Sensitive Data IG
RDA17 Session Report

Distributed 2021-06-21

This report has been added as a record on Zenodo and is contained within the Zenodo RDA Sensitive Data IG community.

Table of Contents (subheadings are hyperlinked)

Overview 2

Links 2

IG overview, and presentation of the charter: Aleksandra Michalewicz 3

Lightning talks 3

- Defining sensitive data: Romain David 3
- Sensitive data case study - setting up for a sensitive data project: Rita Silva 4
- Sensitive data case study - lessons from working with sensitive data: Amy Pienta 5

Discussion of the charter and/or subtopics: Nichola Burton 6

Where to next with the IG: Steven McEachern 9

Next steps 11

Contributors to this session report. Contributors are listed in alphabetical order using the CRediT Contributor framework.

Name	**CRediT**
Dharma Akmon	Writing – review & editing
Nichola Burton	Writing – review & editing
Romain David	Conceptualization, Writing – original draft, Writing – review & editing
Sarah Davidson	Writing – review & editing
Steven McEachern	Writing – review & editing
Aleksandra Michalewicz	Writing – original draft, Writing – review & editing
Priyanka Pillai	Writing – original draft

Rita Silva	Writing – original draft, Writing – review & editing
Kristal Spreadborough	Writing – original draft, Writing – review & editing, Project administration
Frankie Stevens	Writing – review & editing

Overview

The goal of this document is to provide a high level summary of the RDA17 Sensitive Data Interest Group session. It is designed to complement the session recording by highlighting the next steps and future directions from the session. The structure of this report reflects that of the RDA17 session.

Links

Sensitive Data Interest Group poster	https://zenodo.org/record/4690571
RDA Sensitive Data IG Community on Zenodo	https://zenodo.org/communities/rda-sensitive-data-ig
Link to the outputs from the RDA17 session	https://www.rd-alliance.org/group/sensitive-data-interest-group/outcomes/rda17-plenary-session-establishing-sensitive-data
Link to current charter	https://www.rd-alliance.org/group/sensitive-data-interest-group/case-statement/sensitive-data-interest-group-charter
Link to session recording	https://www.youtube.com/watch?v=AoydTRrYSEE
Link to session page	https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual/establishing-sensitive-data-interest-group
Slides from speakers	https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual/establishing-sensitive-data-interest-group
Notes from the session (including the session Chat and Q and A)	https://docs.google.com/document/d/1vrXl1SN868mgSVLcBvMBvWA_ndaQlYecMpY08Vfg6P8/edit

IG overview, and presentation of the charter: Aleksandra Michalewicz

Slides available here: https://zenodo.org/record/4895641
Introduction by Aleksandra Michalewicz.

The Sensitive Data group met for the first time as a Birds of a Feather at RDA16. After this, the current co-chairs met regularly to formalise the interest group and prepare our first RDA session at RDA17. We are currently working towards formal RDA endorsement. We invite people to join our group, and to provide feedback on the draft charter.

Our working definition of Sensitive Data is:

> "Information that is regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations. Sensitive personal data include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an individual. These data could be identifiable and potentially cause harm through their disclosure. For local and government authorities, sensitive data is related to security (political, diplomatic, military data, biohazard concerns, etc.), environmental risks (nuclear or other sensitive installations, for example) or environmental preservation (habitats, protected fauna or flora, in particular). The sensitive data of a private body concerns in particular strategic elements or elements likely to jeopardise its competitiveness."
> Adapted from: David et al., 2020, "Templates for FAIRness evaluation criteria - RDA-SHARC IG" https://zenodo.org/record/3922069#.YCJU7ehKg2w

Lightning talks

Defining sensitive data: Romain David

Slides available here: https://zenodo.org/record/5002758

This lightning talk presented a short overview of sensitive data in the context of ERINHA - a research institute of biocontainment laboratories which is specialized in infectious disease research.
In general, two main kinds of sensitive data are encountered:

- Environmental sensitive data: Endangering coveted and scarce resources (including relocalisable data)

- Personal data: Endangering persons (including re-identifiable data)

Four main risks are associated with this sensitive data:

- Economic risks
- Interference with security programs / tools
- Misappropriation of knowledge and data (for instance, to build a weapon)
- Terrorism (e.g., sharing information about secure access might encourage terrorists to target hospitals, stadiums, etc.)

The sensitive data produced in this context is vulnerable to dual use - where data that is gathered and used with the intention of benefiting society can also be used maliciously by third parties. Different examples of sensitive data which may be Dual Use Research of Concern were discussed with specific reference to the life sciences. Dual use considerations are an important aspect of sensitive data for this Interest Group to address.

Sensitive data case study - setting up for a sensitive data project: Rita Silva

Slides available here: https://zenodo.org/record/4895645

In this lightning talk, Rita Silva presented her personal experience in setting up the project aMILE - *Application of text mining tools for the study of patients with acute myeloid leukemia* at the Portuguese Institute of Oncology of Porto (IPO-Porto), a project still in an early phase of development, involving sensitive data from the electronic health records (EHRs).

The talk started with an introductory presentation of Rita and IPO-Porto, followed by the reasons that motivated the creation of the project. Then, Rita focused on the planned measures to ensure the data privacy and security, and explored some ethical issues that were considered for the project approval by the local Ethics Committee. This part of the talk originated a posterior discussion regarding the need for informed consent from the patients or their relatives, the compliance with the Declaration of Helsinki, and the patients' opinion about the project.

As a medical doctor, Rita works with health data on a daily basis. Health data corresponds to all the information related to the health status of a person, including medical, administrative and financial information. Health data is sensitive data because it may cause discrimination, harm and unintended attention if disclosed. EHRs include information regarding the physical and mental status of patients, along with their laboratorial analysis, imagiology and other exams, treatments and prognosis, but also information regarding their social and cultural contexts. All this data refers to a person's intimate sphere and it is protected by particularly strict rules.

The primary purpose of health data is to be used in the clinical practice and it can only be processed and used by health professionals in their workplaces. However, it is consensual that the health data contained in the EHRs have an enormous potential to foster high quality research in disease prevention, diagnosis and therapeutic innovation. Due to obstacles for the secondary use and reuse of health data, doctors and researchers spend a lot of time and resources to access the data they need to answer important research questions. This results in the majority of health data not being used for research purposes.

The aMile project aims to streamline the access to data for research purposes while being fully compliant with the General Data Protection Regulation (GDPR) and the local requirements. The legal and ethical issues related to the secondary use of health data in research ensuring patients' rights of privacy, confidentiality and data safety should be addressed by the Sensitive Data Interest Group.

Sensitive data case study - lessons from working with sensitive data: Amy Pienta

Slides available here: https://zenodo.org/record/4895648

This lightning talk explored lessons from working with sensitive data. The talk focused specifically on the National Addiction & HIV Data Archive Program. In the social sciences, the benefits of data sharing often outweigh the risks. Sensitive data can often be deidentified while still being very useful for research. Participants often do want their data shared. For this reason, The National Addiction & HIV Data Archive Program has multiple open access datasets that have been de-identified. Restricted datasets are also held by the Archive that do still contain some identifying information for which people must apply for access.

The specific example of Population Assessment of Tobacco and Health was shared. This data is sensitive because:

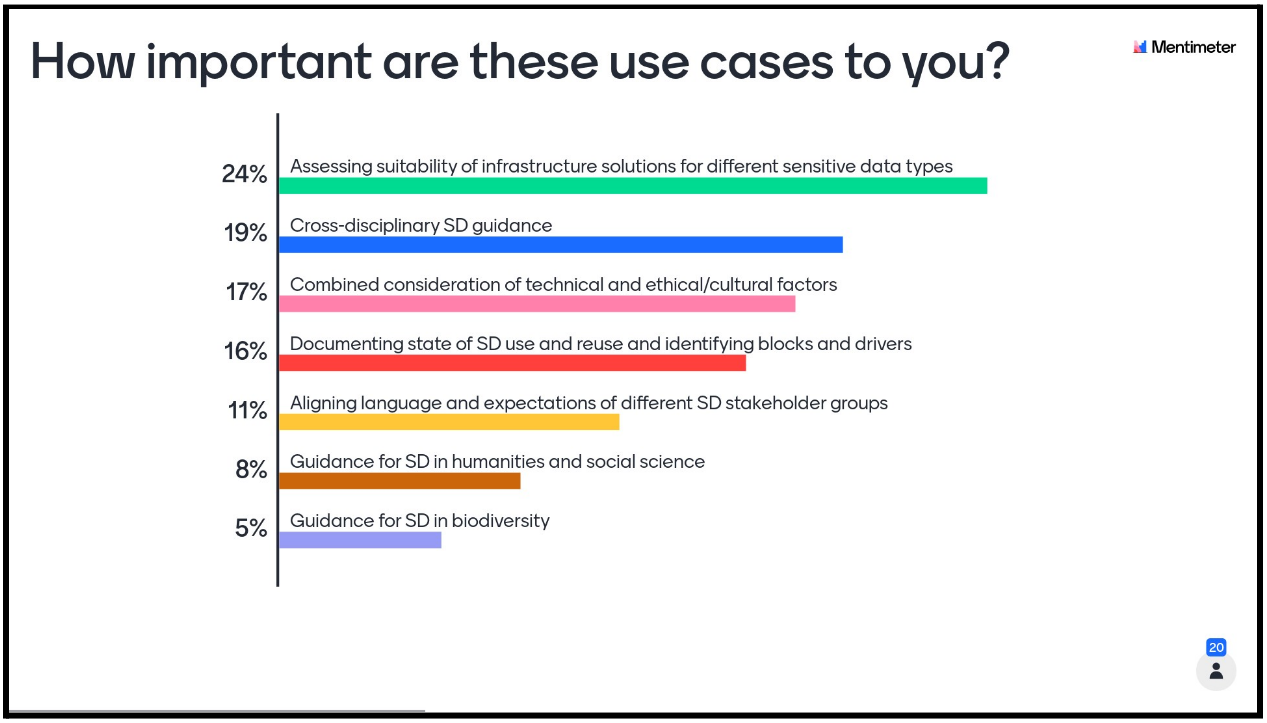
- There is political interest as it touches on the tobacco industry
- These datasets are very disclosive and contain other sensitivities, including also data on biomarkers and the implications of tobacco use across the US
- It contains information from parents and children, so parents could potentially identify the data of their children if parents had access to the dataset.

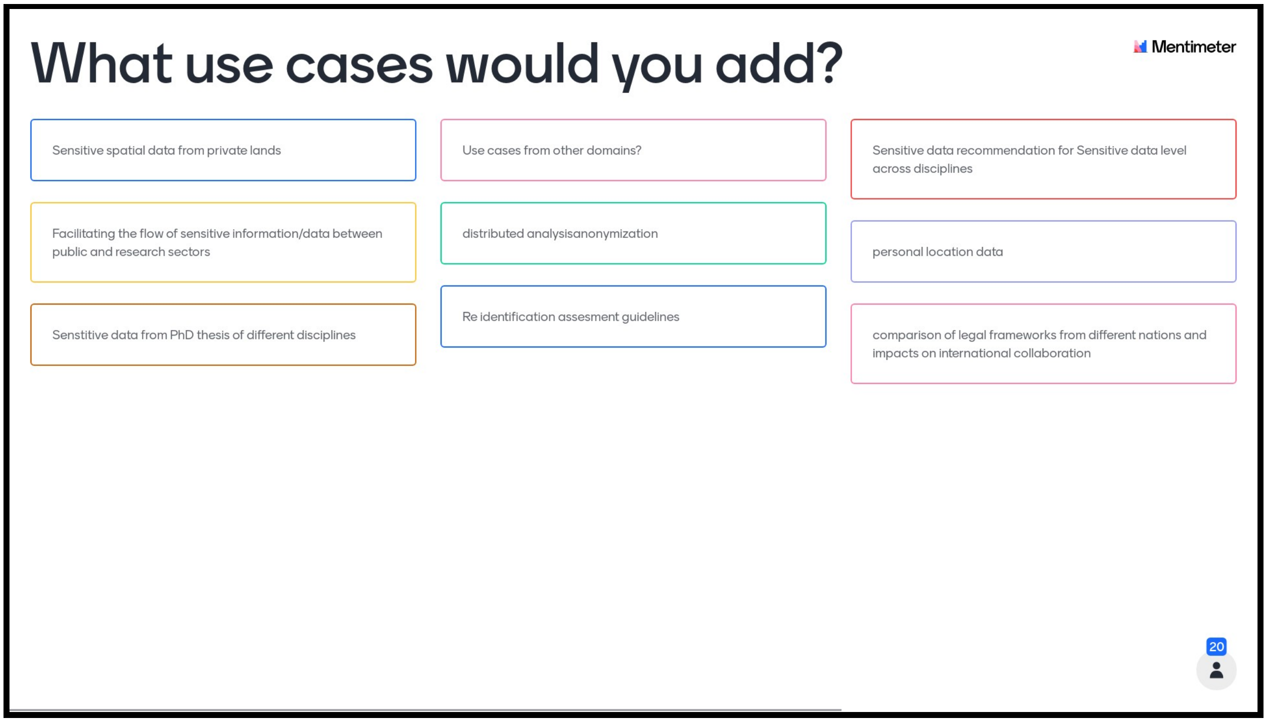
Data is shared in an iterative way and with both research and industry. Analysts are always focused on doing the science fast and there have been breaches, which has led the Archive to improve their training for data use. Breaches and training should be important considerations for this Sensitive Data Interest Group.

Discussion of the charter and/or subtopics: Nichola Burton

Discussion led by Nichola Burton.
In this section, the participants at the session provided their feedback through a Mentimeter interactive survey. Below are the results. Some general comments:

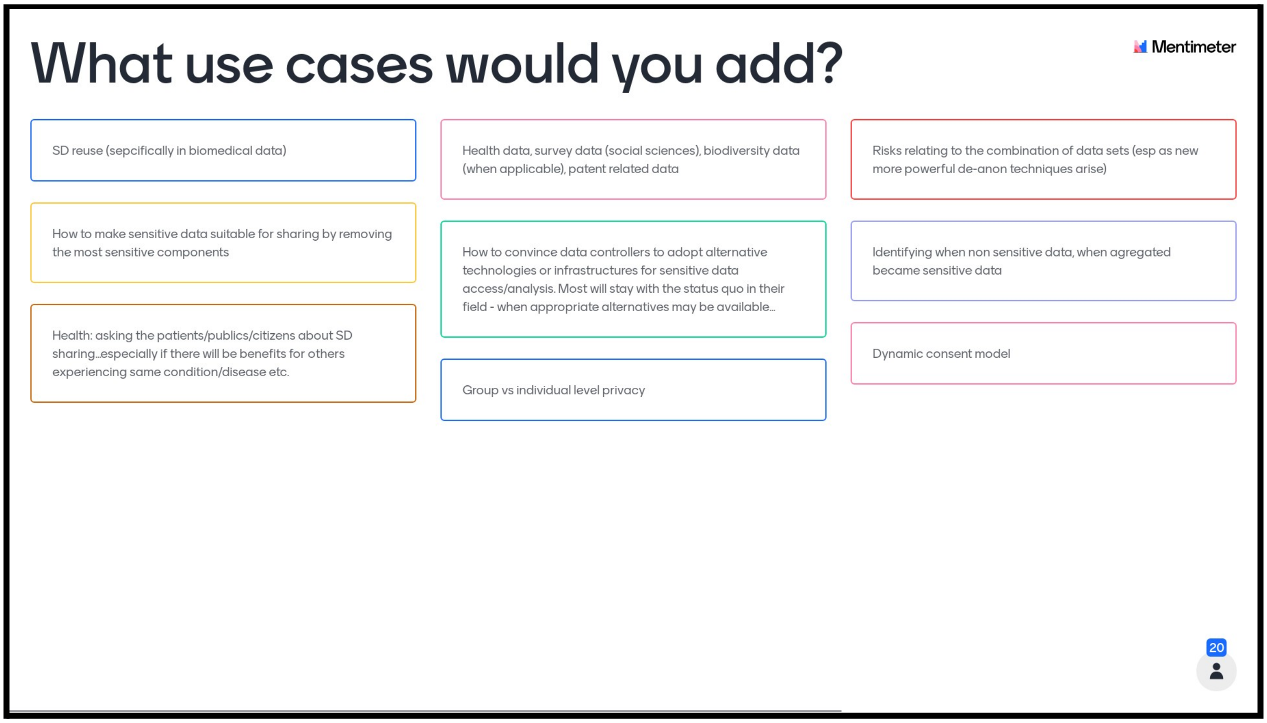
- There is interest in both the technical requirements for sensitive data as well as the governance aspects.
- Sharing sensitive data from PhD theses is a current gap - data can often only be taken by repositories if it is de-identified.
- Some data is not sensitive alone, but when it is aggregated with other data (e.g., linking data) it becomes sensitive.
- It will be important for the Interest Group to work through definitions of sensitive data and how they interrelate.
- What does all the sensitive data we work with have in common? What unifies this data under the umbrella of sensitive data? Some answers provided by the participants:
  - “Someone somewhere cares deeply if that information is exposed.”
  - “Information that potentially harms people or animals”
  - “Commercially sensitive”


How important are these use cases to you?



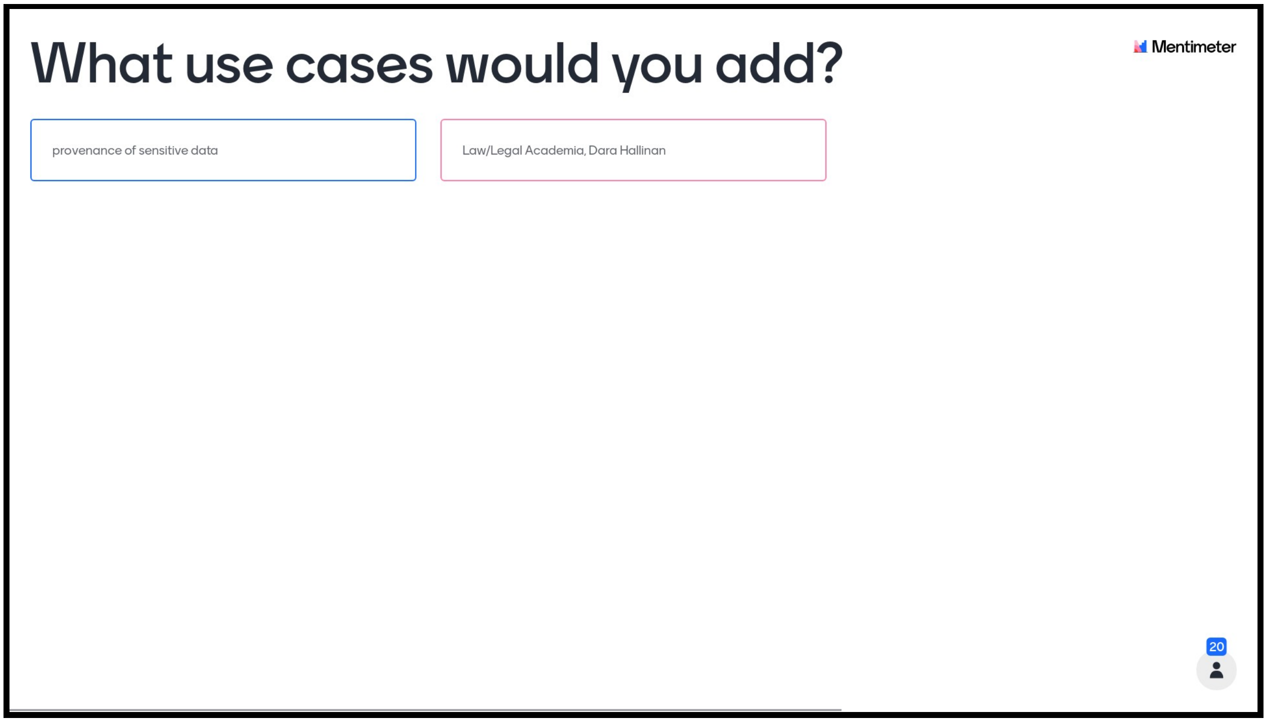
What use cases would you add?

- Sensitive spatial data from private lands
- Facilitating the flow of sensitive information/data between public and research sectors
- Senstitive data from PhD thesis of different disciplines
- Use cases from other domains?
- distributed analysisanonymization
- Re identification assesment guidelines
- Sensitive data recommendation for Sensitive data level across disciplines
- personal location data
- comparison of legal frameworks from different nations and impacts on international collaboration



What use cases would you add?

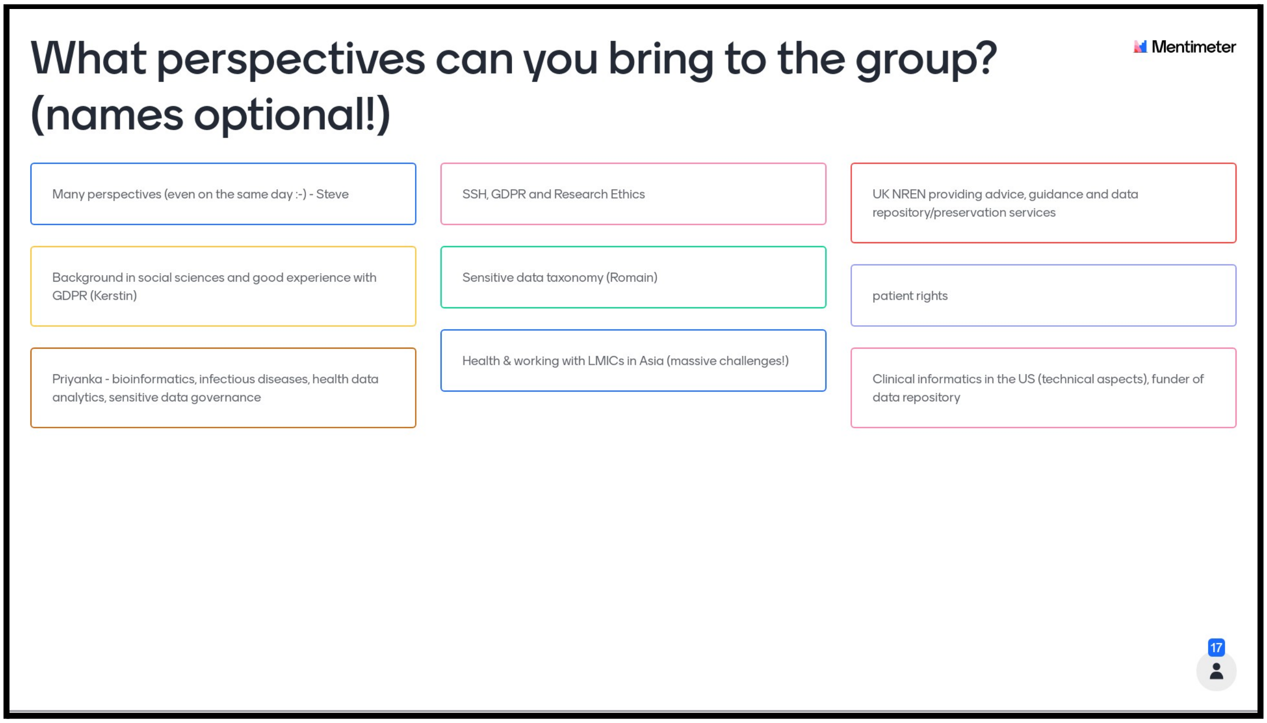
- SD reuse (sepcifically in biomedical data)
- How to make sensitive data suitable for sharing by removing the most sensitive components
- Health: asking the patients/publics/citizens about SD sharing..especially if there will be benefits for others experiencing same condition/disease etc.
- Health data, survey data (social sciences), biodiversity data (when applicable), patent related data
- How to convince data controllers to adopt alternative technologies or infrastructures for sensitive data access/analysis. Most will stay with the status quo in their field - when appropriate alternatives may be available...
- Group vs individual level privacy
- Risks relating to the combination of data sets (esp as new more powerful de-anon techniques arise)
- Identifying when non sensitive data, when agregated became sensitive data
- Dynamic consent model



What use cases would you add?

provenance of sensitive data

Law/Legal Academia, Dara Hallinan



What perspectives can you bring to the group? (names optional!)

Many perspectives (even on the same day :-) - Steve

SSH, GDPR and Research Ethics

UK NREN providing advice, guidance and data repository/preservation services

Background in social sciences and good experience with GDPR (Kerstin)

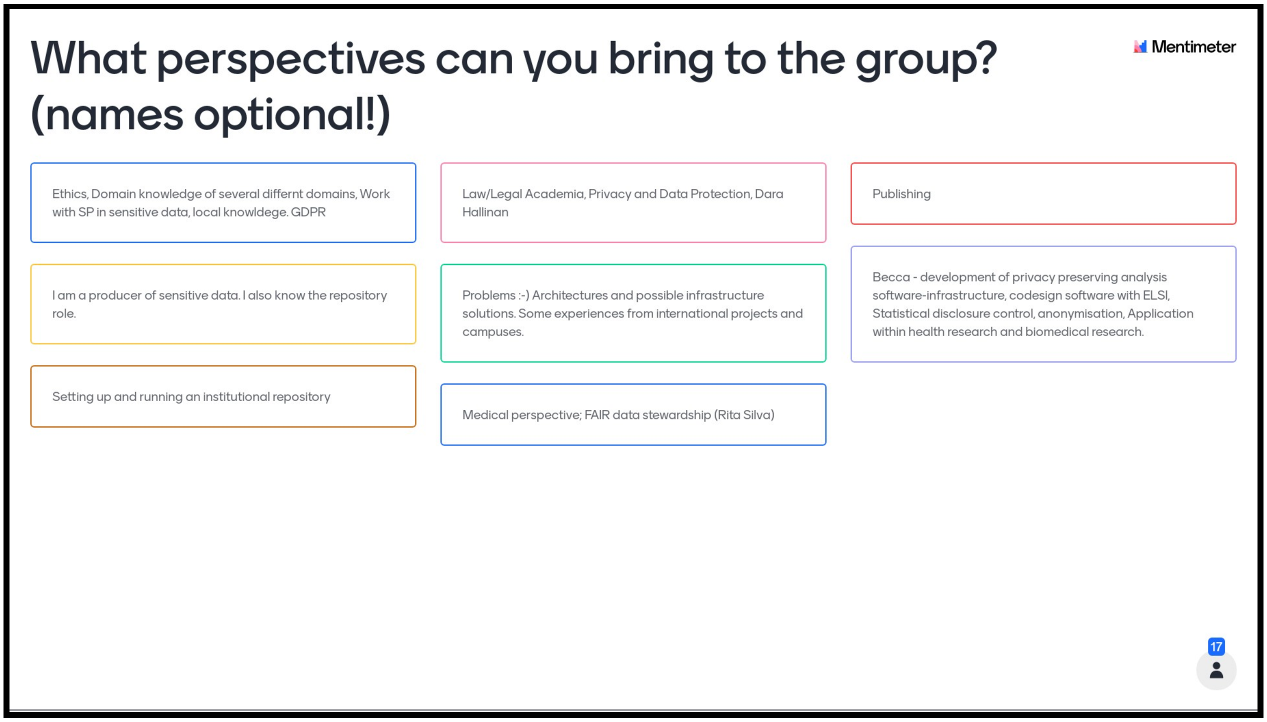
Sensitive data taxonomy (Romain)

patient rights

Priyanka - bioinformatics, infectious diseases, health data analytics, sensitive data governance

Health & working with LMICs in Asia (massive challenges!)

Clinical informatics in the US (technical aspects), funder of data repository



What perspectives can you bring to the group? (names optional!)

Ethics, Domain knowledge of several differnt domains, Work with SP in sensitive data, local knowldege. GDPR

I am a producer of sensitive data. I also know the repository role.

Setting up and running an institutional repository

Law/Legal Academia, Privacy and Data Protection, Dara Hallinan

Problems :-) Architectures and possible infrastructure solutions. Some experiences from international projects and campuses.

Medical perspective; FAIR data stewardship (Rita Silva)

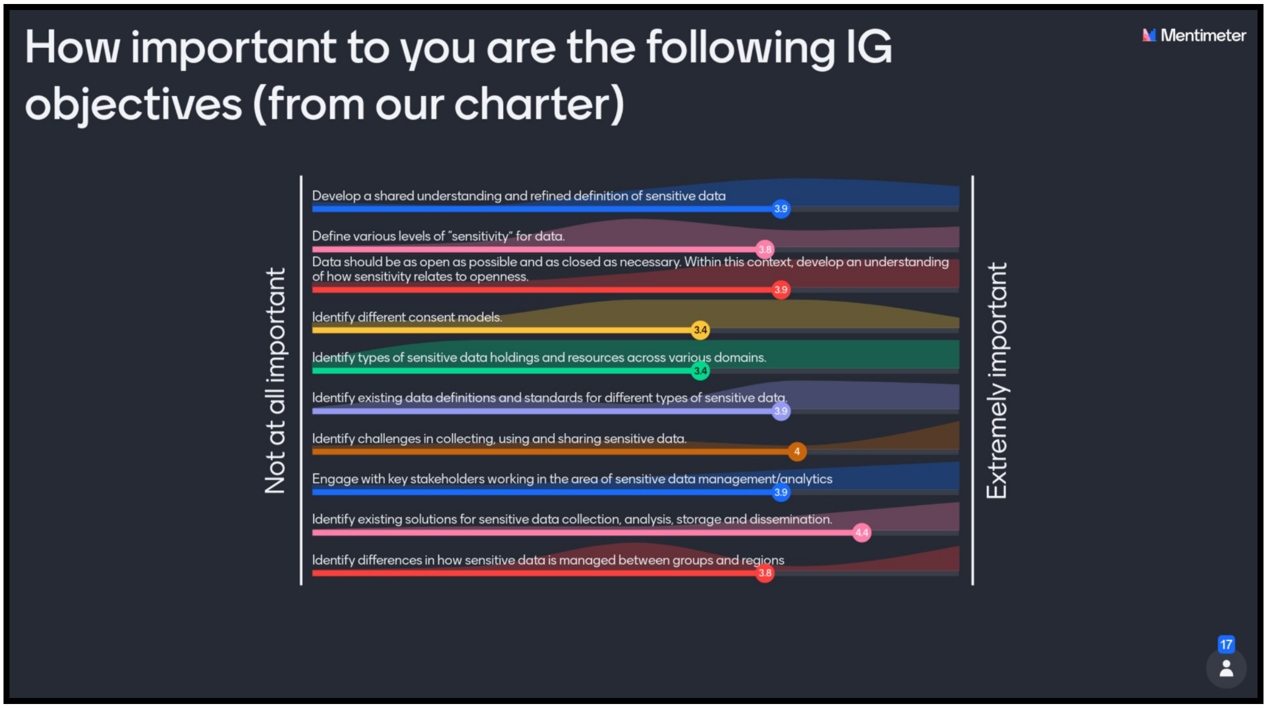
Publishing

Becca - development of privacy preserving analysis software-infrastructure, codesign software with ELSI, Statistical disclosure control, anonymisation, Application within health research and biomedical research.

Where to next with the IG: Steven McEachern

Discussion led by Steven McEachern.
In this section, the participants at the session provided their feedback through a Mentimeter interactive survey. Below are the results. Some general comments:

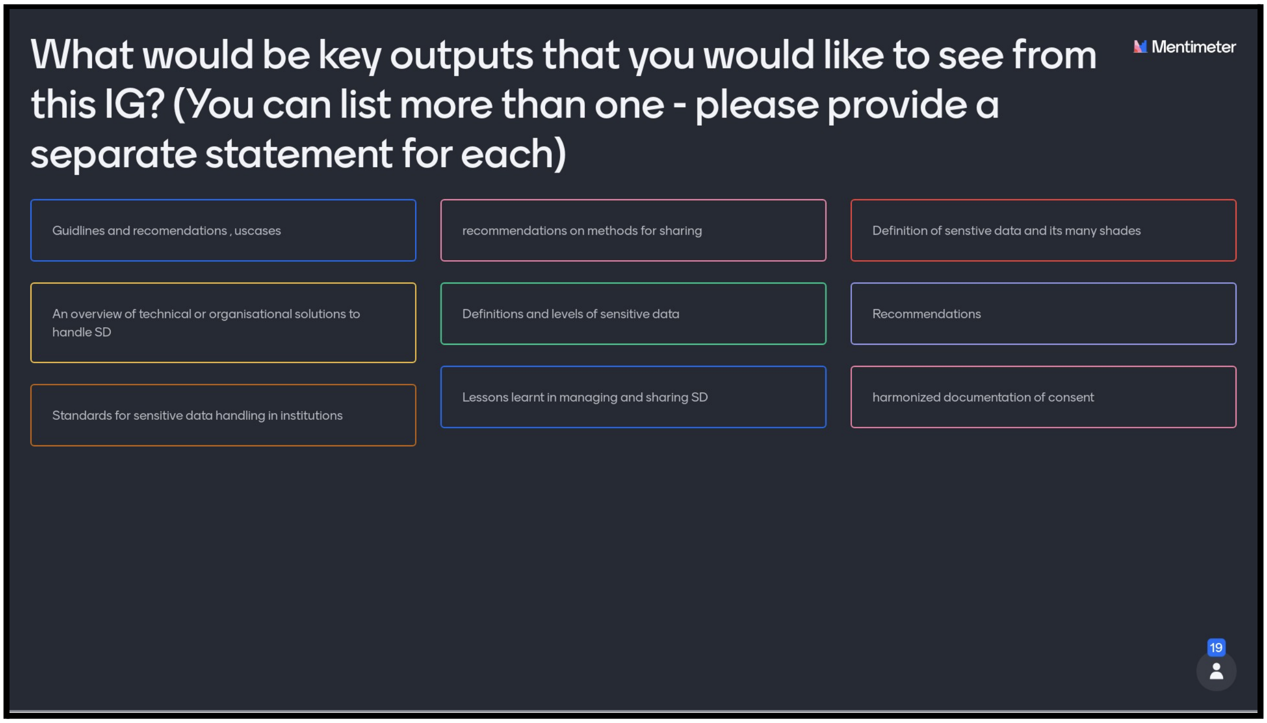
- There is a lot of interest in exploring the existing solutions for working with sensitive data. This is similar to what was observed in the Mentimeter activity outlined above.
- One theme that is common as a desired output are recommendations for de-identification.
- Sharing of real world experience with sensitive data is something that participants would like to see the Interest Group produce/facilitate.



How important to you are the following IG objectives (from our charter)

Mentimeter

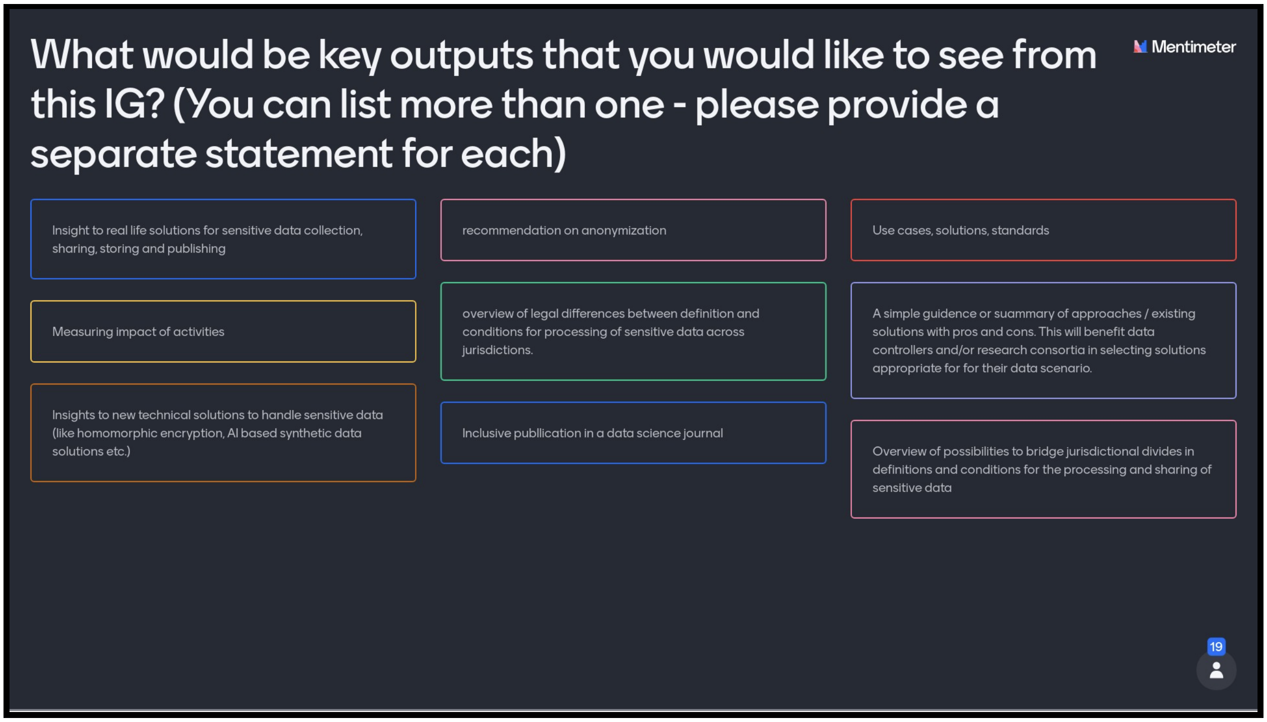
- Develop a shared understanding and refined definition of sensitive data 3.9
- Define various levels of "sensitivity" for data. 3.8
- Data should be as open as possible and as closed as necessary. Within this context, develop an understanding of how sensitivity relates to openness. 3.9
- Identify different consent models. 3.4
- Identify types of sensitive data holdings and resources across various domains. 3.4
- Identify existing data definitions and standards for different types of sensitive data. 3.9
- Identify challenges in collecting, using and sharing sensitive data. 4
- Engage with key stakeholders working in the area of sensitive data management/analytics 3.9
- Identify existing solutions for sensitive data collection, analysis, storage and dissemination. 4.4
- Identify differences in how sensitive data is managed between groups and regions 3.8



What would be key outputs that you would like to see from this IG? (You can list more than one - please provide a separate statement for each)

Mentimeter

- Guidlines and recomendations , uscases
- recommendations on methods for sharing
- Definition of senstive data and its many shades
- An overview of technical or organisational solutions to handle SD
- Definitions and levels of sensitive data
- Recommendations
- Standards for sensitive data handling in institutions
- Lessons learnt in managing and sharing SD
- harmonized documentation of consent



What would be key outputs that you would like to see from this IG? (You can list more than one - please provide a separate statement for each)

Insight to real life solutions for sensitive data collection, sharing, storing and publishing

Measuring impact of activities

Insights to new technical solutions to handle sensitive data (like homomorphic encryption, AI based synthetic data solutions etc.)

recommendation on anonymization

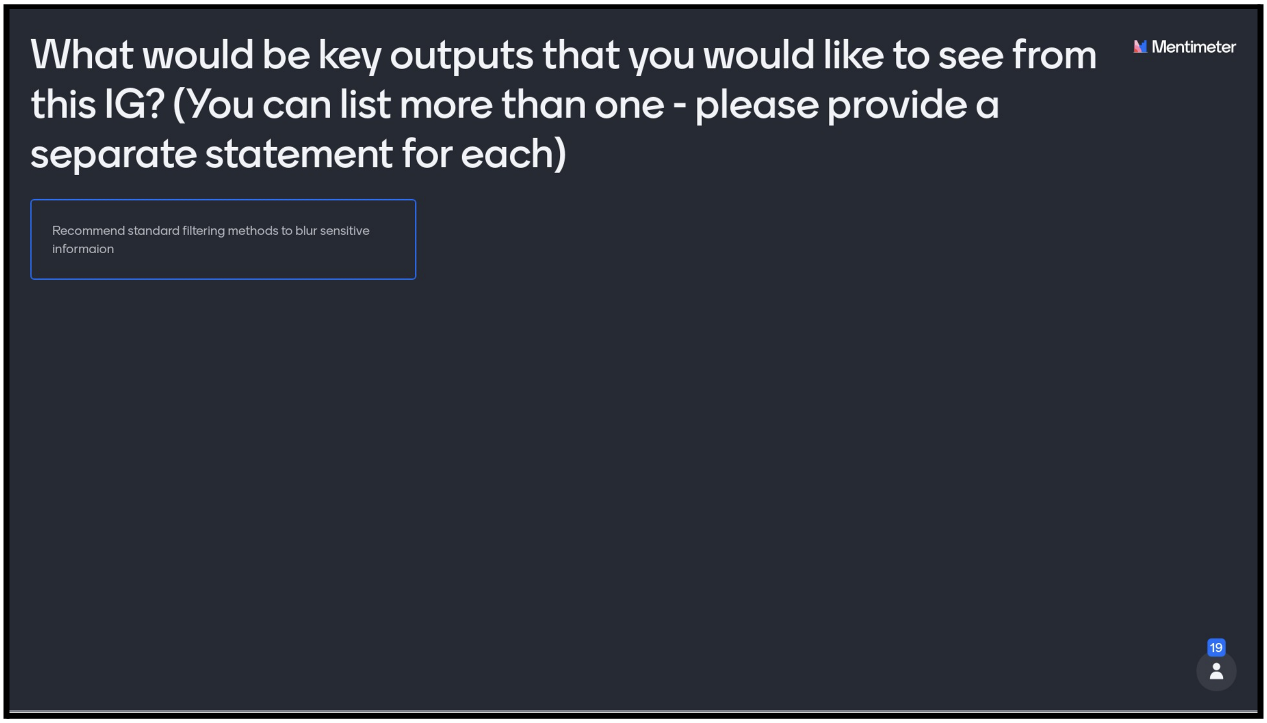
overview of legal differences between definition and conditions for processing of sensitive data across jurisdictions.

Inclusive publlication in a data science journal

Use cases, solutions, standards

A simple guidence or suammary of approaches / existing solutions with pros and cons. This will benefit data controllers and/or research consortia in selecting solutions appropriate for for their data scenario.

Overview of possibilities to bridge jurisdictional divides in definitions and conditions for the processing and sharing of sensitive data



What would be key outputs that you would like to see from this IG? (You can list more than one - please provide a separate statement for each)

Recommend standard filtering methods to blur sensitive informaion

Next steps

In working towards RDA endorsement, next steps include:

1. June 2021: Disseminate this session report to the RDA community
2. July 2021: Submit session proposal for next Plenary (Due 9 July)
3. July 2021: Circulate the revised charter to the TAB and RDA community
4. July 2021: Contact participants who indicated interest in the collaborative about the next interest group meeting.

There is a rolling call for participation, please do get in touch!

Funding Acknowledgments

Romain David is supported by the EOSC-Life european program under grant agreement Nº824087 and ERINHA-Advance european program under grant agreement Nº824061.