Digital Echoes:
Understanding Patterns of Mass Violence with Data and Statistics

Patrick Ball
Director of Research, HRDAG
November, 2023

CDHES

MONUSCO
United Nations Organization Stabilization Mission in the Democratic Republic of the Congo

EAST · WEST MANAGEMENT INSTITUTE

CAMPAIGN for GOOD GOVERNANCE

AHPN
Archivo Histórico de la Policía Nacional
-Guatemala-

ABA CEELI
Promoting the Rule of Law

ICTJ Justice Truth Dignity

Syrian Network For Human Rights
الشبكة السورية لحقوق الإنسان

GUATEMALA MEMORIA DEL SILENCIO
TZ'INIL NA'TAB'AL
CEH

CIIDH
Centro Internacional para Investigaciones en Derechos Humanos

United Nations Human Rights
Office of the High Commissioner for Human Rights

TRUTH AND RECONCILIATION COMMISSION
TRC
TRUTH, JUSTICE AND RECONCILIATION

FAFG
Fundación de Antropología Forense de Guatemala

Ministerio Público
Ciencia • Verdad • Justicia

Comisión de la verdad y reconciliación

AAAS

KOMISYON NASYONAL VERITE AK JISTIS
COMMISSION NATIONALE DE VERITE ET DE JUSTICE
KOMISYON NATIONAL VERITE AK JISTIS
COMMISSION NATIONALE DE VERITE ET DE JUSTICE

HUMAN RIGHTS CENTER
UC Berkeley School of Law | Pursuing justice through science and law

Comisión Colombiana de Juristas

HURIDOCS
HUMAN RIGHTS INFORMATION AND DOCUMENTATION SYSTEMS

INTERNATIONAL CRIMINAL TRIBUNAL FOR THE FORMER YUGOSLAVIA
UNITED NATIONS


Security Force Monitor
UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER
DATACÍVICA
Más datos para más personas
INVISIBLE
INSTITUTE
HUMAN
RIGHTS
WATCH
ITJP
AMNESTY
INTERNATIONAL
COMISSÃO DE ACOLHIMENTO VERDADE E RECONCILIAÇÃO
CAVR
TIMOR LESTE
Carnegie
Mellon
University
CHRS
UNIVERSIDAD
IBEROAMERICANA
Comisión
de la
verdad
y
reconciliación
WITNESS TO
TRUTH:
Report of the
Sierra Leone
Truth & Reconciliation
Commission
equitas
Ensaaf
END IMPUNITY. ACHIEVE JUSTICE.
benetech
المركز السوري للإحصاء و البحوث
Syrian Center for Statistics and Research e.V.
THE
LEGAL AID
SOCIETY
CAE
Dakar, Sénégal
Chambres Africaines
Extraordinaires
FOUNDATION FOR
HUMAN RIGHTS
truth
reconciliation
commission
TRUTH
THE ROAD TO
RECONCILIATION
The National Security Archive
The George Washington University

Post-truth?

That's nothing new for human rights activists

People and institutions that commit mass violence nearly always lie about it. The lies are often grotesque and easily disproven.

Human rights campaigns are successful because we're persistent, we speak with the moral authority of the victims, and because we're committed to the truth.

In human rights, we speak truth to power — so it better be true.

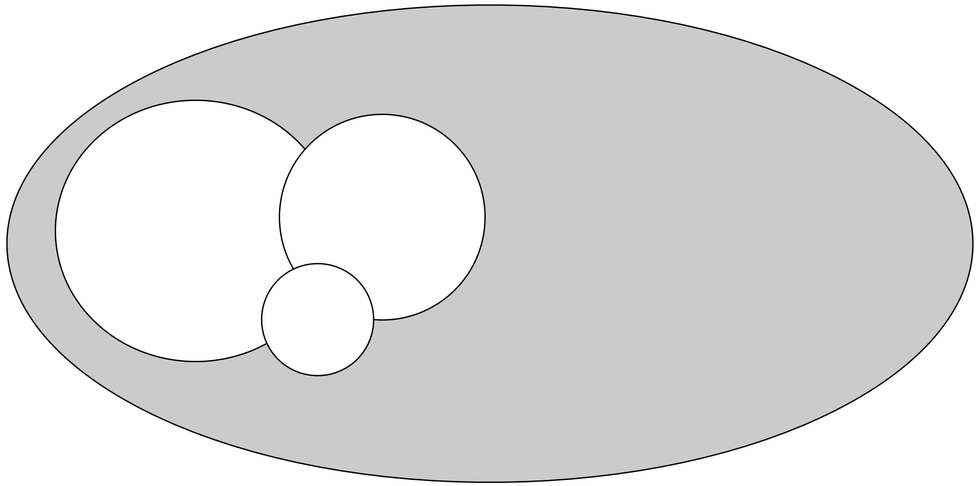
Statistics should be a footnote -- but **the statistics must be right.**

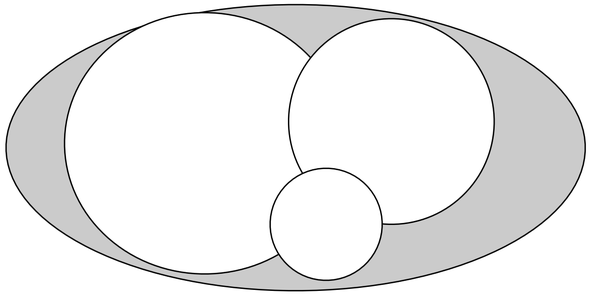
In human rights data collection,
we (usually) don't know what we don't know

In human rights data collection,
we (usually) don't know what we don't know

Which means that we don't know if what we don't know is systematically different from what we do know.

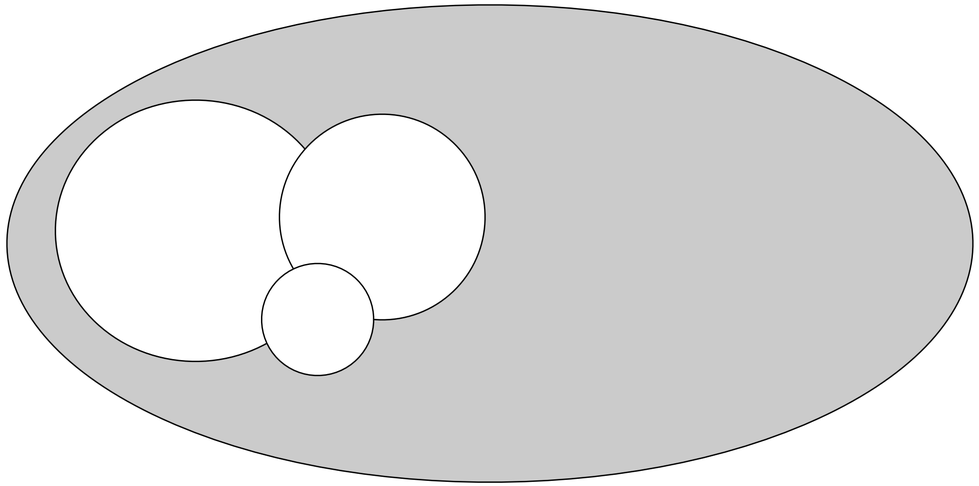
Imagine that you collect and combine three databases:



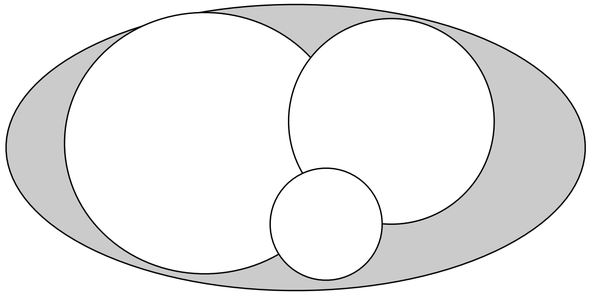


Which picture is accurate?

What is it that we *don't know*?



Sendero Luminoso?



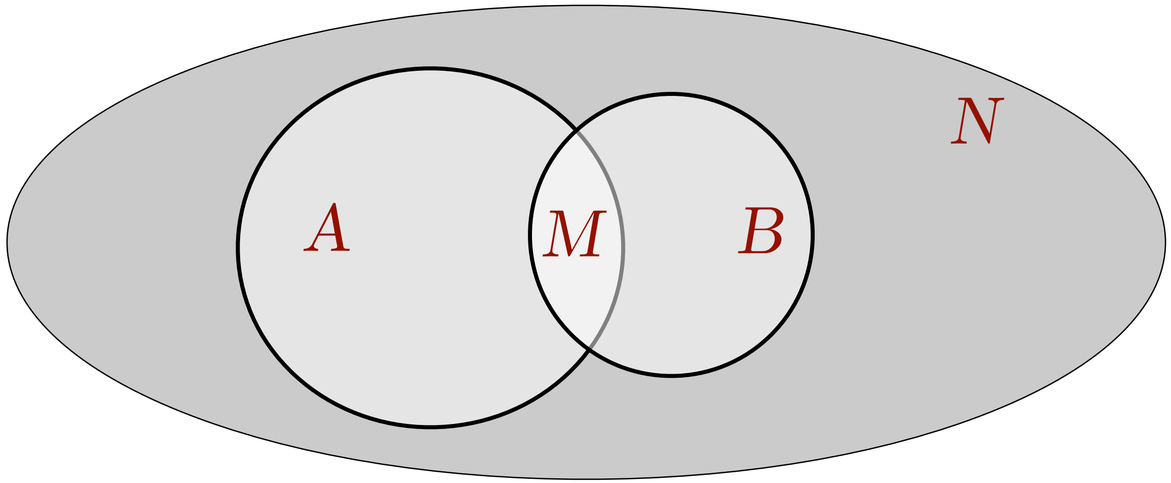
Peruvian Army?

The relationship between <u>what is observed</u> (the sample) and <u>what is true</u> (the population) can be very complicated. Only a mathematical, probability-based model can bridge the gap.

An introduction to MSE

$$p(A) = \frac{A}{N}$$

$$p(B) = \frac{B}{N}$$



$$p(M) = \frac{M}{N} = \frac{A}{N} \cdot \frac{B}{N}$$

$$MN = AB$$

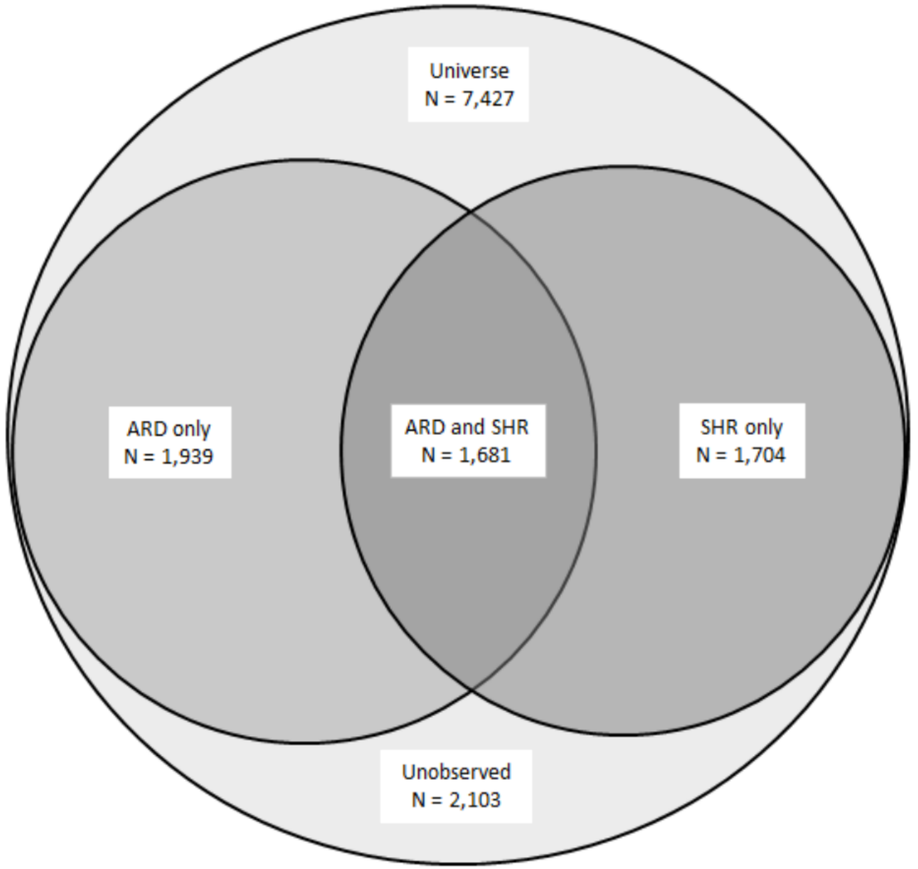
$$\hat{N} = \frac{AB}{M}$$

There are decades of development on this method: see Lum et al. (2013), *The American Statistician*, 67(4):191-200.

Police homicides in the US

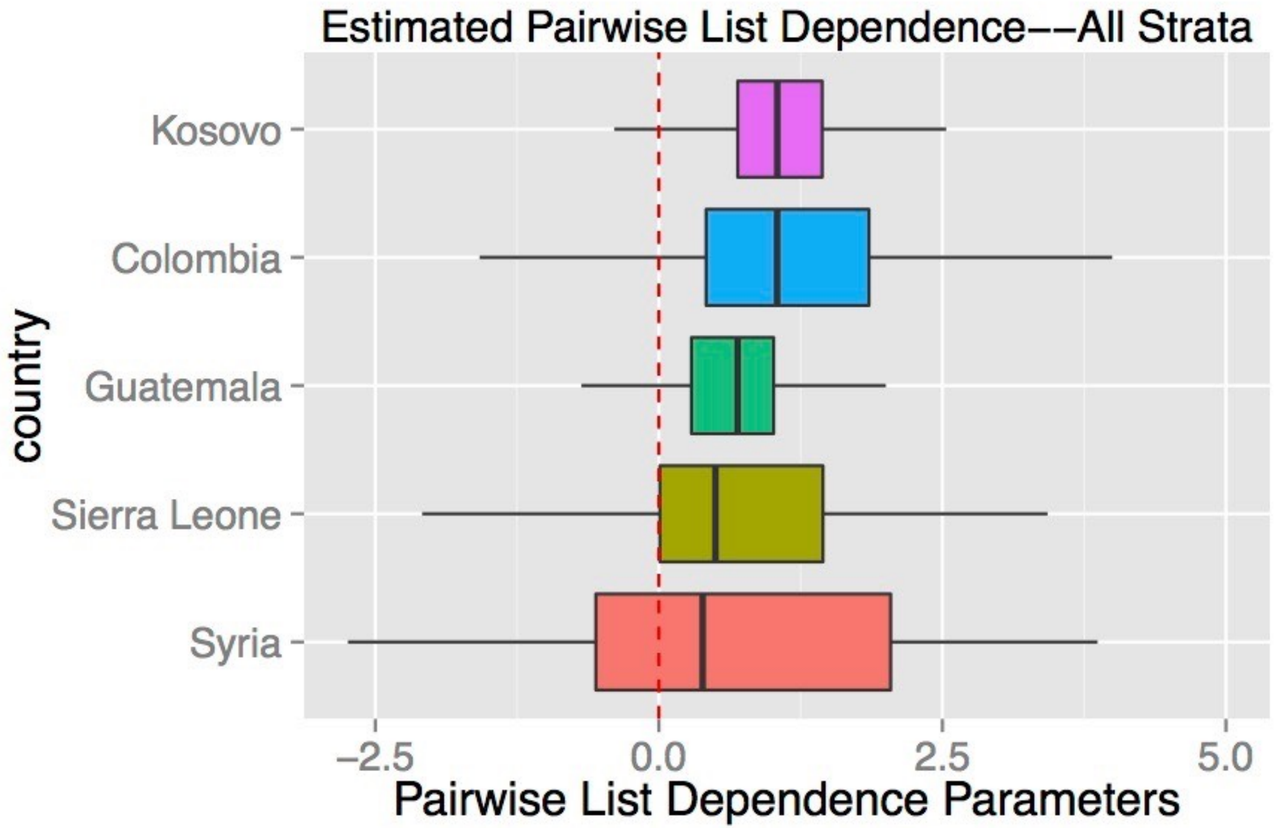
- US Bureau of Justice Statistics used MSE to estimate police homicides.
- Two datasets: ARD (by the Bureau of Justice Statistics) and the SHR (by the FBI)
- Covering 2003-2009 and 2011 (what happened to 2010?)
- Estimate assumes that the part in the middle is independent of the parts on the sides: not true.

Figure 1: ARD and SHR coverage and overlap of the universe of law enforcement homicides in the United States, with no agency adjustment, 2003–09 and 2011



Police homicides in the US

- The BJS estimate was flawed by assuming list independence.
- What list dependence do we see in other countries?

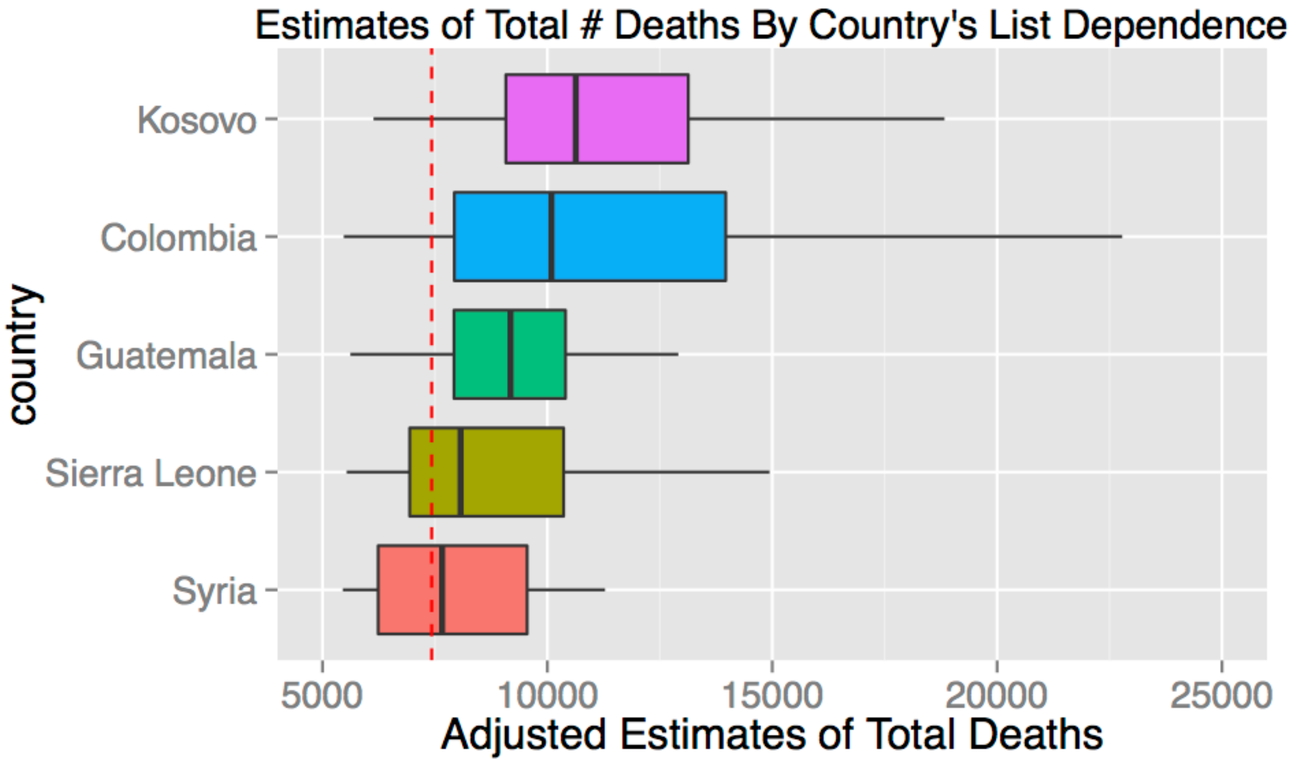


Police homicides in the US

- If list dependence in the US is like list dependence in other countries, what is the total estimated number of police homicides?

- (Colombia is the context most like the US in terms of the lists available)

- The key question in the US is magnitude.



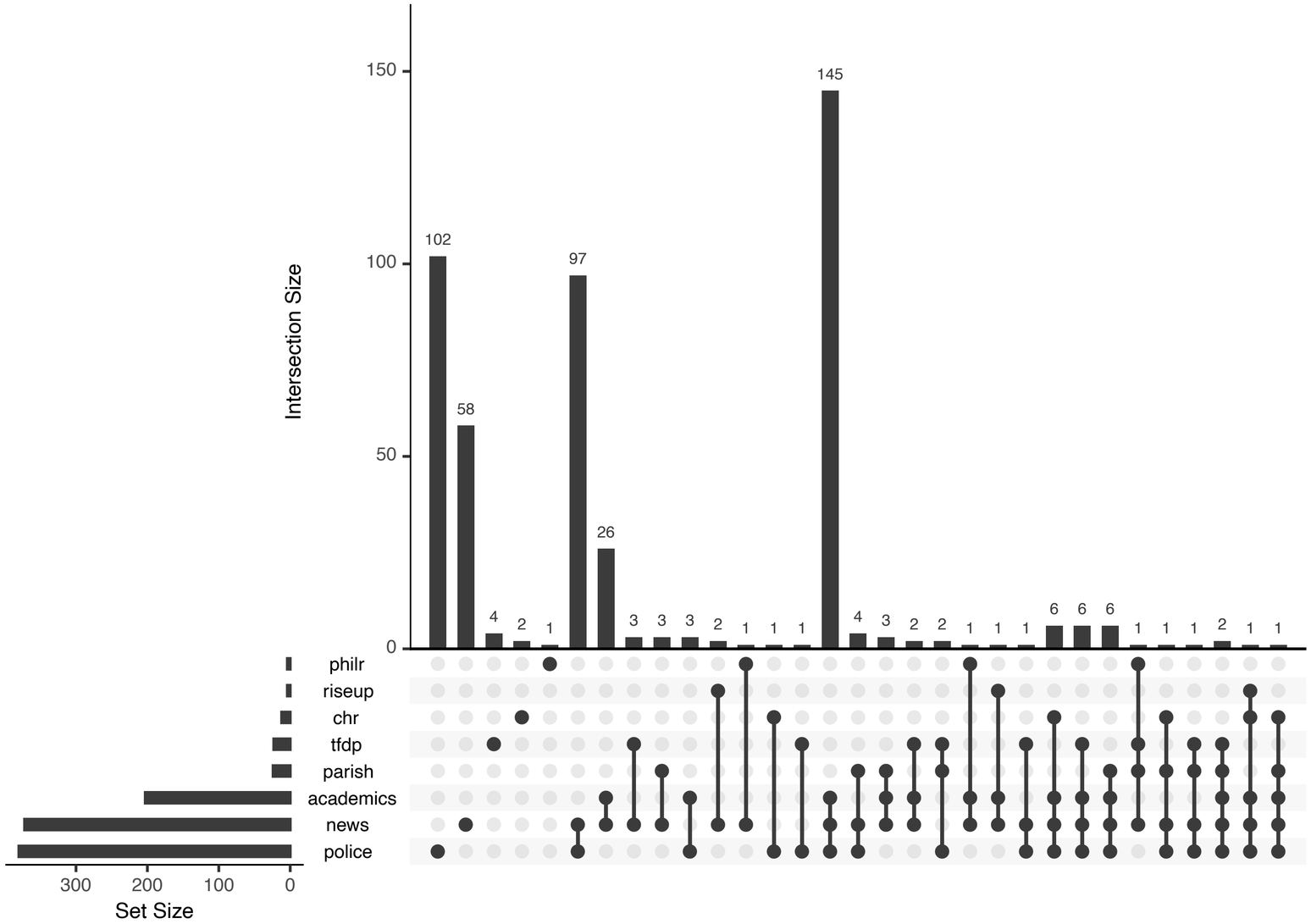
Police killings in the Philippines (2019)

- Project with faculty and students at the Stabile Center for Investigative Journalism at Columbia University
- Data from eight sources (two sources were themselves integrated from more underlying sources)

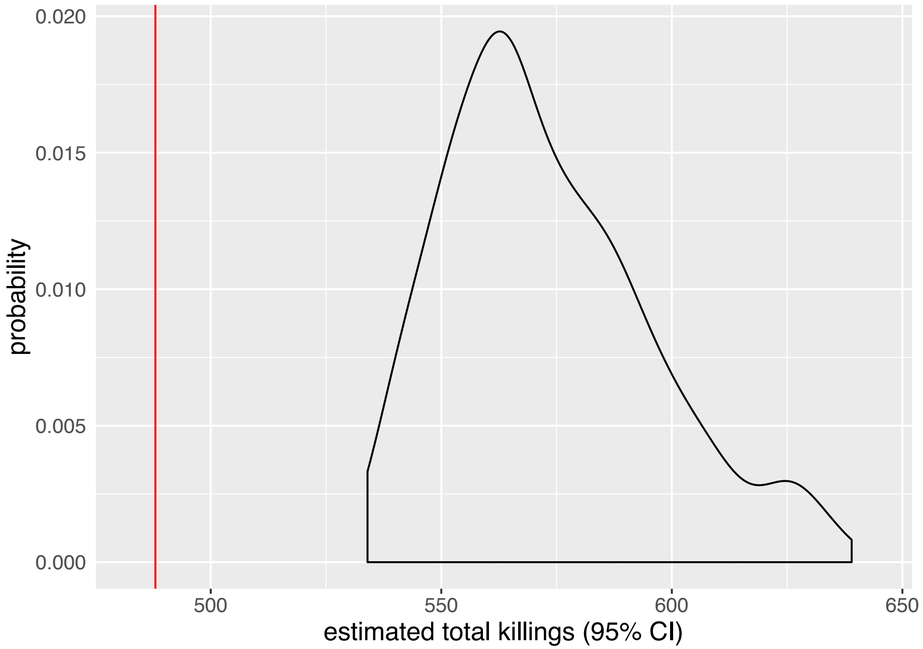
STABILE CENTER
FOR INVESTIGATIVE JOURNALISM
Columbia Journalism School

- Estimates for Manila, Quezon City, and Caloocan covering July 2016 through December 2017
- Main outcome was long-form article in *The Atlantic* magazine; widely discussed in the Philippines, including among legislators
- Nonetheless, violence by police remains popular

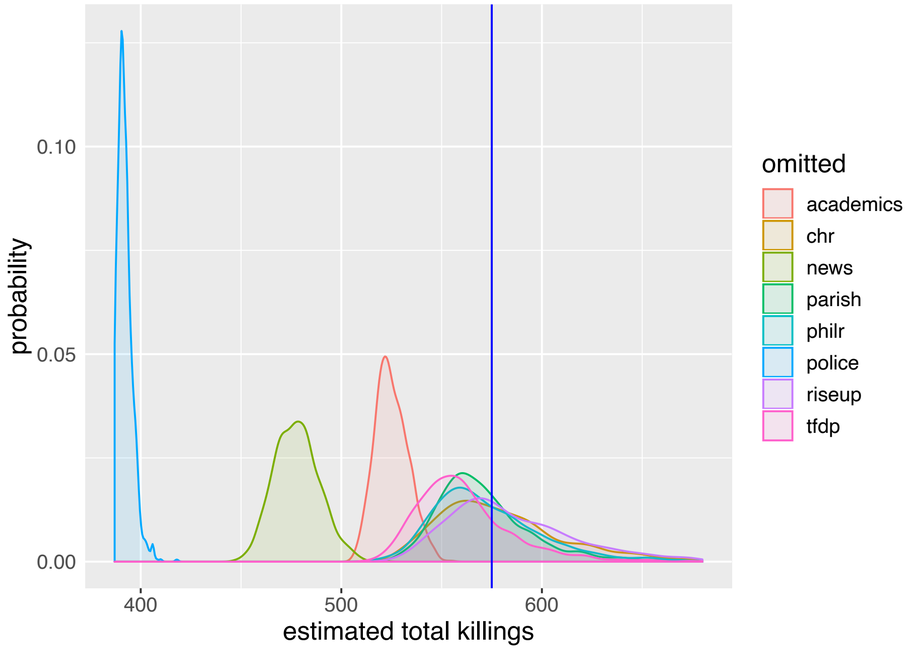
Data from Manila (police only)



Data from Manila (police only)



- density graph is 95% credible interval
- capture heterogeneity is apparent in graph showing estimates with omitted sources
- note: bias tends to be downward



Overall results

Table 2: Documented, Reported, and Estimated, Killings in Three Cities

Region	Perpetrator	Documented	Reported by Police	Estimated	Est/Police
Manila	police	488	313	575	1.8
Manila	unidentified	245	158	414	2.6
Quezon City	police	337	245	348	1.4
Quezon City	unidentified	357	24	428	17.8
Caloocan	police	286	206	331	1.6
Caloocan	unidentified	599	19	745	39.2

- Murders of low-level drug users and sellers were encouraged by commanders and the government
- violence by "unidentified perpetrators" tends to be unreported by police, sometimes entirely
- "unidentified" may just mean violence by police that officers chose not to report to supervisors: with the delegation of violence, there is **always** a principal-agent problem!

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

- https://hrdag.org/2022/09/20/colombia-100-databases/
- Disponible en la página web de la CEV. Licencia CC BY-SA 4.0.

HRDAG
Human Rights Data Analysis Group

Nuestros insumos iniciales

```
pball@sirvjnr:~/projects/CO-S2
 $ find individual/ -type f | grep 'import/output' | grep parquet
individual/JEP/import/output/gaov.parquet
individual/JEP/import/output/sigi.parquet
individual/JEP/import/output/up-caso06.parquet
individual/JEP/import/output/vfarc.parquet
individual/INML/import/output/inml-fatales.parquet
individual/INML/import/output/inml-legacy.parquet
individual/INML/import/output/inml-desaparecidos.parquet
individual/FUNDACION_LAZOS_DE_DIGNIDAD/import/output/excombatientes_familiares_asesin
ados.parquet
individual/FUNDACION_LAZOS_DE_DIGNIDAD/import/output/fundacion_lazos_dignidad.parquet
individual/CEV/import/output/cev.parquet
individual/SOMOS_DEFENSORES/import/output/somos_defensores.parquet
individual/CASO_NORTE_SDER/import/output/caso-norsantander37.parquet
individual/CASO_NORTE_SDER/import/output/caso-norsantander38.parquet
individual/INDEPAZ_CEV/import/output/indepaz_cev.parquet
individual/LIDERES_RELIGIOSOS-CNMH/import/output/lideres_religiosos.parquet
individual/CREDHOS/import/output/credhos.parquet
individual/QUE_FUTURO/import/output/que-futuro.parquet
individual/QUE_FUTURO/import/output/que_futuro.parquet
individual/PGN/import/output/pgn_fallos.parquet
individual/PGN/import/output/pgn_archivos.parquet
individual/PGN/import/output/pgn_activos.parquet
individual/UPH/import/output/uph2-vsx.parquet
individual/UPH/import/output/uph2-f1.parquet
```

Resulta en 125 archivos, 26M de registros
(no todos son víctimas)

HRDAG
Human Rights Data Analysis Group

Sobre el proyecto

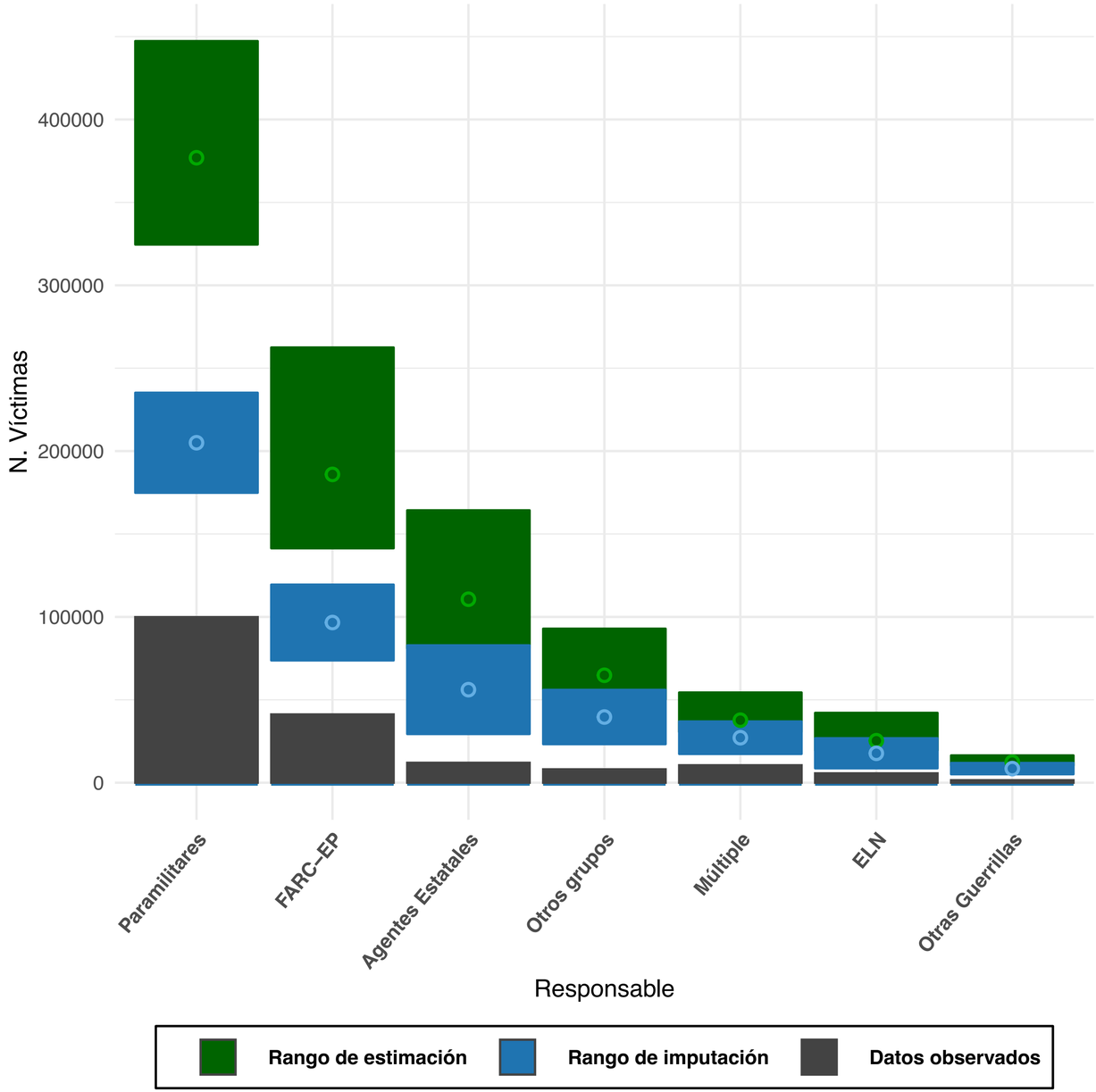
3 años, 25 programadores, 11542 commits en git

Language	archivos	comentario	codigo
YAML	198	244	1347880
R	1695	23480	148694
make	1013	7950	29816
Rmd	85	8138	4489
TeX	32	14	4019
XML	2	0	3394
Python	29	435	2501
Jupyter	21	5501	1309
Julia	7	88	940
bash	9	89	274
Markdown	3	0	67
Lua	1	0	2

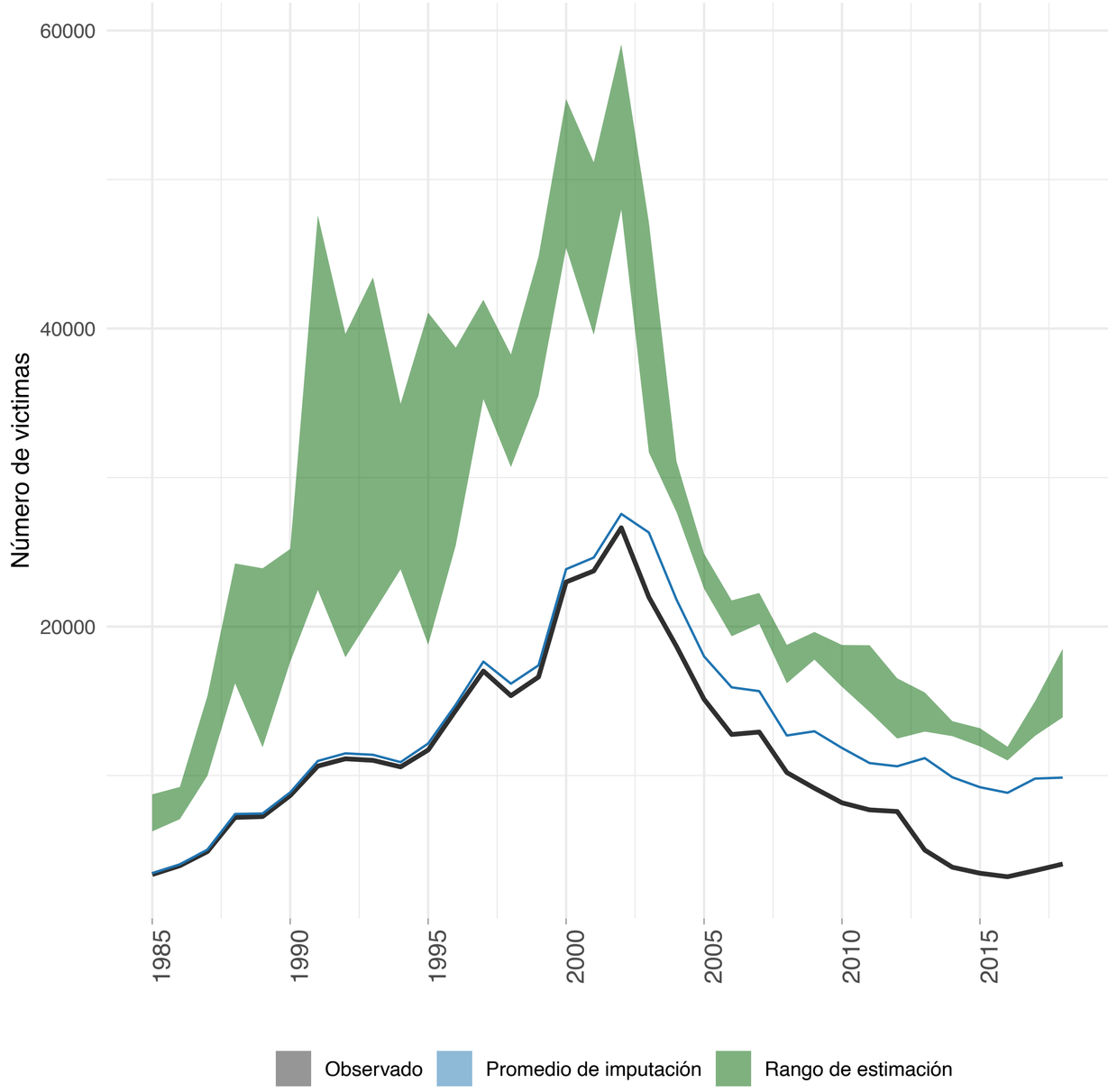
HRDAG
Human Rights Data Analysis Group

Objetivos principales

- Estimar campos faltantes
- Estimar víctimas no observadas
- Propagar la incertidumbre a una única estimación de varianza



HRDAG
Human Rights Data Analysis Group



HRDAG
Human Rights Data Analysis Group

Más detalles en el reporte

HRDAG
Human Rights Data Analysis Group

Three ways to have rigorous statistics (and machine learning predictions, btw)

- A perfect census: if you have all the possible data, you can do anything you like. This is what "big data" *should* mean. There are a very small number of projects in human rights in which we have all the data, but even there, the only proof of complete data is modeling.

- A random sample of the population (or a probability sample of some kind). Very hard to do, and many challenging technical issues.

- Posterior modeling of the sampling process (e.g., capture-recapture, raking). Requires exactly the right data, plus a lot of math and computing capacity.

Questions about human rights statistics

- Do we have all the data? What's missing? Who is excluded, silenced, marginalized? Or by contrast, what groups are overpoliced (i.e., minor offenses treated as serious crimes)?

- If we don't have all the data (and we almost certainly do not), how do we manage the missingness? Can we quantify our uncertainty? (if not, what do we really know?)

- What is the point of a statistic, analysis, or model? Is to to show "how much?" Or are we interested in a pattern? Are all the elements of the pattern equally represented (or estimated) in the analysis, i.e., has the model corrected for -- or amplified and made worse -- selection bias in the data.

Because we *must* be right.

If we get it right:

The accused were sentenced to 40 years in prison.