Designing for Data Rights in the AI Production Pipeline

Facilitators:
Jennifer Ding, The Alan Turing Institute
Anne Lee Steele, The Alan Turing Institute
Yacine Jernite, Hugging Face

To continue the conversation after today's session, join *The Turing Way* Slack or a future community call!
https://the-turing-way.start.page/

⚡Brain Freeze?⚡

Here are a few resources to get you started

Anything you'd like to share on this topic with the group?

Drop it here!

www.adalovelaceinstitute.org

Working with the CARE principles: operationalising Indigenous data governance

Shifting the focus of data governance from consultation to values-based relationships to promote equitable Indigenous participation in data processes.

blog.papareo.nz

OpenAI's Whisper is another case study in Colonisation

Whisper, a speech recognition model trained on 680,000 hours of audio taken from the web, demonstrates the problems with Big Tech and data theft.

dl.acm.org

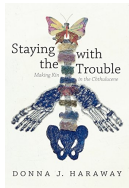
On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible...

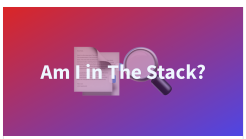
datasociety.net

A Primer on AI in/from the Majority World

A resource by Sareeta Amrute, Ranjit Singh, and Rigoberto Lara Guzmán exploring the presence of artificial intelligence and technology in the Majority World. 160 thematic works, available in English and Spanish, curated in collaboration with Aishatu Gwa...



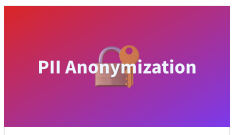
twitter.com



huggingface.co

Am I in The Stack? - a Hugging Face Space by bigcode

Discover amazing ML apps made by the community



huggingface.co

PII Anonymization - a Hugging Face Space by bigcode

Discover amazing ML apps made by the community

arxiv.org

Data Governance in the Age of Large-Scale Data-Driven Language Technology

The recent emergence and adoption of Machine Learning technology, and specifically of Large Language Models, has drawn attention to the need for systematic and transparent management of language data. This work proposes an approach to global language da...

the-turing-way.netlify.app

Case Studies: Choosing an ML License

Below are two hypothetical case studies based on concerns raised through the license selection process during BigScience, a one-year long research workshop on large multilingual models and datasets. Cidney, a hypothetical ML researcher, is working for a...

www.licenses.ai

The BigScience OpenRAIL-M License - Responsible AI Licenses (RAIL)

To help the AI community reuse the BigScience BLOOM RAIL License more broadly for distributing models, we adapted the terms to make the license applicable to any associated AI Model.

deliverypdf.ssrn.com

The Steep Cost of Capture

6 Pages Posted: 20 Jun 2022 Date Written: 2021 In considering how to tackle this onslaught of industrial AI, we must first recognize that the "advances" in AI celebrated over the past decade were not due to fundamental scientific breakthroughs in AI tec...

agenda

welcome & intros	00 — 05
activity 1: who & what of ai production	05 — 15
the BigCode approach	15 — 25
activity 2: re-imagining the pipeline	25 — 50
closing	50 — 59

ground rules

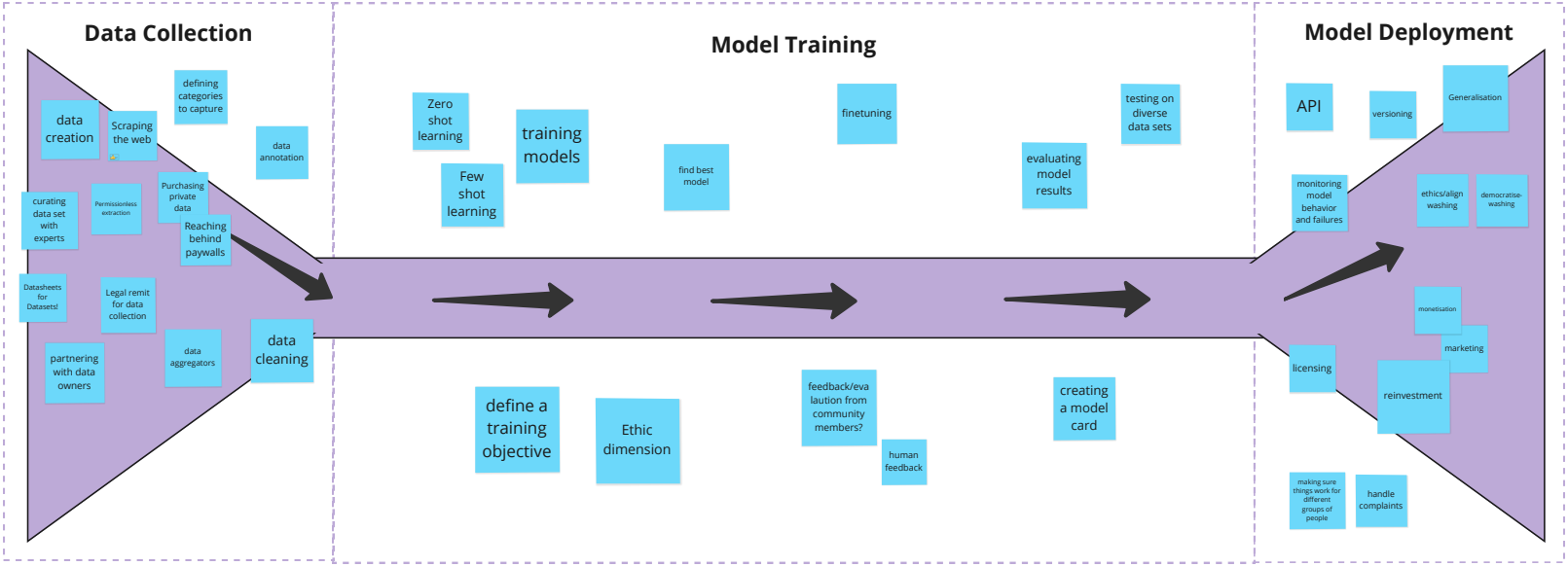
1. Treat yourselves and each other with respect and kindness, per Mozilla's Community Participation Guidelines
2. Please raise your hand or use the chat for any questions or comments you have during the presentation
3. Have fun! 🤗

welcome!

Double click on a sticky and share with the group **something about yourself** (e.g. name, pronouns, location)

What kinds of activities are a part of the AI production pipeline?

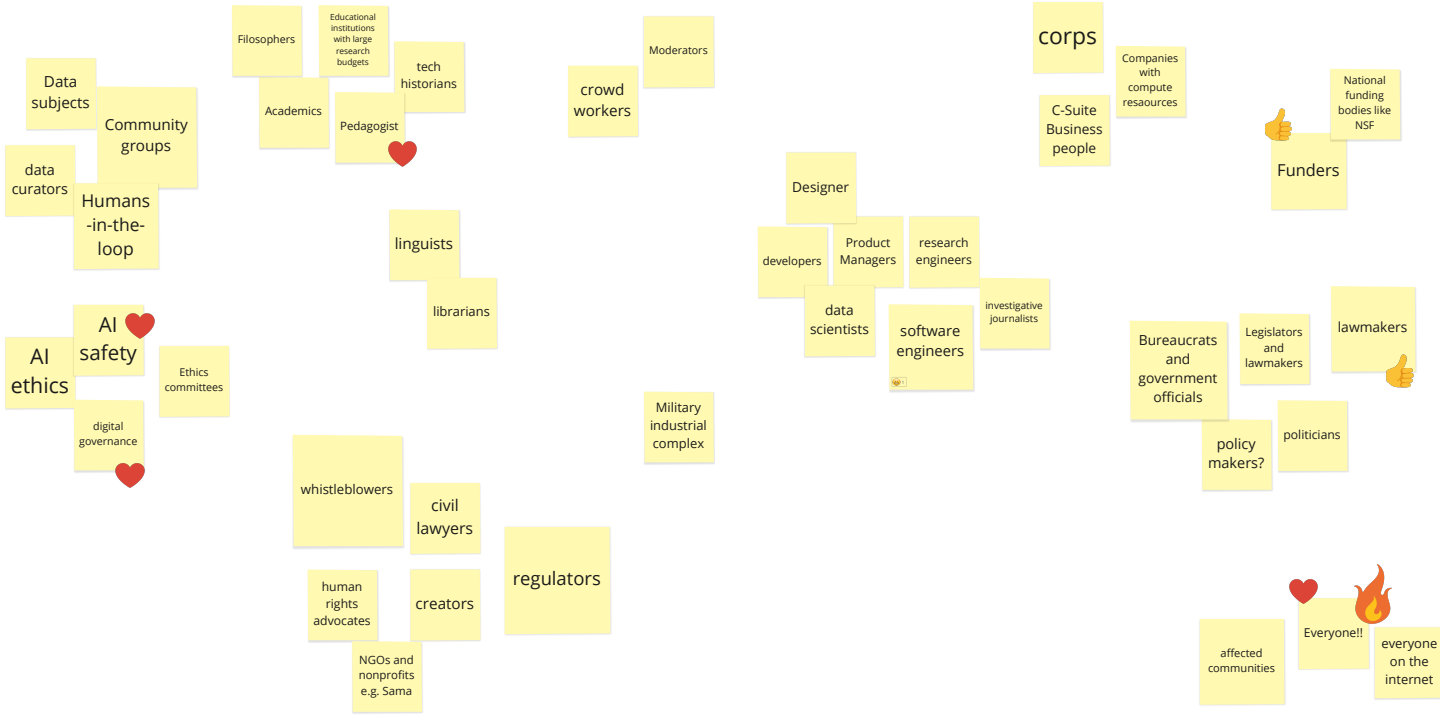
ai production pipeline



Drag and duplicate if you like what someone else wrote

Who is part of or can shape the AI production pipeline?

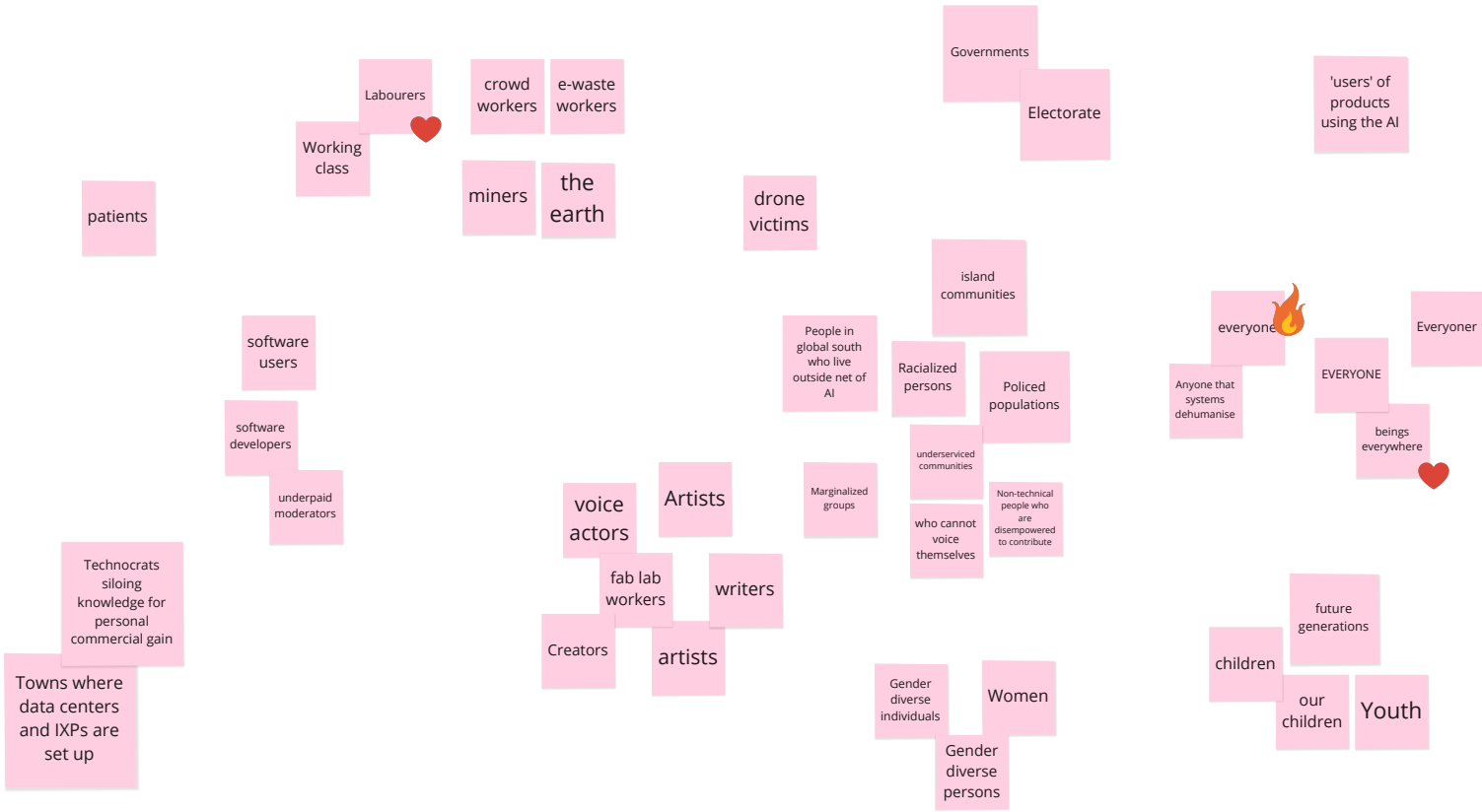
who



Drag and duplicate if you like what someone else wrote

Who is impacted by the AI production pipeline?

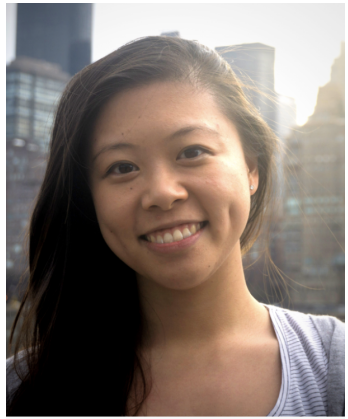
who



Drag and duplicate if you like what someone else wrote

intros

Who are we?



Position
Research Application Manager

Position
Community Manager, The Turing Way

Position
Research Scientist, Hugging Face

resources

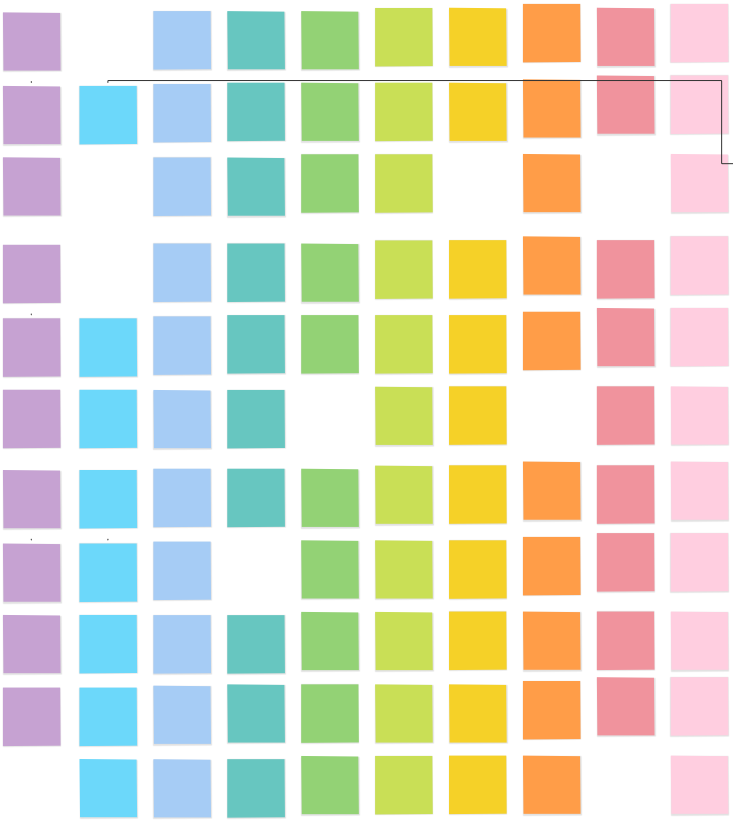
Other Resources

Padlet: https://pad.sfconservancy.org/p/ttw-mozfest-data-pipeline

Session info: https://schedule.mozillafestival.org/session/KAS9YF-1

Sticky-landia

Grab a sticky if you need one!



BigCode

Hugging Face servicenow

Open and responsible research & development of large language models for code

What is a language model?

My ___ → name
My name ___ → is
My name is ___ → Loubna
My name is Loubna ___ → and
My name is Loubna and ___ → I
My name is Loubna and I ___ → work
My name is Loubna and I work ___ → at
My name is Loubna and I work at ___ → Hugging Face

The Stack

A dataset with **6.4TB** of **permissively licensed** code in **358 programming languages** with an **Opt-Out** mechanism

Other Data Governance efforts:

PII redaction

- Removal of Emails, Keys and IP addresses with regexes

- Next pipeline: PII dataset annotation + train NER model on 7 PII entities for code.

Model license

- CodeML OpenRAIL-M v0.1

BigCode: Open collaboration

We are building LLMs for code in a collaborative way:

- 500+ participants
- 30+ countries

Language models for code generation

Sequence to sequence, encoders, and **decoders**.

```
def add_numbers(a, b):
    """add two numbers"""
    return
```

→ Decoder →

```
def add_numbers(a, b):
    """add two numbers"""
    return a + b
```

The Stack: Am I In The Stack

https://huggingface.co/spaces/bigcode/in-the-stack

definitions

ai production: process of building an ML model

ai pipeline: steps that form ai production, along which data moves and takes different forms

data rights: rights afforded to an individual or community relating to data they create or data that is about them

BigCode Pipeline

case study: BigCode

Data Collection

Procuring and storing dataset

Where does data come from?

How does this work in a scraping dataset process?

Processing and filtering data

How do we get consent from our people's data?

Creating a public data sheet for the dataset

Model Training

Who is the public?

And how can non-technical populations give access to the process?

Defining the model architecture & training objectives

Obtaining compute infrastructure to train the model

How are the models evaluated?

Model evaluation

Create model demo for user access

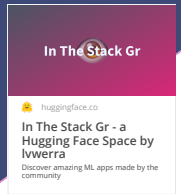
Model integration into product or platform for user access

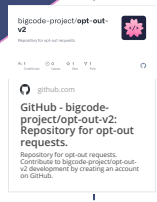
Model Deployment

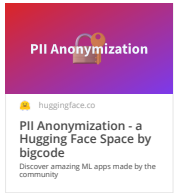
Create public model card

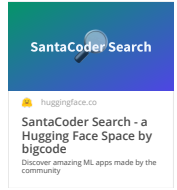
Monitoring real world use cases

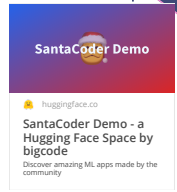
Assessing compliance with model license

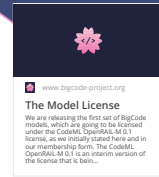








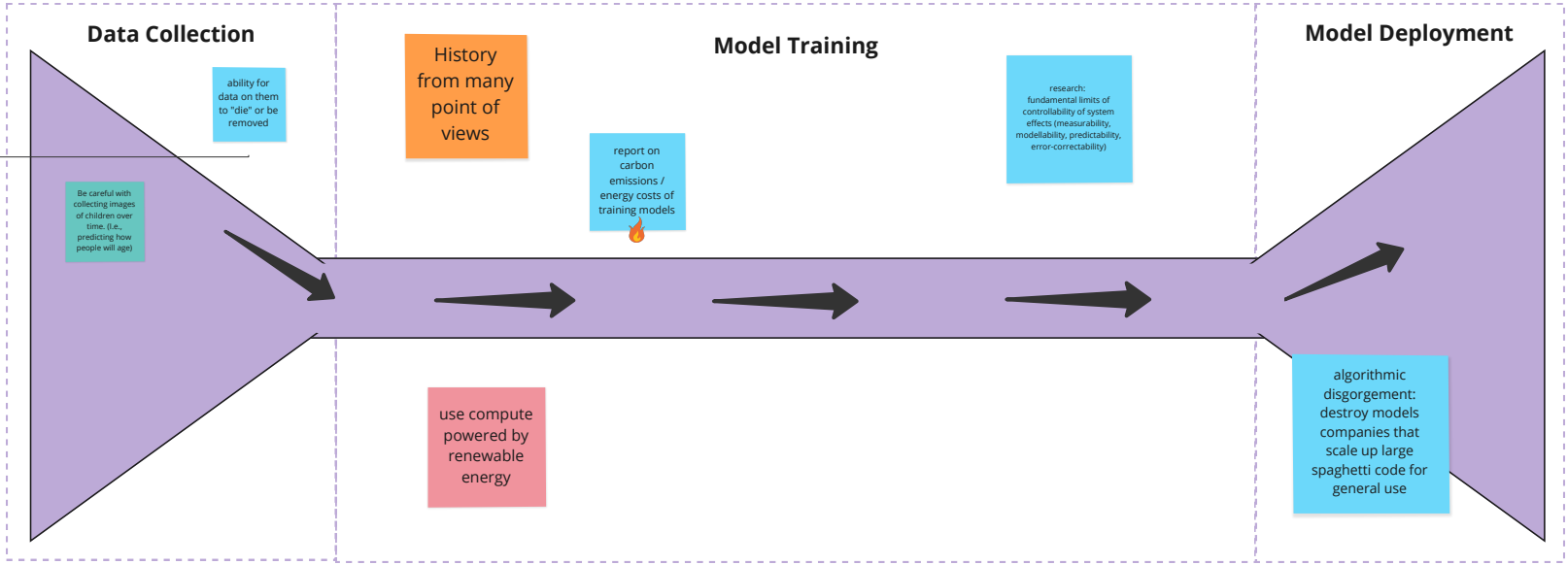




What questions, comments, or concerns do you have about the BigCode Pipeline?

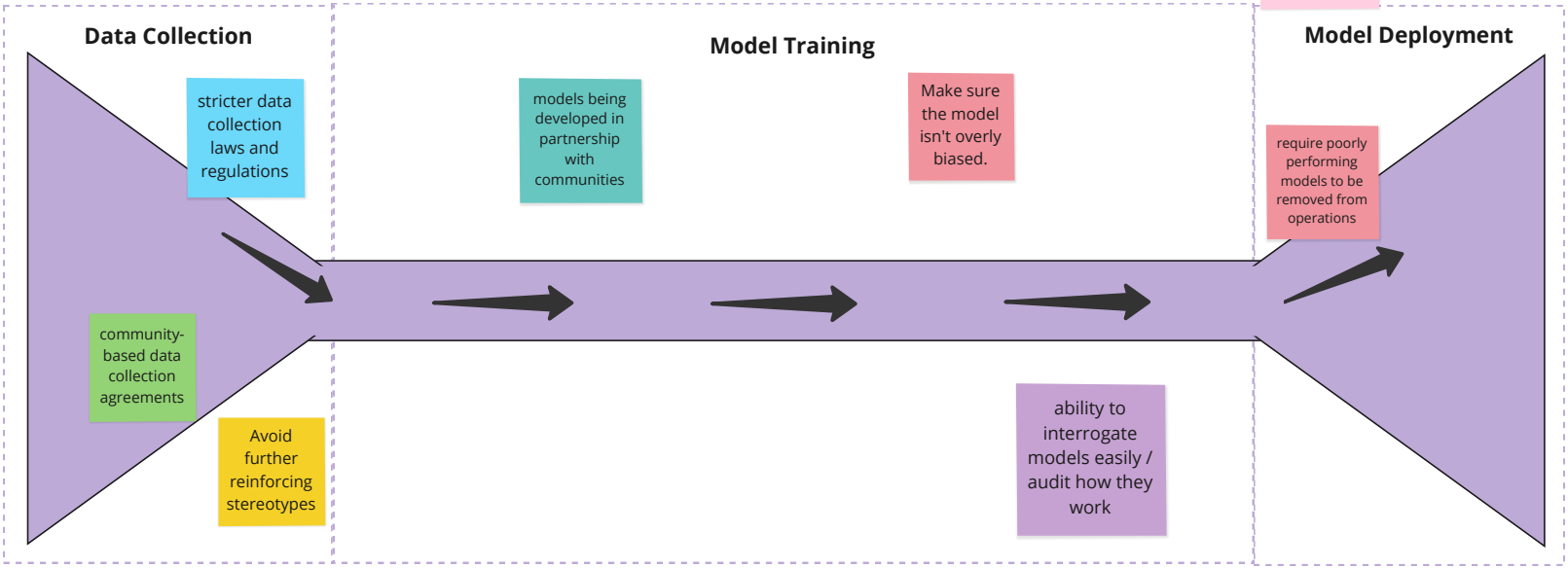
Re-imagining the AI Production Pipeline

What kinds of activities *should be* a part of the AI production pipeline for/with Future generations ?



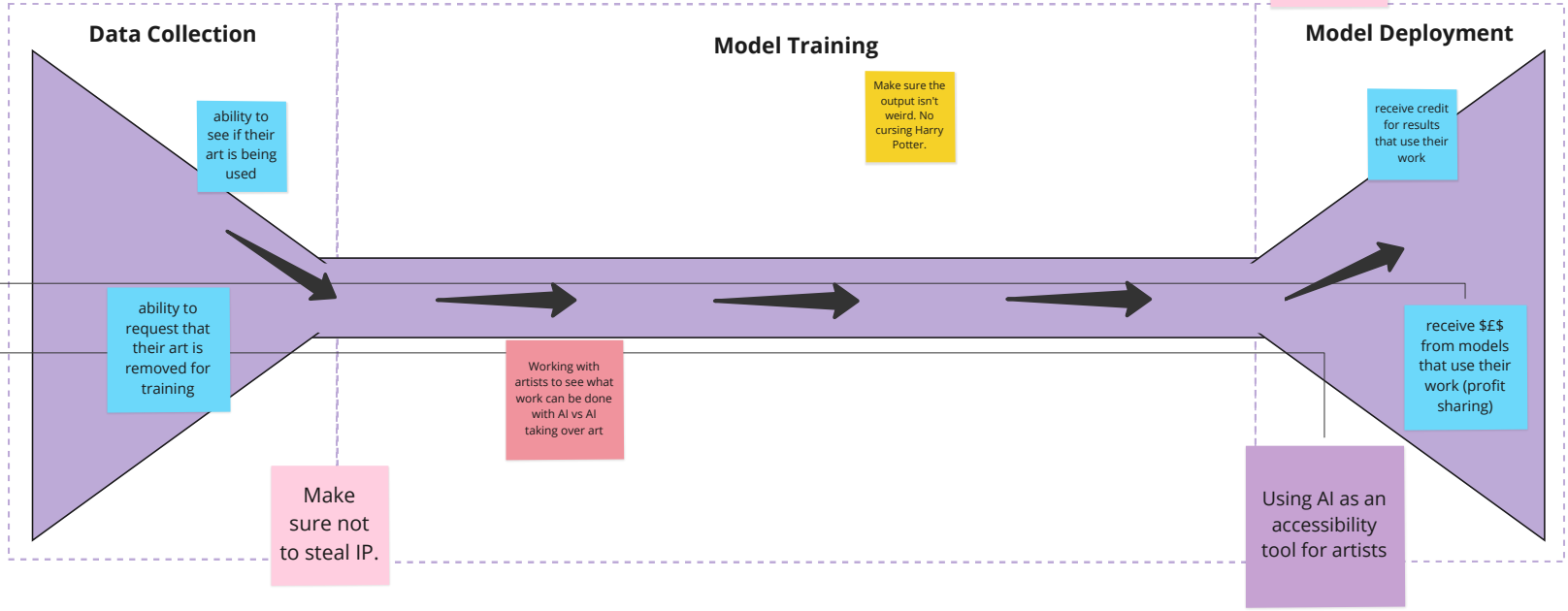
Re-imagining the AI Production Pipeline

What kinds of activities *should be* a part of the AI production pipeline for/with Policed populations ?



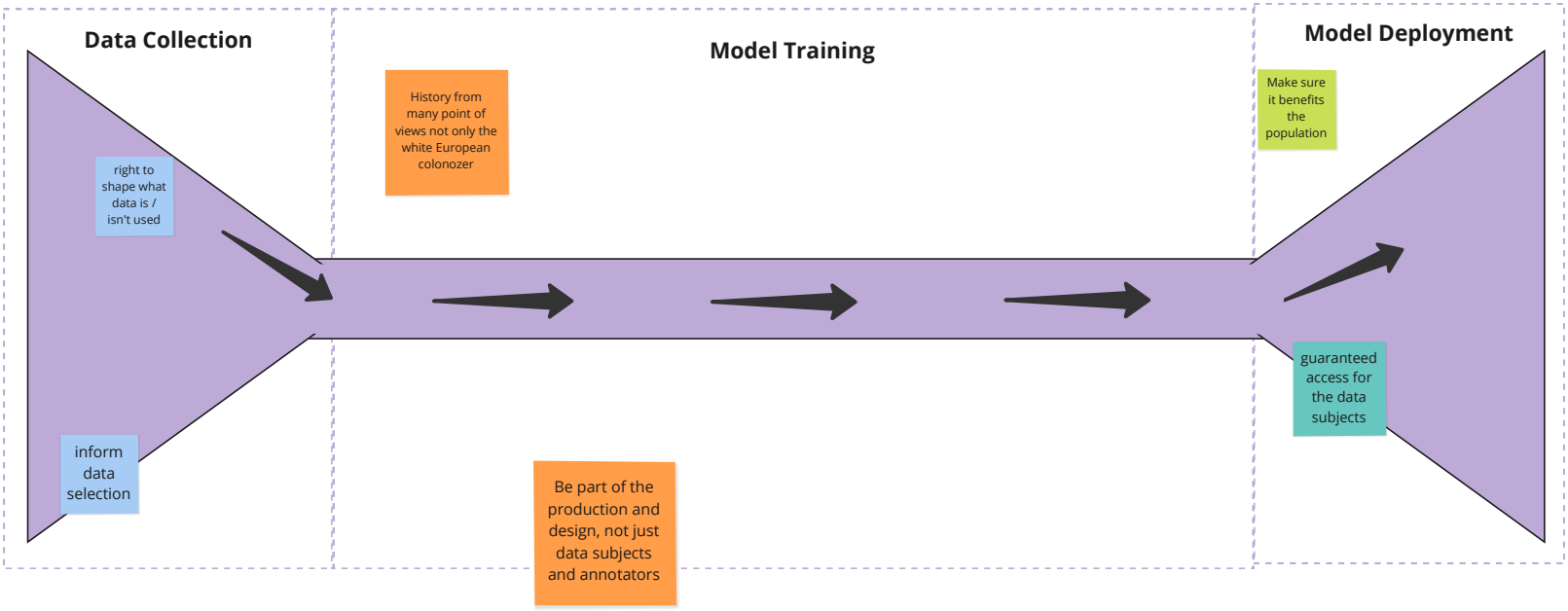
Re-imagining the AI Production Pipeline

What kinds of activities *should be* a part of the AI production pipeline for/with Artists ?



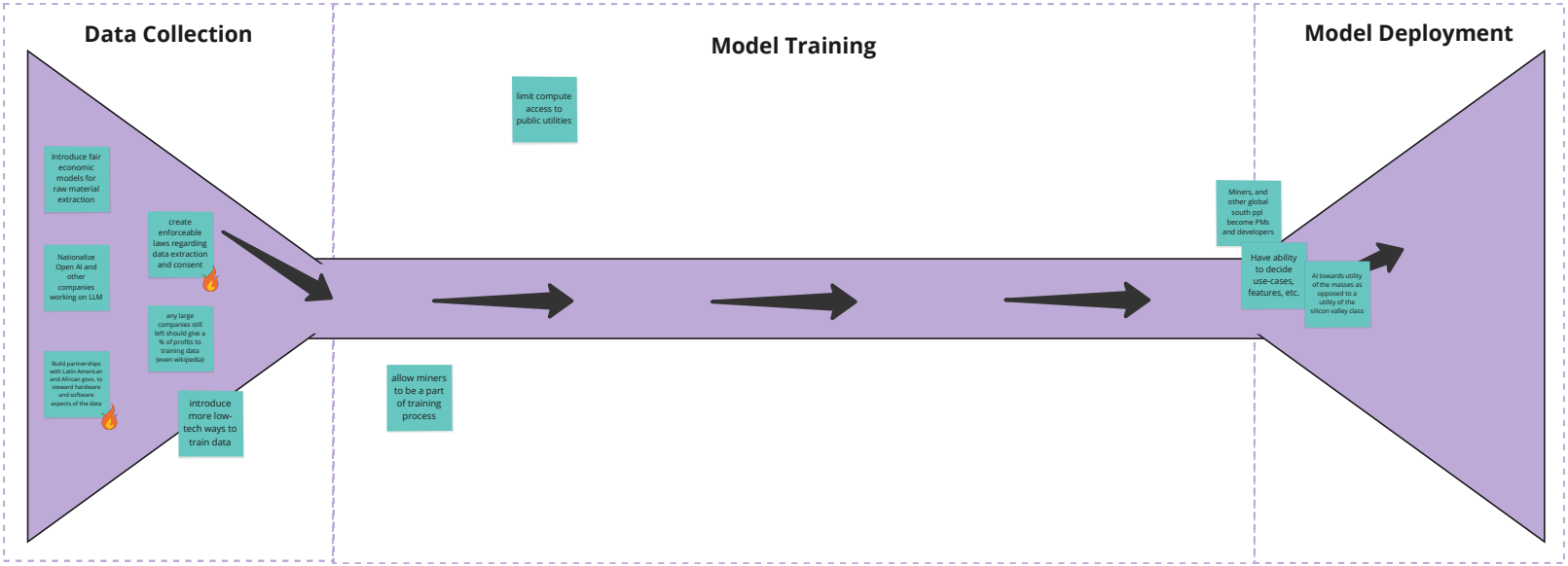
Re-imagining the AI Production Pipeline

What kinds of activities *should be* a part of the AI production pipeline for/with People in global south who live outside net of AI ?

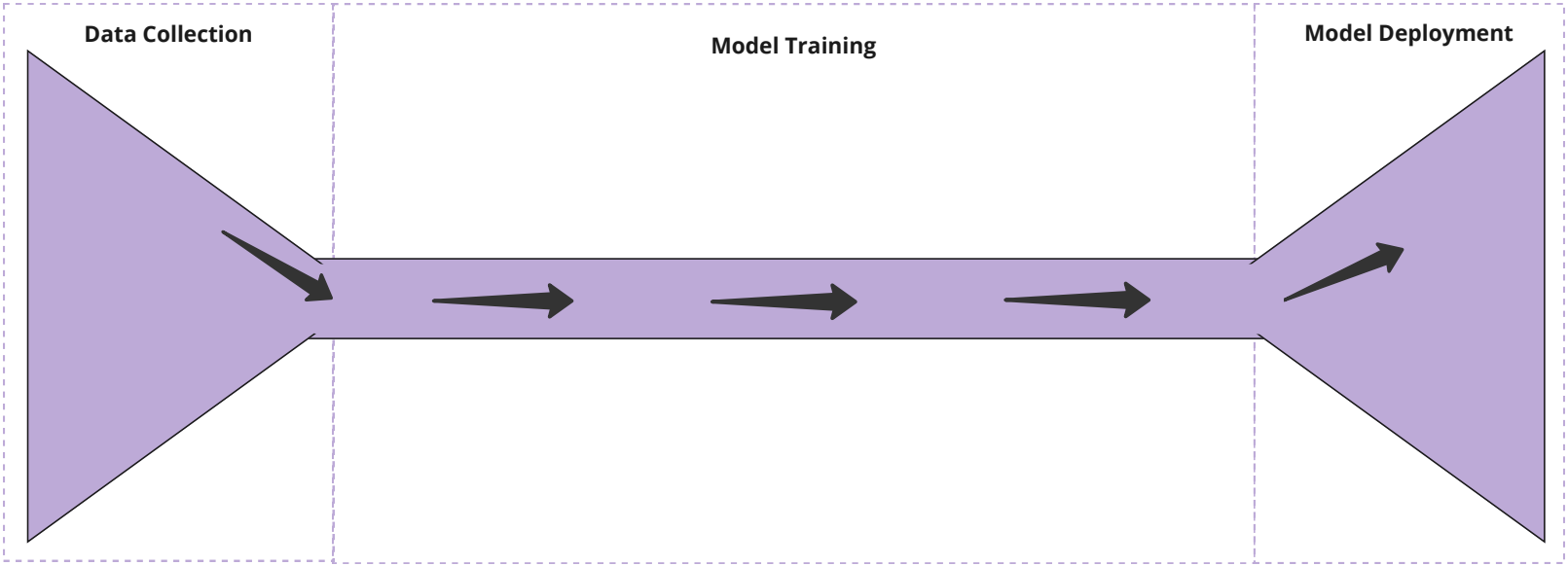


Re-imagining the AI Production Pipeline

What kinds of activities *should be* a part of the AI production pipeline for/with miners ?



What kinds of activities *should be* a part of the AI production pipeline for/with XXX **?**

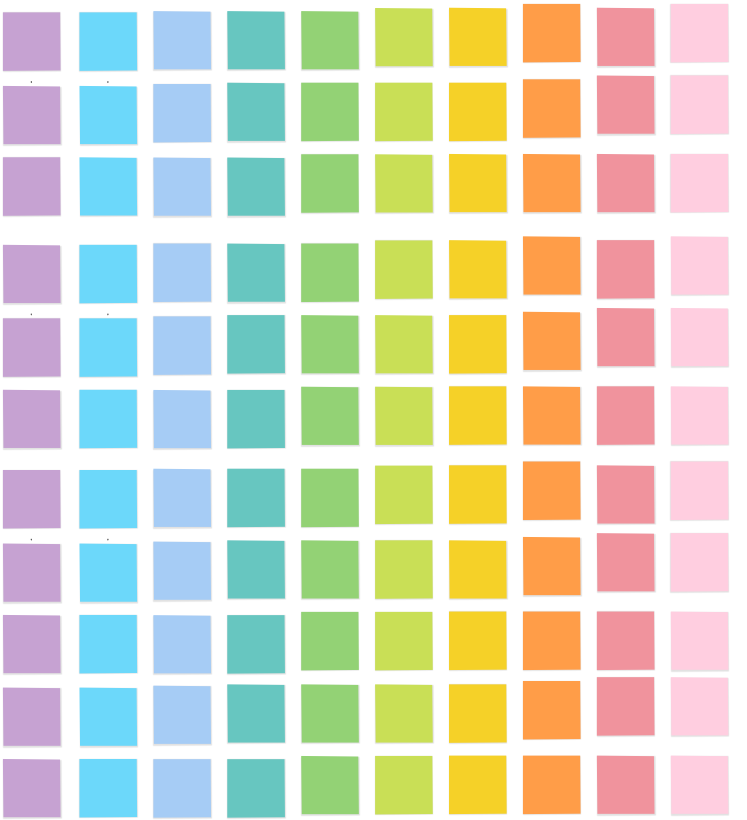


Sticky-landia

Grab a sticky if you need one!

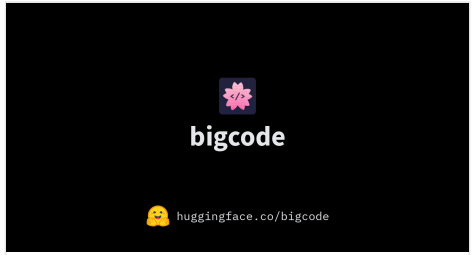
Sticky-landia

Grab a sticky if you need one!



final announcements

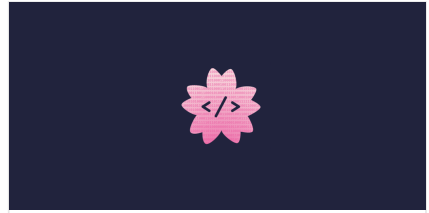
bigcode



huggingface.co

bigcode (BigCode)

We're on a journey to advance and democratize artificial intelligence through open source and open science.



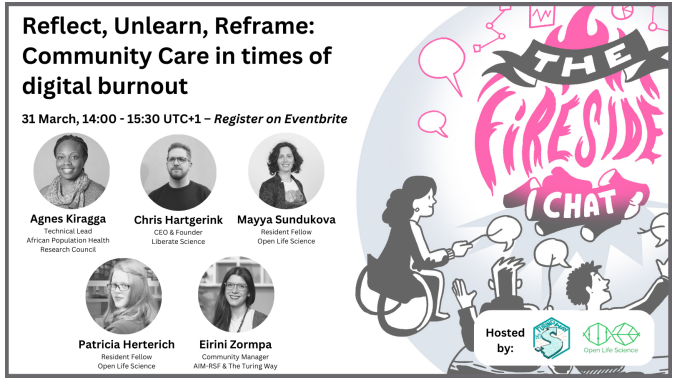
www.bigcode-project.org

Open and responsible development of LLMs for code

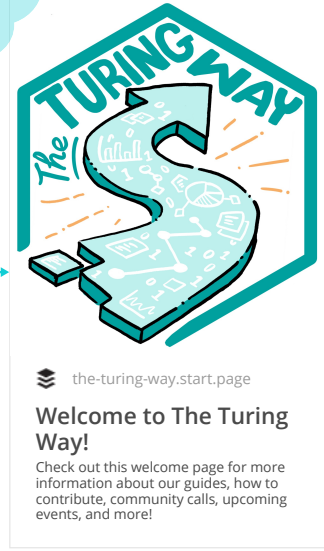
BigCode is an open scientific collaboration working on the responsible development of large language models for code

final announcements

the turing way



Click here to see links and & resources about the project



Join us at our upcoming Fireside Chat!

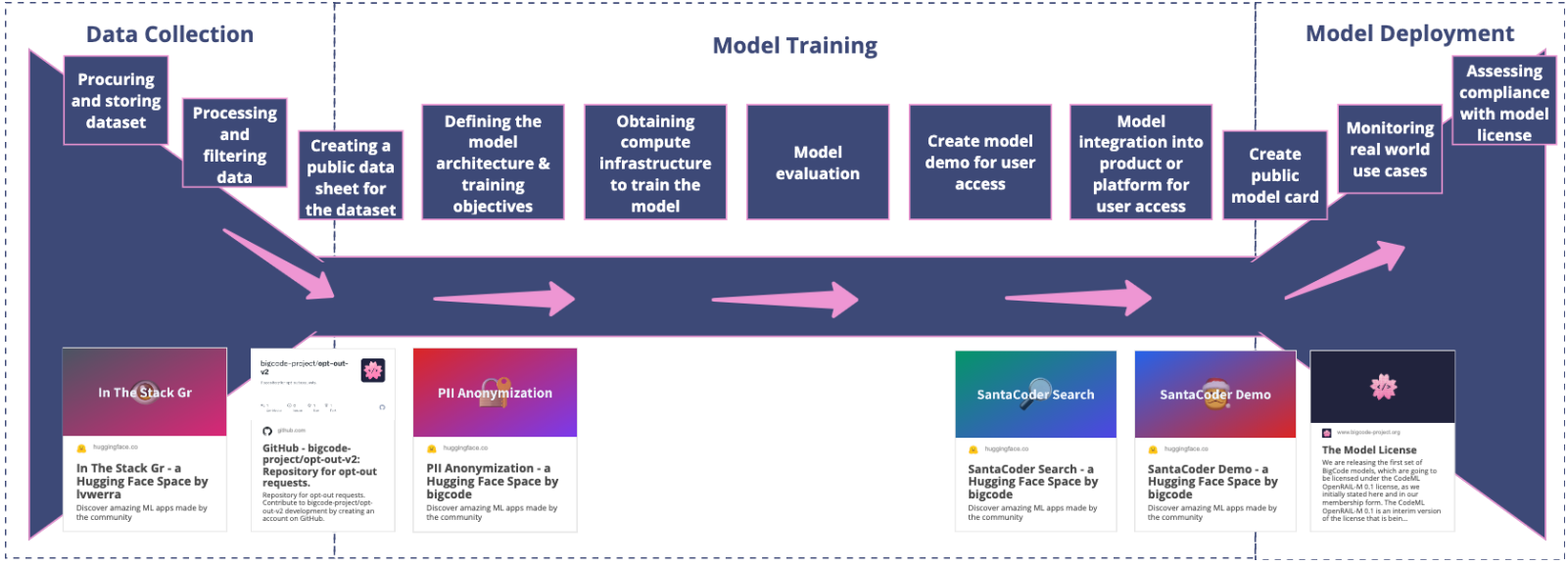
...Or at a Collaboration Cafe!

keep in touch

Designing for Data Rights in the AI Production Pipeline

Facilitators:
Jennifer Ding, The Alan Turing Institute jding@turing.ac.uk
Anne Lee Steele, The Alan Turing Institute asteele@turing.ac.uk
Yacine Jernite, Hugging Face yacine@huggingface.co

To continue the conversation after today's session, join *The Turing Way* Slack or a future community call!
https://the-turing-way.start.page/


Data Collection
Model Training
Model Deployment
Procuring and storing dataset
Processing and filtering data
Creating a public data sheet for the dataset
Defining the model architecture & training objectives
Obtaining compute infrastructure to train the model
Model evaluation
Create model demo for user access
Model integration into product or platform for user access
Create public model card
Monitoring real world use cases
Assessing compliance with model license
In The Stack Gr
huggingface.co
In The Stack Gr - a Hugging Face Space by lvwerra
Discover amazing ML apps made by the community
bigcode-project/opt-out-v2
github.com
GitHub - bigcode-project/opt-out-v2: Repository for opt-out requests.
Repository for opt-out requests. Contribute to bigcode-project/opt-out-v2 development by creating an account on GitHub.
PII Anonymization
huggingface.co
PII Anonymization - a Hugging Face Space by bigcode
Discover amazing ML apps made by the community
SantaCoder Search
huggingface.co
SantaCoder Search - a Hugging Face Space by bigcode
Discover amazing ML apps made by the community
SantaCoder Demo
huggingface.co
SantaCoder Demo - a Hugging Face Space by bigcode
Discover amazing ML apps made by the community
www.bigcode-project.org
The Model License
We are releasing the first set of BigCode models, which are going to be licensed under the CodeML OpenRAIL-M 0.1 license, as we initially stated here and in our membership form. The CodeML OpenRAIL-M 0.1 is an interim version of the license that is bein...