Observing the Long Tail of Research

#AltmetricsForAll #InAllLanguages

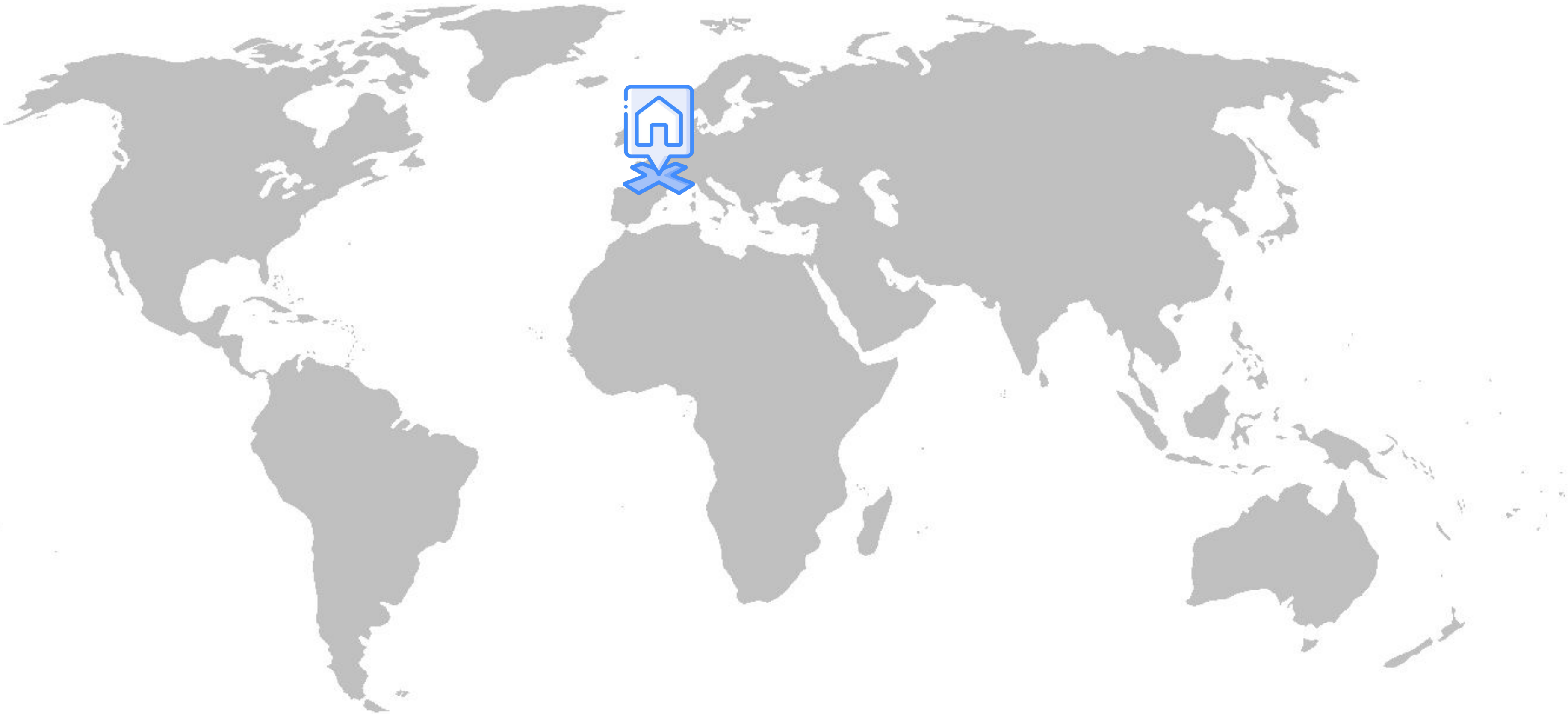
Luc Boruta — Thunken
luc@thunken.com — @thunkenizer

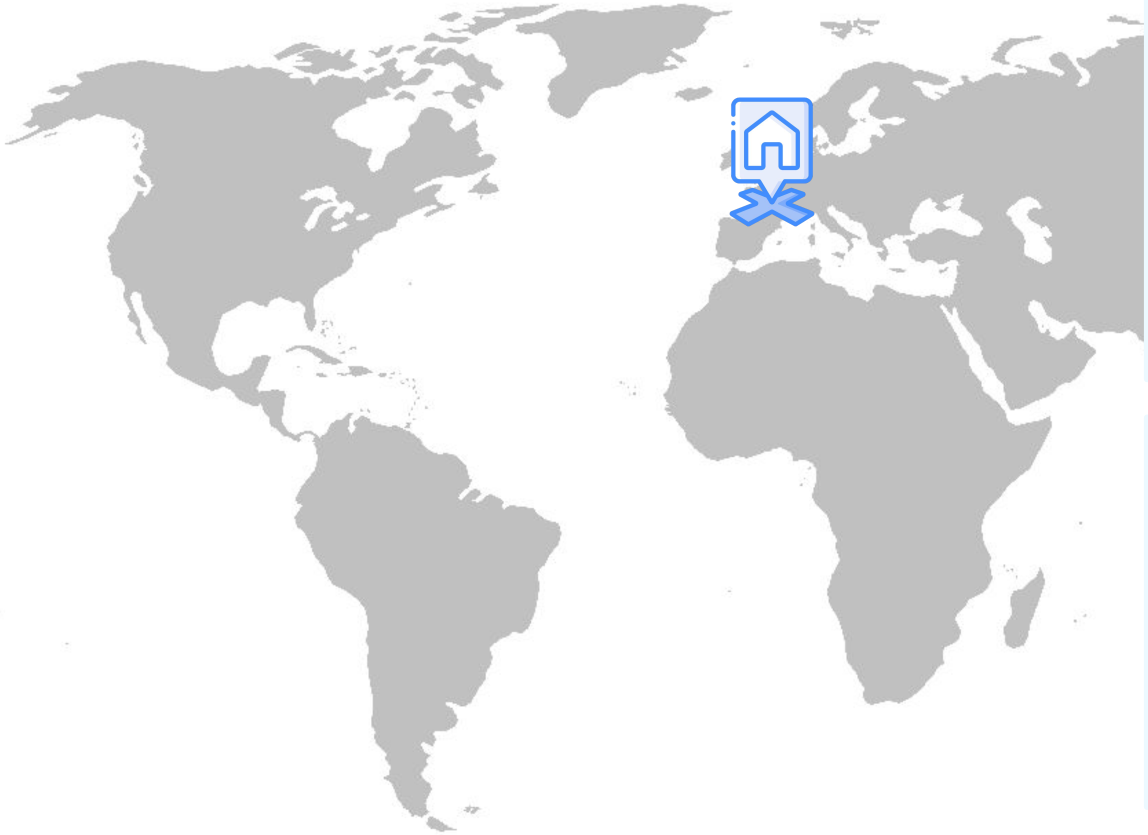
LATmetrics, Cusco, 2019/11/06

https://doi.org/10.6084/m9.figshare.10060088

THUNKEN

cobaltmetrics.com





https://w.wiki/BWt + https://www.rtl.fr/actu/politique/7795681059

Are your metrics alt- enough?

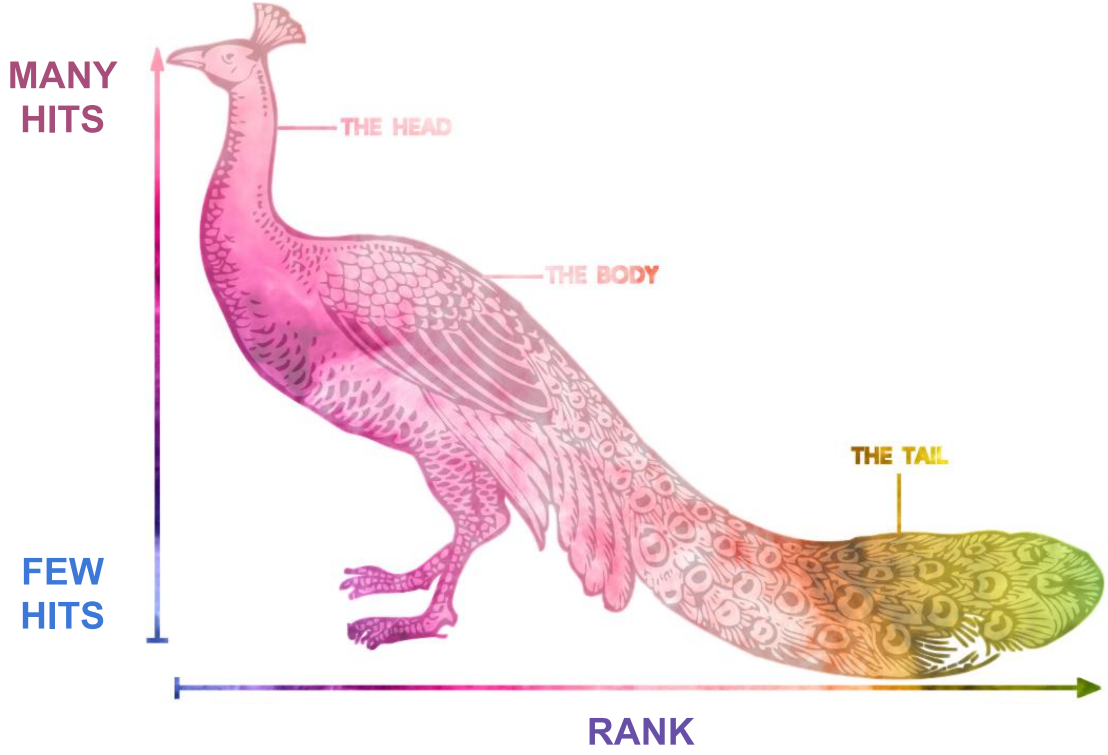
NO.

Program

1. The long tail of research
2. Latent discrimination in scientometrics
3. Introduction to Cobaltmetrics and URI transmutation
4. Group discussion: let's make altmetrics genuinely alt-!
5. Final remarks

The Long Tail of Research

The long tail of research



http://tiny.cc/vtfsfz

The long tail of research

- RDA interest group: http://doi.org/10.15497/RDA00023
- e-IRG task force: http://tiny.cc/iwmtfz
- Talk by Chuck Humphrey: http://tiny.cc/50mtfz
- 6:AM do-a-thon: http://tiny.cc/2ma5dz

Attention vs. Impact

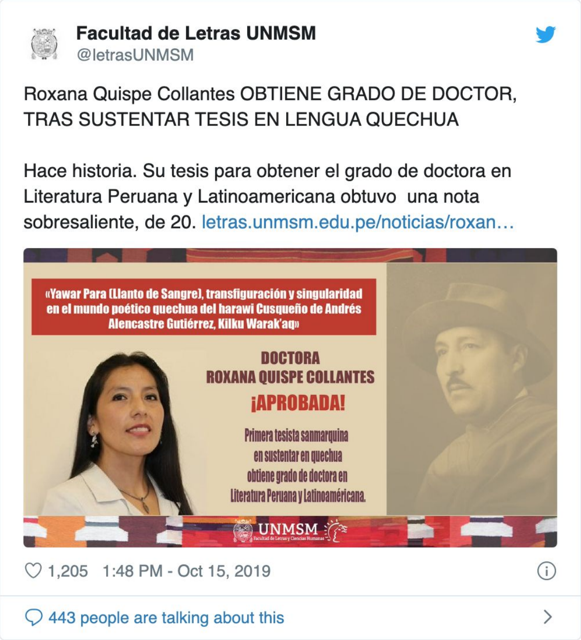
Citations and altmetrics are proxies for impact.

Citations and altmetrics measure attention.

Attention correlates w/ impact. So do influence and privilege.

See also Sugimoto’s “Attention is not impact” (2015).

Attention vs. Impact



Facultad de Letras UNMSM
@letrasUNMSM

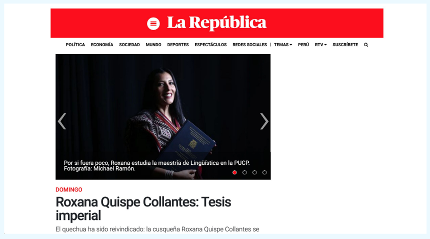
Roxana Quispe Collantes OBTIENE GRADO DE DOCTOR, TRAS SUSTENTAR TESIS EN LENGUA QUECHUA

Hace historia. Su tesis para obtener el grado de doctora en Literatura Peruana y Latinoamericana obtuvo una nota sobresaliente, de 20. letras.unmsm.edu.pe/noticias/roxan…

«Yawar Para (Llanto de Sangre), transfiguración y singularidad en el mundo poético quechua del harawi Cusqueño de Andrés Alencastre Gutiérrez, Kilku Warak'aq»
DOCTORA
ROXANA QUISPE COLLANTES
¡APROBADA!
Primera tesista sanmarquina en sustentar en quechua obtiene grado de doctora en Literatura Peruana y Latinoamericana.
UNMSM
Facultad de Letras y Ciencias Humanas

♡ 1,205 1:48 PM - Oct 15, 2019

443 people are talking about this



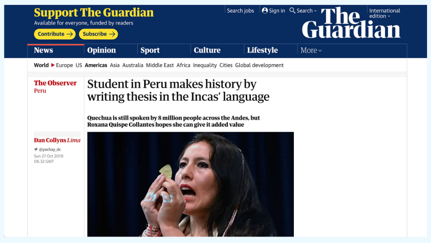
La República
POLÍTICA ECONOMÍA SOCIEDAD MUNDO DEPORTES ESPECTÁCULOS REDES SOCIALES | TEMAS PERÚ RTV SUSCRÍBETE

Por si fuera poco, Roxana estudia la maestría de Lingüística en la PUCP. Fotografía: Michael Ramón.

DOMINGO

Roxana Quispe Collantes: Tesis imperial

El quechua ha sido reivindicado: la cusqueña Roxana Quispe Collantes se



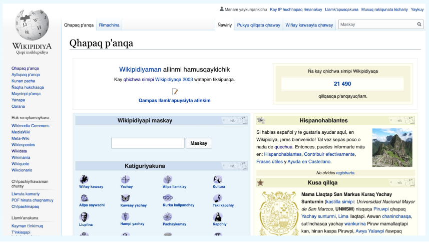
Support The Guardian
Available for everyone, funded by readers
Contribute →
Subscribe →
Search jobs
Sign in
Search
The Guardian
International edition
News
Opinion
Sport
Culture
Lifestyle
More
World ▸ Europe US Americas Asia Australia Middle East Africa Inequality Cities Global development

The Observer
Peru

Student in Peru makes history by writing thesis in the Incas' language

Quechua is still spoken by 8 million people across the Andes, but Roxana Quispe Collantes hopes she can give it added value

Dan Collyns *Lima*
@yachay_dc
Sun 27 Oct 2019
06.32 GMT



Qhapaq p'anqa

Wikipidiyaman allinmi hamusqaykichik

Kay qhichwa simipi Wikipidiyaqa 2003 watapim tiksipusqa.

Qampas llamk'apuysiyta atinkim

Ña kay qhichwa simipi Wikipidiyaqa
21 490
qillqasqa p'anqayuqñam.

Wikipidiyapi maskay

Maskay

Katiguriyakuna

Hispanohablantes

Si hablas español y te gustaría ayudar aquí, en Wikipidiya, ¡eres bienvenido! Tal vez sepas poco o nada de quechua. Entonces, puedes informarte más en: Hispanohablantes, Contribuir efectivamente, Frases útiles y Ayuda en Castellano.

No olvides registrarte.

Kusa qillqa

Mama Llaqtap San Markus Kuraq Yachay Sunturnin (kastilla simipi: *Universidad Nacional Mayor de San Marcos*, **UNMSM**) nisqaqa Piruwpi qhapaq Yachay sunturmi, Lima llaqtapi. Aswan chaninchasqa, sut'inchasqa yachay wankurina Piruw mamallaqtapi kan, hinan kaspa Piruwpi, Awya Yalawpi ñawpaq

Cobaltmetrics

Event Data

Altmetric

PLUM ANALYTICS

Latent Discrimination in Scientometrics

A partial landscape of citation aggregators

- Journal to journal: Web of Science, Scopus
- DOI to DOI: OpenCitations
- URL to DOI: ALM/Lagotto, Crossref Event data
- URL to URL: Altmetric, Plum, **Cobaltmetrics**

Common issues with citation aggregators

- Imbalanced datasets
  - Predefined lists of supported research outputs
  - Predefined lists of supported languages
- Irreproducible indicators
  - Dependency on 3rd party servers (short URLs, APIs)

Why should we care?

Metrics are a sampling game.

Imbalanced datasets reinforce **discrimination**.

What have we really gained by changing the statistic if **the sample remains biased**?

Why should we care?

It is not up to citation aggregators to decide what is citable.

Our role is to **observe all citation patterns on the web**.

We are interested in **low-frequency phenomena**,
and in distinguishing **structural zeros** from **sampling zeros**.

Linguistic diversity

>7k languages are spoken today.

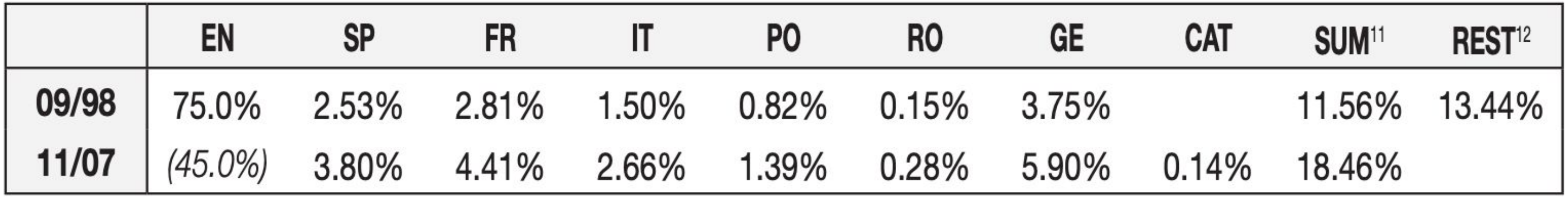
23 languages account for >50% of the world population.

L1-English accounts for <5% of the world population.

https://www.ethnologue.com/guides/how-many-languages

Linguistic diversity on the web

UNESCO report by Pimienta et al. (2009):



	EN	SP	FR	IT	PO	RO	GE	CAT	SUM[11]	REST[12]
09/98	75.0%	2.53%	2.81%	1.50%	0.82%	0.15%	3.75%		11.56%	13.44%
11/07	*(45.0%)*	3.80%	4.41%	2.66%	1.39%	0.28%	5.90%	0.14%	18.46%	

Wikipedia: **303 languages**, **49M articles**

English Wikipedia: 5.7M articles (<12%)

https://w.wiki/7iw +
http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

Selection biases: Wikipedia languages

Not all indicators are sensitive to linguistic diversity, but...

- Altmetric: 3 languages (en, fi, sv)
- Plum: 3 languages (en, es, pt)
- ALM: 25 most popular languages
- **Cobaltmetrics: 180+ languages!**

Latent discrimination, real consequences

Systematic bias in stats on cross-linguistic citation practices.

Systematic exclusion of some contributors from metrics-based evaluations.

Reinforces discrimination in other parts of the community, cf. Nylenna et al. (1994) and Lazarev & Nazarovets (2018).

Selection biases: document types

Strong focus on traditional peer-reviewed publications. Preprints are still treated as **second-class documents**.

What about patents, clinical trials, law articles, etc.?

What about **non-textual objects**, e.g. datasets or software?

Selection biases: PIDs vs. URLs

Nothing lasts forever on the web:

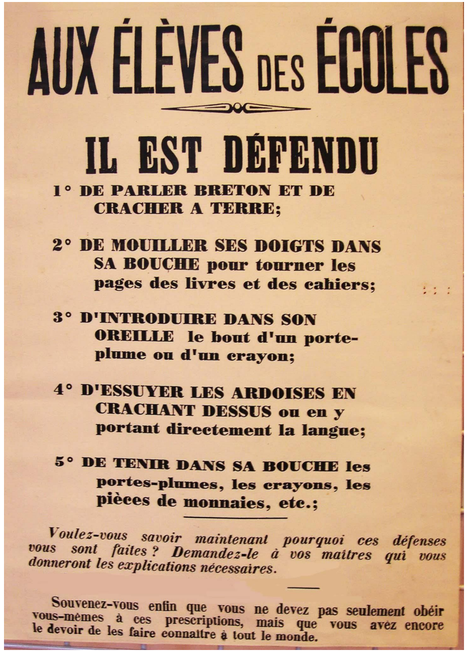
- Link rot!
- Content drift!
- Outages!

Weapons of math destruction

“There is a moral obligation to **challenge machine biases**.”
— Heather Staines, PIDapalooza’19

Algorithmic bias reflects the values of the humans involved in designing the algorithm and/or collecting the data.

From la vergonha to math destruction



AUX ÉLÈVES DES ÉCOLES

IL EST DÉFENDU

1° DE PARLER BRETON ET DE CRACHER A TERRE;

2° DE MOUILLER SES DOIGTS DANS SA BOUCHE pour tourner les pages des livres et des cahiers;

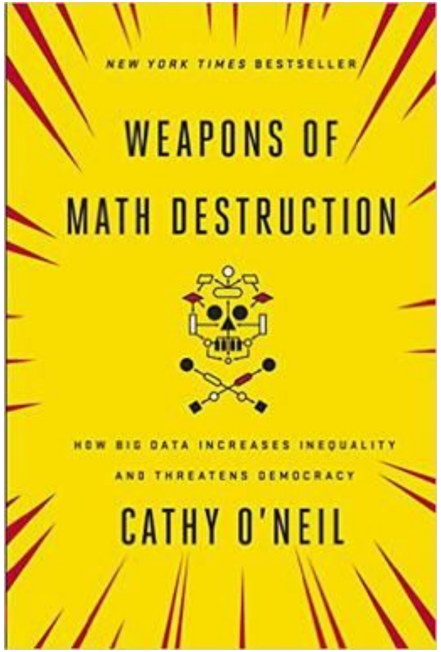
3° D'INTRODUIRE DANS SON OREILLE le bout d'un porte-plume ou d'un crayon;

4° D'ESSUYER LES ARDOISES EN CRACHANT DESSUS ou en y portant directement la langue;

5° DE TENIR DANS SA BOUCHE les portes-plumes, les crayons, les pièces de monnaies, etc.;

Voulez-vous savoir maintenant pourquoi ces défenses vous sont faites? Demandez-le à vos maitres qui vous donneront les explications nécessaires.

Souvenez-vous enfin que vous ne devez pas seulement obéir vous-mêmes à ces prescriptions, mais que vous avez encore le devoir de les faire connaitre à tout le monde.

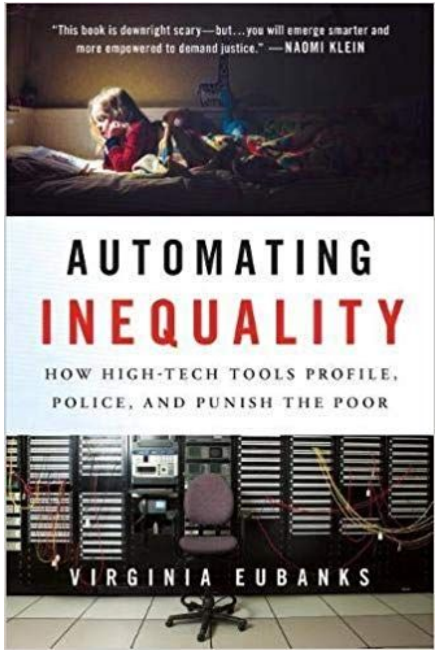


NEW YORK TIMES BESTSELLER

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL



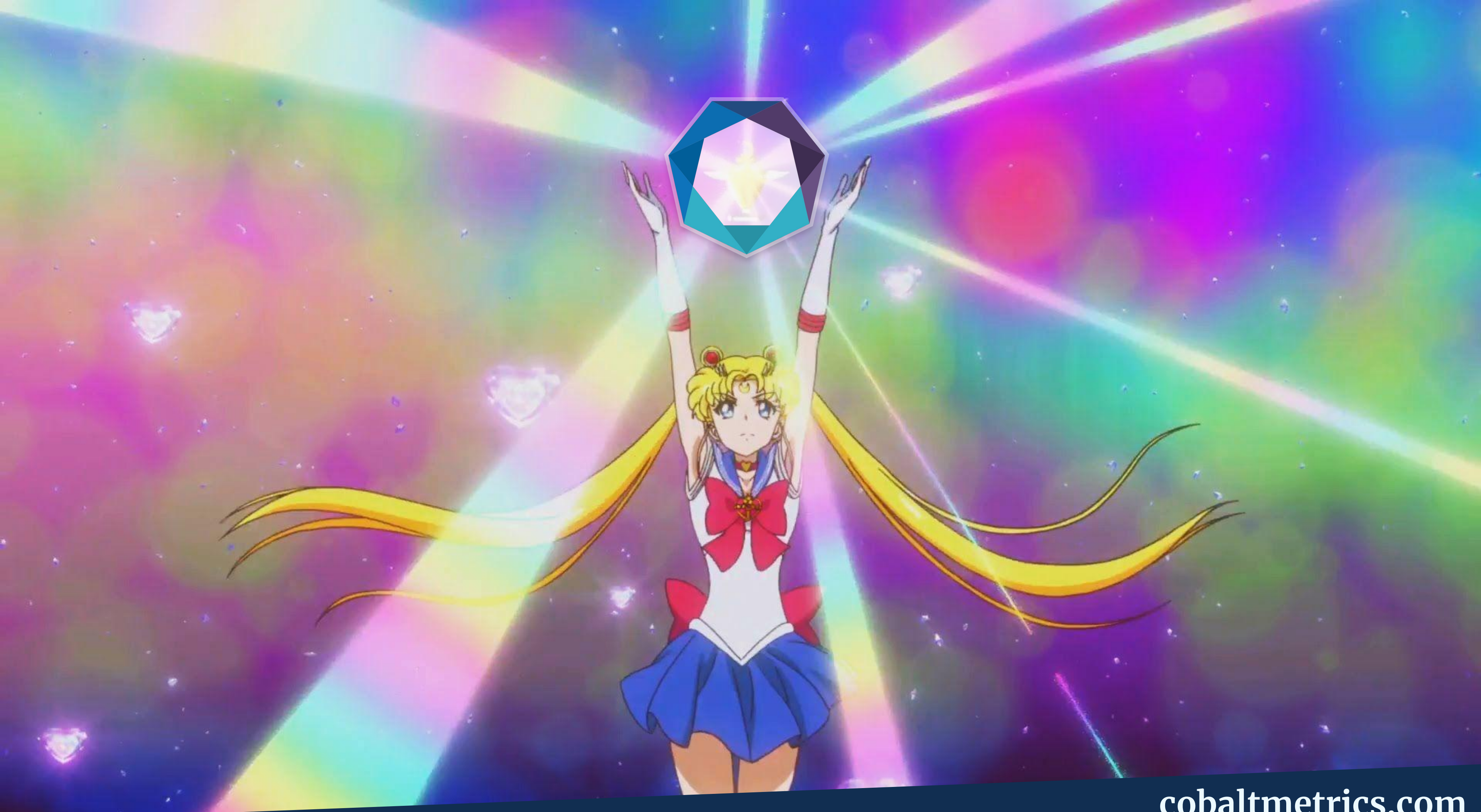
"This book is downright scary—but...you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN

AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR

VIRGINIA EUBANKS

Introduction to Cobaltmetrics



Cobaltmetrics

Cobaltmetrics crawls the web to index **hyperlinks and PIDs as first-class citations**.

The web is our corpus, and our **URI transmutation API** collates citations to all known versions of a document.

Design rationale

Cobaltmetrics tracks all URIs, URLs, and typed PIDs.

Cobaltmetrics can only be queried by URIs.

Cobaltmetrics will never create new identifiers.

Cobaltmetrics will never create new metrics.

Web-scale citation tracking

- Wikimedia (all projects, all languages)
- StackExchange/StackOverflow (all projects, all languages)
- US legal opinions (via CourtListener)
- Hypothes.is annotations
- Usenet posts (via the Internet Archive)
- **CommonCrawl (3.1 billion webpages)**

https://cobaltmetrics.com/docs/page/data-sources

Introduction to URI Transmutation

PIDs are not silver bullets

There are **billions of documents**
that will never get DOIs or any other fancy PID:
old documents, grey literature, and **the rest of the web**.

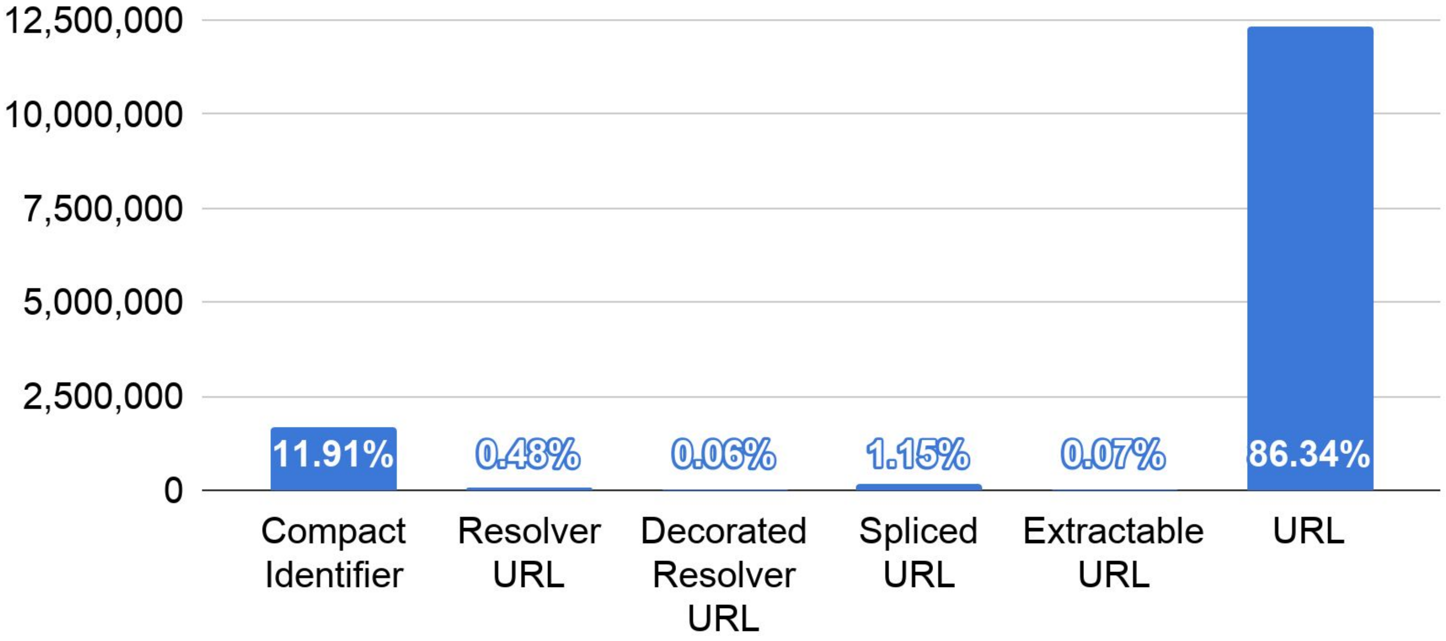
There are tons of documents with PIDs that are cited
with no mention of their PIDs.

Compact IDs vs. good old URLs

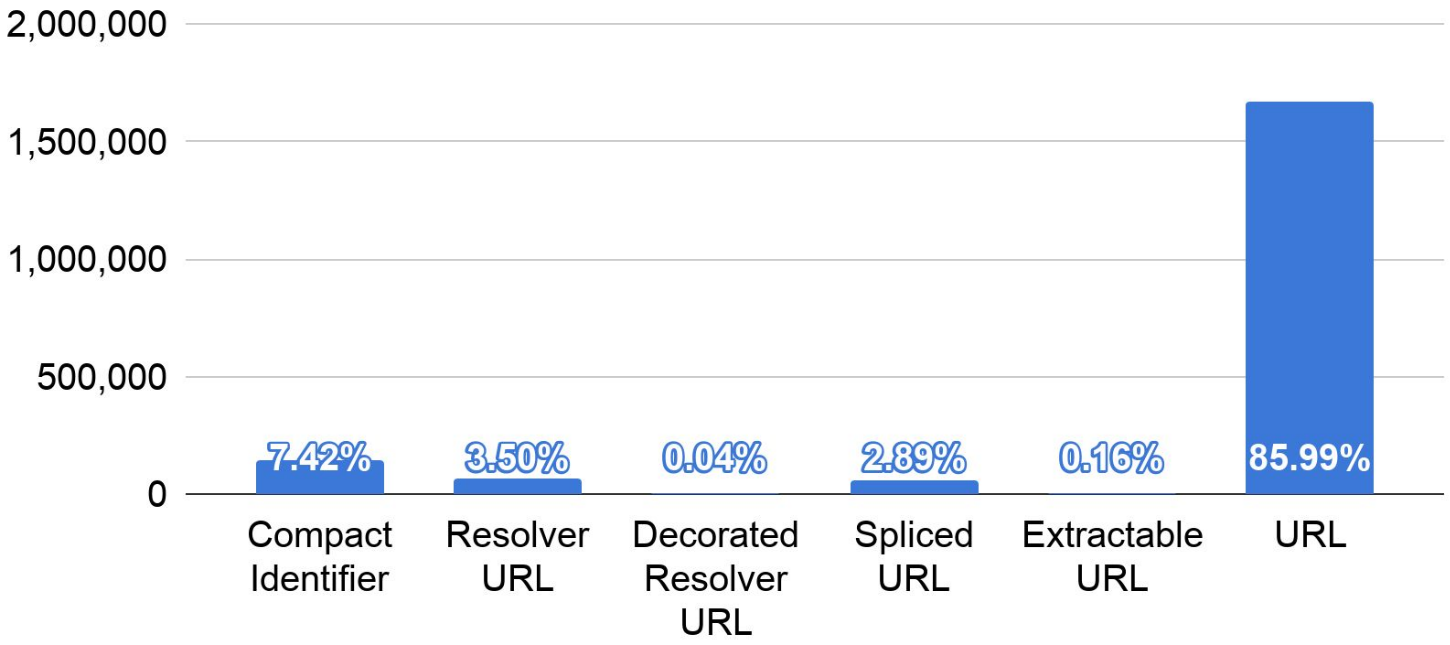
1. Compact identifier
2. Resolver URL
3. Decorated resolver URL
4. Spliced URL
5. Extractable URL
6. URL

https://doi.org/10.6084/m9.figshare.9970265

source-dataset:wikimedia



source-dataset:wikimedia language:eng

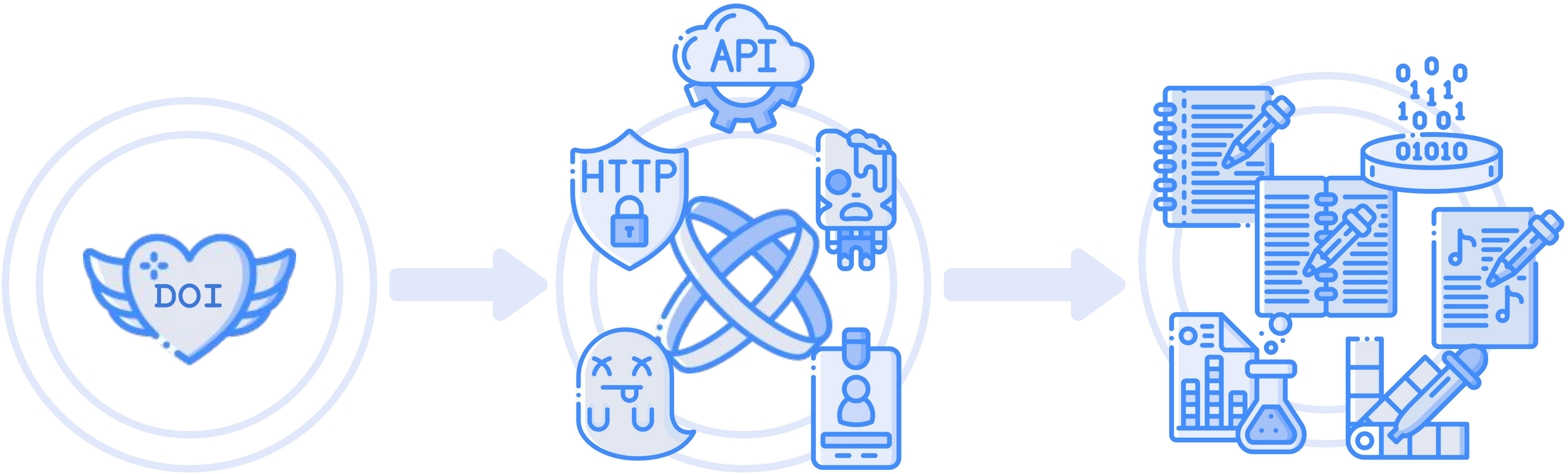


URI Transmutation

DOI
Query
01010
Citation Index

https://www.flaticon.com/authors/smashicons

URI Transmutation



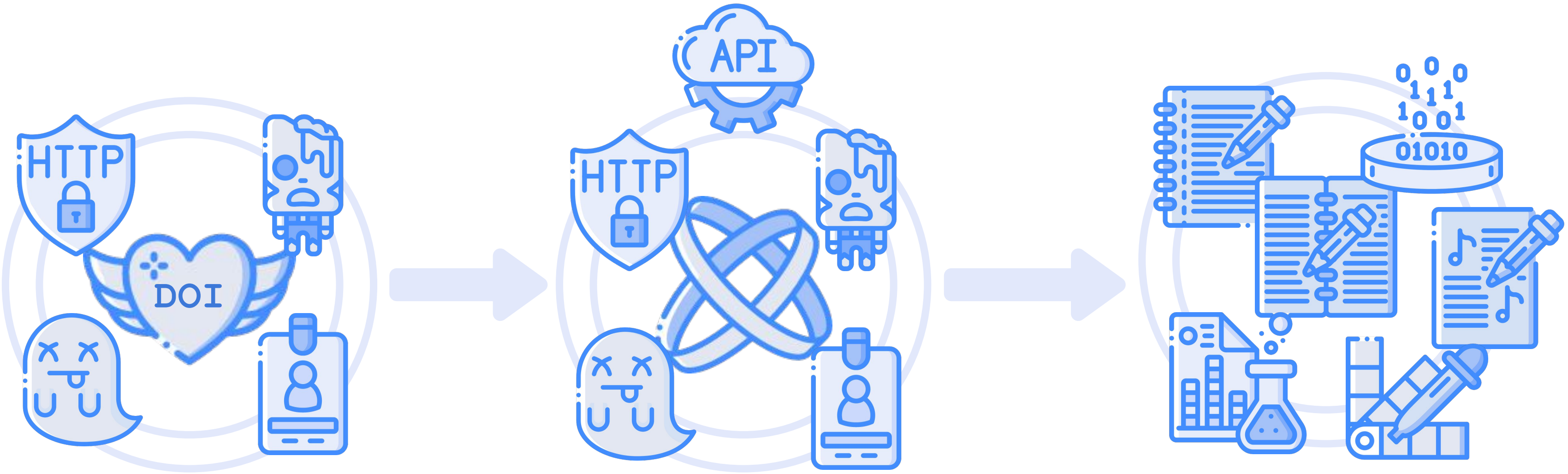
Query

Transmutation API

Citation Index

https://www.flaticon.com/authors/smashicons

URI Transmutation



Query

Transmutation API

Citation Index

https://www.flaticon.com/authors/smashicons

URI Transmutation

Transmutation = normalization + conversion

- Equivalencies we can compute (e.g. ORCID⇄ISNI)
- Equivalencies we must learn (e.g. short URL⇄URL)

Our transmutation API is open and free, let’s try it out!

Cobaltmetrics

Cobaltmetrics

https://cobaltmetrics.com/

Email: latmetrics@thunken.com
Password: 2latmetrics

Y'all got any more of that?

https://cobaltmetrics.com/docs/api

Let's Make Altmetrics Genuinely Alt–!

Let's Make Altmetrics Genuinely Alt-!

- Diversity of **sources**
  - Are there parts of the web that we haven't looked at yet? Or types of documents that we still ignore?
- Diversity of **voices**
  - Are there communities that are still underserved by altmetrics providers?

Final Remarks

Cobaltmetrics

Cobaltmetrics is built based on **community feedback**, customer demand, market observation, and our own ideas.

If you think your community is underserved by Cobaltmetrics, or if you think there are values that we need to develop, please reach out and **contribute to our open roadmap**.

https://medium.com/thunken/90b9faae912d

6:AM Altmetrics Providers Survey

https://tinyurl.com/6am-panel

- What is the next big thing in altmetrics?
- How can we make the data richer?
- Which communities are still underserved by altmetrics?
- Which values do altmetrics providers need to develop?

Helsinki Initiative on Multilingualism

1. Support dissemination of research results for the full benefit of the society
2. Protect national infrastructures for publishing locally relevant research
3. Promote language diversity in research assessment, evaluation, and funding systems

https://www.helsinki-initiative.org/

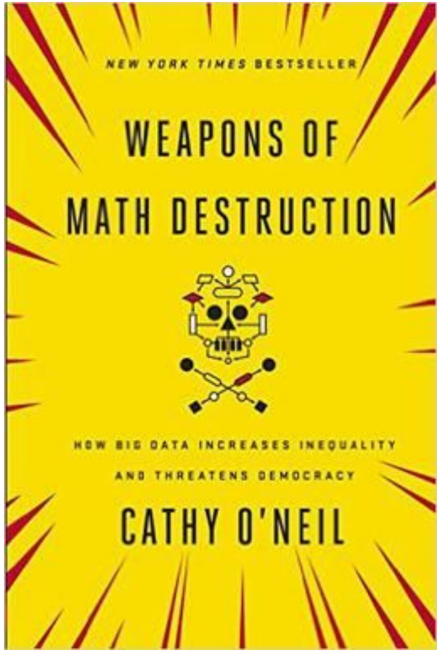
Conclusion

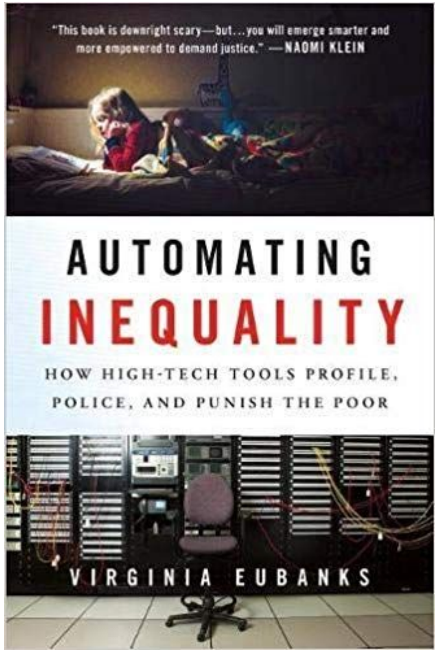
Imbalanced datasets reinforce discrimination.

Algorithmic bias reflects the values of the humans involved in designing the algorithm and/or collecting the data.

It is not up to citation aggregators to decide what is citable.

Recommended readings


NEW YORK TIMES BESTSELLER
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL


"This book is downright scary—but...you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN
AUTOMATING INEQUALITY
HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR
VIRGINIA EUBANKS

THUNKEN

cobaltmetrics.com