

Contracting for Transparency: Artificial Intelligence and the Need for a Contractual Commons

Erik Stallman and Sonia Katyal

Intellectual Property Scholars Conference
August 9, 2019

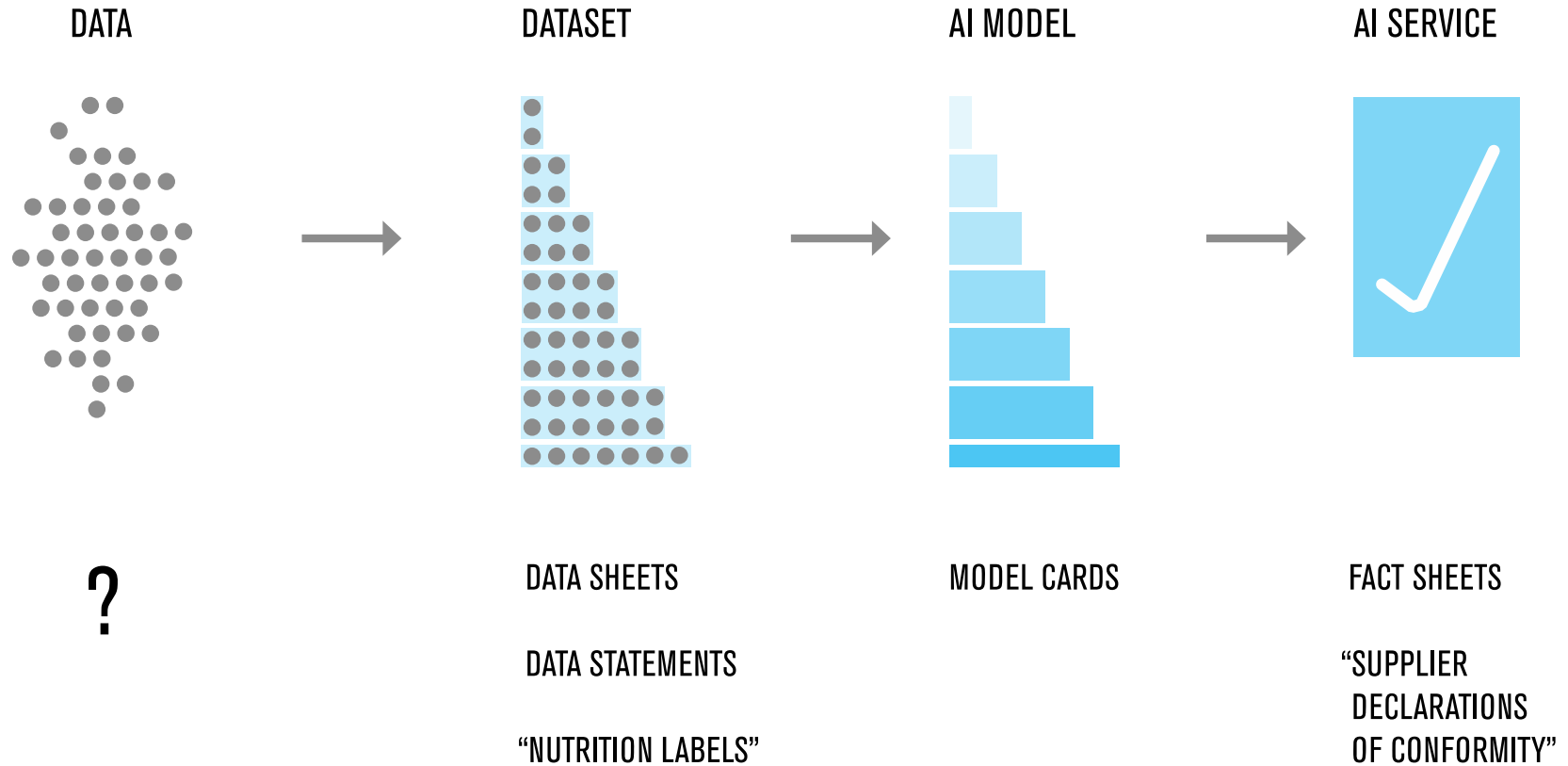


Problem statement.

Assertions of proprietary interests in the data and code used in AI systems are impeding transparency and accountability just as new model disclosures and accountability tools are emerging.

Can proprietary interests in datasets flip this equation, enabling terms that reinforce and build on model disclosures? Can this be accomplished without unduly encumbering shared resources and open innovation?

Model disclosures for components in an AI system.



Licensing regimes are not intended to address model disclosures or data used in AI generally.

IBM

IBM Research Blog Topics Labs About

AI

IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems

THE VERGE

TECH

REVIEWS

SCIENCE

CREATORS

ENTERTAINMENT

VIDEO

MORE



TECH \ IBM \ ARTIFICIAL INTELLIGENCE \

IBM didn't inform people when it used their Flickr photos for facial recognition training

Food for the algorithms

By Shannon Liao | @Shannon_Liao | Mar 12, 2019, 7:14pm EDT



SHARE

[Creative Commons](#) > [Blog](#) > [Legal tools / licenses](#) > [Use and Fair Use: Statement on shared images in facial recognition AI](#)

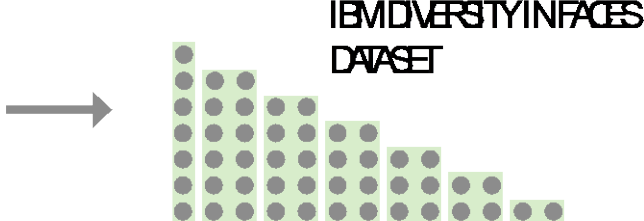
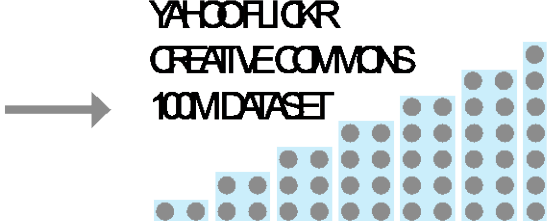
Use and Fair Use: Statement on shared images in facial recognition AI



Ryan Merkle

March 13, 2019

The chain of licenses and terms of use in the DiF dataset.



Data Sharing Agreement

Please read and agree to the Data Sharing Agreement below.

YAHOO FLICKR CREATIVE COMMONS 100M DATASET
TERMS OF USE

1. PARTIES & AGREEMENT TO TERMS

1.1. Yahoo! Inc. ("Yahoo") provides you access to and use of the Yahoo Flickr Creative Commons 100M dataset made available from the Yahoo website located at <http://webscope.sandbox.yahoo.com> ("Dataset") subject to the terms and conditions of this Terms of Use ("TOU"). The term "Dataset" includes the selection and arrangement of information along with the data found at the URL above.

1.2. This TOU is a binding agreement between Yahoo and you and, if applicable, the

I have read and agree to the Data Sharing Agreement.

IBM RESEARCH DIF DATASET TERMS OF USE

(updated April 02, 2019)

PART 1 - General Terms

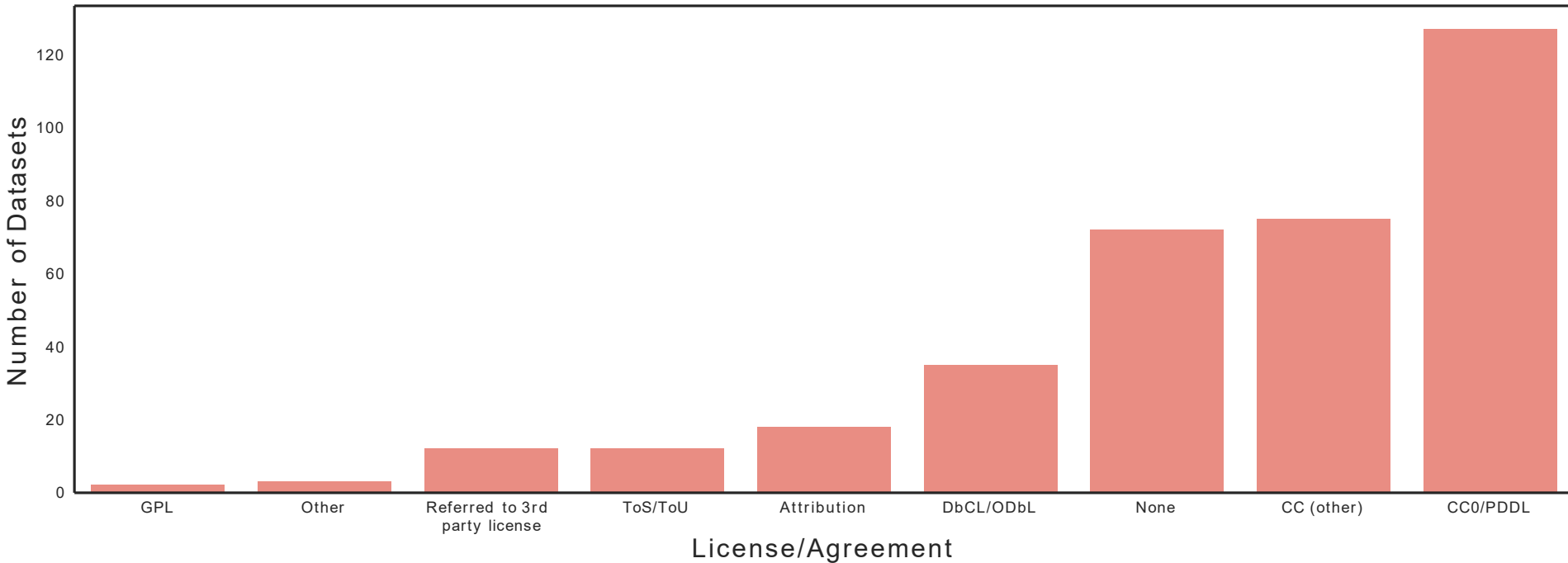
BY DOWNLOADING, INSTALLING, COPYING, ACCESSING, CLICKING ON THE "ACCEPT" BUTTON, OR OTHERWISE USING THE WEBSITE, YOU AS "LICENSEE" AGREE TO THESE TERMS OF USE AND COVENANT THAT YOU, AS LICENSEE, ARE EIGHTEEN (18) YEARS OR OLDER. IF YOU ARE ACCEPTING THESE TERMS ON BEHALF OF LICENSEE, YOU REPRESENT AND WARRANT THAT YOU HAVE FULL AUTHORITY TO BIND LICENSEE TO THESE TERMS OF USE. IF YOU DO NOT AGREE TO THESE TERMS, DO NOT DOWNLOAD, INSTALL, COPY, ACCESS, CLICK ON THE "ACCEPT" BUTTON, OR USE THE WEBSITE.

* ACCESS TO THE IBM DIF DATASET IS NOT AVAILABLE TO ANYONE UNDER THE AGE OF EIGHTEEN (18).

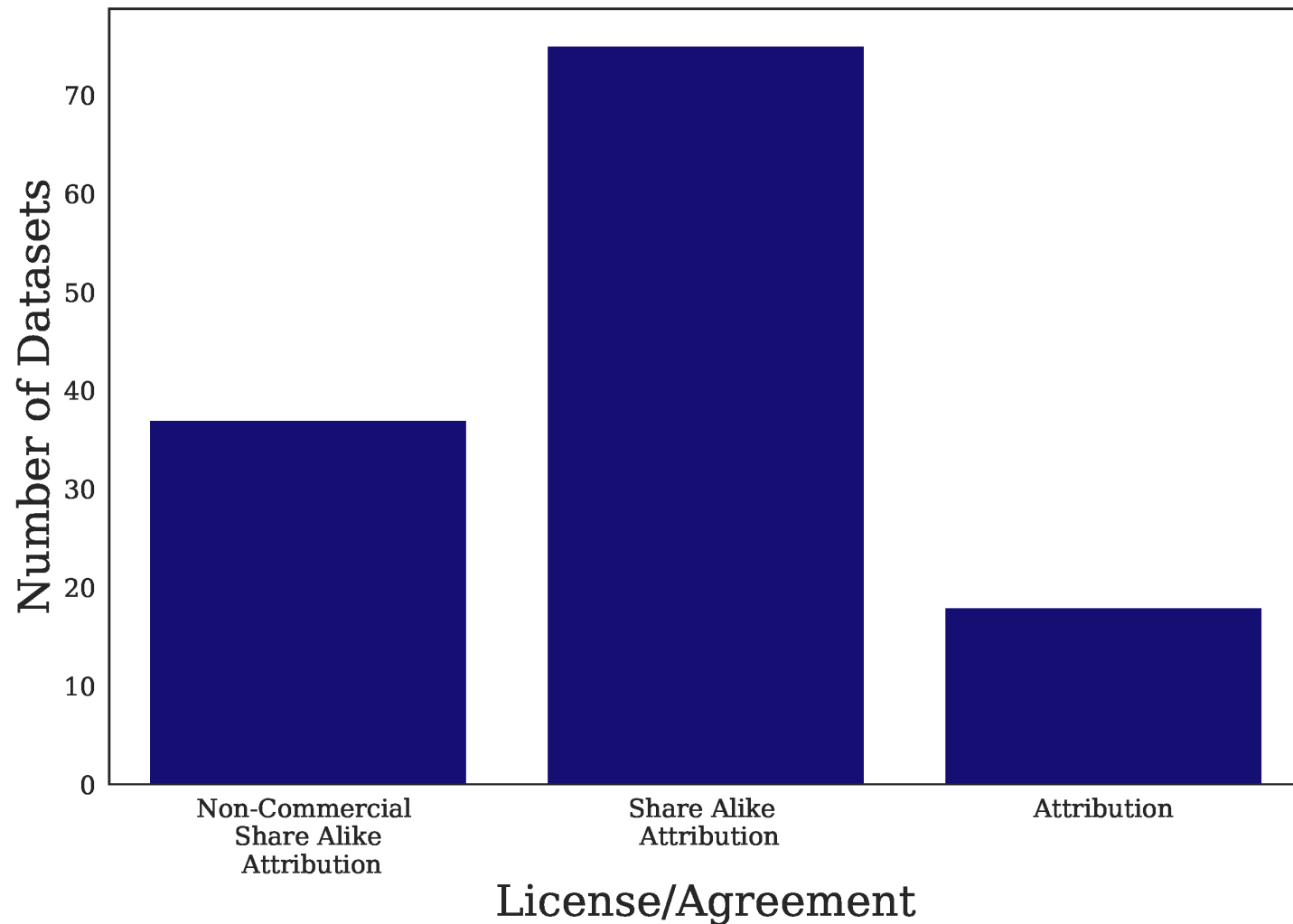
These Terms of Use relate only to the IBM Research DIF Dataset; the images referenced in the IBM Research DIF Dataset are governed by other terms as detailed in Sec. 2 below.

PART 2

Distribution of dataset licenses and agreements used by a subset of the open AI community (Kaggle).



Distribution of Creative Commons and Open Data Commons licenses by term.



Can we draw from these terms to support transparency and accountability efforts?

Attribution/Citation:

- Require preservation of metadata and transparency disclosures (e.g., data sheets, nutrition labels).
- Extend the attribution condition to require disclosing whether the dataset was altered.

Share-alike:

- Require like disclosures (model cards, fact sheets).
- What to do about proprietary data?

Address legal as well as technological protection measures: The CFAA likely poses a greater obstacle to transparency than the DMCA.

Doctrinal and practical challenges in addressing transparency through licenses and terms of use.

Imposing new servitudes: Information costs, lack of notice, and salience.

Networks effects and interoperability: Less reliance on common or benchmark datasets and greater difficulty combining different data sources.

Open source culture and ethics: Open innovation communities may not accept further encumbering shared resources.

Standardization: Datasets and models are as diverse as data types. And any standard needs buy-in from both open innovation and AI communities.

Conclusion.

This is a work in progress.

It very well may turn out that licenses or terms of use are not the right answer.

But curators of datasets need to think about their options, which repositories should respect and provide.

Thanks!