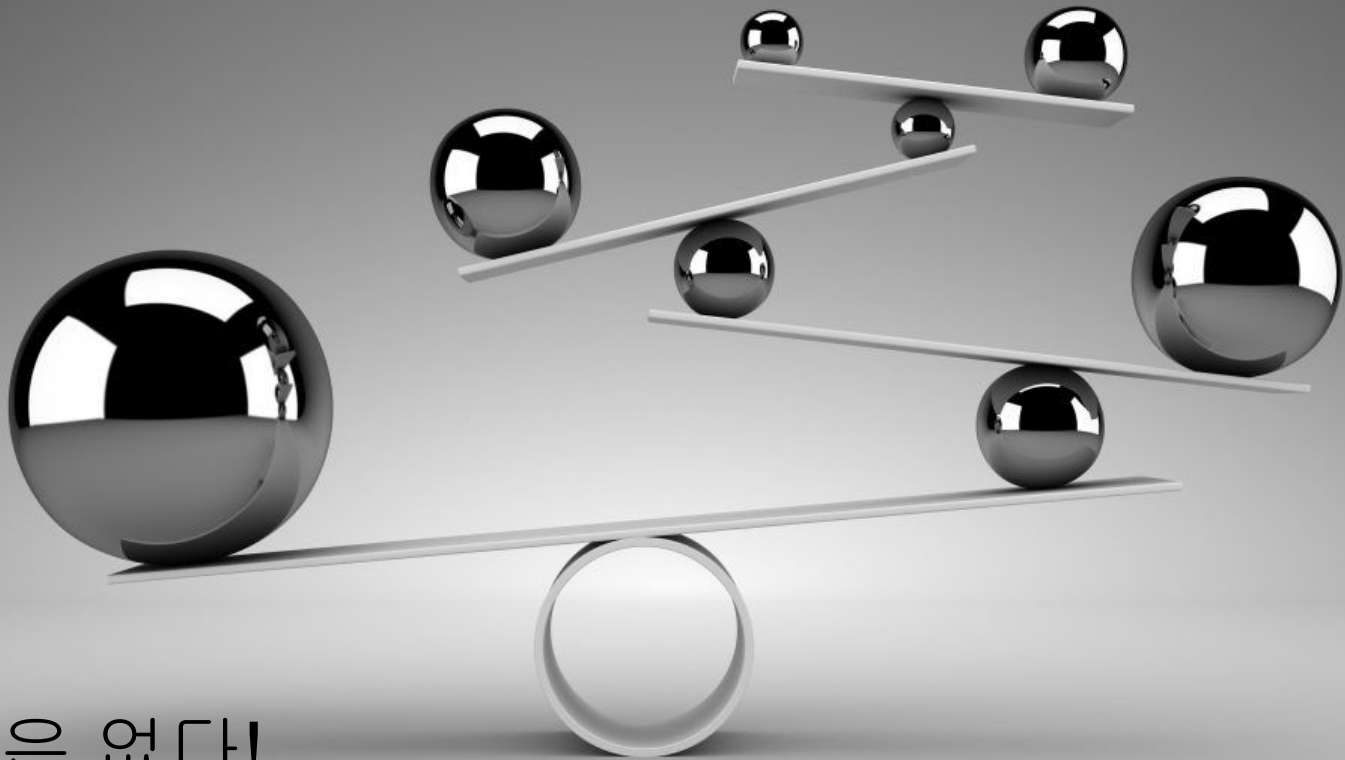


# Balance 잡기

시간  
돈



귀찮음



정답은 없다!

빨리 빨리 기준을 잡을수 있는것: 바로 경력

# Project Life Overview

- Problem Definition (문제)
- Data (데이터)
  - Collection
  - Annotation (Optional)
  - Processing



- Modelling (AI)
  - Train
  - Evaluate
- Deploy
  - Server
  - UI
- Monitor



AI Memes for Artificially Intelligent Teens  
@ai\_memes



Machine learning pipelines



1:45 AM · Apr 15, 2021 · Twitter Web App

9,119 Retweets 1,985 Quote Tweets 34.2K Likes

# Project Life Overview

- Problem Definition (문제)
- Data (데이터)
  - Collection
  - Annotation (Optional)
  - Processing
- Modelling (AI)
  - Train
  - Evaluate
- Deploy
  - Server
  - UI
- Monitor

%85



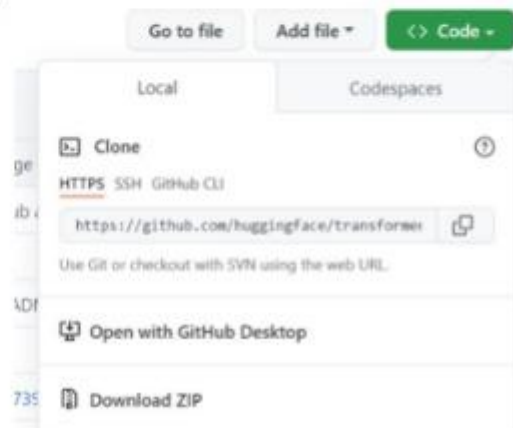
( + 영원히 반복 )

%15 or less if you use the right tools!

Can we stop and get some **AI** ?

Mom: We have **AI** at home

**AI** at home:



# Academia vs Industry

- Human In-the-loop에 대한 요구 (Human Computer Interaction (HCI))
- Interpretability의 필요성!= 수학적 이해
- Model Drift - 모델 계속 관리를 해야되요
  - (다른 카메라로 찍힌 사진, 센서의 저하)
- Ethics, law
- Paper / Patent





ML



Environment 관리

Reproduce

Log  
Monitoring

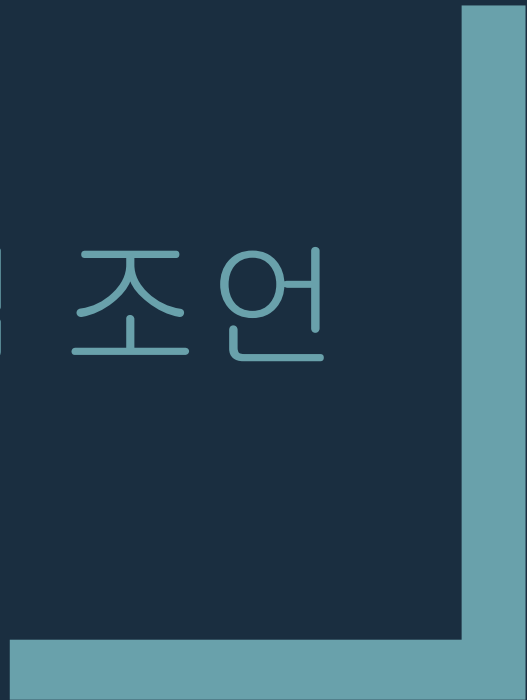
Data  
Wrangling

Hyper  
Parameter  
Tuning

Model Drift

Security  
OAuth

# 신입 조언





\*made

**RESEARCHER**



**42**

**IS THIS A RANDOM SEED**

# 💩 I wish I knew & 💩 I'm still learning

당연하다고 얘기 안하고 넘어가면 손해!



Github - git사용법/원리 (PR, private fork, branch 관리)  
(요즘 data & model 마저 version control 하는 세상이다)



Markdown 익숙해지기 (Notion 학생계정 무료)



Community 참요, Communication능력 상승

- 이런것은 CV 그 이상의 중요성을 가짐! (이런 사람 찾기 힘들)





python<3.6



python>=3.7

Version	Released	Security Support	Release
3.10	4 months and 2 weeks ago (04 Oct 2021)	Ends in 4 years and 7 months (04 Oct 2026)	<a href="#">3.10.2</a>
3.9	1 year and 4 months ago (05 Oct 2020)	Ends in 3 years and 7 months (05 Oct 2025)	<a href="#">3.9.10</a>
3.8	2 years and 4 months ago (14 Oct 2019)	Ends in 2 years and 7 months (14 Oct 2024)	<a href="#">3.8.12</a>
3.7	3 years and 7 months ago (27 Jun 2018)	Ends in 1 year and 4 months (27 Jun 2023)	<a href="#">3.7.12</a>
3.6	5 years ago (23 Dec 2016)	Ended 1 month and 3 weeks ago (23 Dec 2021)	<del><a href="#">3.6.15</a></del>

출처: <https://endoflife.date/python>



# 가상 환경 / 컨테이너 알아두기





# IDE 쓰세요

- VS Code 장점
  - 여러 서버 Remote dev 가능
  - Github copilot
  - Github codespaces 쓸때의 익숙함
  - 무거울수 있다
- Jupyter Lab 장점
  - Web으로 접속 가능

mat3e.github.io/brains

Notepad



Syntax Highlighting



Auto Complete



Code Linking





\*made

## What is JAX?

JAX is [Autograd](#) and [XLA](#), brought together for high-performance machine learning research.

With its updated version of [Autograd](#), JAX can automatically differentiate native Python and NumPy functions.



vopani / [jaxton](#)

100 exercises to learn JAX

 Jupyter Notebook  300  21 Updated 11 days ago

**EASY IN MY TIME**



**IT WAS NOT**



말이야

라때는



# Experiment Tracking with Weights & Biases



- Metrics
- Resource Usage / Time
- Parameter Sweep
- Report Generation
- Artifacts

> System 14



Create report

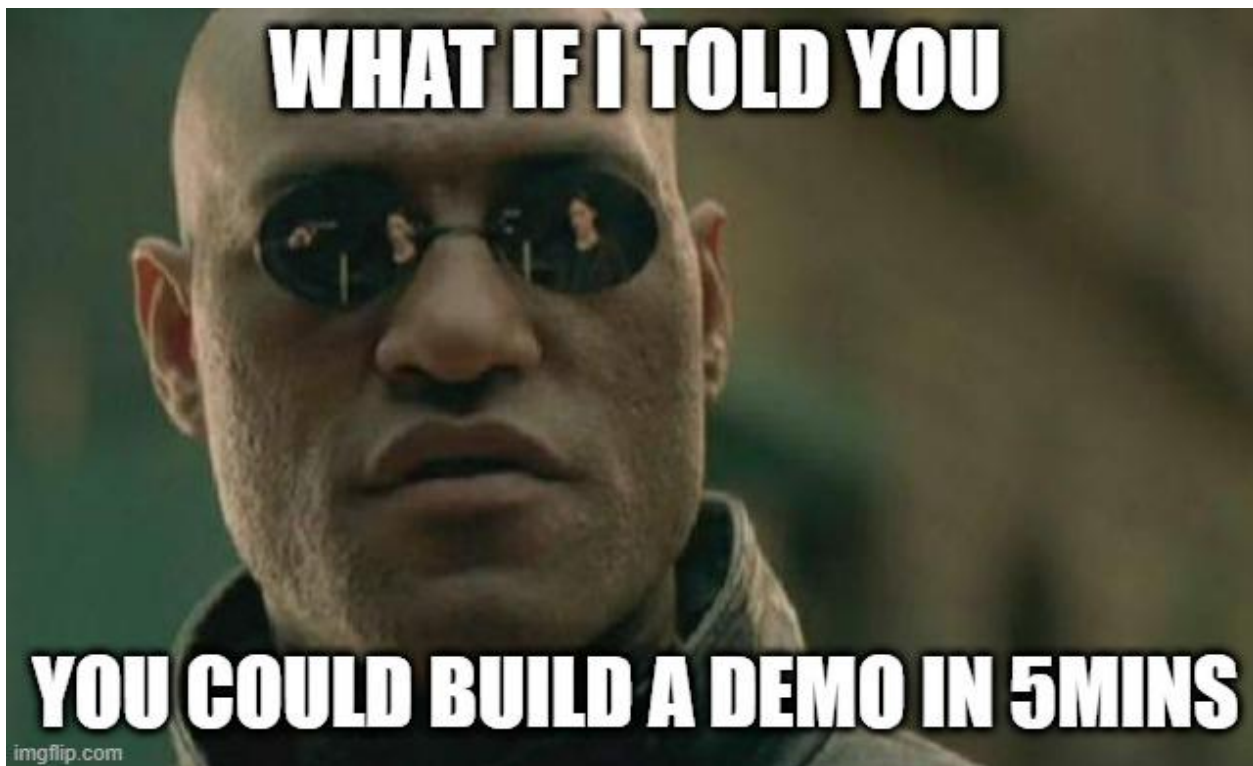
eval\_performance\_measure.FN

```
charmed-energy-7 0  
/mnt/outputs/ner_bert/fashion_ner_test_en 1  
/mnt/outputs/ner_bert/kr/fashion_ner_test_kr 2  
/mnt/outputs/ner_bert/kr/fashion_ner_test_kr 3  
/mnt/outputs/ner_bert/fashion_ner_test_en 4
```



\*made

\*Javascript필요 없음!



\*made



Streamlit



gradio



# Streamlit Build Web Apps with Python

\*Affiliated

```
1 from spacy import displacy
2 import os
3 import streamlit as st
4 from utils import hf_ents_to_displacy_format, make_color_palette
5 import httpx
6
7 HTML_WRAPPER = """<div style="overflow-x: auto; border: 1px solid #e6e9ef; border-radius: 0.25rem; padding:
8 1rem; margin-bottom: 2.5rem">{</div>"""
9
10 @st.cache()
11 def load_model():
12     labels = ["B-BRAND", "B-COLOR", "B-FABRIC", "B-KEYWORD", "I-BRAND", "I-COLOR", "I-FABRIC", "I-KEYWORD", "O"]
13     color_map = make_color_palette([l.split("-")[-1] for l in labels])
14     return color_map
15
16 def predict(input_):
17     # confirm input str
18     res = httpx.post("http://127.0.0.1:7863/predictions/fashioner", data=input_)
19     return res.json()
20
21 def display(bert_ents):
22     bert_doc = hf_ents_to_displacy_format(bert_ents, ignore_entities=["O"])
23     html = displacy.render(bert_doc, manual=True, style="ent", options={"colors": color_map})
24
25     html = html.replace("\n", " ") # Newlines seem to mess with the rendering
26     st.write(HTML_WRAPPER.format(html), unsafe_allow_html=True)
27
28 color_map = load_model()
29
30 st.header("FashionER")
31 input_ = st.text_input("Input", "silk blue mini skirt by Designovel")
32 bert_ents = predict(input_)
33 display(bert_ents)
```

## FashionER

Input

silk blue mini skirt by Designovel

silk FABRIC blue COLOR mini skirt KEYWORD by Designovel BRAND

## Korean FashionER

Input

레오파드 패턴 치마

레오파드 FABRIC 패턴 치마 KEYWORD

Just 30~50 lines

```
import gradio as gr
```

```
def sketch_recognition(img):
```

```
    # Implement sketch recognition model here...
```

```
    # Return labels and confidences as dictionary
```

```
iface = gr.Interface(fn=sketch_recognition, inputs="sketchpad", outputs="label").launch()
```



Who wants to play Pictionary? Draw a common object like a shovel or a laptop, and the algorithm will guess in real time!



Clear

Submit

OUTPUT

0.0s

**umbrella**

umbrella 34%

helmet 34%

mushroom 26%

dumbbell 2%

table 1%



# PAPERS

읽을때 & 쓸때 & 사용할때





## 논문 읽는 법 (의견)

단어 하나하나 다 읽지 않는다!

1. Abstract 부터 시작
2. Conclusion 먼저- 어떤 주장?
3. 표/사진 보고, 어떤 dataset/s 썼는지
4. 어떤 방법 (un/semi/supervised)
5. Code-Data 공개인지
6. 기관/저자 (첫-끝)

New paper 따라가기...



## 논문 쓸때



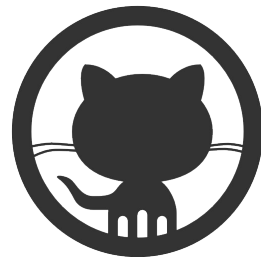
- 같은 주제의 Survey 논문 읽고 시작
- Citation tree
- 옵션이 아닌 필수!
  - 예뻐야 됨 (graph가독성, picture quality, alignment)
  - Github + [repo citation기능](#)
  - 데모
  - 홍보!
- 기억에 남을만한 약자



Papers With Code



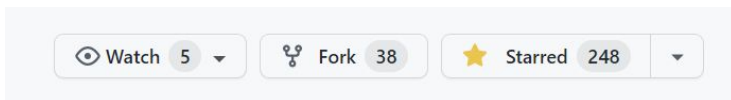
# 논문 사용 할 때



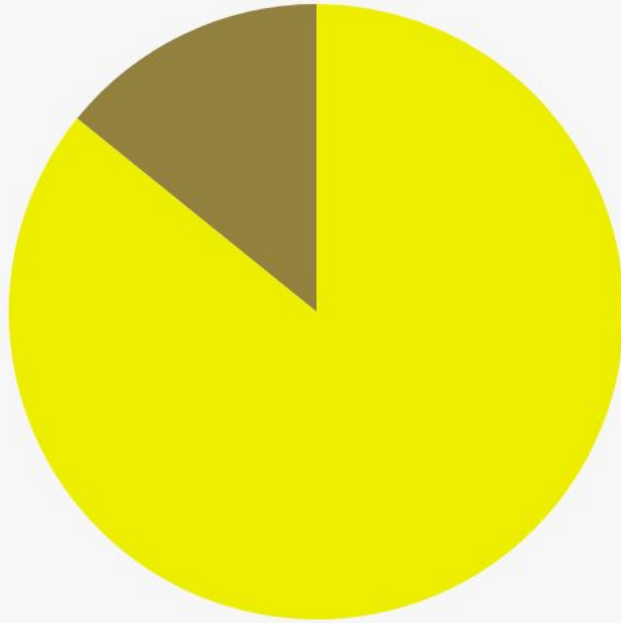
Official Repo인지 Check

제가 좋아하는 framework(pytorch/tensorflow 등등) 보전 찾기

- Star개수
- Issue개수
- Latest update / created 날짜
- License check
- Library/Python version (fork이 있나)



■ meme만든 시간  
■ 발표 내용 준비 시간



# 일에 재미를 +덤

## CLIP React Demo

[Model Card](#)



Researching fun

Motivation

Reaction GIFs became an integral part of communication. They convey complex emotions with many levels, in a short compact format.

If a picture is worth a thousand words then a GIF is worth more. A lot of people would agree it is not always easy to

### 1. 🏆 Winning Projects and Mentions 🏆

We want to congratulate the top 3 projects selected by the jury:

- **First Place:** DALL-E mini
- **Second Place:** CLIP+NeRF: Fewshot Learning, Putting NeRF on a Diet
- **Third Place:** Fine-tune CLIP on satellite images+captions

The jury was quite impressed with the projects, so there are a couple of additional special nominees they would like to recognize as well:

- BERTIN: PreTrain RoBERTa-large from scratch in Spanish
- CLIP like contrastive vision-language models for Italian with pre-trained text and vision models
- **Generate GIF reply to English text with VQGAN + CLIP** my project
- Sentence Embeddings

And finally the jury gave an **honorary mention** to the Chef Transformer (Recipe Generation Model). You can find all comments from the jury in this [document](#) and find all 15 top projects [here](#). We'll follow up with the teams with next steps.

### 2. Results

This has been the largest Hugging Face event, and we're extremely excited by the results. Almost 800 members joined Slack, people were very active in Discord as well, and had **almost 100 projects, 170 models and 36 Spaces!** 🤖 That is super impressive given the timeframes of the event!

## Search Reaction GIFs with CLIP

Example Queries: ?

- OMG that is disgusting
- I'm so scared right now
- I got the job 🎉
- Congratulations to all the flax-community week teams
- You're awesome
- I love you ❤️

Write text you want to get reaction for

You're awesome

Found these images within validation set:

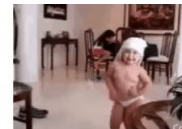
Play GIFs ?



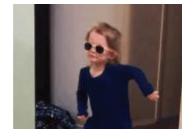
0.8692463636398315



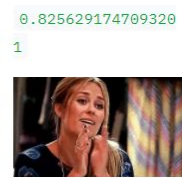
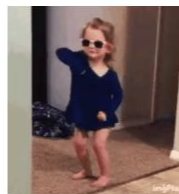
0.829218864440918



0.825629174709320



0.822228193283081



Searches among the validation set images if not specified

(There may be non-exact duplicates)

(optional) Upload images to rank them



Drag and drop files here

Limit 200MB per file • JPG, JPEG, GIF

Browse files

<https://huggingface.co/spaces/flax-community/clip-reply-demo>



# Summary & Advice

- Use Open Source
- Contribute
- Don't be afraid to try new things
- 석사 is discovery time
  - 자신의 workflow
  - 자신의 발표 style
  - Academy vs work
  - 대기업 vs 스타트업

## Me listening to my own advice



@cceyda



@ceyda\_cinarel



<https://namecard.kakao.com/ceyda>



Have a nice week 🍷💫

