UNITED STATES
PATENT AND TRADEMARK OFFICE

uspto

Recent Developments in AI and USPTO Open Data

PatentSemTech Workshop
2022 July 15

UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

USPTO Open Data

Hundreds of terabytes spanning two centuries of scientific, technical, and commercial data.

Patents: 11M+

Trademarks: 5M+

- Applications
- Prosecution history
- Appeals
- Contested matters
- IP assignments
- Artifact deposits

uspto

Open Data as fuel for AI

- USPTO Open Data is useful for:
  - Patent analytics
  - Economics
  - Commercial tools for practitioners
- But *also* serves as a substrate for AI:
  1. AI and NLP techniques ➔ the patent domain
  2. Patent data ➔ frontiers of AI and NLP research

uspto

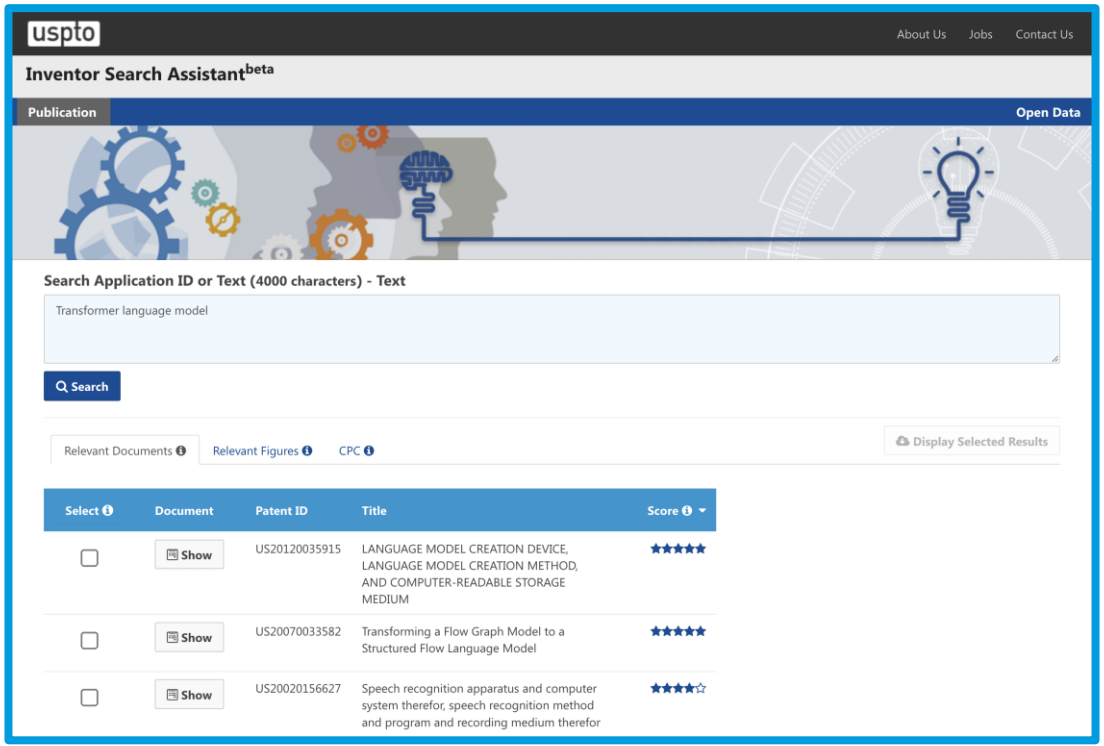
AI and NLP techniques for the patent domain

AI & NLP for IP administration

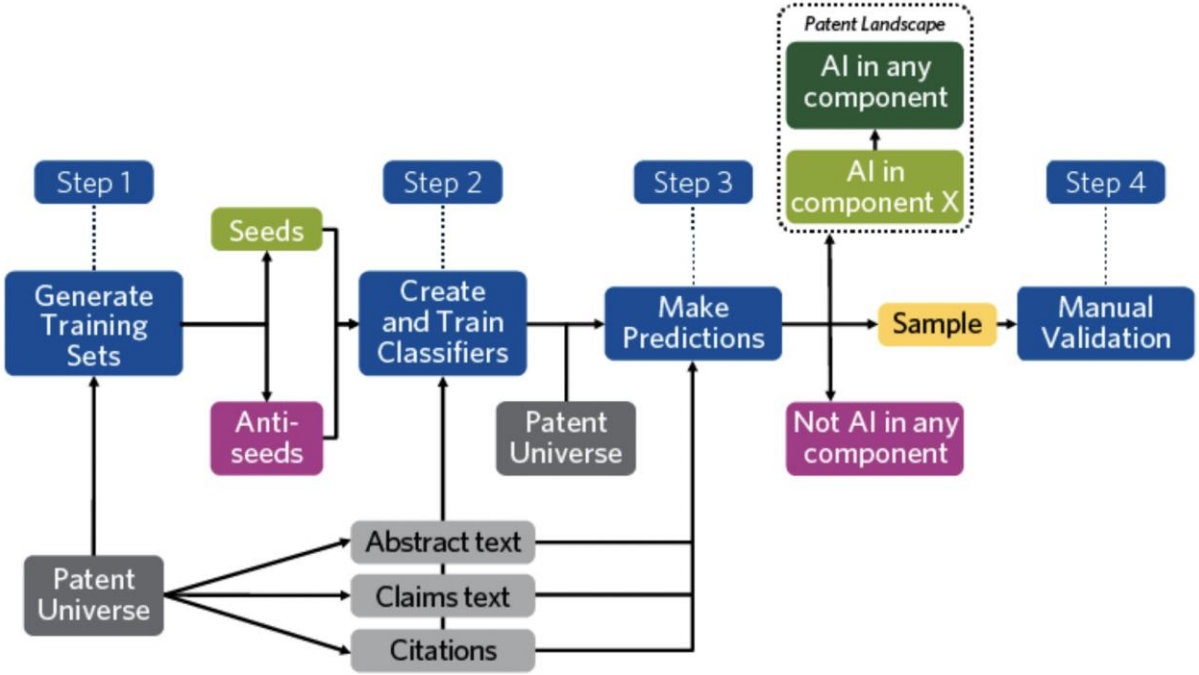
- Prior art search
  - From keywords to representation learning
- Patent classification
  - From claim indicators to full autoclassification

uspto

AI & NLP tools for IP practitioners & inventors



AI-powered empirical research & analytics



Patent data toward advancing SotA in AI research

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Recent work has demonstrated that **increased training dataset diversity improves general cross-domain knowledge** and downstream generalization capability for large-scale language models. With this in mind, we present The Pile: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from **22 diverse high-quality subsets** -- both existing and newly constructed -- many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, **models trained on the Pile improve significantly** over both Raw CC and CC-100 on all components of the Pile, while **improving performance on downstream evaluations**. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.

Abstract of Gao et al. 2020

uspto

The Pile: impact

- Large AI language models trained on The Pile ➔ increased data diversity.
- New research customers of USPTO data via The Pile:
  - UC Berkeley
  - UT Austin
  - Oxford
  - Google
  - DeepMind
  - Microsoft

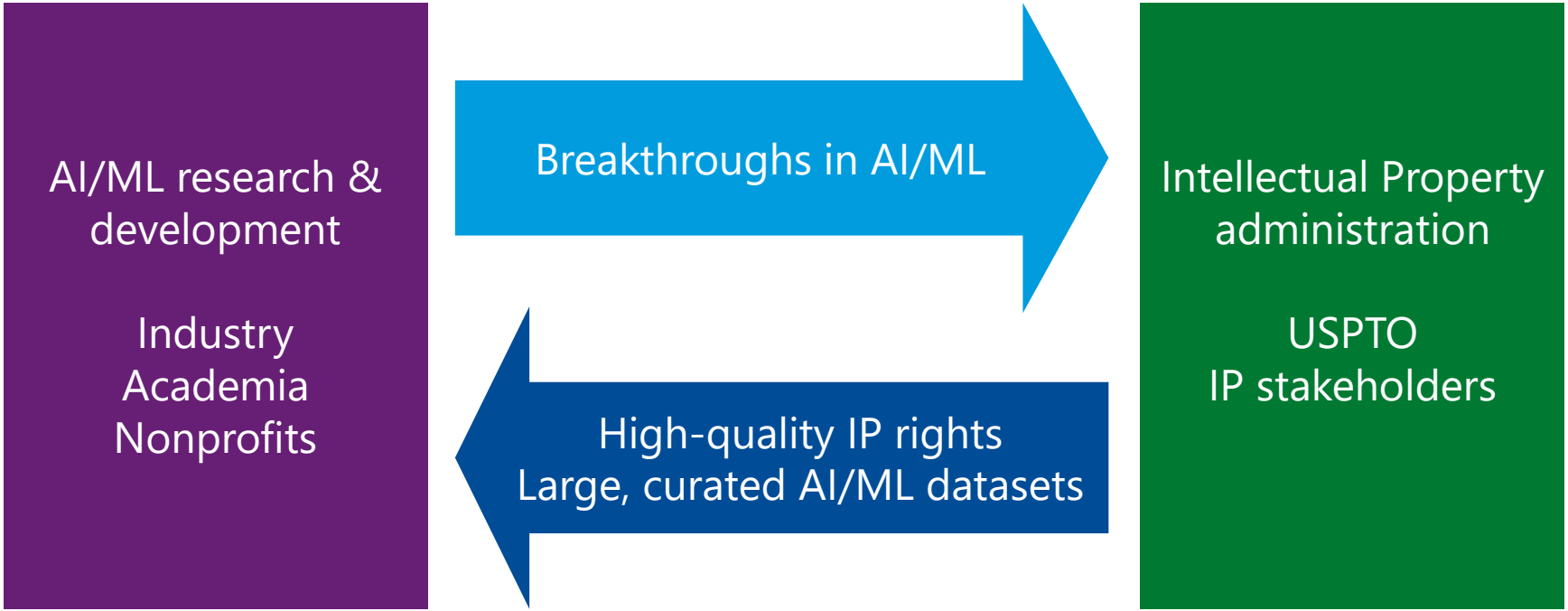
uspto

Patent-sourced research benchmarks

- Public USPTO data + CPC annotations ➔ text classification benchmarks.
  - Now used in evaluation of general-purpose large language models.
- Public USPTO data + third-party annotations ➔ specialized benchmark datasets.
  - Test the ability of novel AI and NLP models to penetrate complex technical concepts.
  - Semantic similarity—Aslanyan & Wetherbee 2022.
  - Kaggle contest: 2.3K participants, 42K code submissions, 13 gold medalist teams from 9 countries.

uspto

AI & IP: a collaborative data ecosystem



Thank you!

Scott Beliveau

Scott.Beliveau@uspto.gov

Jerry Ma

Jerry.Ma@uspto.gov

www.uspto.gov