BerkeleyLaw
UNIVERSITY OF CALIFORNIA

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

Erik Stallman and Sonia Katyal

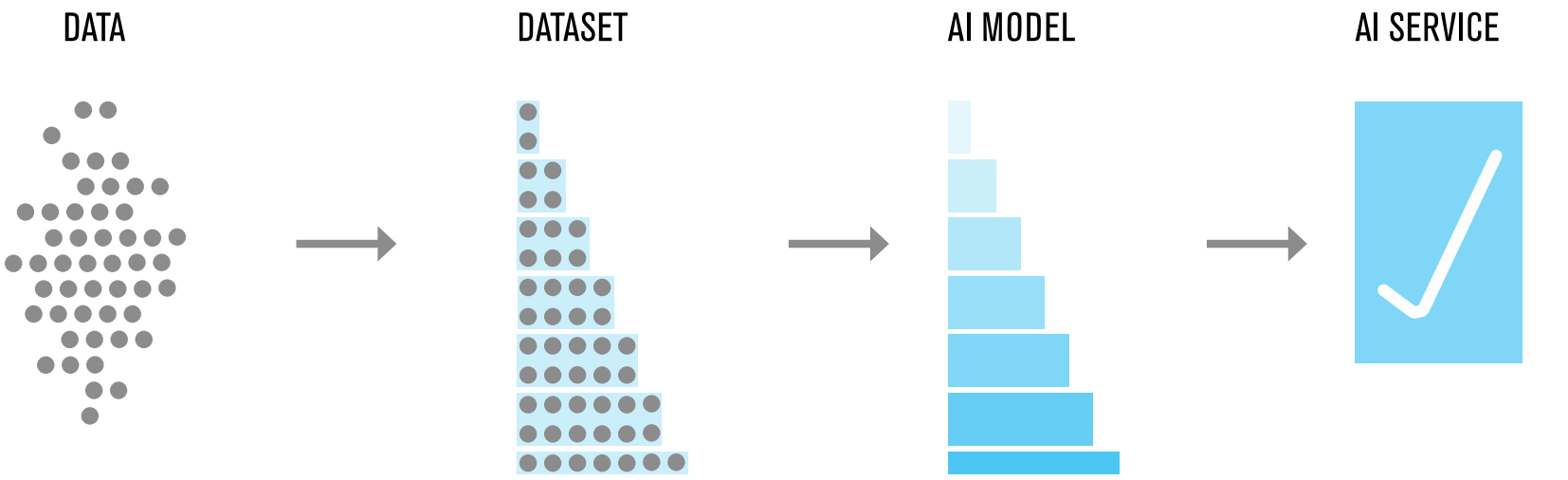
Intellectual Property Scholars Conference
August 9, 2019

Problem statement.

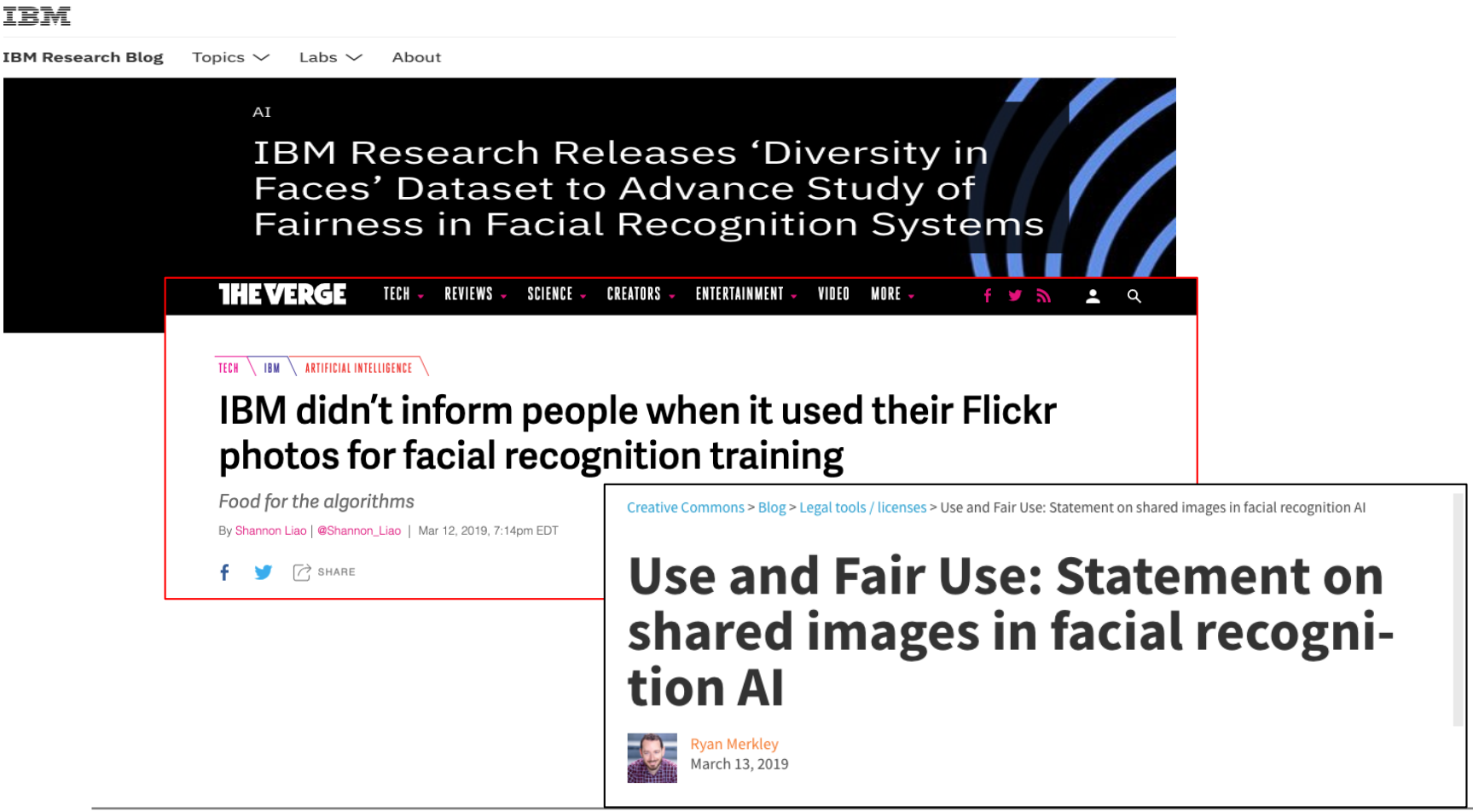
Assertions of proprietary interests in the data and code used in AI systems are impeding transparency and accountability just as new model disclosures and accountability tools are emerging.

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?

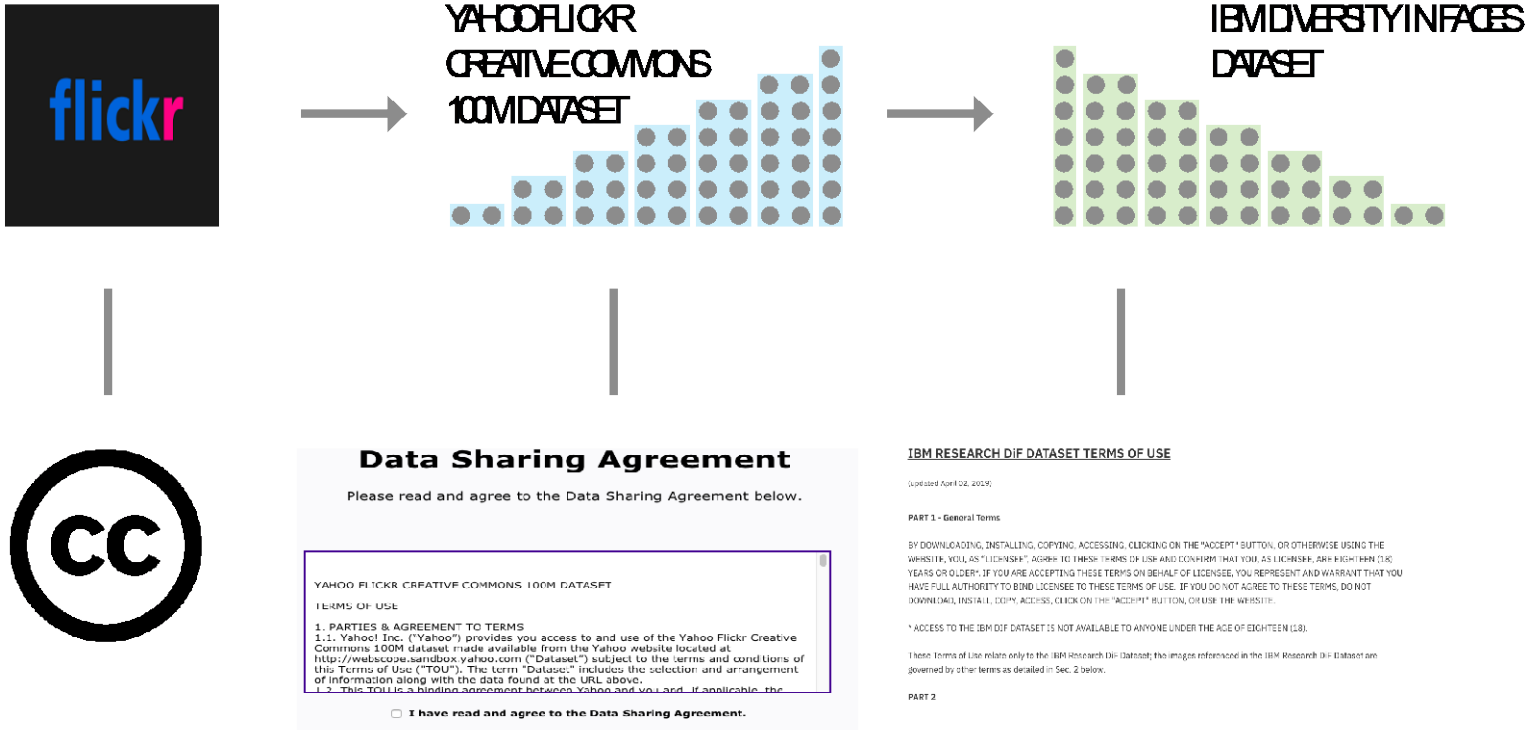
Model disclosures for components in an AI system.



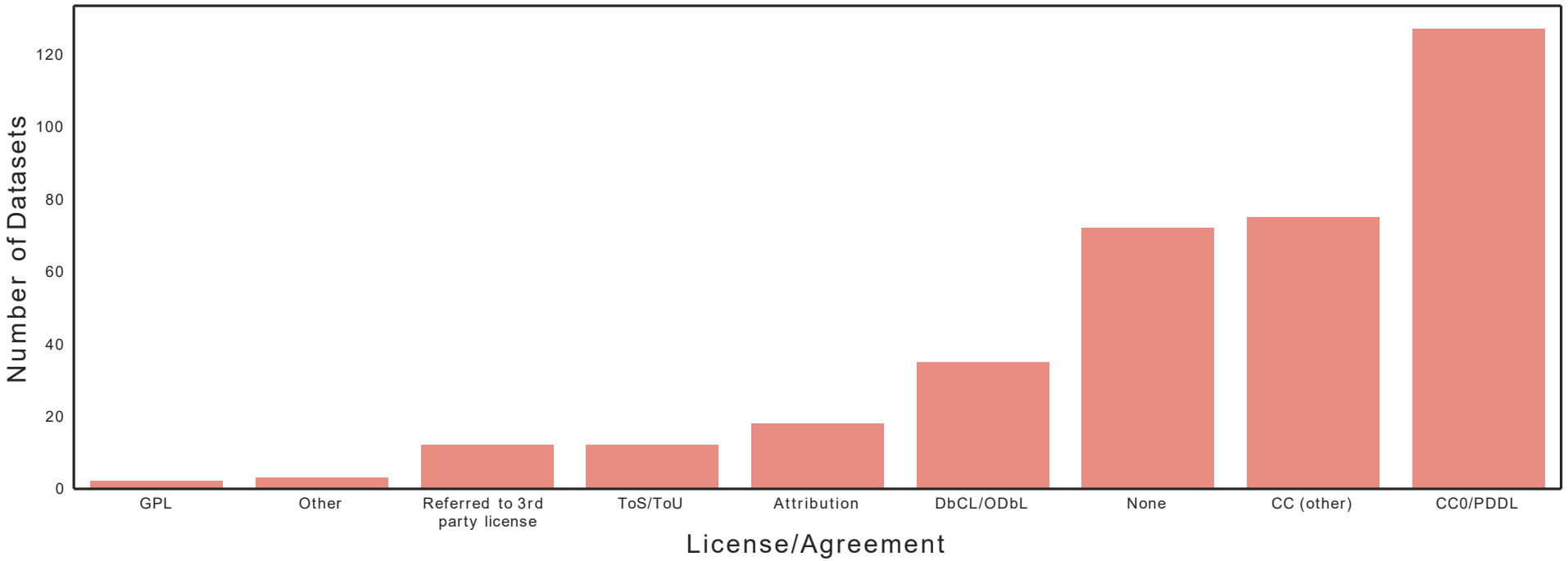
Licensing regimes are not intended to address model disclosures or data used in AI generally.



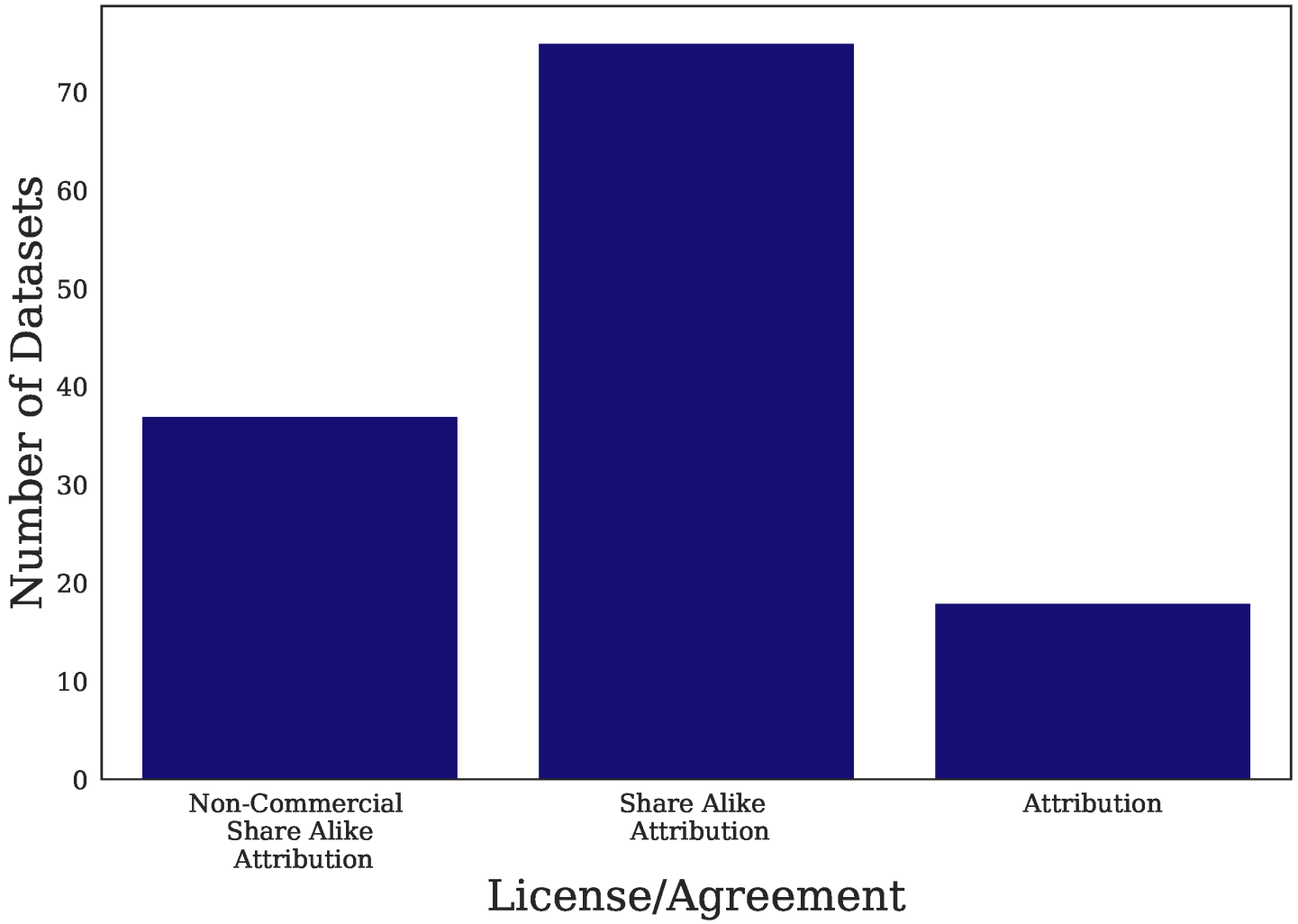
The chain of licenses and terms of use in the DiF dataset.



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!