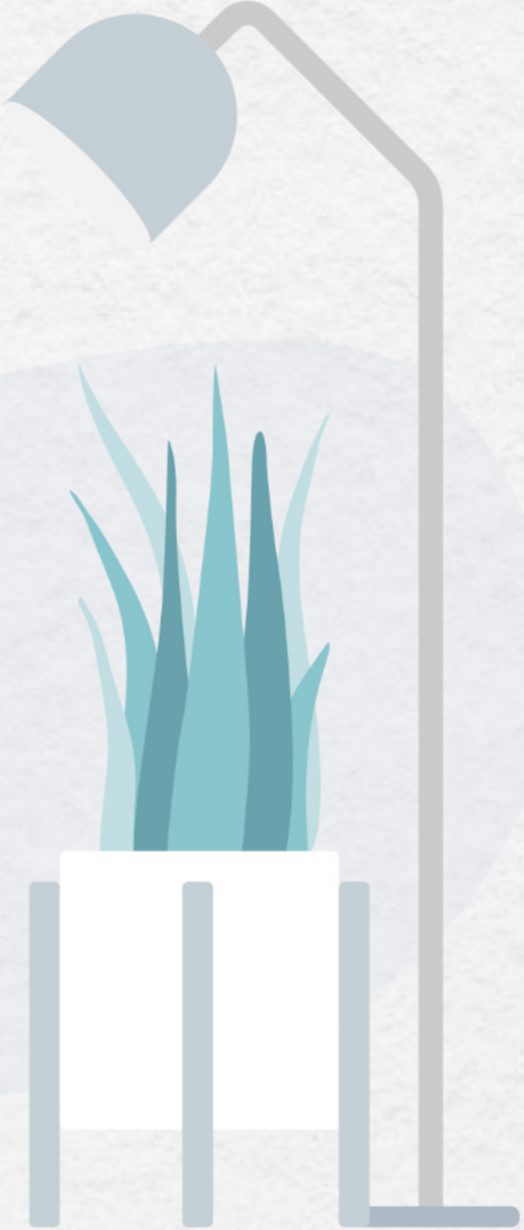
Evaluation of Statistical and Machine Learning Systems

(Two Challenging Problems)

JSM 2022
Washington, DC
August 10, 2022

Olivier Binette
Duke University / American Institutes for Research



Overview

Two challenging evaluation problems:

1. the reliability of **multiple systems estimation**, and
2. the accuracy of **entity resolution** algorithms.

Where and what is our science of statistical evaluation?

(it often seems fragmented or neglected in favor of modeling)

What is Evaluation?

Systematic assessment of a model's performance and properties for the purpose of:

1. **choosing** the best model,
2. **using** models appropriately, and
3. **understanding** real-world effects.

Evaluation studies need to answer specific questions using appropriate methodology.

1. Reliability of Multiple Systems Estimation

The Problem

How many victims of human trafficking?

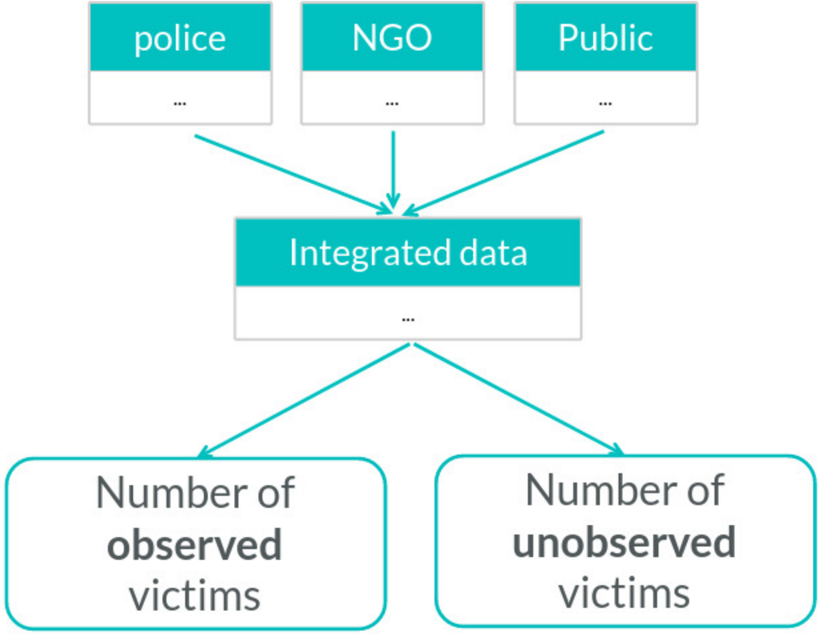
- Victims are **hidden and hard to reach.**
- **Organizations** like the police and NGOs only **reach a small proportion** of the victims.

How can we get a representative picture?

Multiple Systems Estimation (MSE)

How it works:

- **Integrate data** (observed victims) from multiple sources through record linkage.
- Perform a **missing data analysis** to estimate the number of unobserved victims.



police

...

NGO

...

Public

...

Integrated data

...

Number of **observed** victims

Number of **unobserved** victims

Does MSE Work?

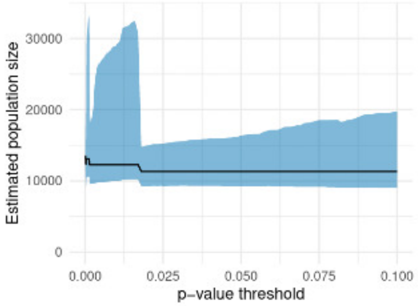
Contentious question.

- **200 years** of controversy!
- No **ground truth** to check results.

It's all about:

- Missing data assumptions
- Data sufficiency and robustness
- Inductive biases

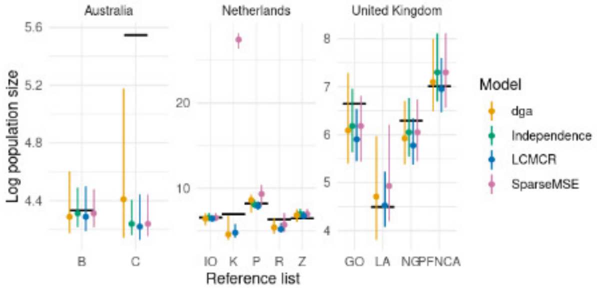
Our Evaluation Proposal



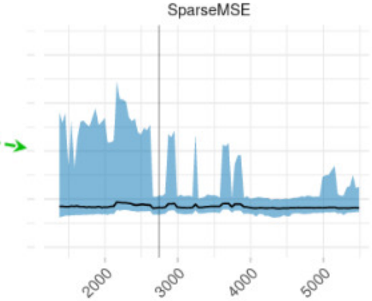
Drop simulation studies that can give any result you like.

Instead:

1. Perform **sensitivity analyses.**
2. Dig through data for **pseudo ground truths.**
3. **Quantify** the consequences of model assumptions.
4. Generate visual & meaningful **assessments of robustness.**



$$\lim_{N\to\infty} \frac{\hat{N} - N}{N} = p_{\mathbf{0}}(e^{\gamma} - 1).$$



https://github.com/**OlivierBinette/MSETools**

Conclusion Regarding MSE

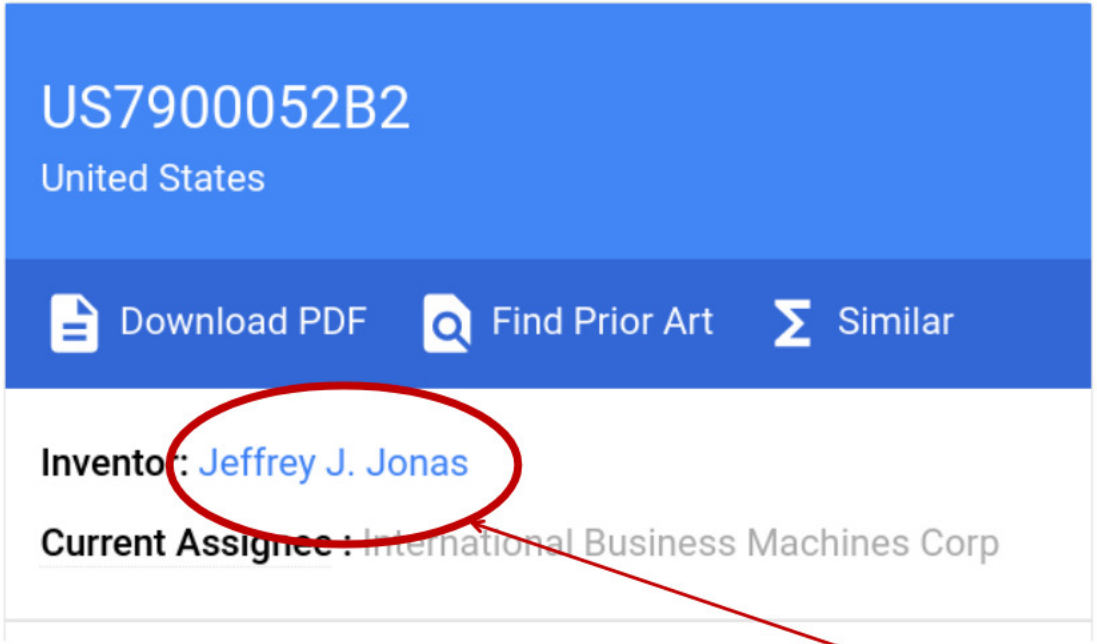
I don't think we've closed the discussion, but these evaluation tools provide significant practical instights.

I wish I had known more about the **science of evaluation** when going into this project.

- I **feel** like this science is or has been **neglected**. What do you think?

2. Evaluation of Entity Resolution Algorithms

Inventor Disambiguation at PatentsView.org

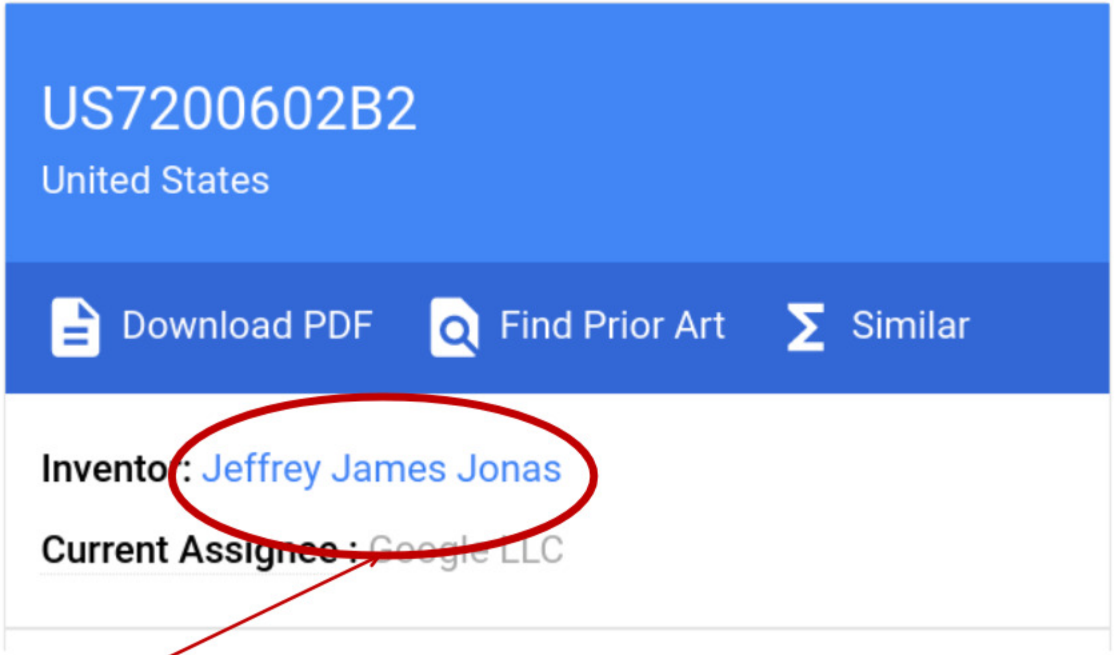


US7900052B2

United States

Inventor: Jeffrey J. Jonas

Current Assignee : International Business Machines Corp



US7200602B2

United States

Inventor: Jeffrey James Jonas

Current Assignee : Google LLC

Are they the same?

Inventor Disambiguation

Goal:

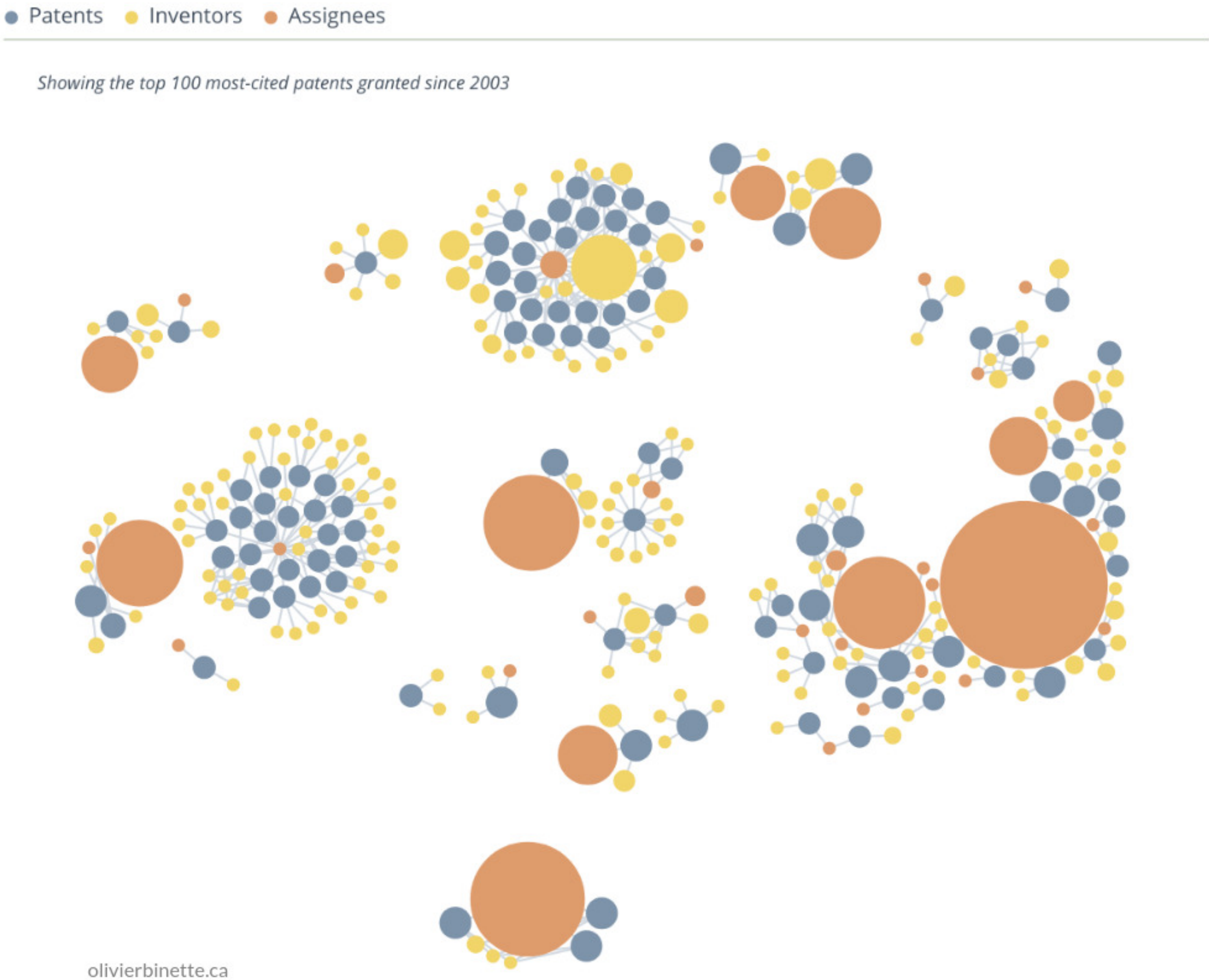
- **Cluster inventor mentions** that refer to the same real-world person.

Evaluation metrics:

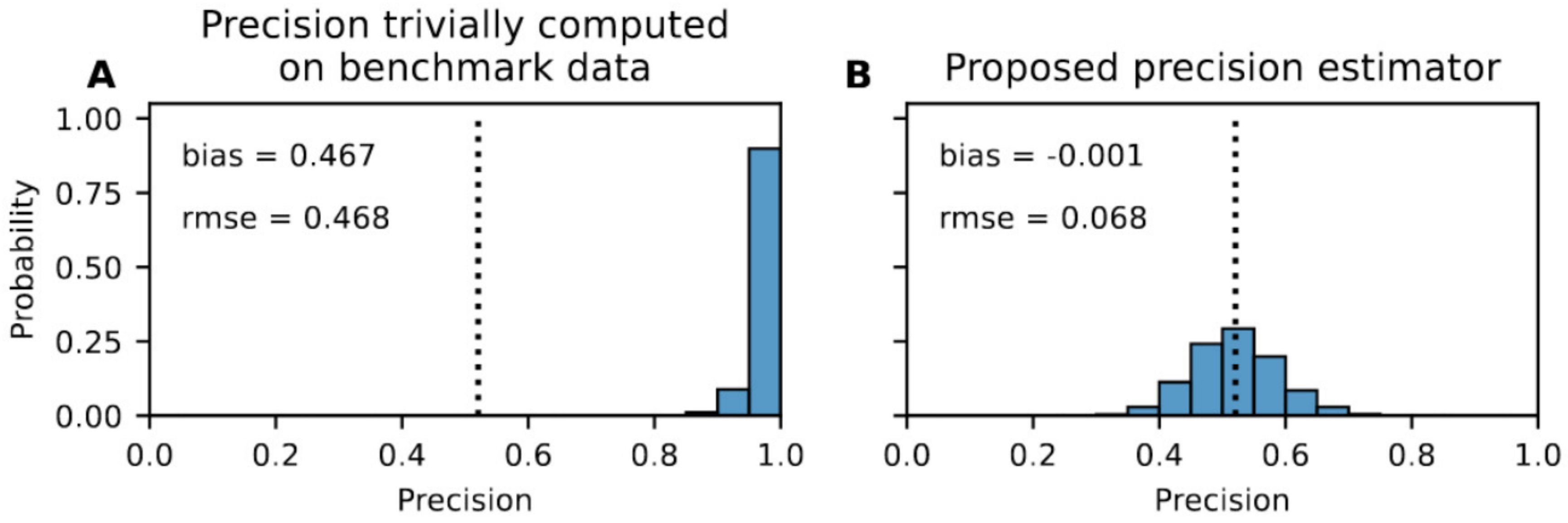
- Precision and recall

Benchmark datasets:

- Hand-disambiguated subsets of the data


● Patents ● Inventors ● Assignees
Showing the top 100 most-cited patents granted since 2003

Evaluation should be straightforward... right!?



A Precision trivially computed on benchmark data

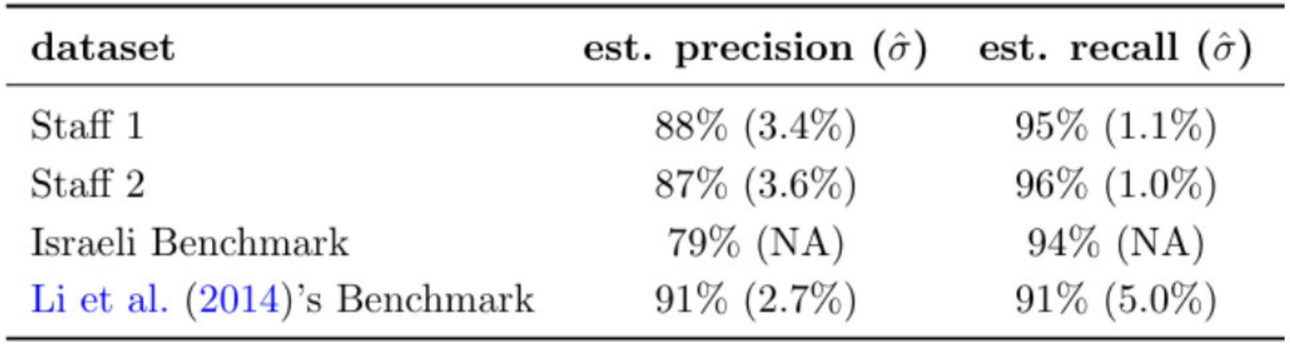
B Proposed precision estimator

Figure 1: *Distribution of precision estimates versus the true precision of 52% (shown as a dotted vertical line). Panel **A** shows the trivial precision estimates computed for sampled records. Panel **B** shows our proposed precision estimates which accounts for the sampling mechanism. Sample bias and root mean squared error (rmse) are reported in each figure.*

Evaluation is not straightforward.

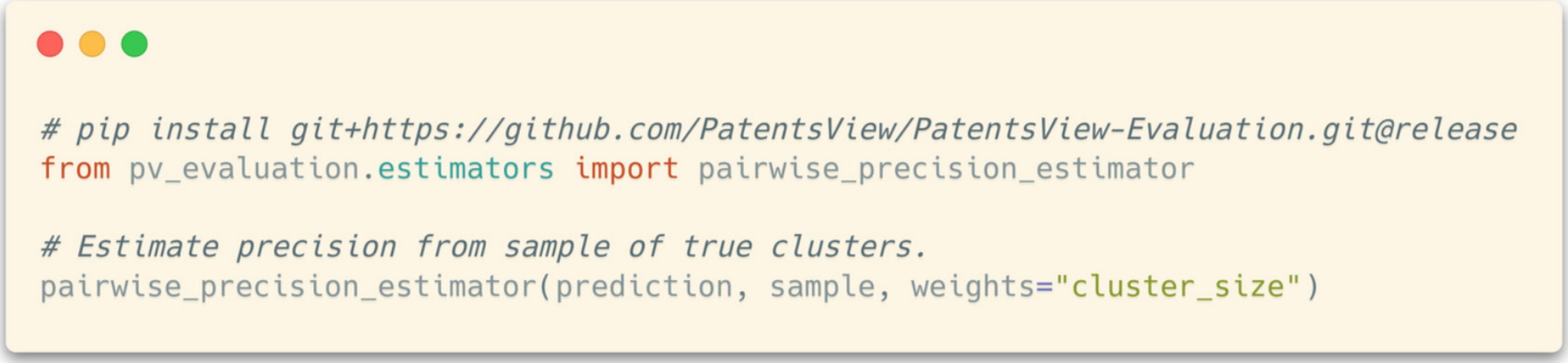
We proposed new methodology for **unbiased performance estimation** based on **sampling ground truth clusters**.

- Representative performance estimates **for the first time** at PatentsView.org
- More **cost-effective and practical** (for PatentsView) than sampling record pairs or other approaches.



dataset	est. precision ($\hat{\sigma}$)	est. recall ($\hat{\sigma}$)
Staff 1	88% (3.4%)	95% (1.1%)
Staff 2	87% (3.6%)	96% (1.0%)
Israeli Benchmark	79% (NA)	94% (NA)
Li et al. (2014)'s Benchmark	91% (2.7%)	91% (5.0%)

Usage



```
# pip install git+https://github.com/PatentsView/PatentsView-Evaluation.git@release
from pv_evaluation.estimators import pairwise_precision_estimator

# Estimate precision from sample of true clusters.
pairwise_precision_estimator(prediction, sample, weights="cluster_size")
```

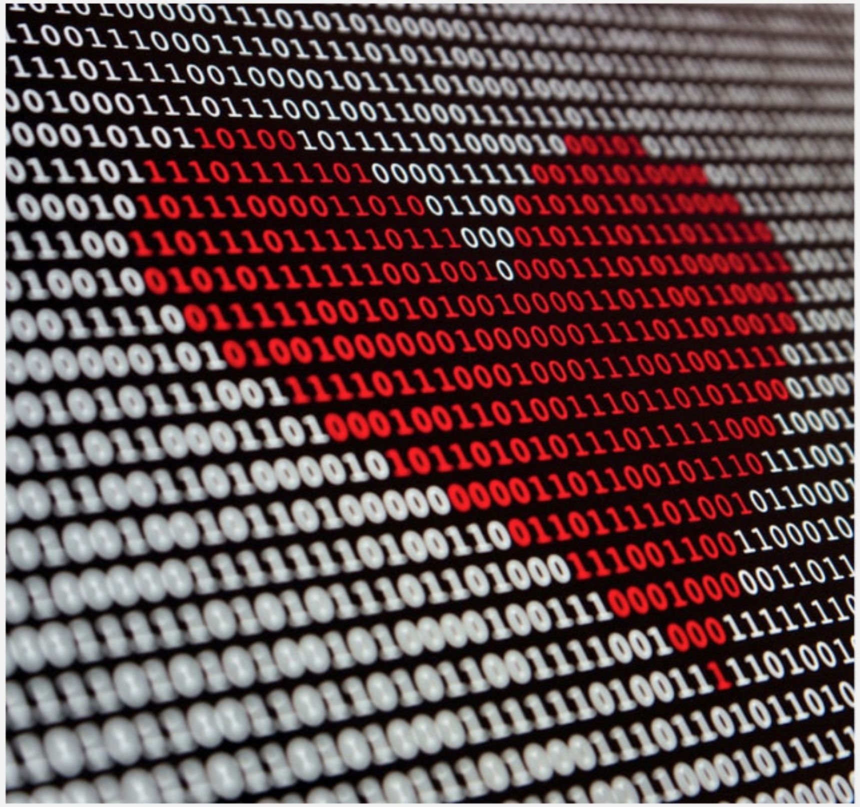
https://github.com/**PatentsView/PatentsView-Evaluation**

Leave a star!

PatentsView is releasing new data!

New training data (n > 150,000) to support methodological research

olivier@olivierbinette.ca



Conclusion

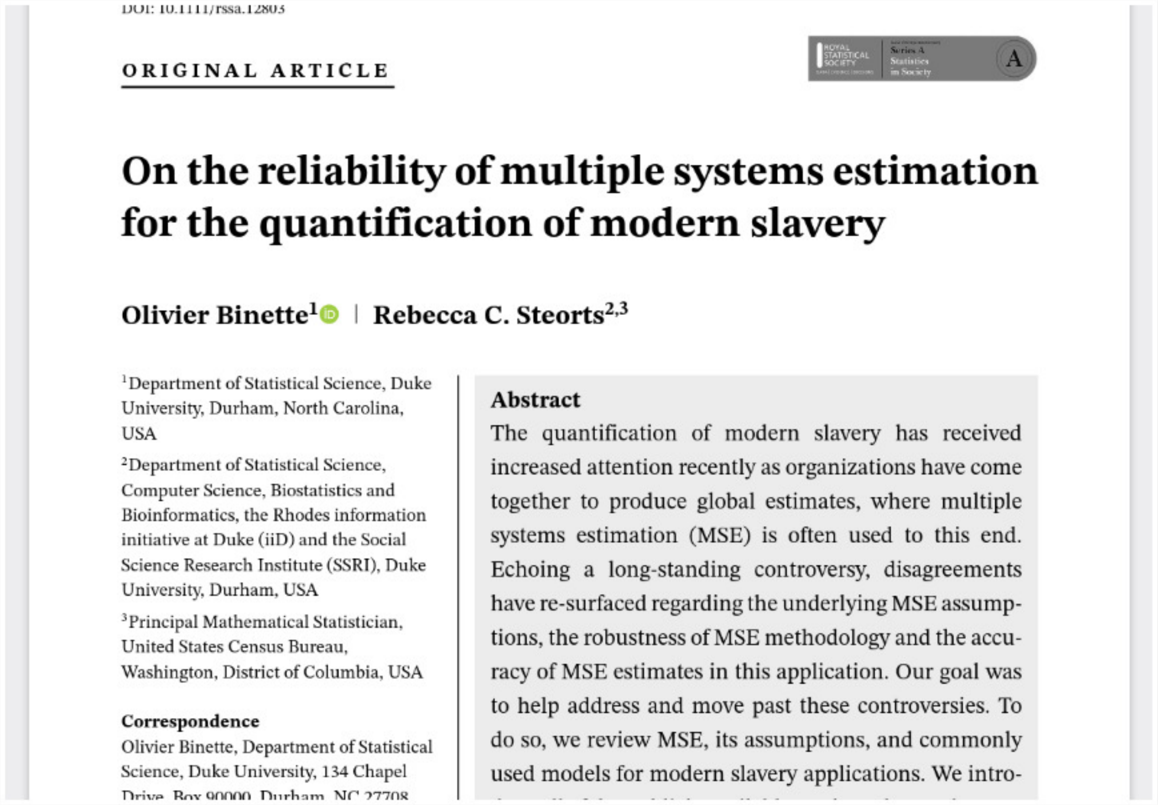
Concluding Thoughts

- Evaluation is often **not straightfoward**.
- It is often **neglected.**
- We need to **value it more**.
- **Where** is our science of evaluation? Help me find it!

I want to hear your stories and thoughts.

olivier@olivierbinette.ca

Papers



DOI: 10.1111/rssa.12803

ORIGINAL ARTICLE

On the reliability of multiple systems estimation for the quantification of modern slavery

Olivier Binette[1] | **Rebecca C. Steorts**[2,3]

[1] Department of Statistical Science, Duke University, Durham, North Carolina, USA

[2] Department of Statistical Science, Computer Science, Biostatistics and Bioinformatics, the Rhodes information initiative at Duke (iiD) and the Social Science Research Institute (SSRI), Duke University, Durham, USA

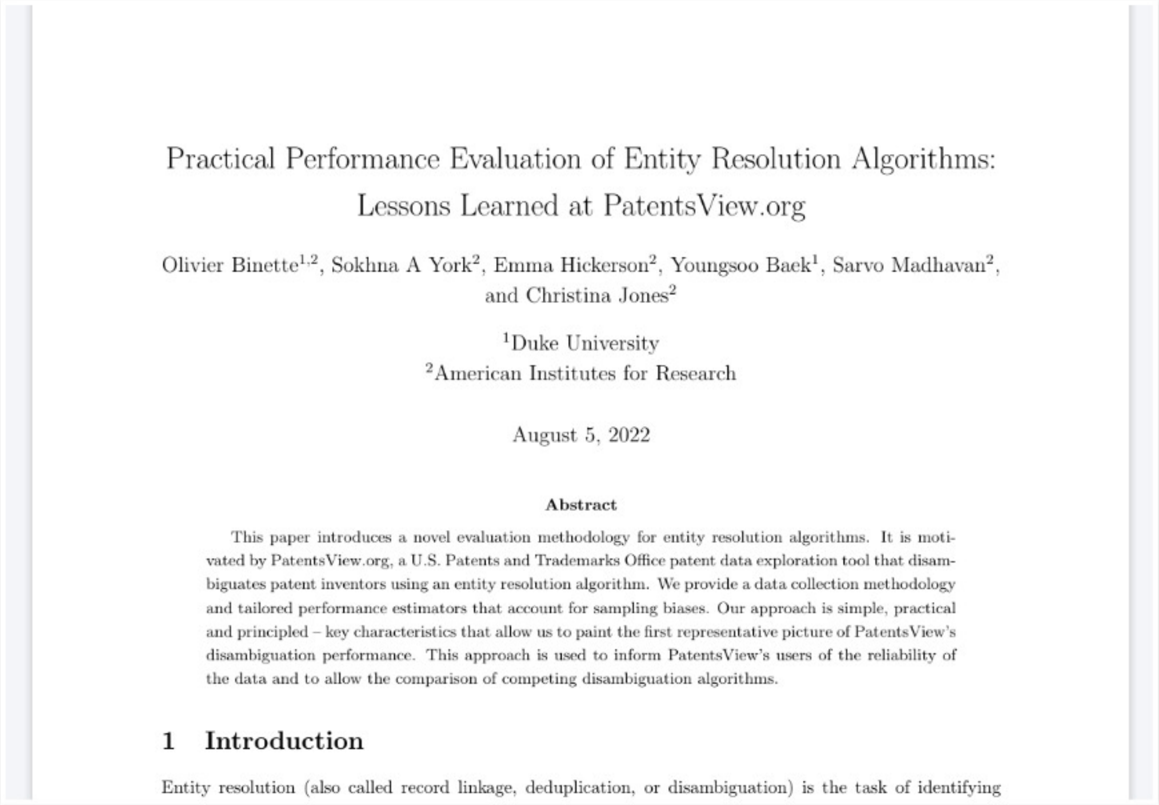
[3] Principal Mathematical Statistician, United States Census Bureau, Washington, District of Columbia, USA

Correspondence
Olivier Binette, Department of Statistical Science, Duke University, 134 Chapel Drive, Box 90000, Durham, NC 27708

Abstract

The quantification of modern slavery has received increased attention recently as organizations have come together to produce global estimates, where multiple systems estimation (MSE) is often used to this end. Echoing a long-standing controversy, disagreements have re-surfaced regarding the underlying MSE assumptions, the robustness of MSE methodology and the accuracy of MSE estimates in this application. Our goal was to help address and move past these controversies. To do so, we review MSE, its assumptions, and commonly used models for modern slavery applications. We intro-

arXiv:2112.01594



Practical Performance Evaluation of Entity Resolution Algorithms: Lessons Learned at PatentsView.org

Olivier Binette[1,2], Sokhna A York[2], Emma Hickerson[2], Youngsoo Baek[1], Sarvo Madhavan[2], and Christina Jones[2]

[1] Duke University
[2] American Institutes for Research

August 5, 2022

Abstract

This paper introduces a novel evaluation methodology for entity resolution algorithms. It is motivated by PatentsView.org, a U.S. Patents and Trademarks Office patent data exploration tool that disambiguates patent inventors using an entity resolution algorithm. We provide a data collection methodology and tailored performance estimators that account for sampling biases. Our approach is simple, practical and principled – key characteristics that allow us to paint the first representative picture of PatentsView's disambiguation performance. This approach is used to inform PatentsView's users of the reliability of the data and to allow the comparison of competing disambiguation algorithms.

1 Introduction

Entity resolution (also called record linkage, deduplication, or disambiguation) is the task of identifying

Soon on arxiv. Available on my website

Thank you!

Funding:

- American Institutes for Research (USPTO)
- NSERC Canada Graduate Scholarship
- NSF CAREER Award (Rebecca Steorts)
- ASA Travel award
- Github sponsors (individual contributors)
- G-Research PhD grant