BigSurv20

BIG DATA MEETS SURVEY SCIENCE

NOVEMBER 6
TO DECEMBER 4
VIRTUALLY
EVERY FRIDAY

www.BigSurv20.org

4. Dec. 2020

Closing Plenary on Data & Privacy

- **Jennifer Hunter Childs**, Assistant Center Chief for Emerging Methods and Applications at the U.S. Census Bureau
- **Dr. Chipo Dendere**, Assistant Professor of Africana Studies at Wellesley College & Chair of Social Science One Africa
- **Dr. Radha Iyengar Plumb**, Director for Trust and Safety Research and Insights at Google
- **Bianca Marcu**, the Senior Advocacy and Standards Programmes Coordinator at ESOMAR, a not-for-profit organisation that promotes the value of market, opinion, and social research and data analytics
- **Dr. Amelia Burke-Garcia**, Program Area Director, NORC at the University of Chicago (moderator)

Not pre-viewable. Tune in to live session to watch this talk.

Sponsored by: European Survey Research Association

ESRA

Privacy/Accuracy Trade-off and the 2020 Census

Jennifer Hunter Childs
Assistant Center Chief, Emerging Methods and Applications
Center for Behavioral Science Methods
U.S. Census Bureau

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

Our Commitment to Privacy and Confidentiality

Honoring privacy and confidentiality is key to public trust, which is critical to our ability to collect and produce high-quality statistics about the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.

Our commitment must extend to threats of tomorrow.

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you “leak” a small amount of private information.

If you release too many statistics, with accuracy, you will eventually reveal the entire underlying confidential data source.

As data proliferates and computers get faster, the risks increase.

Dinur, Irit and Kobbi Nissim (2003) “Revealing Information while Preserving Privacy” PODS, June 9-12, 2003, San Diego, CA

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

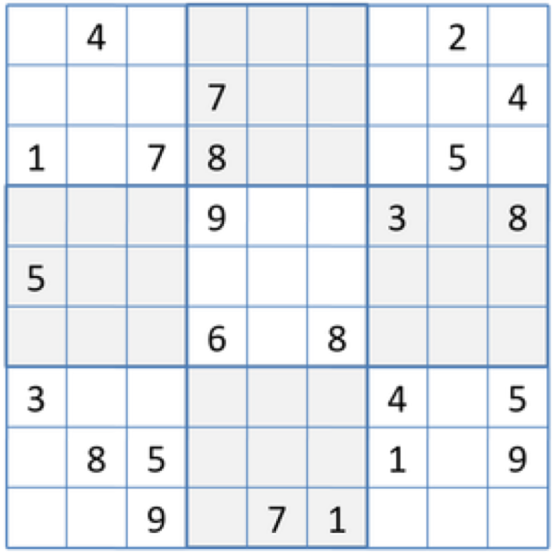
Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.



	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex
Jane Smith	66	Female
Joe Public	84	Male
John Citizen	30	Male

External Data

+

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married

Confidential Data

Reconstructing the 2010 Census

- 2010 Census collected information on the age, sex, race, ethnicity, and relationship status for ~309 million individuals (1.9 billion confidential data points)
- 2010 Census data products released over 150 billion statistics
- Using only a small portion of the 2010 public data products, Census Bureau researchers were able to:
  - accurately **reconstruct** individual-level data for all 6 million inhabited blocks in the U.S.
  - **reconstruct** detailed individual level information, including sex, age (to within a year), race and ethnicity for 71% of the entire U.S. population
  - confirm accurate **re-identifications** on all variables for 52 million individuals with publicly available commercial records.

Differential Privacy

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of “acceptable risk,” and to calibrate where on the privacy/accuracy spectrum the resulting data will be.

Data Quality | Bnae Kegouqe
Dada Qualitg | Vrkk Jzcfkdy
Data Qaality | Dncb PrhvBln
Dzte Qvality | Dncb Prtnavy
Dfha Quapyti | Tgta Ppijacy
Tgta Qucjity | Dfha Pnjvico
Dncb Qhulitn | Dzhe Njivaci
Ntue Quevdto | Dzte Privecy
Vrkk Zuhnvry | Dada Privacg
Bnaq Denorbe | Data Privacy

Establishing a Privacy-loss Budget

This measure is called the “Privacy-loss Budget” (PLB) or “Epsilon.”

- $\varepsilon=0$ (perfect privacy) would result in completely useless data
- $\varepsilon=\infty$ (perfect accuracy) would result in releasing the data in fully identifiable form

Epsilon

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it allows for measurement of the privacy risk.

Implications for the 2020 Census

The switch to Differential Privacy does not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy-loss budget (epsilon).

The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) will be making decisions about the privacy-loss budget for the 2020. This includes allocation across different 2020 Census data products.

Additional Resources

Disclosure Avoidance and the 2020 Census Website

https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

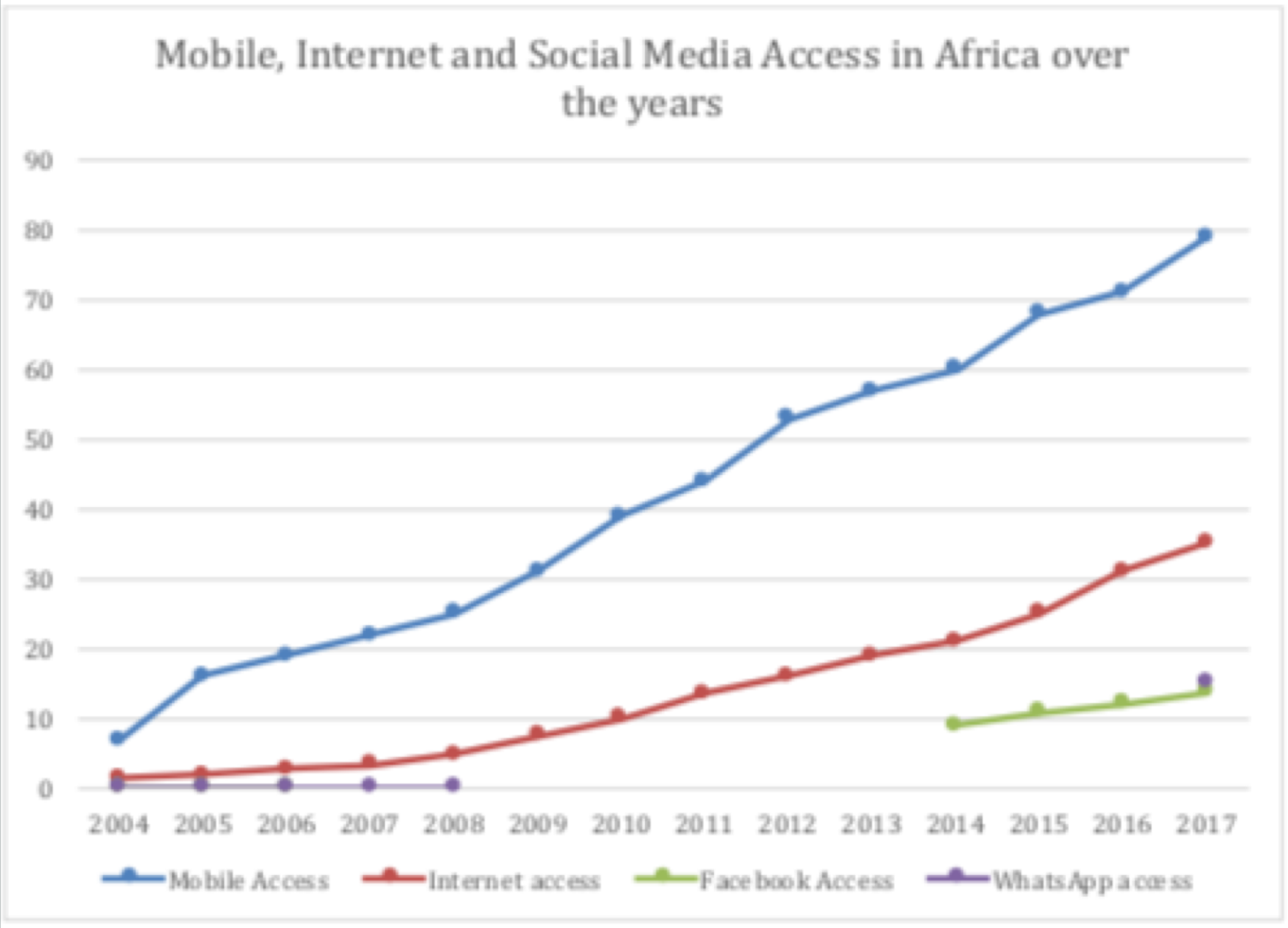
Questions? Suggestions?

Email them to 2020DAS@census.gov

DATA
THE GOOD, BAD & UGLY BEYOND WESTERN BORDERS

Dr. Chipo Dendere

WELLESLEY COLLEGE


Mobile, Internet and Social Media Access in Africa over the years
90
80
70
60
50
40
30
20
10
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Mobile Access
Internet access
Facebook Access
WhatsApp access

INFLUENCE OF SOCIAL MEDIA ON POLITICS

Social media platforms can foster political mobilization and contributing positively to democratic growth

Mobile phones and internet access shield citizens (authoritative regimes) from eyes of the state

The success of #feesmustfall campaign ignited hashtag movements across the African continent

GOVERNMENT RESPONSE TO IMPACT OF TECHNOLOGY

- **Authoritarian governments have developed sophisticated tools to censor online engagement**

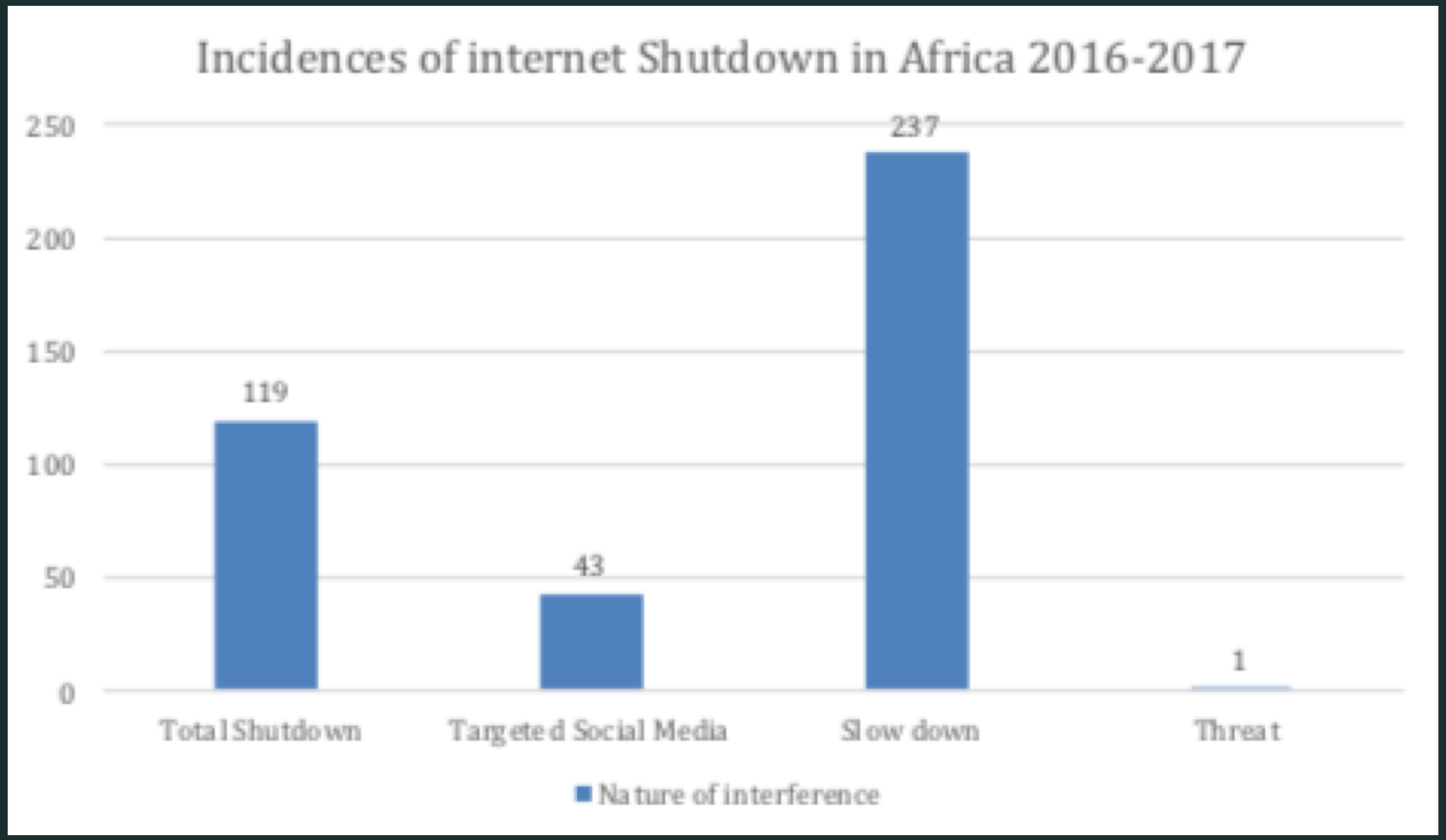
  Co-option

  Hiking internet costs

  Tightening and creating repressive laws to curb social media access

  Complete and partial shutdowns

  - **with not physical violence**
  - **With physical violence (e.g. Zimbabwe 2019, Nigeria 2020, Tanzania 2020)**

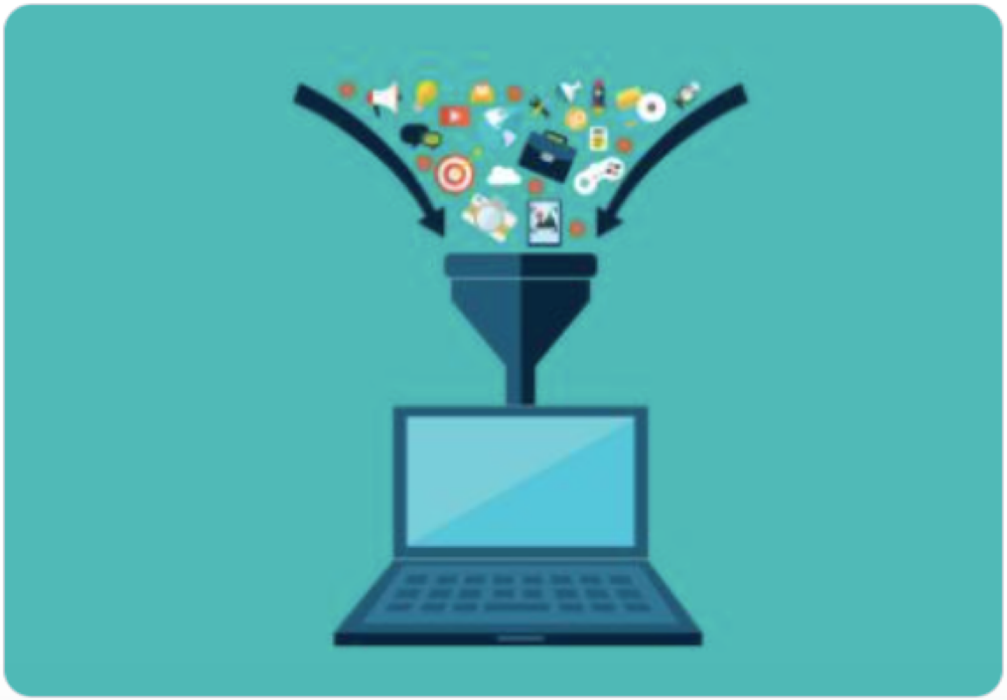

Incidences of internet Shutdown in Africa 2016-2017
250
200
150
100
50
0
119
43
237
1
Total Shutdown
Targeted Social Media
Slow down
Threat
Nature of interference

African Institute for Investigative Journalism
@AfricanIIJ

In 2017, data overtook oil to become the world's most valuable commodity, but it remains hard to define.

In Europe, 96% of countries currently have data protection laws in place, only 50% in Africa.

#AIIJ



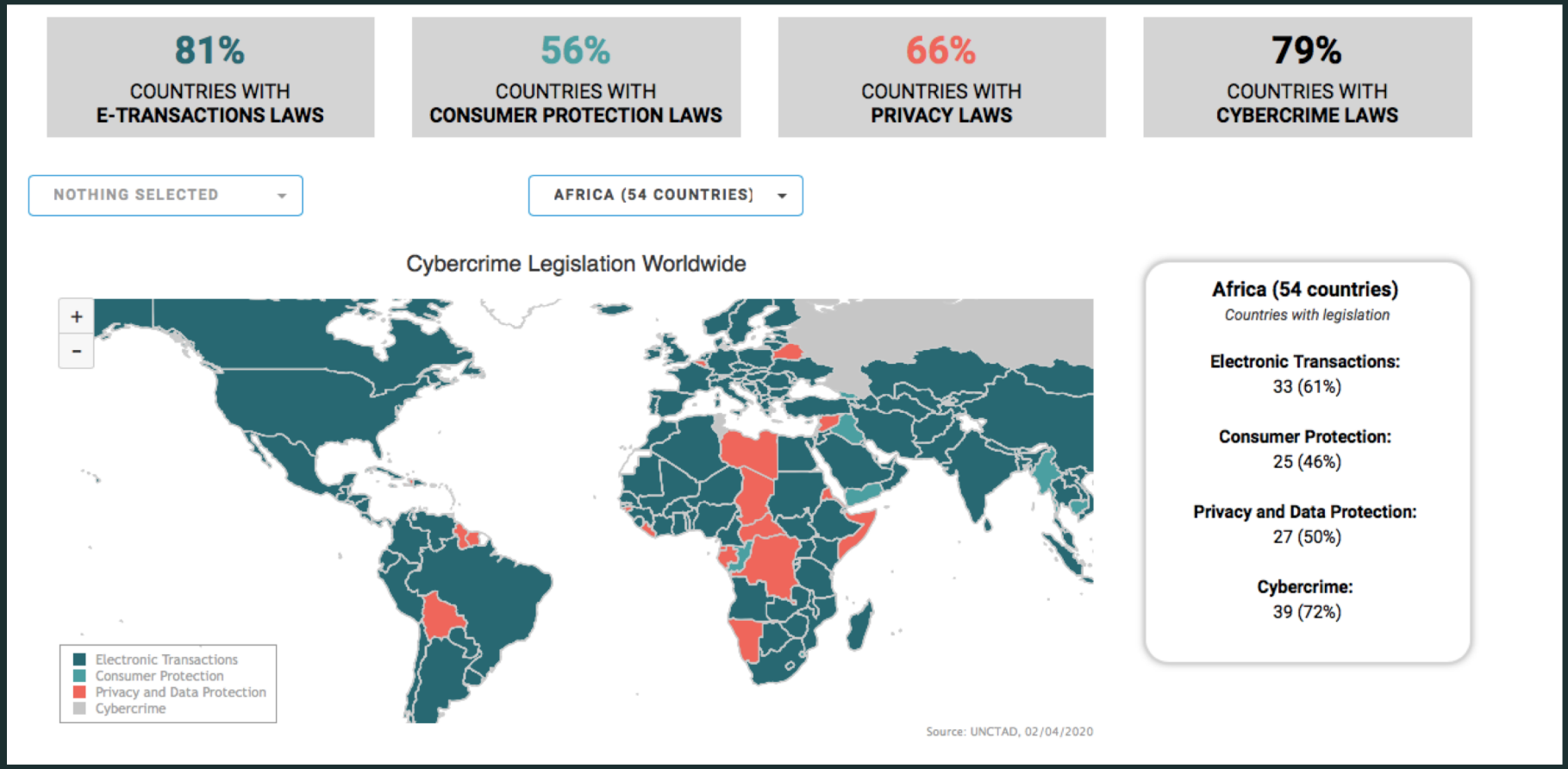
AFRICA IS GROUND ZERO FOR DATA COLONIALISM

- **Citizen data is the most sought-after natural resource**

- **Africans are used as testing ground for global technologies - providing financial benefit to the government at the expense of citizens**

- **Elections in Africa are vulnerable to data manipulation- nearly every election is using biometric data systems run by for profit companies**

CYBER-CRIME LEGISLATION



By 2020, 24 African countries, out of 53, adopted laws and regulations to protect personal data, and the number is slowly rising. The 2016 EU General Data Protection Regulation stands as a model for many.

WHERE DO WE STAND

- Poverty as a threat to cyber security

  Governments desperate for money are likely to continue selling citizen data even as they implement stronger measures to curb access to the internet

SOLUTIONS

- Support for pro-citizen activist groups
- Support for citizen education
- Calling out for-profit data companies
- Support home-grown African innovation companies
- Continent wide policies that push for local data protections – localized data capture and preservation

Research on Data Transparency

Radha Iyengar Plumb

BigSurv Plenary Session

December 2020

Key Questions we'll Discuss today

1. **What is meaningful transparency?**
   Key components and how they fit together

2. **How to meet the needs of different use cases?**
   Balancing different audiences to meet policy making, research, and user needs

3. **Why is it hard?**
   Inherent tensions and trade-offs in being transparent with data

Approach presented here: Summary findings from research (legal, social science, HCI) and analysis across industries (tech, health care, accounting, finance) assess key components, use cases, and tradeoffs

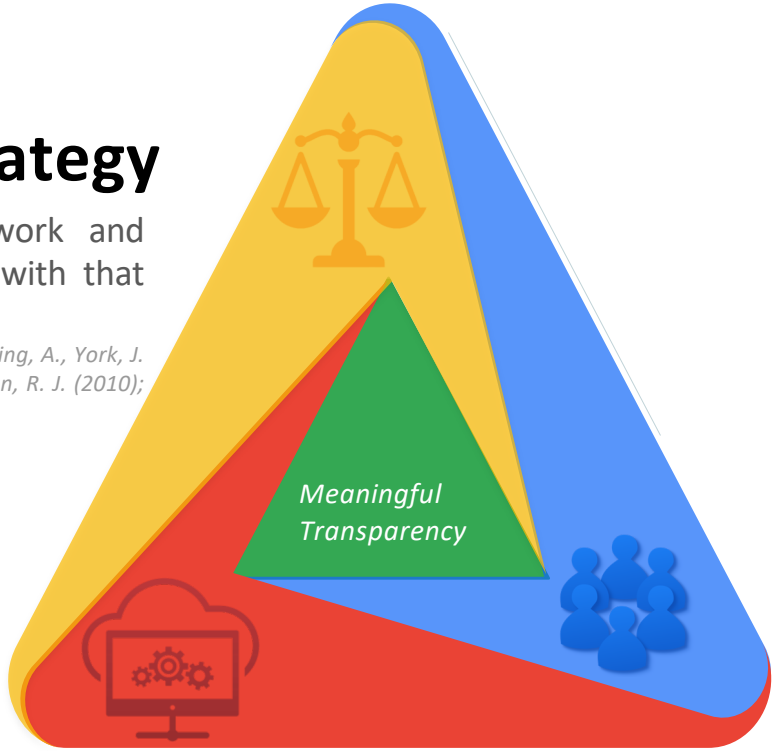
What is meaningful transparency?

Key components and how they fit together based on existing research

Principles & Strategy

A clear, consistent framework and specific outputs associated with that approach.

References: Suzor, N.P., West, S. W., Quodling, A., York, J. (2019); Granados, N., Gupta, A. & Kauffman, R. J. (2010); McGee and Javenta (2010);



External Engagement

Develop and deliver content that explains policies, products, and processes tailored for specific interests and concerns of policy makers, journalists, researchers, civil society groups, and general users.

References: Tworek, 2019; McBride (2014); Turilli and Floridi (2009)

Technical Information

Accessible, relevant information that provides precise, consistent, and reliable explanations and quantification of policies, products, and processes.

References: Bradford, et al (2019); Cho(2009); Karr (2008); Santa Clara Principles (2020);

How to meet the needs of use cases?

Balancing different audience to support policy making, research, user needs

Analysis of transparency approaches across industries and government approaches combines diversity of Information with different delivery modes

Deep Dives

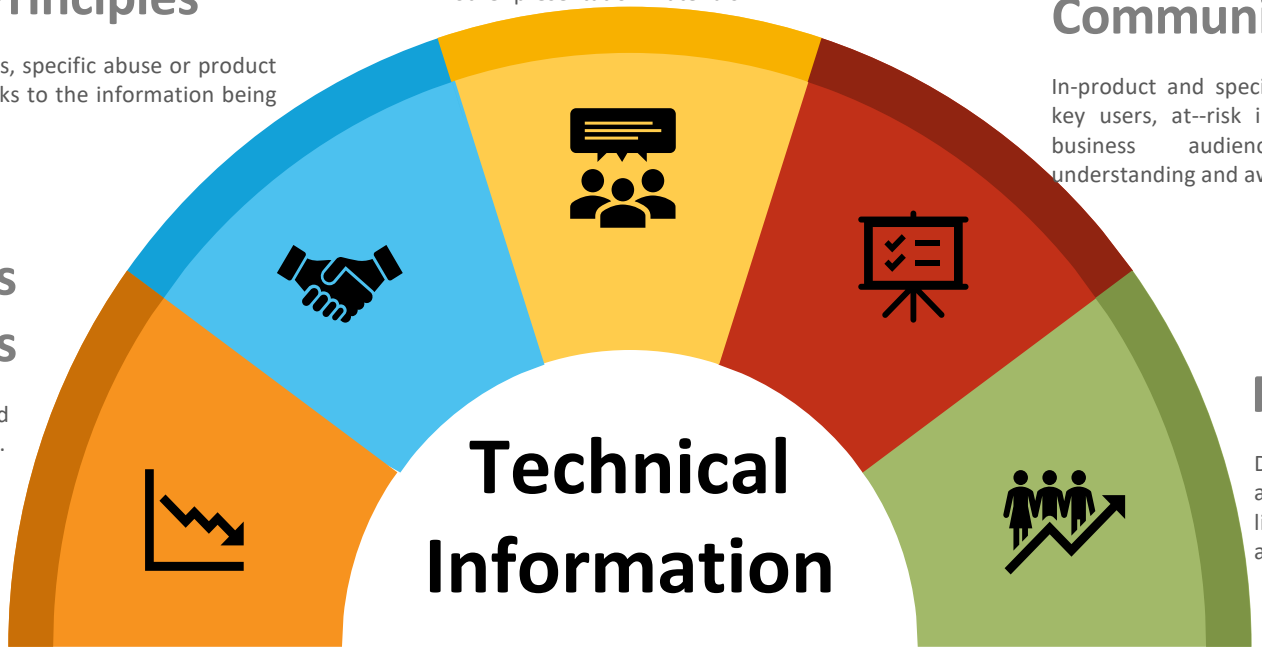
Additional details provided to key audiences including presentation, data access, and other presentation materials

Policies & Principles

Clarity on policy principles, specific abuse or product policies, and how that links to the information being shared.

Direct Communications

In-product and specific communications to key users, at--risk individuals, or broader business audiences to increase understanding and awareness



External Metrics & Reports

Reports and metrics published externally and is publicly available.

Process Maps

Descriptions and graphics on how decisions and operations are enacted with clear linkages to data sourcing and generation and metric relevance.

TRUST & SAFETY

Why is it hard?

Inherent tensions and trade-offs in being transparent with data

Key issues identified in industry and academic research on data transparency

- Comprehensive data and metrics that are still comprehensible and accessible
- Granularity to ensure it is useful but still ensures the protection of private/sensitive data
- Meaningful details about process but doesn't enable bad actors to abuse platforms.
- Being clear around different level of detail delivered for metrics and data based on needs of different audiences
- Providing timely access to metrics and information without lowering data and information quality
- Delivering ongoing consistent iterative metrics that can also adapt to changing circumstances and innovation in approaches.

References

Cho, Charles H., et al. “Media Richness, User Trust, and Perceptions of Corporate Social Responsibility.” *Accounting, Auditing & Accountability Journal*, vol. 22, no. 6, Emerald Group Publishing, Limited, 2009, p. 933.

Bradford, Ben, et al. *Report Of The Facebook Data Transparency Advisory Group*. 2019.

Granados, N., Gupta, A. & Kauffman, R. J. (2010). Research Commentary: Information Transparency in Business-to-Consumer Markets: Concepts, Framework and Research Agenda. Information Systems Research, 21(2), pp. 207-226

Karr, Alan F. "Citizen access to government statistical information." *Digital Government*. Springer, Boston, MA, 2008. 503-529.

McGee, Rosemary, and J. Javenta. *Synthesis Report: Review of Impact and Effectiveness of Transparency and Accountability Initiatives*. 2010, https://www.transparency-initiative.org/blog/607/synthesis-report-review-of-impact-and-effectiveness-of-transparency-and-accountability-initiatives/.

Kenneth McBride, Neil. “ACTIVE Ethics: An Information Systems Ethics for the Internet Age.” Journal of Information, Communication and Ethics in Society, vol. 12, no. 1, Emerald Group Publishing Limited, Mar. 2014, pp. 21–44, doi:10.1108/JICES-06-2013-0017.

Suzor, N.P., West, S. W., Quodling, A., York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication 13 (2019), 1526-1523.*

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105-112.

Tworek, Heidi. “How Transparency Reporting Could Incentivize Irresponsible Content Moderation.” *Centre for International Governance Innovation*, 2019, https://www.cigionline.org/articles/how-transparen

ESOMAR

ESOMAR is the global voice of the data, research and insights community, since 1947

ESOMAR

A Quick Introduction

ESOMAR is the International Associations for Market, Opinion and Social Research and Data Analytics since 1947

Amsterdam, the Netherlands

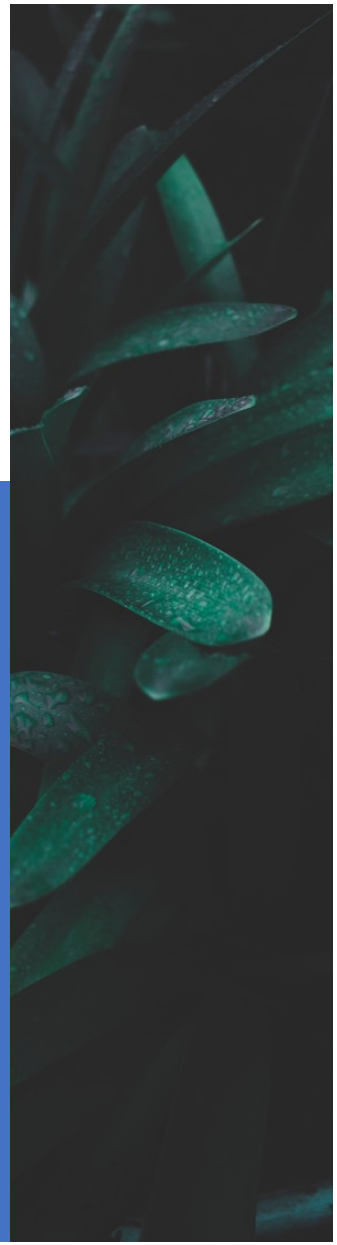
One of the biggest challenges is not only for a global industry to agree on and undersign to abide by specific ethical principles, but also to find ways in which those principles come to life in an organisation

A starting point for the sector: the ICC/ESOMAR International Code of Conduct

ESOMAR

The Context

- We live in a data-driven economy
- Laws regulating data use have lagged behind technological developments and innovation giving rise to a growing ‘grey area’
- Ethics can help us better operate in this grey area by pushing us to ask the right questions and rethink our data strategies



ESOMAR

The Key Principles

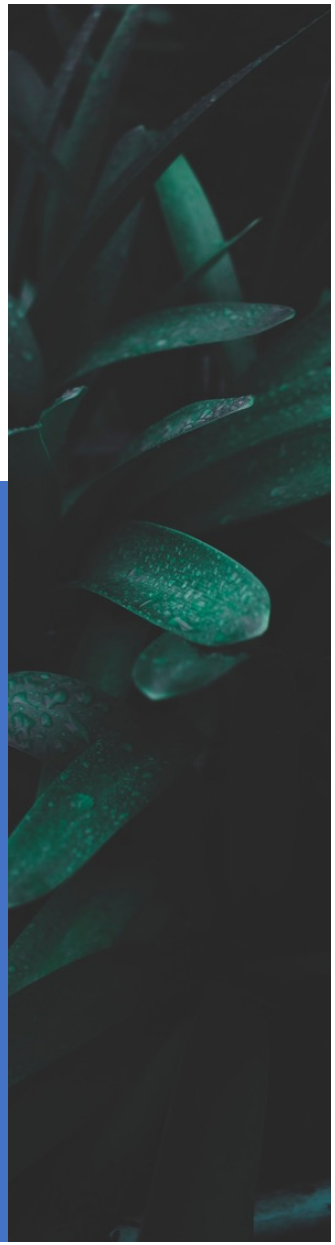
1. **Fairness** → data minimisation and data quality
2. **Respect** → for the people behind the data
3. **Transparency** → making clear the scope of the processing
4. **Accountability** → codify ethics into business decisions

ESOMAR

The Data Counselling Approach

Data Ethics are part of an organisation's overall data strategy

An organisation depends on a **data strategy** that sets the moral compass for the organisation in order to better leverage it.

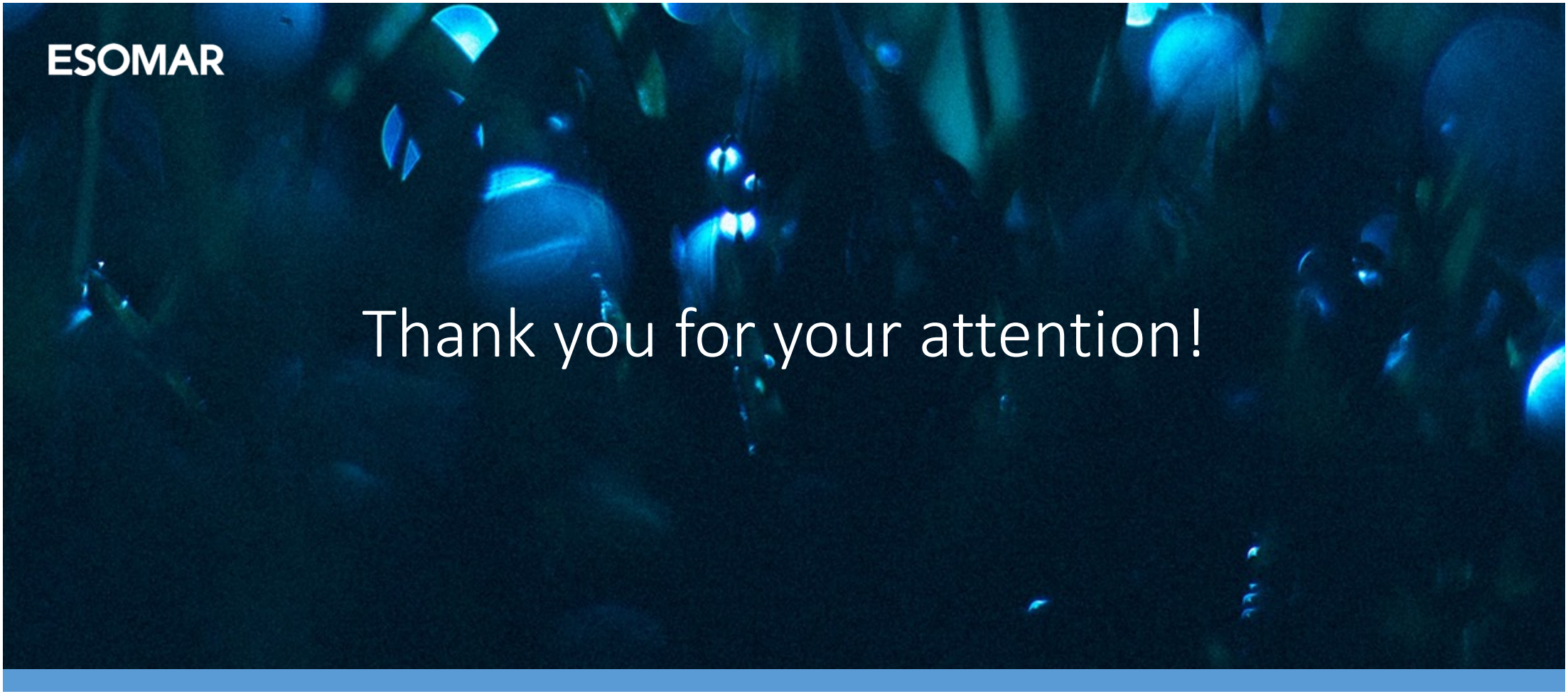


The Data Counselling Approach

ESOMAR

Some final points…

- Taking inspiration from data protection & privacy **laws**
- Embedding a sound data strategy **organisation-wide**
- **Training** and awareness-raising
- **Routine, routine, routine**



Let’s stay in touch at professional.standards@esomar.org

ESOMAR

ESOMAR is the global voice of the data, research and insights community, since 1947

BigSurv20

BIG DATA MEETS SURVEY SCIENCE

NOVEMBER 6
TO DECEMBER 4
VIRTUALLY
EVERY FRIDAY

www.BigSurv20.org

4. Dec. 2020

Panel Discussion

Not pre-viewable. Tune in to live session to watch this talk.

Sponsored by: European Survey Research Association

ESRA