Croissant-RAI

Standardized Machine-readable Dataset Documentation Format for Responsible AI

Dr Nitisha Jain

Informatics Department, King's College London

September 17, 2024

Motivation

- Data is a crucial element in emergent AI technologies, plays a central role in training and evaluating AI models.
- Yet, quality and documentation remain significant challenges, leading to adverse downstream effects (e.g., potential biases) in applications*.
- Working with data is time-consuming and painful
  - Wide variety of data formats
  - Lack of interoperability between tools
  - Difficulty of discovering and combining datasets

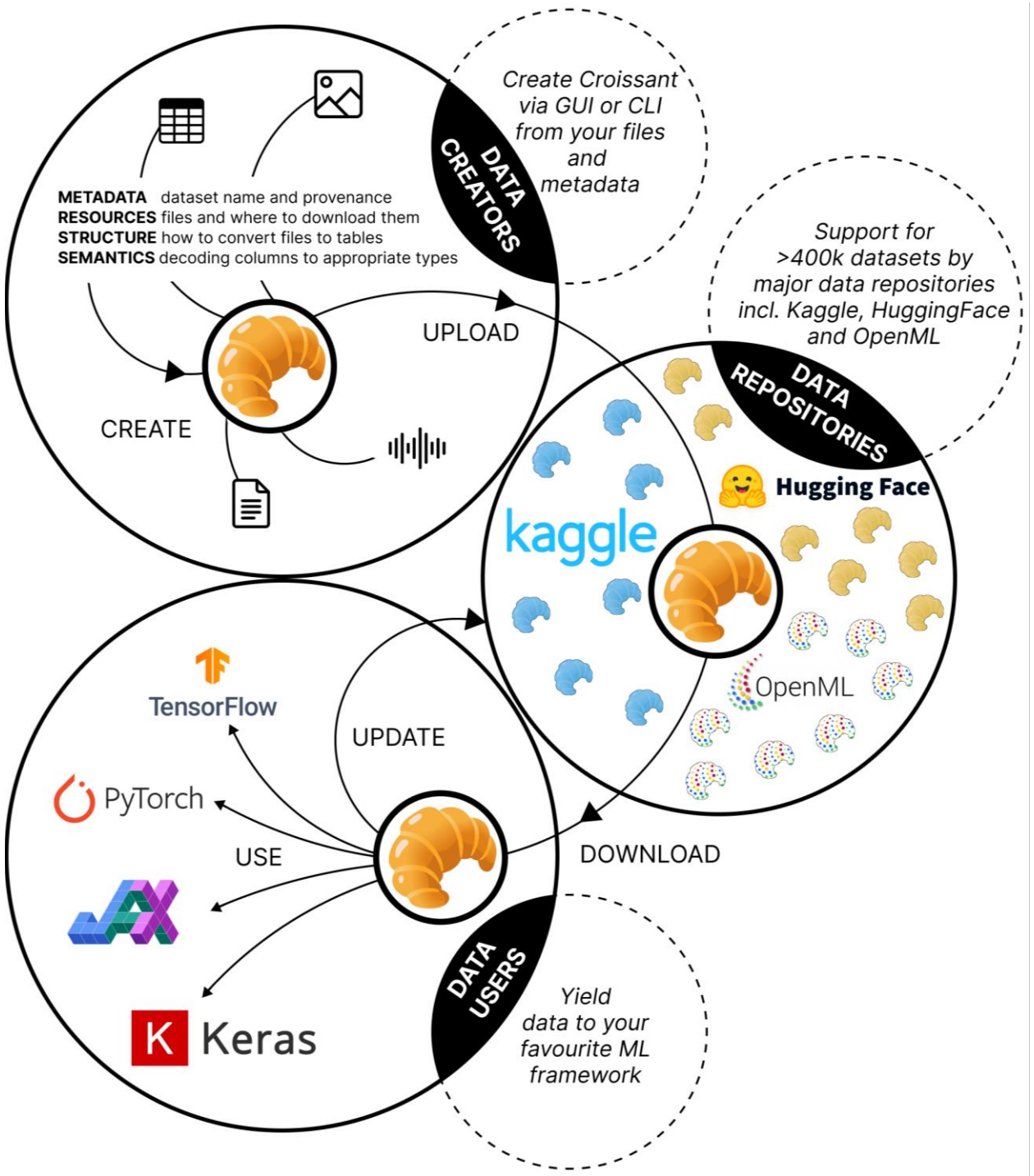
*Larrazabal et al. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis.

- Metadata format to help standardize machine learning datasets.
- Make datasets easily discoverable and usable across tools and platforms, no need for reformatting.
- MLCommons Working Group
- Industry support from HuggingFace, Google Dataset Search, Kaggle, OpenML, TFDS amongst others.
- First release – March 6, 2024
    - Documentation, open-source library, visual editor.

- Extension to schema.org, a machine-readable standard to describe structured data, used by over 40M datasets on Web.
- Allows datasets to be discoverable through dataset search engines such as Google Dataset Search.
- Adds metadata layer to represent dataset contents in standardized way, describing key attributes, properties.
- Popular ML frameworks like TensorFlow, JAX and PyTorch can already load Croissant datasets via TensorFlow Datasets library.

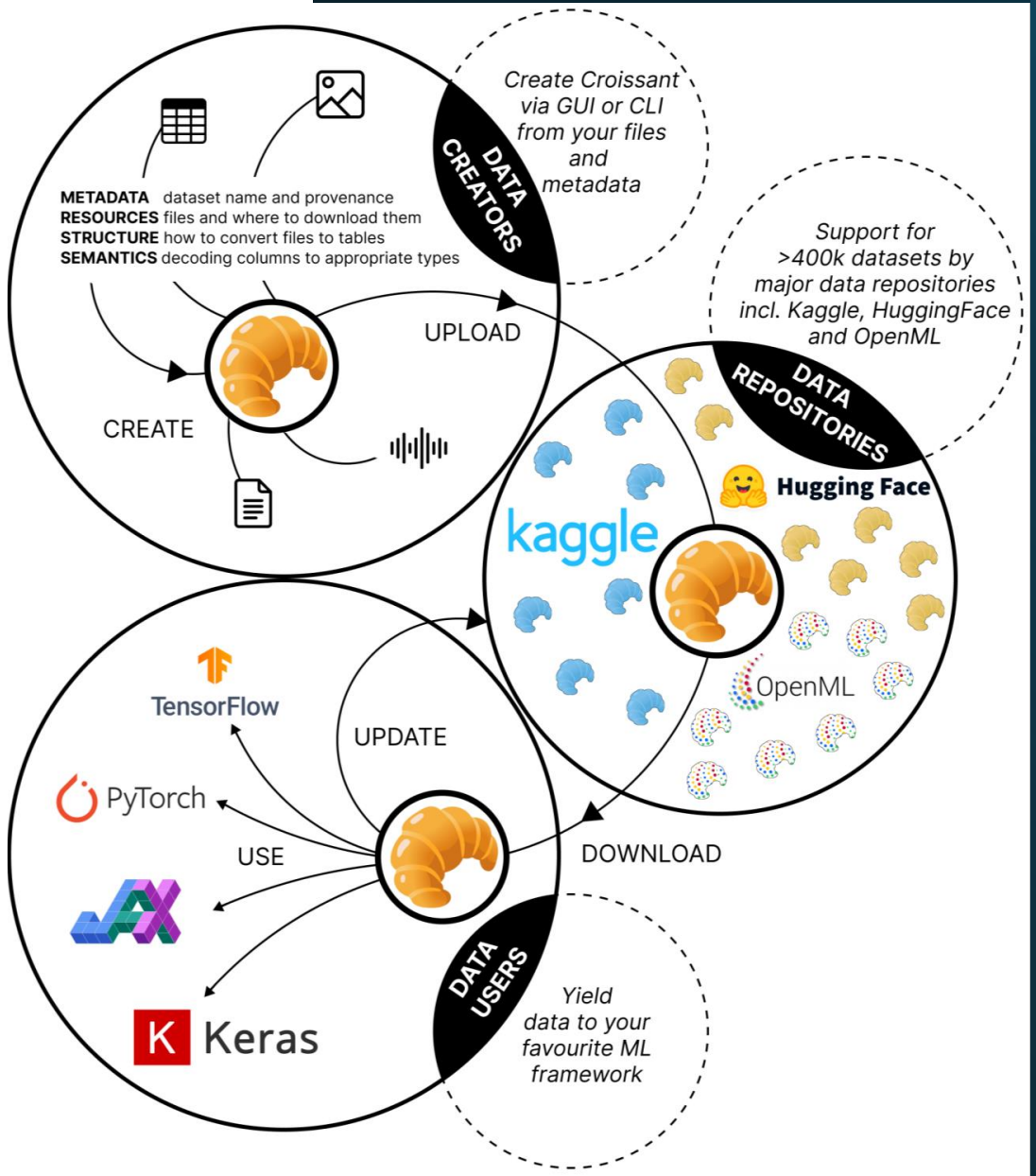
Croissant

ML ready dataset format



Layers of Croissant

Dataset Metadata Layer	Contains relevant information such as name, description, and version.
Resource Layer	Describes the source data used in the dataset.
Structure Layer	Describes and organizes the structure of the resources.
Semantic Layer	Provides ML-specific data interpretation and semantics.



Operationalizing Responsible AI

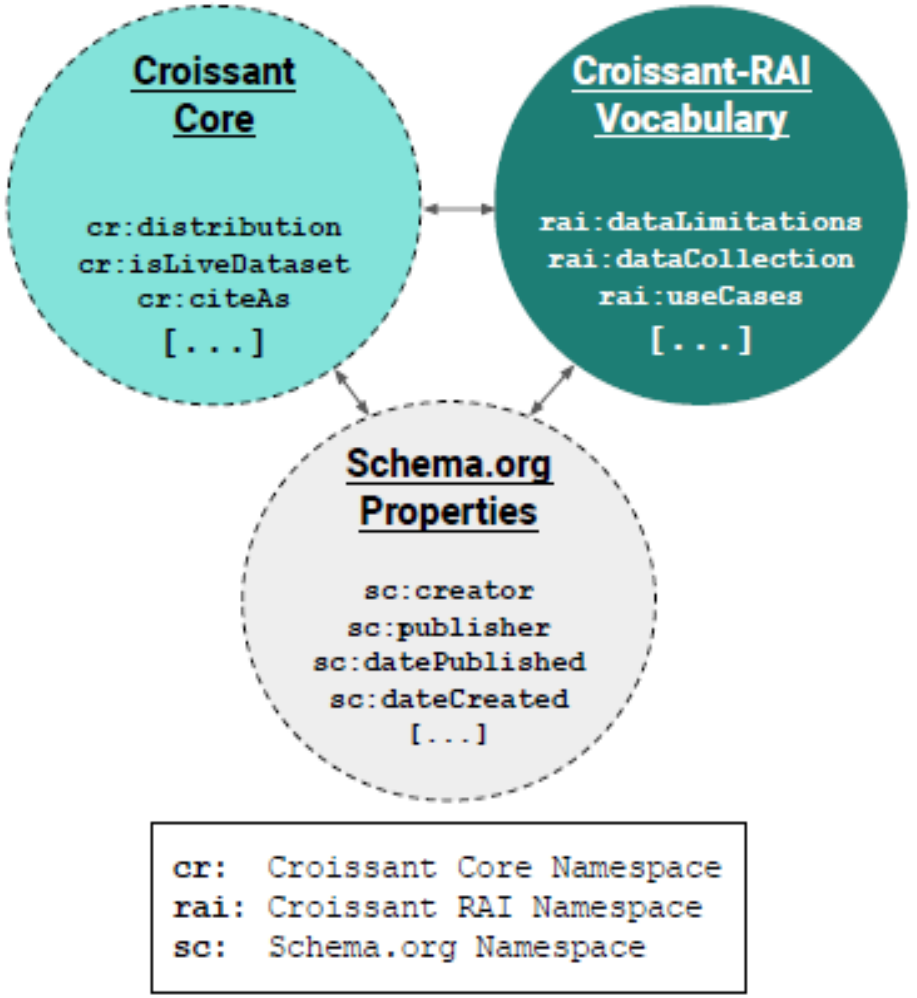
- Responsible AI (RAI) community has identified data documentation work as critical to the development of trustworthy AI systems*.
- Limitations of current documentation proposals
  - Overlapping formats
  - Requirement of data documentation in natural language
  - Lack of standard structure
  - No integration with widely used tools and ML frameworks
- Hence, do not support machine readability for automation.

*Datasheets for Datasets (Gebru et al. 2021), Data statements (Bender and Friedman 2018)

Croissant-RAI

- Machine-readable metadata format to enhance the discoverability, interoperability, and trustworthiness of AI datasets.
- Extends the Croissant metadata format.
- Consists of set of attributes organized around RAI use cases such as AI safety and regulatory compliance.
- Builds on and complements existing RAI dataset documentation proposals,
- Aim to make publishing, discovering, and reusing existing RAI documentation easier.

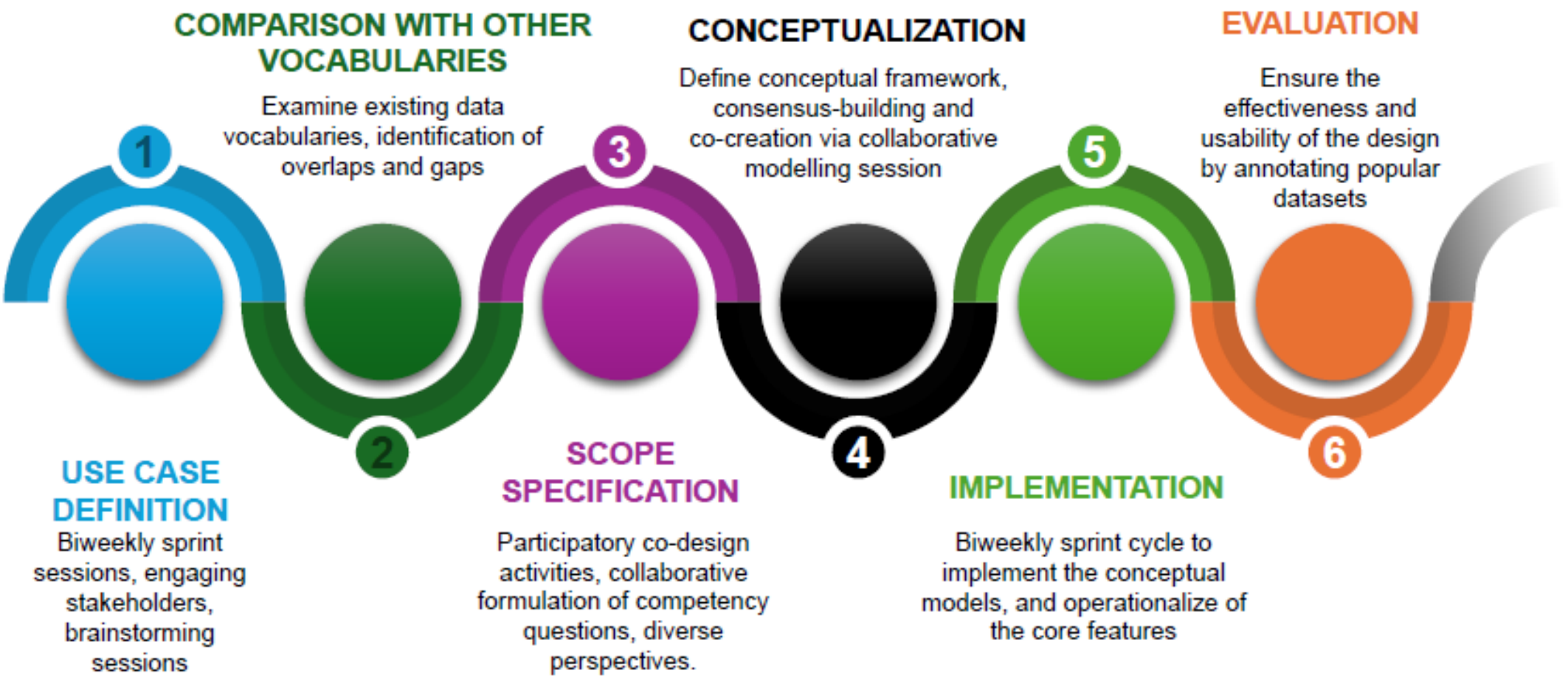
Croissant-RAI Overview



Base Dataset documentation toolkits

- HuggingFace Dataset Cards (https://huggingface.co/docs/hub/en/datasets-cards)
- Kaggle metadata (https://github.com/Kaggle/kaggle-api/wiki/Dataset-Metadata)
- Data Nutrition Labels (Holland et al. 2018)
- Data Cards (Pushkarna, Zaldivar, and Kjartansson 2022)
- Croissant (Akhtar et al. 2024)
- Crowdworksheets (Diaz et al. 2022)
- Fairness Datasets Ontology (https://fairnessdatasets.dei.unipd.it/schema/)
- DescribeML (Giner-Miguelez, Gomez, and Cabot 2023)

Vocabulary Engineering Process for Croissant-RAI



Use Cases for Croissant-RAI

Data Life Cycle

Data Labeling

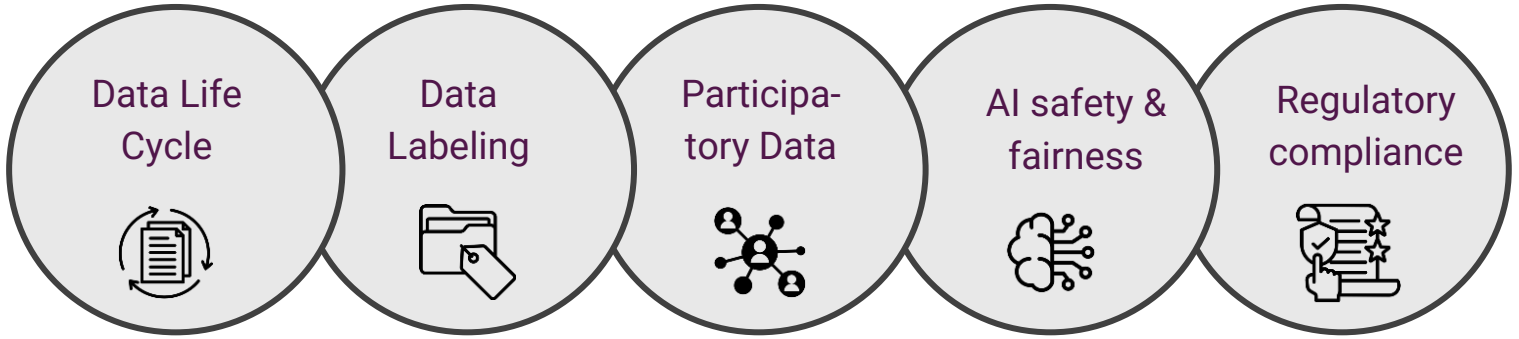
Participa-tory Data

AI safety & fairness

Regulatory compliance

Data Life Cycle

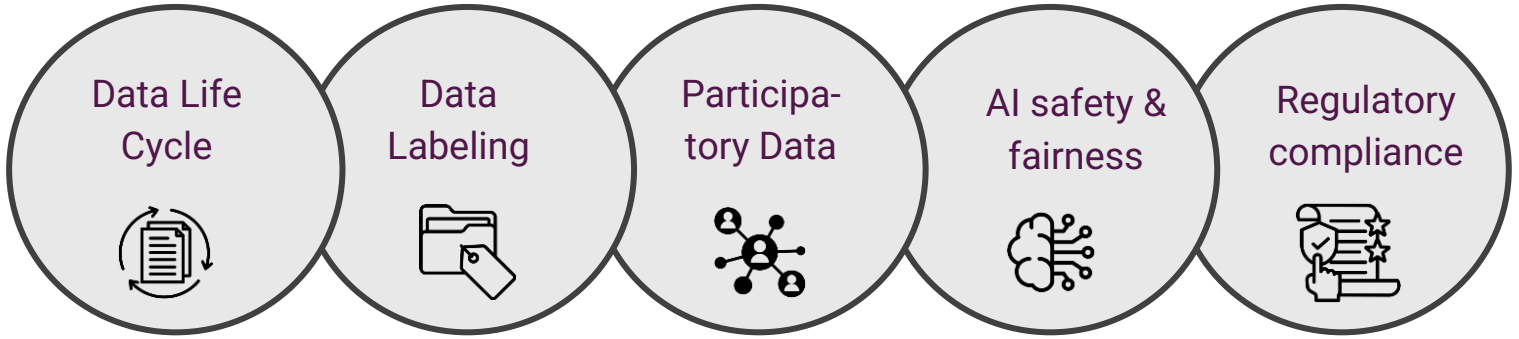
cr:distribution
cr:isLiveDataset
cr:citeAs
sc:creator
sc:publisher
sc:datePublished
sc:dateCreated
sc:dateModified
sc:version
sc:license
sc:maintainer



rai:dataLimitations
rai:dataCollection
rai:useCases
rai:dataReleaseMaintenance
rai:Data Processing
rai:Data Selection
rai:Data Inclusion/Exclusion
rai:Relationship to Source

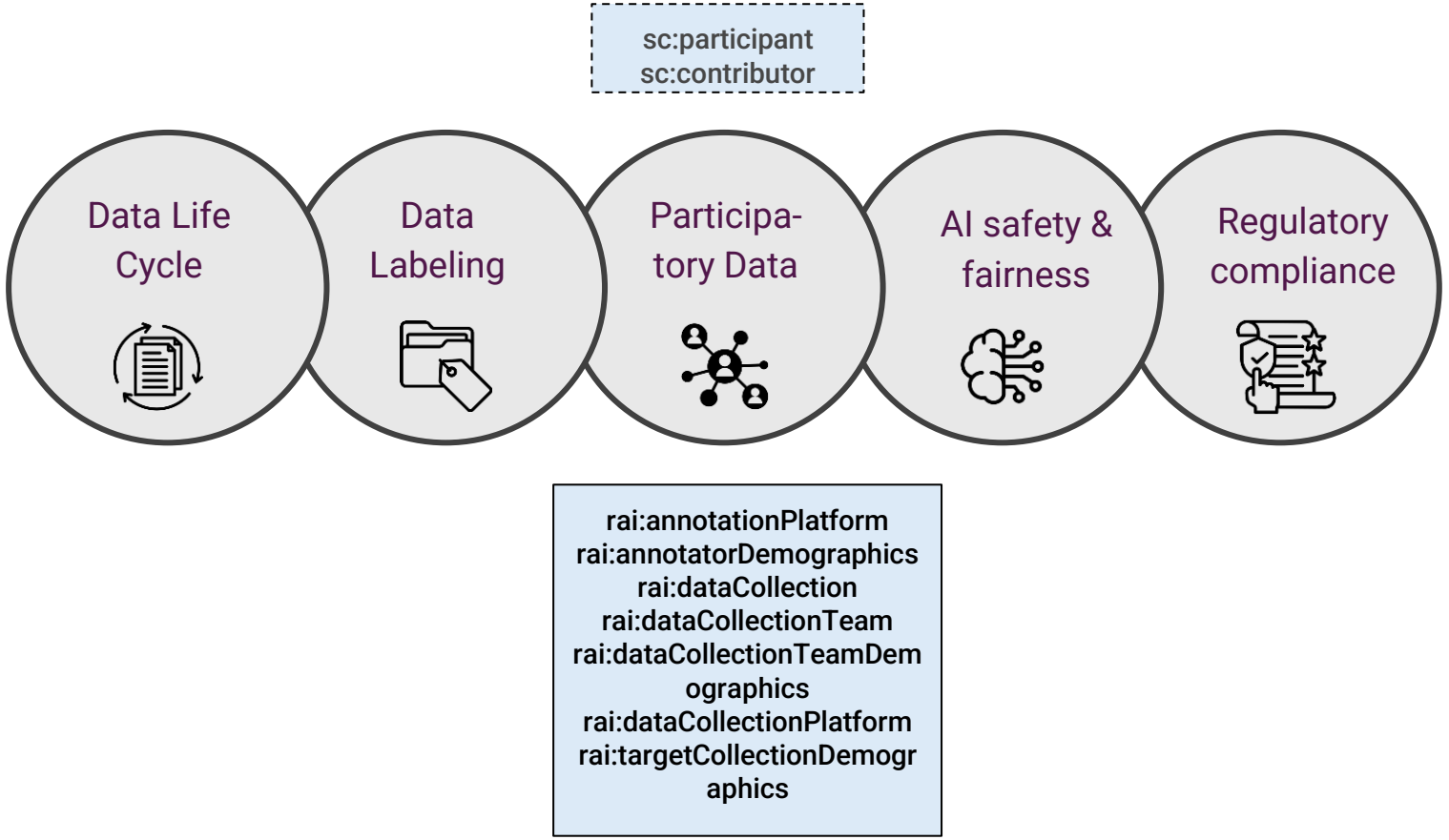
Data Labeling

rai:annotationPlatform
rai:annotatorDemographics
rai:machineAnnotationTools
rai:Annotation Workforce Type
rai:Annotation Characteristic(s)
rai:Annotation Description(s)
rai:Annotation Guidelines
rai:Annotation Distribution(s)

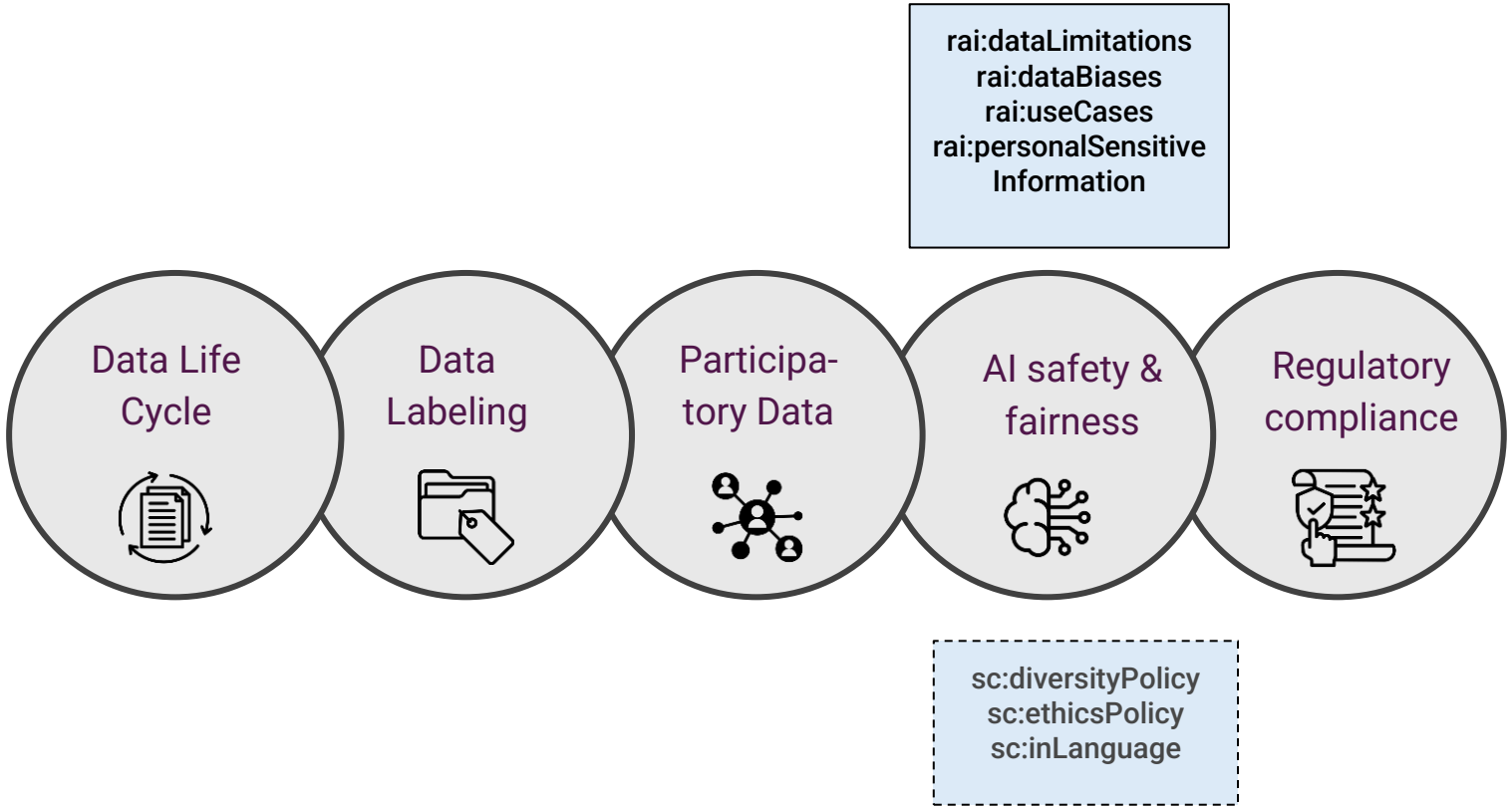


cr:distribution
cr:isLiveDataset
cr:isAnnotationOf

Participatory Data

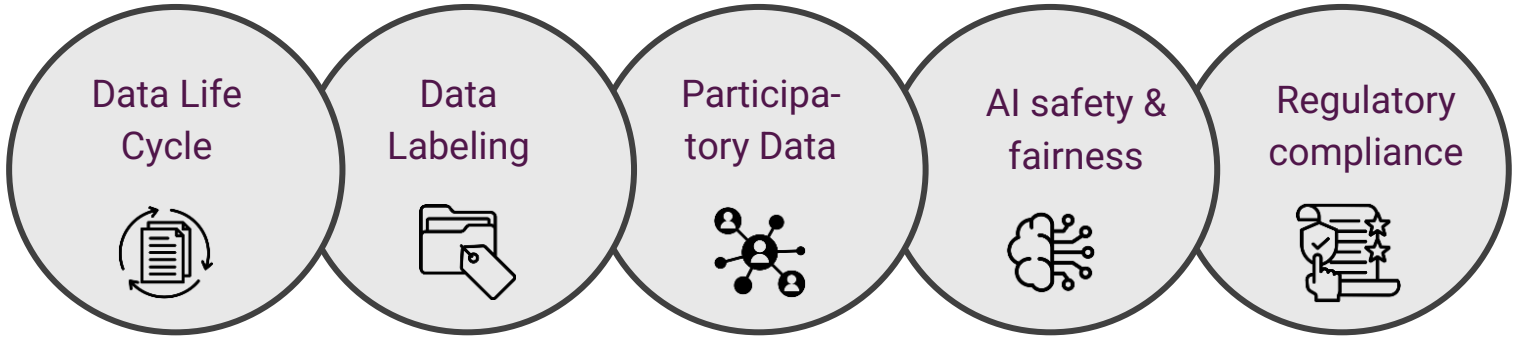


AI Safety and Fairness



Regulatory Compliance

rai:personalSensitiveInformation
rai:useCases
rai:dataReleaseMaintenance
rai:dataImputationProtocol
rai:dataManipulationProtocol
rai:dataSharingAgreements
rai:dataGovernanceProtocol



Croissant is for

Creators and maintainers of ML datasets

Croissant makes datasets more widely available, across repositories and ML frameworks.

Designed to be modular and extensible for domain, modalities.

ML researchers and practitioners

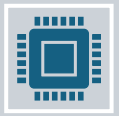
Users of Croissant-enabled datasets have access to dataset documentation to understand the data and contribute to it.

Enable loading of data into ML platforms, no transformations needed.

RAI researchers and practitioners

Machine-readable summary of important attributes captured in a variety of data cards and similar approaches,

Portable and discoverable, promoting better documentation practices.



Policy makers

Standardized way to collect core information about datasets,

Facilitating the development of data-centric AI audit and assurance tools such as transparency indexes.

Tool Support

Croissant Editor

Home

OVERVIEW | METADATA | EXTENSION: RESPONSIBLE AI | RESOURCES | RECORD SETS

Provenance

Data Collection

Explanation

DICES-350 consists of 350 adversarial multi-turn conversations,annotated by a pool of annotators along 16 safety criteria.

Data Collection Type

Define the data collection type.

Secondary Data … × User-generated c… ×

Raw Data.

The input data for this data collection was sampled from an 8K multi-turn conversation corpus (comprising 48K turns in total) generated by human agents interacting with a generative AI-chatbot.

Data collection timeframe:

Start and end the collection process

Start: **range** https://schema.org/DateTime

2022/02/01

End **range**: https://schema.org/DateTime

2023/02/01

Ongoing efforts

Next steps

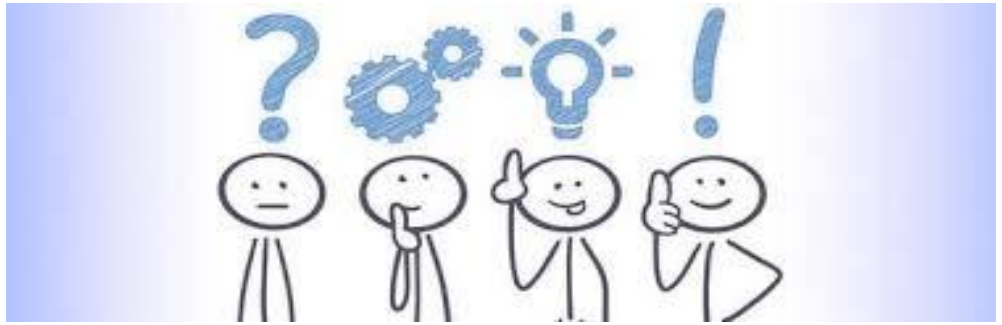
Track the community's uptake of the Croissant-RAI vocabulary, offering valuable insights into its real-world application.

Collaborate with both public and private partners, regulators and corporations, knowing the socio-technical dimensions of RAI practices.

Extensions for domains – geospatial datasets, life sciences, digital humanities.

Refine Use cases further, talk to stakeholders and potential adopters.

Look out for the next version ☺



Thank you for your attention!

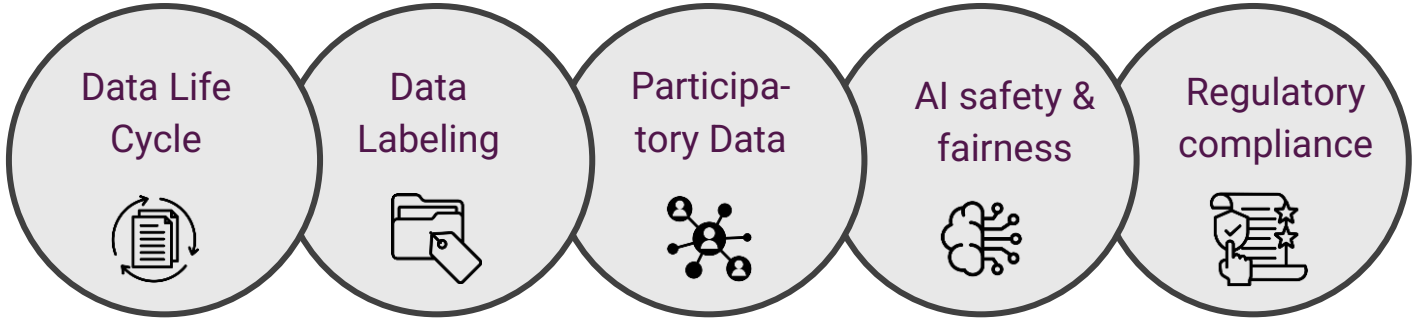
Dr Nitisha Jain
Postdoctoral Researcher
Informatics Department
King's College London (KCL)
https://www.linkedin.com/in/nitisha-jain/
https://nitishajain.github.io

Contact us !

Join our working group !

Data Life Cycle

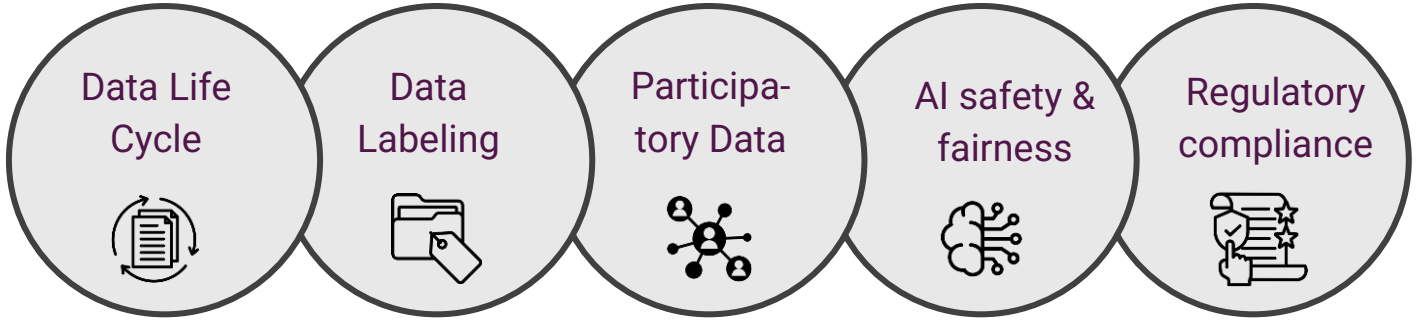
cr:distribution
cr:isLiveDataset
cr:citeAs
sc:creator
sc:publisher
sc:datePublished
sc:dateCreated
sc:dateModified
sc:version
sc:license
sc:maintainer



rai:dataLimitations
rai:dataCollection
rai:useCases
rai:dataReleaseMaintenance
rai:Data Processing
rai:Data Selection
rai:Data Inclusion/Exclusion
rai:Relationship to Source

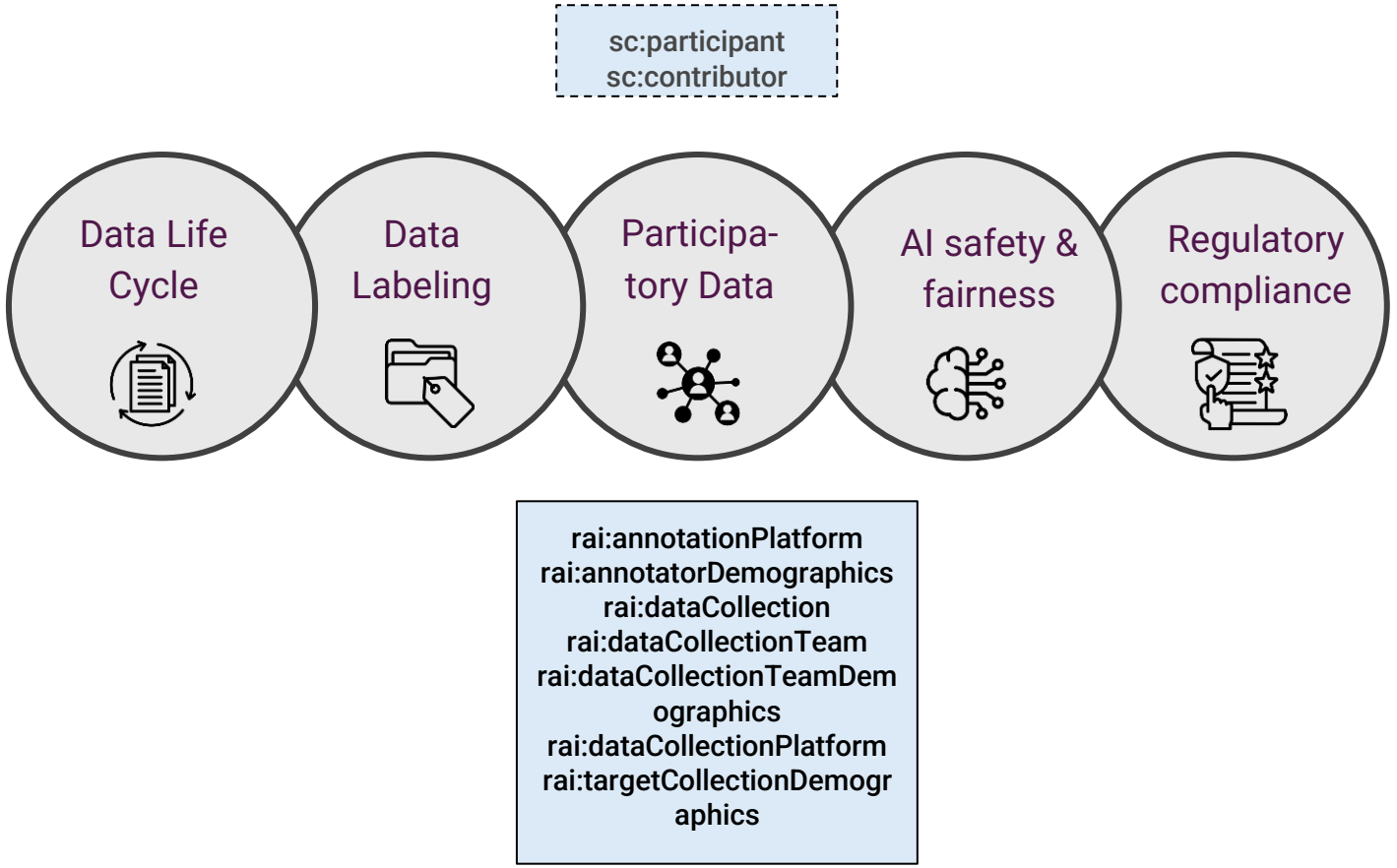
Data Labeling

rai:annotationPlatform
rai:annotatorDemographics
rai:machineAnnotationTools
rai:Annotation Workforce Type
rai:Annotation Characteristic(s)
rai:Annotation Description(s)
rai:Annotation Guidelines
rai:Annotation Distribution(s)

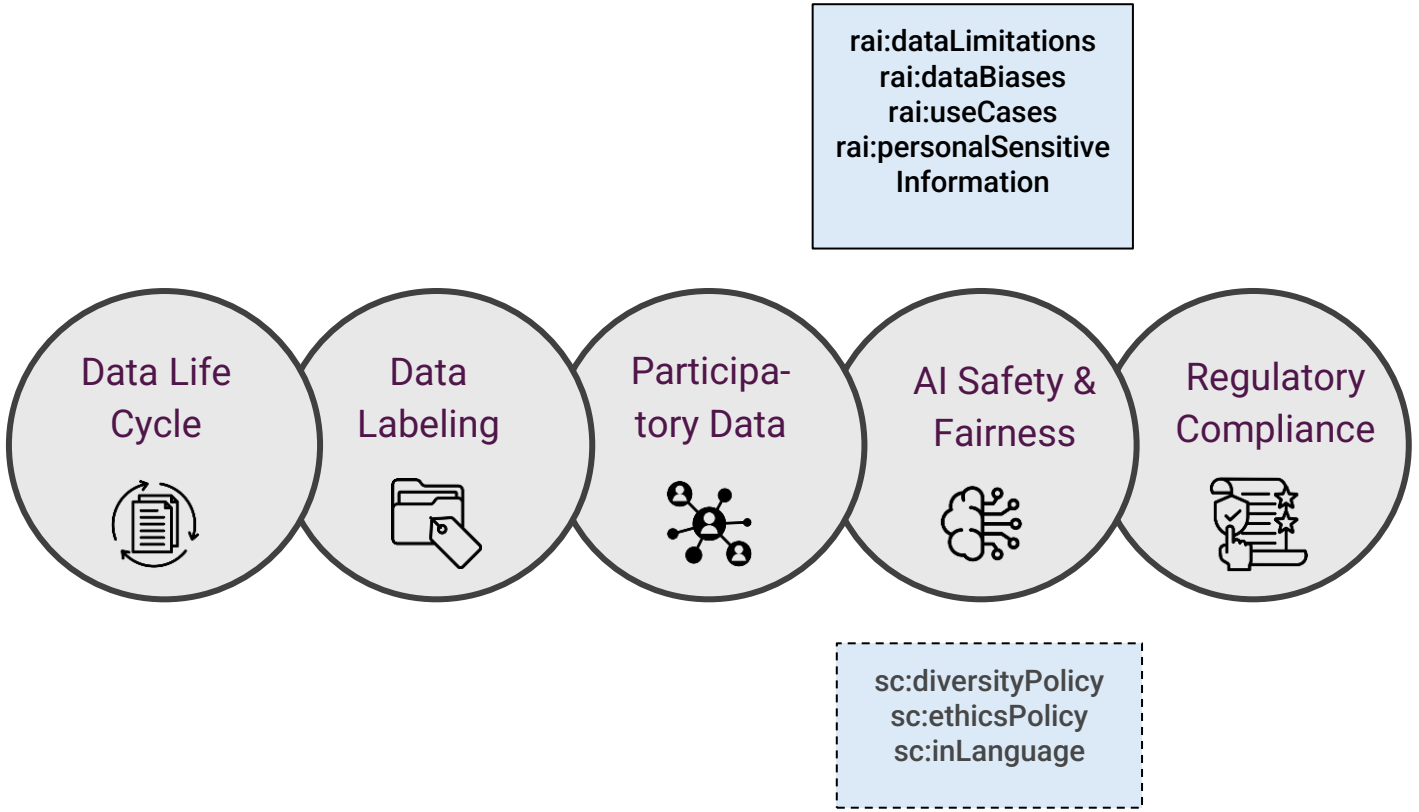


cr:distribution
cr:isLiveDataset
cr:isAnnotationOf

Participatory Data

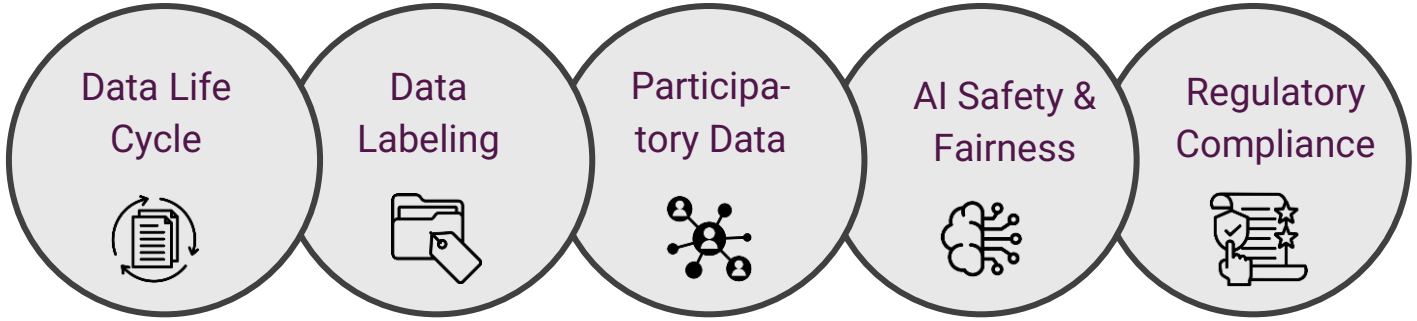


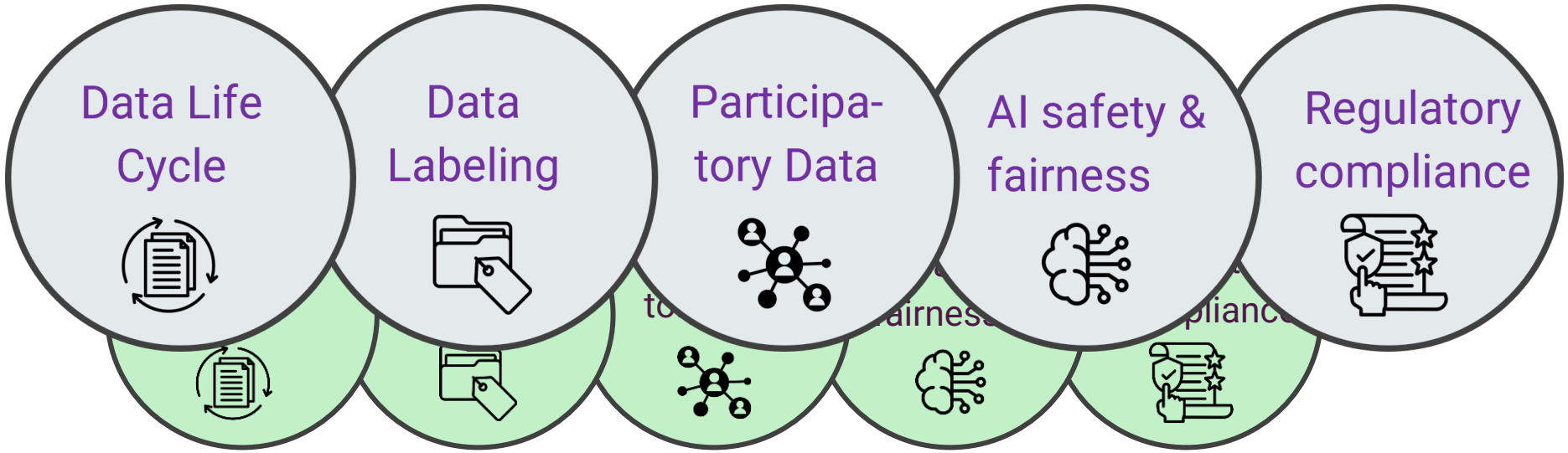
AI Safety and Fairness



Regulatory Compliance

rai:personalSensitiveInformation
rai:useCases
rai:dataReleaseMaintenance
rai:dataImputationProtocol
rai:dataManipulationProtocol
rai:dataSharingAgreements
rai:dataGovernanceProtocol




Data Life Cycle
Data Labeling
Participa-tory Data
AI safety & fairness
Regulatory compliance