Genoveva Vargas-Solar (she/her)

CNRS, Database Group

LIRIS, France

Jérôme Darmont (he/his)

Univ. Lumière Lyon 2

ERIC, France

Alejandro Adorjan (he/his)

Univ. ORT, Uruguay

Javier A. Espinosa-Oviedo (he/his)

CPE, Univ. Lyon

LIRIS, France

Carmem S. Hara (she/her)

Univ. Federal do Paranà, Brazil

Sabine Loudcher (she/her)

Univ. Lumière Lyon 2

ERIC, France

Regina Motz (she/her)

Univ. de la República, Uruguay

Martin Musicante (he/his)

Univ. Federal Rio Grande do Norte

DIMAP, Brazil

José-Luis Zechinelli-Martini (he/his)

CPE, Univ. Lyon

LIRIS, France

Reimagining Experimentation: Integrating Quantitative and Qualitative Data, Content and Knowledge production processes from Experiments in the Sciences and Humanities

Genoveva Vargas-Solar, CNRS, Database Group, LIRIS, France

Fédération Informatique de Lyon

Knowledge production model today

- Results oriented: Socratic school "*nonchalance*" (problem understanding) vs speed in productivity (problem solution application DS methods today?)

- Cyborg[1] like knowledge production: human, machine and algorithms

- We need to keep track of problems, solutions and of the production process

→ **to understand, share and reproduce (even if post-mortem)**

[1] Donna J. Haraway, Manifesto Ciborgue, https://edisciplinas.usp.br/mod/resource/view.php?id=1085302&forceview=1

Quantitative and qualitative methodologies for dealing with curated content

Unlocking deeper insights, making informed decisions through the power of mutualised knowledge

5

Seminario Proyecto ANII MENTOR, Mayo, 2025, Montevideo

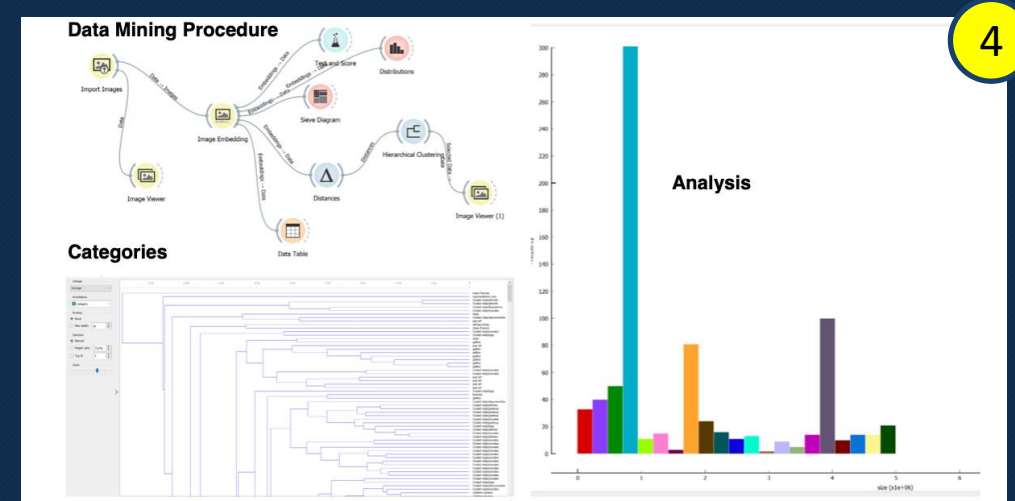# Quantitative & qualitative research methods in a nutshell
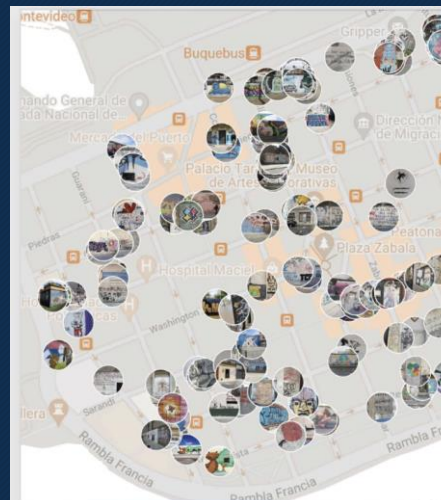
Is it possible to identify the type of political content in graffiti in Montevideo? (1)

(2) **Quantitative**:
Phenomena observations' collected a priori and promote a **deductive approach** for driving conclusions

(3) *geo-location of images, apply classification methods for defining clusters according to content features*

(4)

Quantitative & qualitative research methods in a nutshell

Is it possible to identify the type of political content in graffiti in Montevideo? **1**

2 **Qualitative:**
Explore complex phenomena promoting **inductive techniques** relying on nonstatistical and nonnumerical data collection methods, analysis, and evidence production

3 *criteria to select*
- *urban images that are graffiti*
- *areas of a city to collect images*

define classification concepts & exclusion/inclusion criteria → dataset

From data curation to process curation

It is a matter of locating the knowledge ..

Problem statement (1/2)

- Given a problem stated as a research question:
    - How to go from raw to insights
    - How is data transformed to derive quantitative models?

- How to keep track of the journey of the data and its transformations along an experiment and across different versions of an experiment?

- Who and which processes are accountable of this transformation?

Active & ongoing data collections **management** through their **lifecycle of interest** & **usefulness**

- Enable data **discovery & retrieval**
- Maintain data quality
- Add value
- Provide for **re-use over time**

Done with "artisanal" or semi artisanal procedures using data analytics tools …

Define a meta-data extraction model for reducing the artisanal operations & supporting decision making related to data collections exploitation

- Given a problem stated as a research question:
  - ✅ How to go from raw to insights?
  - `</>` How is data transformed to derive quantitative models?   ← **Explainable IA, …**
- ✅ How to keep track of the journey of the data and its transformations along an experiment and across different versions of an experiment?
- `</>` Who and which processes are accountable of this transformation?   ← **Provenance …**

Problem statement (2/2)

- Given a hypothesis related to a context/discipline:
  - How to build a theoretical framework that has addressed it?
  - How to define the different steps of an experiment?

- How to keep track of the content produced within decision experiment phases and log its transformations along an experiment and across different versions of an experiment?

- Who and which processes are accountable of this transformation?

Data-driven process (i.e., experiment) curation

Process of identifying, managing and versioning research content produced along qualitative experimental processes and keeping track of experimental flow driven by researchers decision-making that produced it (i.e., content)

Process curation problem stated as a modelling problem

- Process of qualitative experiments → **qualitative workflow**

- Process of content construction → **versioned artefacts**

- Research team members → **researcher in the loop**

- Model the production process of content → **consensus**

Adorjan, A., Vargas-Solar, G., & Motz, R. (2022, December). Towards a human-in-the-loop curation: A qualitative perspective. In 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.

Problem statement (2/2)

✅ Given a hypothesis related to a context/discipline:

- How to build a theoretical framework that has addressed it?
- How to define the different steps of an experiment?

⚠️ How to keep track of the content produced within decision experiment phases and log its transformations along an experiment and across different versions of an experiment?

⚠️ • Who and which processes are accountable of this transformation?

Dataverse

- Data Repository Platform
  - Specific software for publishing, citing, and discovering datasets.
  - The most prominent example is the open-source Dataverse project developed by the Institute for Quantitative Social Science at Harvard University
- Platforms and ecosystems for managing, analysing, and sharing data

University of Michigan:
https://www.icpsr.umich.edu/web/about/cms/2365

University of Hamburg:
https://www.fdm.uni-hamburg.de/en/fdm.html

University of Arizona:
https://arizona.figshare.com

Grenoble Dataverse:
https://entrepot.recherche.data.gouv.fr/dataverse/univ-grenoble-alpes
https://scienceouverte.couperin.org/cellule-data-grenoble-alpes/
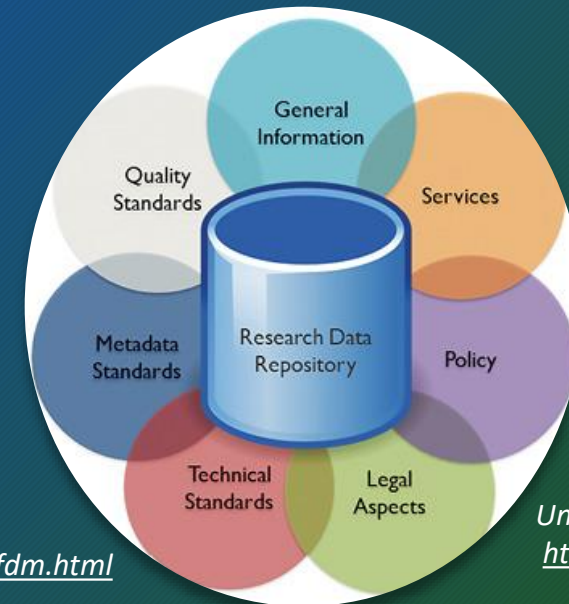
Seminario Proyecto ANII MENTOR, Mayo, 2025, Montevideo

Building a processes universum for Earth & Life Sciences

The compass guiding us towards a deeper understanding of our planet's phenomena

Seminario Proyecto ANII MENTOR, Mayo, 2025, Montevideo

Data-centred experimentation in Life & Earth Sciences

- *Research Questions:*

- *Is there content about a Caravela observation event?*
  - *Which criteria determine whether an event concerns a Caravela Portuguesa observation?*

- *What is the behaviour of Caravela Portuguesa jellyfish along Brazilian coasts?*

- *Can we predict their arrival and their possible migration to other coasts in South America?*

Seminario Proyecto ANII MENTOR, Mayo, 2025, Montevideo

17

Wrangling media data to track Carabela Portuguesa Jellyfish: quali-quanti perspective

Harvesting

- Social media-based data
- Press Media

Question:
Which criteria determine whether an event concerns a Caravela observation?

Qualitative Filtering

Integrated social, press media dataset with associated meta-data & explanation, categories

Biodiversity team
- Inclusion criteria
- Discussion defining features to declare a content valid (representing a Caravela observation event)

Natural or induced earthquakes?

Answered by analyzing & exploring seismic data !

✸ Epicentro
▲ Estacao

| DATA	HORA –UTC	COORDENADAS +/- 5 km lat; +/- 5 km long	MAG
08/04/2022	02:31	-08.52 S ; -35.88 W	1.7 mR
13/04/2022	15:25	-05.75 S ; -37.53 W	1.5 mR
14/04/2022	09:03	-09.84 S ; -39.92 W	2.1 mR
14/04/2022	15:07	-10.99 S ; -39.35 W	2.1 mR
16/04/2022	10:23	-09.88 S ; -36.87 W	1.5 mR
17/04/2022	06:04	-05.68 S ; -36.14 W	1.6 MLv
19/04/2022	15:10	-11.29 S ; -40.52 W	1.3 mR
22/04/2022	00:35	-03.36 S ; -40.19 W	1.0 MLv
22/04/2022	19:14	-11.29 S ; -40.52 W	1.2 mR
25/04/2022	17:12	-11.24 S ; -40.49 W	2.5 mR
27/04/2022	15:25	-09.63 S ; -35.84 W	1.8 mR
29/04/2022	21:43	-10.03 S ; -36.89 W	2.0 mR

Row seismic data

- [Seismic Analysis Code](#) (**SAC**) binary/text format
    - 1 file per waveform
    - Produced by ADAGEO partners

1

2

seismic event to report in a bulletin
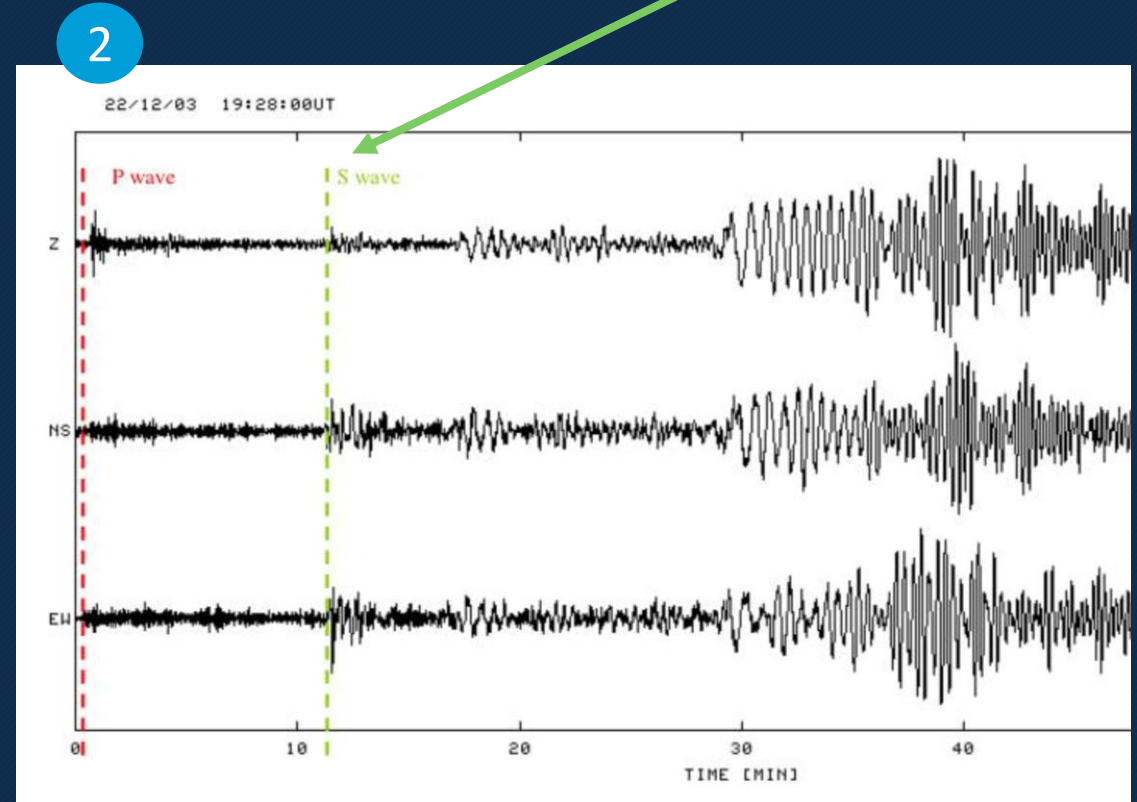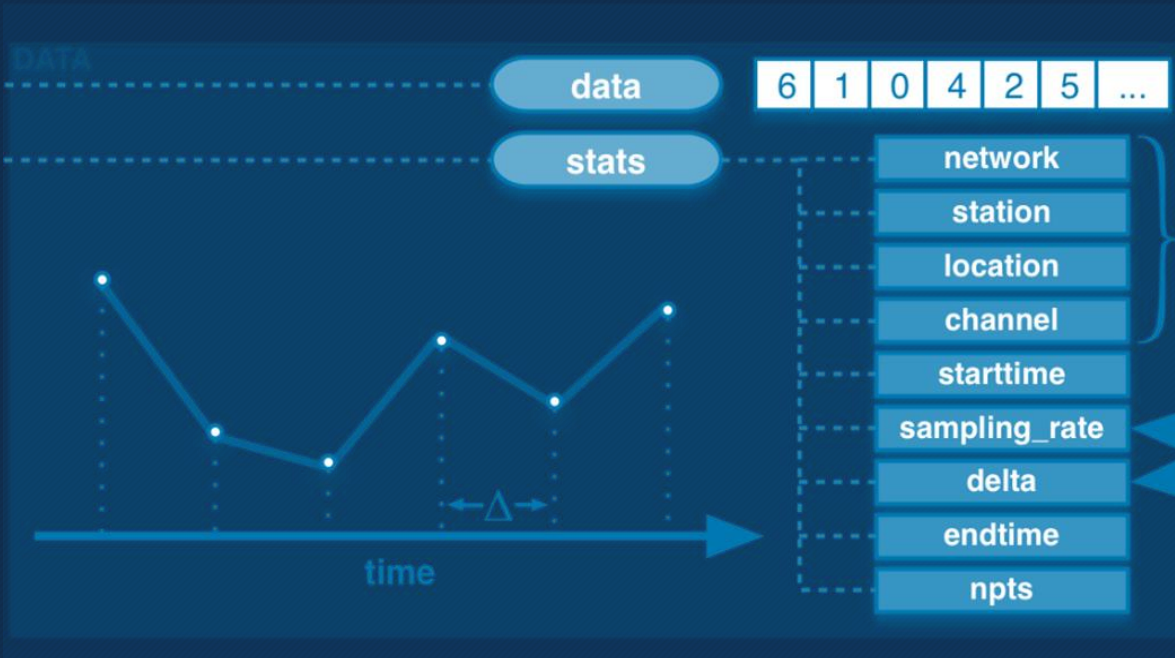
LETITIA Project on curating experimental knowledge production

Vision: Key elements of our solution

Building a experimental universum

Quantitative Filtering

Life & Earth Sciences
Data Harvesting

Classification (O/1)
According to
- which type of content?
- which variables?

Integrated social,
press media dataset with
associated meta-data &
annotation, categories

Life & Earth Sciences
Data Lake

Qualitative Filtering

Life & Earth Sciences
Experimental Universum

Scientific team
- Inclusion criteria
- Discussion defining features to declare a
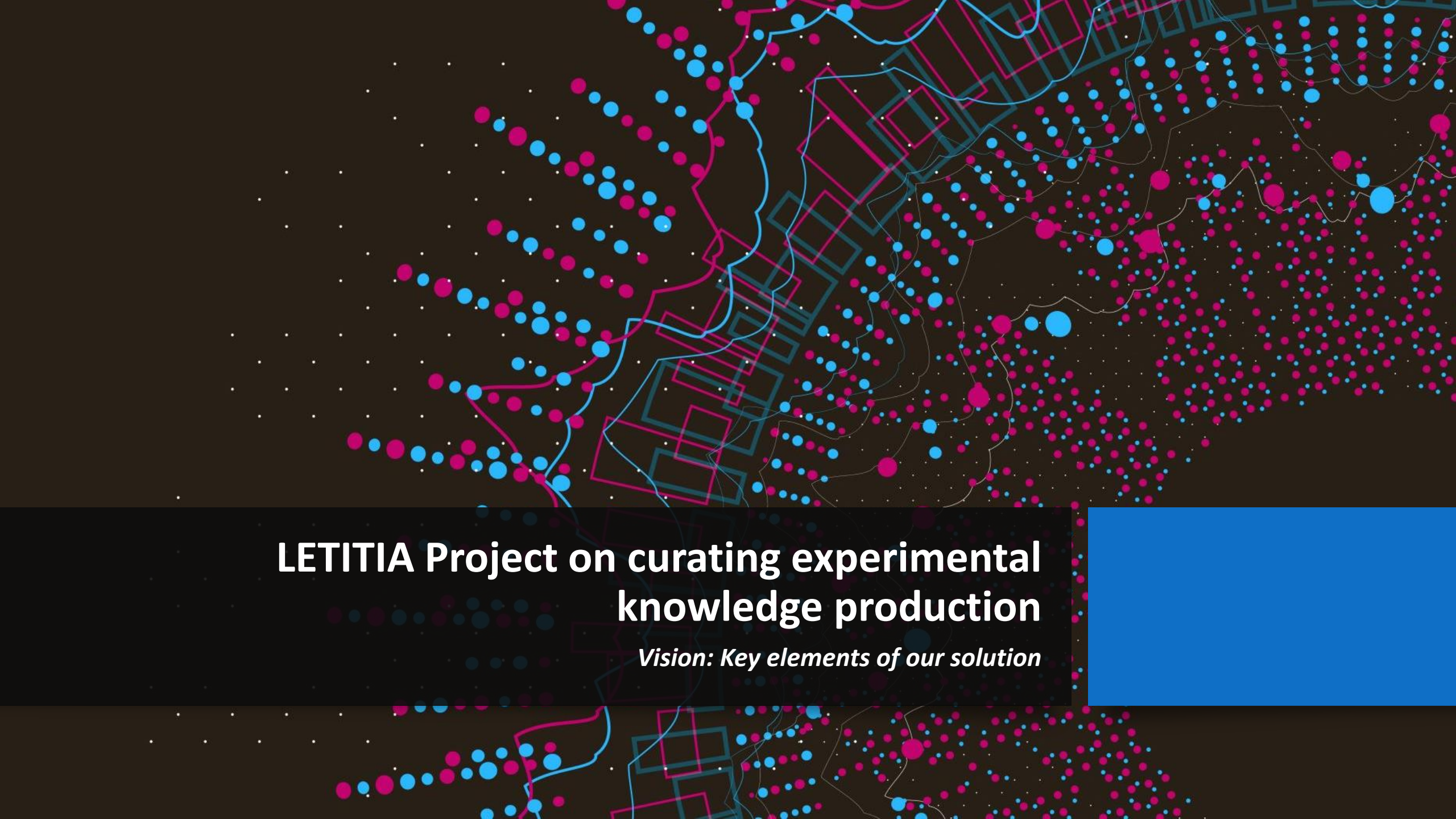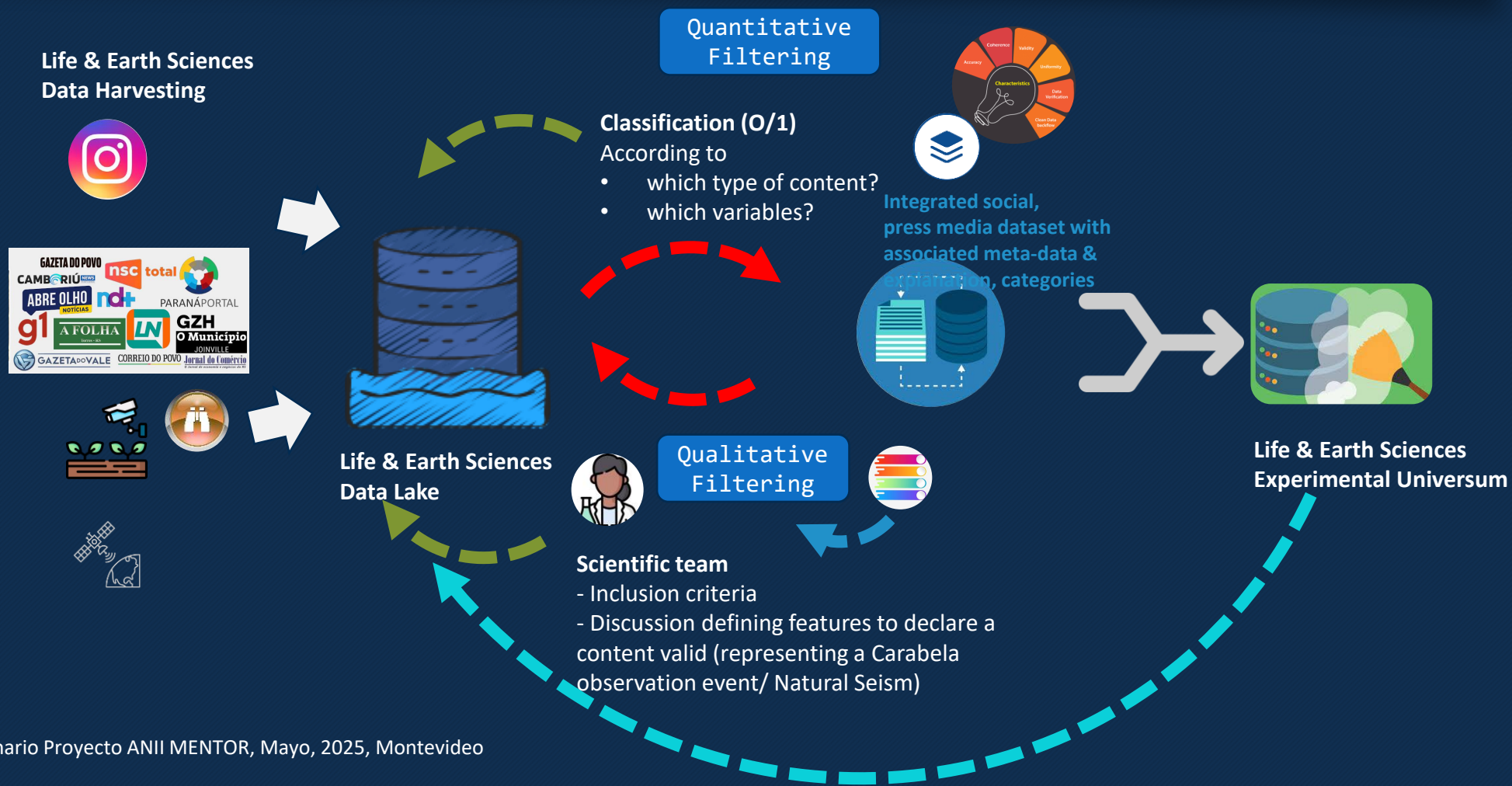content valid (representing a Carabela
observation event/ Natural Seism)

Quanti-qualitative experiment curation[1]

3 HITL decision making

Phase I - Problem Statement

1

Content

Research Questions
Hypothesis
Methods
Review Theories
Construction of Theoretical Framework

2

Methodological Rules

Phase II Data Acquisition

Collect data
Explore
Clean Data
Verify Data Reliability

Phase III Data Management

Generate | Evaluate
Meta-Data
Contextualize Data

Phase IV Analysis

Round of Experiments
Debate
Reflection

Phase V Report

Visualize
Evaluate the Pro
Editorial Voice
Publish

A Adorian, Towards a Researcher-in-the-loop Driven Curation Approach for Quantitative and Qualitative Research Methods,European Conference on Advances in Databases and Information Systems, 647-655

Content

Phase I -
Problem
Statement

Research Questions
Hypothesis
Methods
Review Theories
Construction of
Theoretical
Framework

Phase II
Data
Acquisition

Collect data
Explore
Clean Data
Verify Data Reliability

Phase III
Data
Management

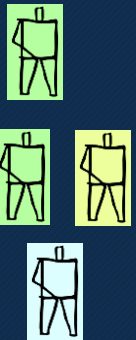Generate | Evaluate
Meta-Data
Contextualize Data

Phase IV

Analysis

Round of
Experiments
Debate
Reflection

Phase V
Report

Visualize
Evaluate the Pro
Editorial Voice
Publish

Artefact

Content: artefact model

g1
Caravelas-portuguesas assustam banhistas e são confundidas com camisinhas no litoral de SP glo.bo/3mSfYL4 #G1
Translate post

5:45 PM · Dec 27, 2020

430   1.6K   9.3K   57

①

Carabela Portuguesa attack

This is the image of a Carabela Portuguesa in the coast

This is not an acceptable evidence because the post is not geo-located

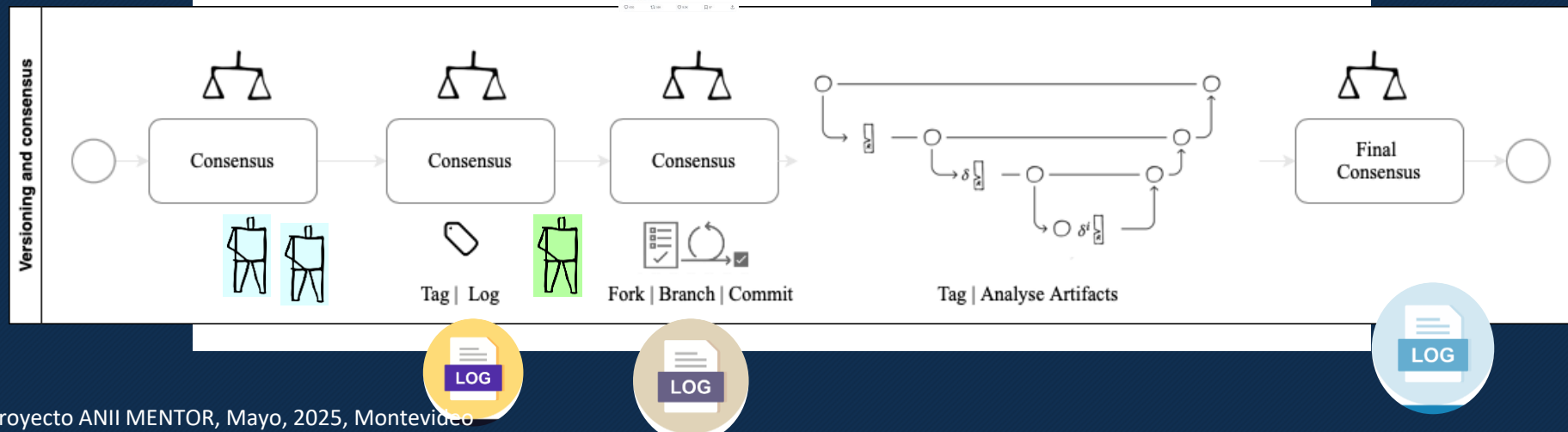# Quanti-qualitative experiment curation: versioning based on consensus

Phase II
Data Acquisition

Collect data
Explore
Clean Data
Verify Data Reliability

Phase III
Data Management

Generate | Evaluate
Meta-Data
Contextualize Data

Exploring curated data driven research processes

- How did the classification criteria of Caravela Portuguesa images proposed by biologists and data scientists evolved

- After how many consensus votes did the Biologists group converged in a final list of categories?

- Which content was produced by junior members of the team?

- How long did it take to converge into a dataset that could be useful to move from the data collection to the data tagging phase?

Final take aways: Curated Data-Centric Experiments in Sciences

27

Seminario Proyecto ANII MENTOR, Mayo, 2025, Montevideo

Final Takeaways

- In the era of high-speed production of cyborg knowledge provide a systematic track of the conditions of production
  - Open science, accountability, reproducibility, explainability, authorship revindication
  - Insights, scientific conclusions backup up by data & process logs
- Challenging content and process maintenance problem
  - Harvesting, modelling, storage, versioning, consensus …
- Method for performing transdisciplinary science: quantitative and qualitative methods
  - Understand, reproduce, assess, share results and best practices, ease and systematise manual and artisanal practices