Leibniz Institute
for the Social Sciences

Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

Leibniz
Association

gesis Leibniz Institute for the Social Sciences

8: Documenting Pitfalls

Photo: Katrin Weller

gesis Leibniz Institute for the Social Sciences

Today's Schedule

Monday, 12.12.	
9:30-11:00	Introduction to documentation approaches for social media data
11:00-11:15	*Break*
11:15-12:00	Group work: designing a social media study and identifying errors
12:00-12:30	*Break*
12:30-14:00	Group work: documenting an example case; discussion and conclusions

optional: starting 13:30 - GESIS CSS Seminar:
Deborah Nozza - Roadmap to universal hate speech detection

Leibniz Association

Agenda

- Session 1: Introduction to Research with Social Media Data (SMD)
- Session 2: SM Data Collection
- Session 3: SMD Preprocessing and Analysis
- Session 4: Potential Pitfalls of SMD
- Session 5: Identifying Pitfalls with help from surveys
- Session 6: Identifying Pitfalls in SMD
- Session 7: Mitigating Pitfalls
- Session 8: Documenting Pitfalls
- Session 9: Recap and Conclusions

Documentation Approaches and Their Application

Documentation / Frameworks as Guidelines

Documentation standards and frameworks…

a. explicitly pinpoint what the important aspects are that need to be considered
   (E.g. source/funding of your data, impact on stakeholders,)
b. advise what potential pitfalls are, connected to each aspect
   (E.g. data collection → biased queries used for retrieval, model building → improper parameter selection)

In this way, they set the direction on what needs to be focused on, documented and avoided/mitigated.

Other Error Frameworks

Error Frameworks: Total Twitter Error

- Twitter as the model organism for social media studies

Hsieh, Yuli Patrick, and Joe Murphy. "**Total twitter error.**" *Total survey error in practice* (2017)

- TTE —> "a general error framework for Twitter opinion research…"

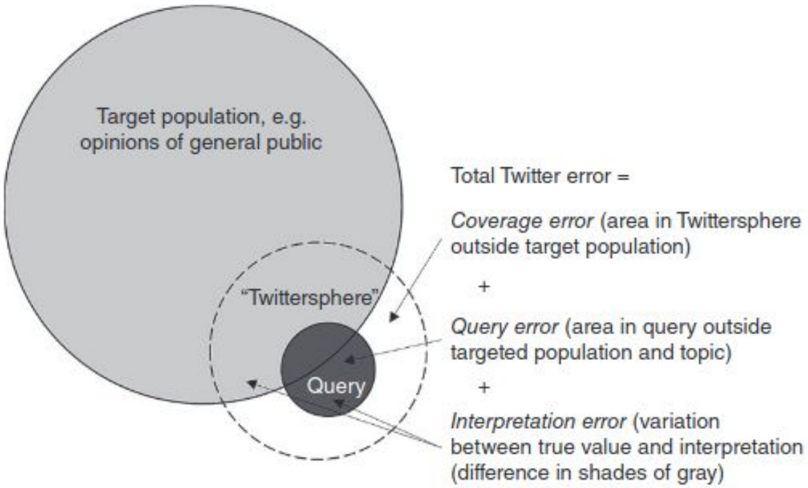


Figure 2.1 Theoretical spaces of Twitter data error.

Error Frameworks: Total Twitter Error

- **Coverage error:** over- and under-coverage of both Twitter users and posts; difference between the target population and Twitter units

- **Query error:** when a researcher mis-specifies the search queries for data collection

- **Interpretation error:** when a researcher uses human or machine methods to infer the construct of interest

Error Frameworks: Total Error Framework for Big Data (TEF)

Amaya, Ashley, Paul P. Biemer, and David Kinyon. "**Total error in a big data world: Adapting the TSE framework to big data.**" *Journal of Survey Statistics and Methodology* 8, no. 1 (2020)

- "evaluate the quality of Big Data using an approach similar to the total survey error (TSE) framework"
- Maps errors in the TSE to errors in Big Data (including web and social media data)
- Utilizes a pipeline of "ETL" --- Extract, Transform, and Load

Error Frameworks: Total Error Framework for Big Data (TEF)

Error Name	**Definition**
Coverage Error	errors that arise due to the difference between the target population and the population under study due to the platform and queries used
Sampling Error	errors that arise as a result of analyzing a (typically random) subset of the population of interest rather than the entire population (census)
Specification Error	when the concept (or construct) needed to address a research question does not precisely align with the concept implied by the dataitem
Non-response / Missing error	A consequences of missing items or units, or undercoverage

gesis Leibniz Institute for the Social Sciences

Error Frameworks: Total Error Framework for Big Data (TEF)

Error Name	Definition
Measurement / Content Error	a consequence of a number of factors including the measurement process, transcription errors, data conversion errors, false readings from mechanical devices, outdated information
Processing Error	data entry, coding, editing, disclosure limitation, and variable conversions or transformations
Modeling / Estimation Error	deficiencies in missing data and coverage error weighting adjustments, as well as imputation for item missing data.
Analytic error	errors made by data users and clients in analyzing and interpreting the results.

Leibniz Association

From Error Frameworks to Documentation?

Documentation along the TED-On Framework

Following the Errors listed in the TED-On Framework can help to identify potential pitfalls along a research design…

Example: Presidential Approval (PA) on Twitter

Specification sheets for studying PA from tweets

Construct: Presidential Approval	**Target Population:** American adults	**Ideal Measurement:** Daily posts with positive or negative sentiment towards Trump	**Platform(s):** Twitter

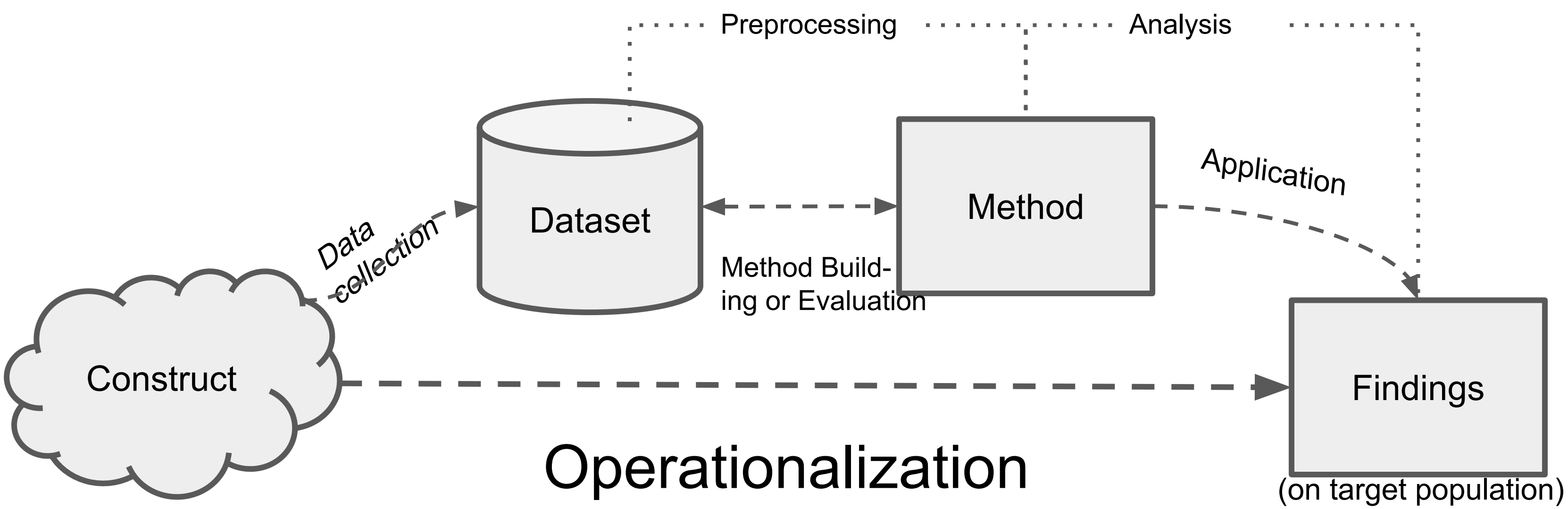
Stage	**Measurement Error**	**Explanation**	**Representation Error**	**Explanation**
Construct Definition	Validity	Tweets about Trump may not be about his Presidential role		
Platform Selection	Platform Affordances	Recommendations by Twitter, Twitter TOS	Platform Coverage	Twitter Population not the same as Target population
Data Collection	Trace Selection	‘Trump’ keyword may include tweets about extended Trump family	User Selection	We miss tweets by groups who may use certain nicknames

Specification sheets for studying PA from tweets (contd…)

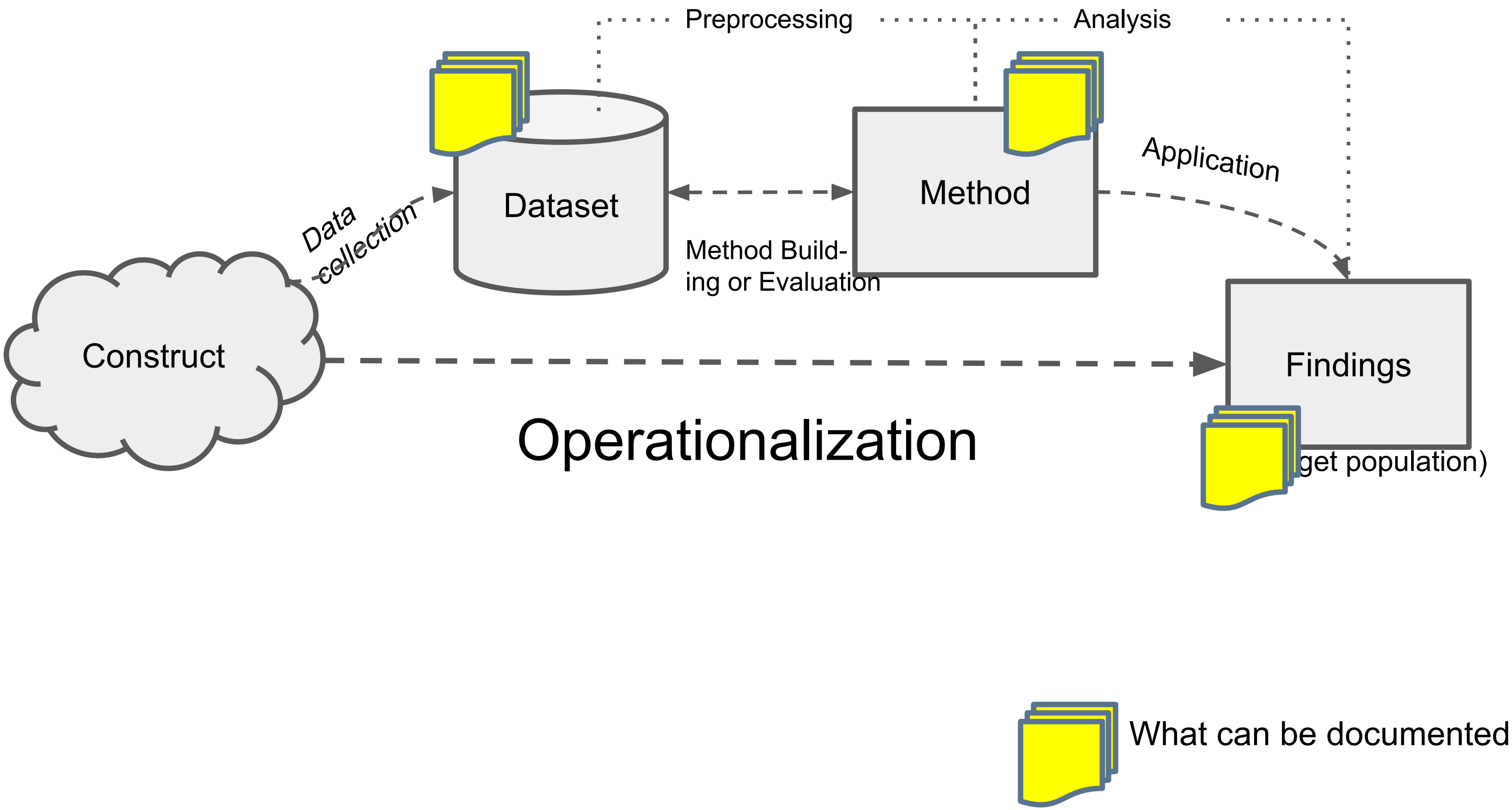
Construct: Presidential Approval	**Target Population:** American adults	**Ideal Measurement:** Daily posts with positive or negative sentiment towards Trump	**Platform(s):** Twitter	
Stage	**Measurement Error**	**Explanation**	**Representation Error**	**Explanation**
Data Preprocessing	Trace Augmentation	sentiment lexicon for annotating approval --- social media vocabulary mismatch, target-independent lexicon	User Augmentation	Use of self-reported age, gender and ethnicity may include misreports
	Trace Reduction	removing non-textual content. Might disregard information in images	User Reduction	Remove users who are not that active
Data Analysis	Trace Measurement	Averaging sentiment of a user on a single day --- may provide mixed traces for particularly vocal users	Adjustment	using age, gender and ethnicity may not sufficiently capture the self-selection of users

Different Documentation Approaches?

Prototypical Pipeline - Artifacts and Steps



Prototypical Pipeline - Artifacts and Steps



Documentation for Datasets

Documentation for Datasets

- Datasets are vital and ensuring their quality is of high importance
- Biased data = biased models and biased measures
- Several issues in data quality especially when using social media and web data
  - Demographic biases
  - Observational and unsolicited data
  - Access

Documentation for Datasets

Practices for data sharing, data management and data documentation are more advanced in other research fields, but approaches are not easily transferable to specific requirements of social media data (work in progress).

Examples for relevant (ongoing) work in the social sciences:

- Data Documentation Initiative (DDI) https://ddialliance.org/
- CESSDA's activities related to social media data
- Specialized archives starting to think about integrating social media data

In the following we introduce some current initiatives with relation to the computer science / social media research community.

gesis Leibniz Institute for the Social Sciences

Documentation for Datasets: *Datasheets for Datasets*

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." *arXiv preprint arXiv:1803.09010* (2018).

- Inspired by the electronics industry where components come with **specification sheets**
- Analogously datasheets describe datasets' many components and facets - focused on datasets used in machine learning
- Intended for dataset creators and consumers, but might also benefit other stakeholders like policy-makers
- Improve reproducibility and accountability - should be consulted before dataset creation
- Ethical considerations integrated with main categories, not separately.

Documentation for Datasets: Datasheets for Datasets

Contains 7 main fields with several subfields:

Motivation
(for what purpose was the dataset created?)

Collection process
(including time frame, sampling, consent)

Uses
(existing and potential use cases, non-suitable use cases)

Distribution
(availability, copyright, other restrictions)

Composition
(What does the dataset consist of - text, images? Does it relate to people/populations/demographics?)

Preprocessing, cleaning, labeling
(raw vs. processed data, software used)

Maintenance
(updates, errata, contact information)

Documentation for Datasets: Datasheets for Datasets

Contains 7 main fields with several subfields:

Uses
(existing and potential use cases, non-suitable use cases)

Motivation
(for what purpose was the dataset created?)

Collection process
(including time frame, sampling, consent)

Distribution
(availability, copyright, other restrictions)

Composition
(What does the dataset consist of - text, images? Does it relate to people/populations/demographics?)

Preprocessing, cleaning, labeling
(raw vs. processed data, software used)

Maintenance
(updates, errata, contact information)

Documentation for Datasets: Data Statements

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587-604.

- Targeting NLP datasets (speech or writing, potential annotations)
- Analogous to documentation developed in medicine and psychology for explicitly reporting the populations under study
- ‘A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize...’
- Engage with issues of exclusion, overgeneralization, underexposure, generalizability, reproducibility
- Includes two example applications (one of them Twitter data)

gesis Leibniz Institute for the Social Sciences

Documentation for Datasets: Data Statements

Proposed Schema:

Curation Rationale
(which texts are included and what were the goals in selecting them? Sub-selection? Automated processes?)

Language Variety
(the language variety, such as dialect, of the text)

Speaker Demographic (e.g. race, gender, native language, socioeconomic status)

Text Characteristics
(genre or topic which might affect the vocabulary or register of the text)

Annotator Demographic
(e.g. crowdworkers? trained experts? Race, native language, training/expertise details)

Provenance Appendix
(For datasets built out of existing datasets, a listing of these existing datasets)

Recorded Quality
(for data with audiovisual components)

Other
(e.g. info about the curators of the dataset)

Speech Situation
(The context, such as time, place, platform in which the text was generated, intended audience, oral/signed/written? transcriptions?)

Leibniz Association

Documentation for Datasets: Data Statements

Proposed Schema:

Curation Rationale
(which texts are included and what were the goals in selecting them? Sub-selection? Automated processes?)

Text Characteristics
(genre or topic which might affect the vocabulary or register of the text)

Recorded Quality
(for data with audiovisual components)

Language Variety
(the language variety, such as dialect, of the text)

Annotator Demographic
(e.g. crowdworkers? trained experts? Race, native language, training/expertise details)

Other
(e.g. info about the curators of the dataset)

Speaker Demographic
(e.g. race, gender, native language, socioeconomic status)

Provenance Appendix
(For datasets built out of existing datasets, a listing of these existing datasets)

Speech Situation
(The context, such as time, place, platform in which the text was generated, intended audience, oral/signed/written? transcriptions?)

Documentation for Datasets: Others

- **Dataset Nutrition Labels**
  - Holland, Sarah, et al. "*The dataset nutrition label: A framework to drive higher data quality standards*."
  - Chmielinski, Kasia S., et al. "*The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence.*"

- **Data Cards** Lighter version adapted from Datasheets

- **Factsheets** Arnold, Matthew, et al. "*FactSheets: Increasing trust in AI services through supplier's declarations of conformity.*"

Documentation for Models

Documentation for Models

- AI models are deployed in several high-stakes decision making processes --- criminal justice systems, hiring, content moderation
- In computational social sciences and social computing, models are either developed for a particular research scenario (**custom**) or re-used from a different context (**off-the-shelf**)
- Often these models are used for analysing social media and web data to find impactful results --- the extent of harassment faced by politicians, voting preferences of different demographic groups

Documentation for Models: Model Cards

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. 2019.

- “specifically aims to standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical, inclusive, and fair considerations.”
- Focuses on ML models --- NLP and Computer vision methods
- Intended to help different stakeholders such as model developers, ML practitioners, policy makers and impacted individuals
- Model cards: model reporting, approx 1-2 pages

Documentation for Models: Model Cards

Sections (with different sub-sections):

- Model details: incl. creators, date, type (Naive Bayes classifier, Convolutional Neural Network, etc.), citation details. Also: Information about training algorithms, parameters, fairness constraints
- Intended use: primary use cases and users, out-of-scope use cases
- Factors: reflecting on groups as categories eg based on demographics, while considering privacy implications. Instrumentation and Environment (e.g. camera quality, external noises). Reflection on factors that may influence performance.
- Metrics: e.g. model performance measures, decision thresholds.

gesis Leibniz Institute for the Social Sciences

Documentation for Models: Model Cards

Sections (continued):

- Evaluation Data: What dataset were used, how were they created or preprocessed [potential link to dataset documentation schemes]
- Training Data: ideally same level of detail as evaluation data
- Quantitative Analyses: for each factor: results of evaluating the model according to the chosen metrics, confidence interval values.
- Ethical Considerations: sensitive data? Human life impact (e.g. health and safety)? Risks and harms? Mitigation strategies?
- Caveats and Recommendations: e.g. any groups that were not represented? Ideal characteristics of datasets.

Other Approaches

- Other helpful frameworks for reflecting on issues with web and social media data --- Measurement Theory [Jacobs and Wallach], Social Biases [Olteanu et al], Internal Algorithmic Auditing [Raji et al]

- Future work: bringing it all together, also including experiences from other research areas.

Summary and Takeaways

- Some of these documentation examples are not prescriptivist
- Users are encouraged to add new fields if needed
- Often the act of filling out the documentation is supposed to be generative --- help researchers and designers reflect on their decisions

- Future work: bringing it all together, also including experiences from other research areas.

Applying the Documentation

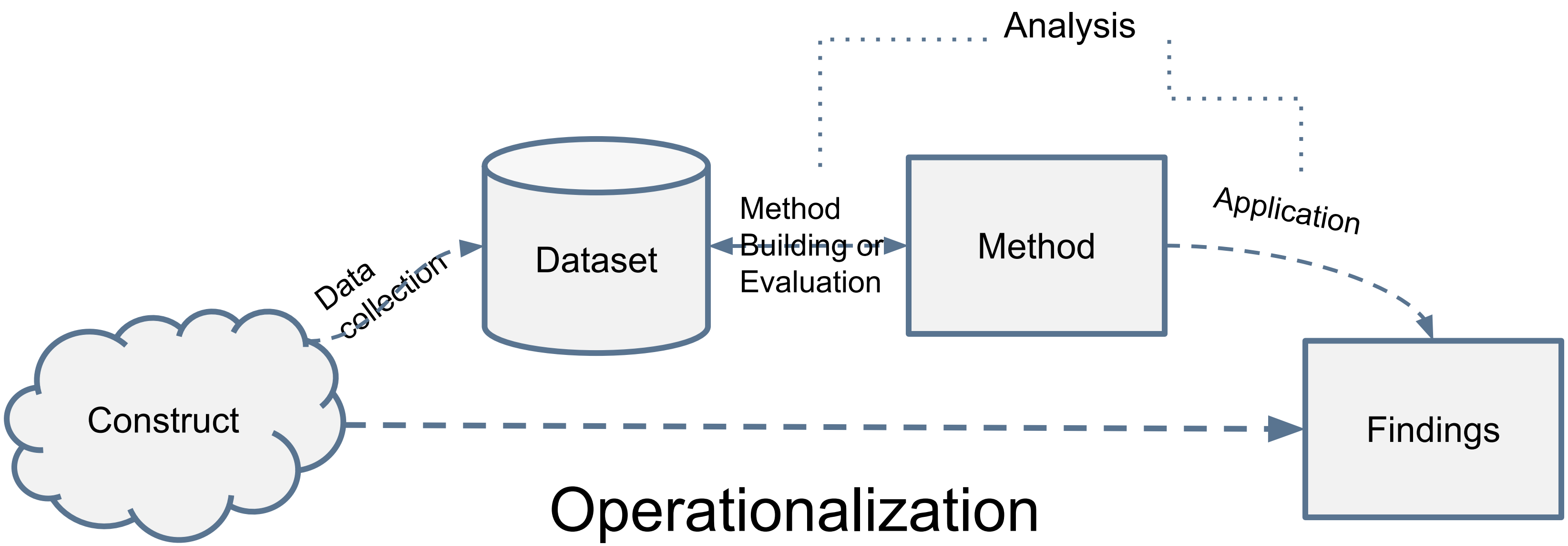
Example case study:
Xenophobic Attitudes towards migrants on Whatsapp

Leibniz Institute for the Social Sciences

Applying the Documentation

- We use an example case study leveraging web and social media data
- Meant to give examples of *some* errors, issues, and pitfalls
- Not comprehensive --- we do not include all issues possible but demonstrate how the frameworks can help us unearth issues

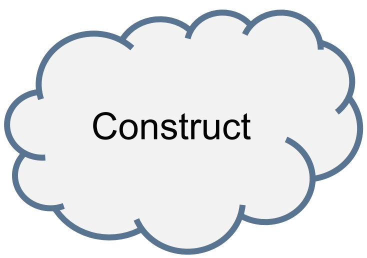
The Pipeline



gesis Leibniz Institute for the Social Sciences

The Pipeline

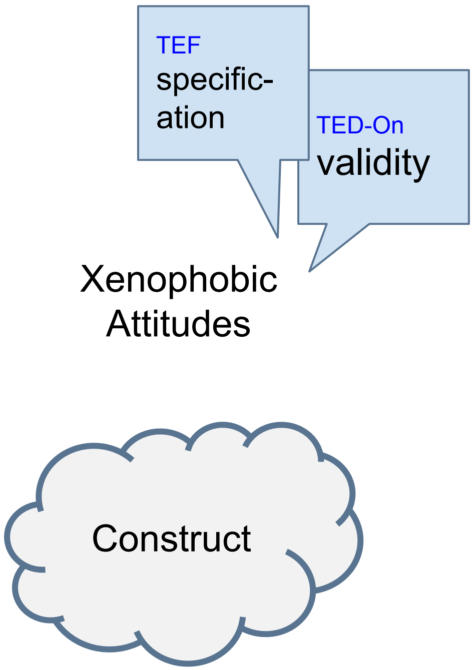
Xenophobic
Attitudes



Leibniz Association

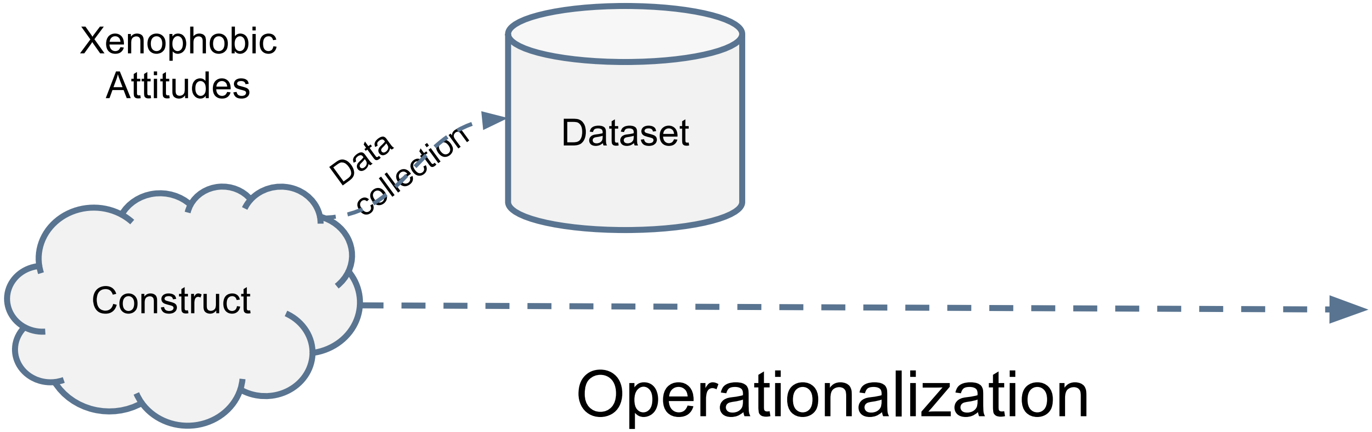
gesis Leibniz Institute for the Social Sciences

The Pipeline



The Pipeline

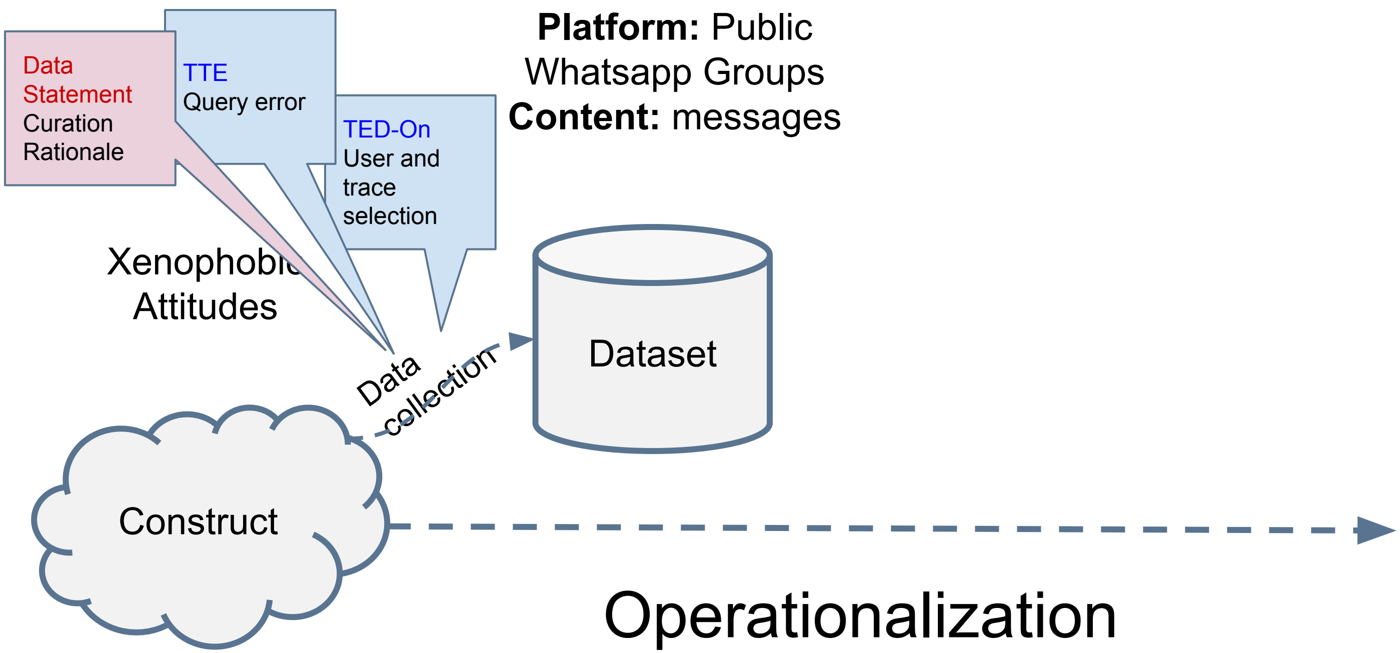
Platform: Public Whatsapp Groups
Content: messages



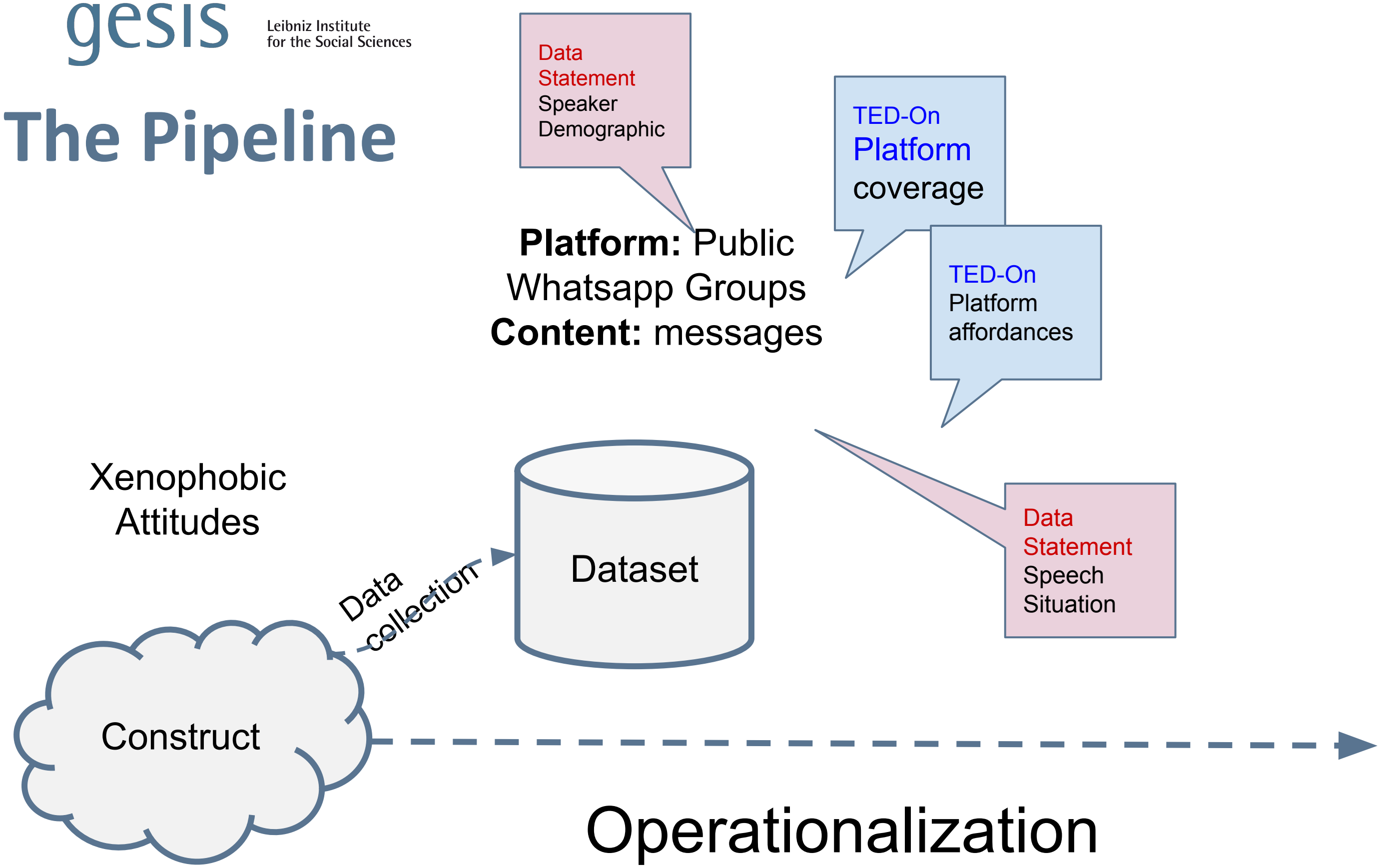
Operationalization

gesis Leibniz Institute for the Social Sciences

The Pipeline

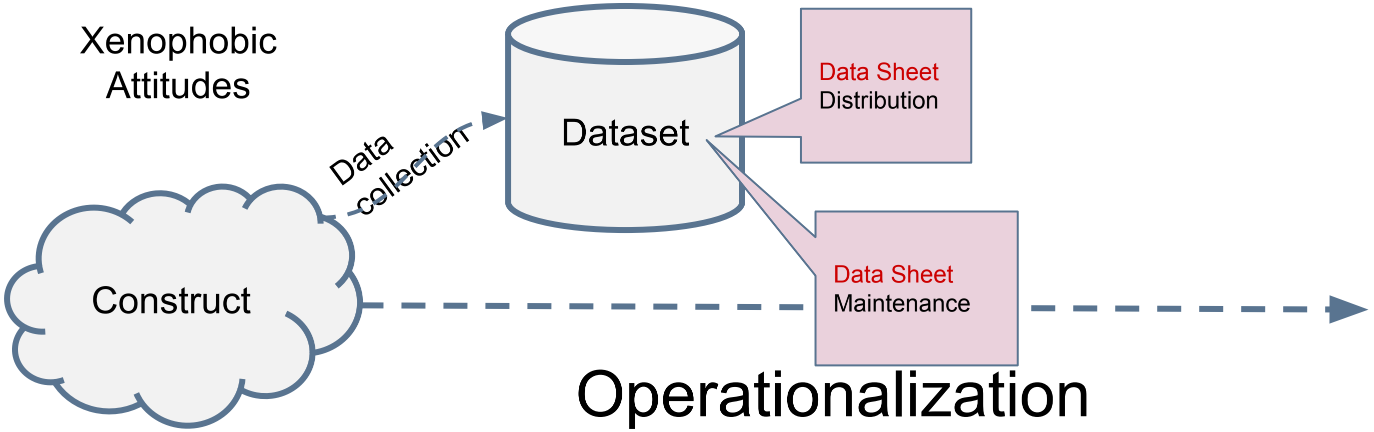


The Pipeline

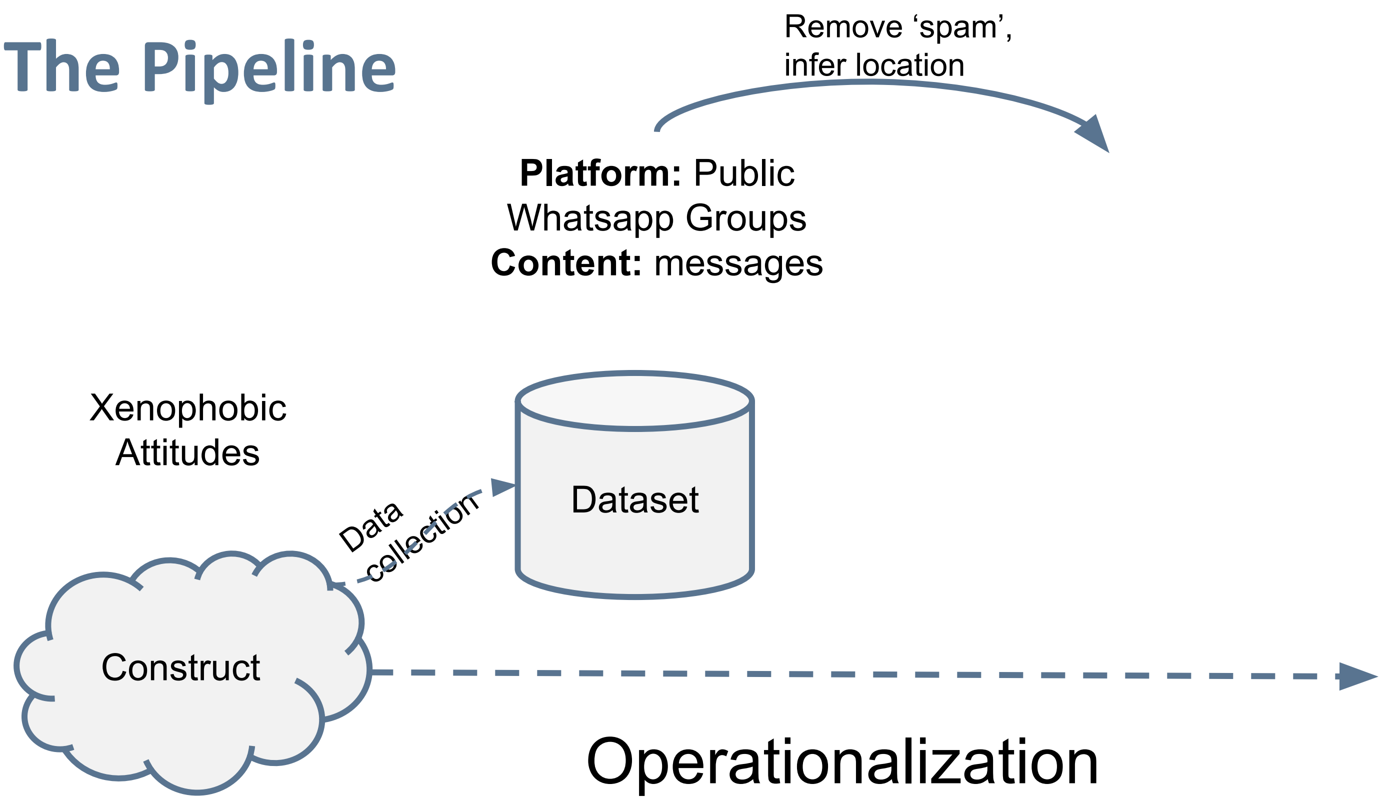


The Pipeline

Platform: Public Whatsapp Groups
Content: messages

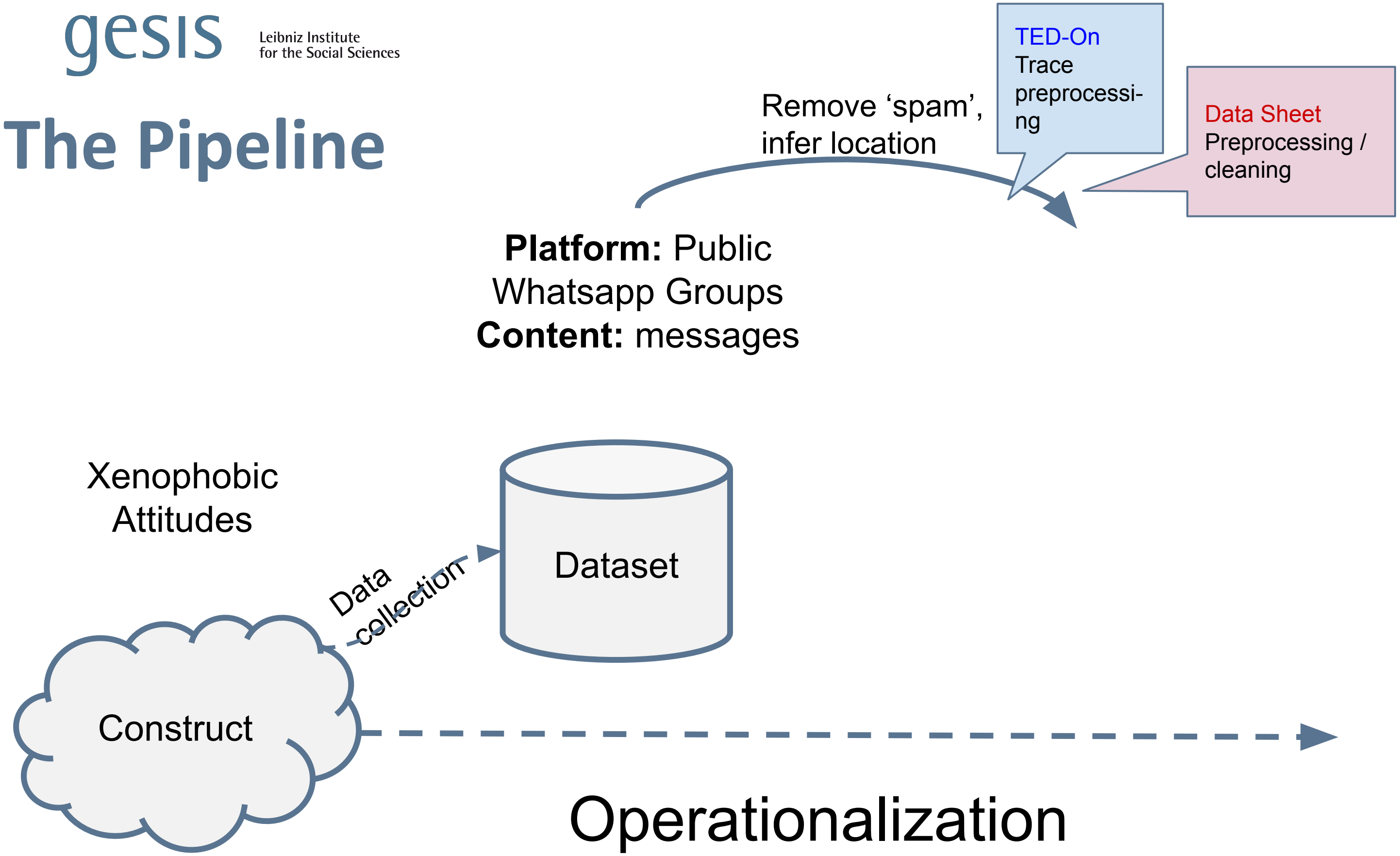


The Pipeline



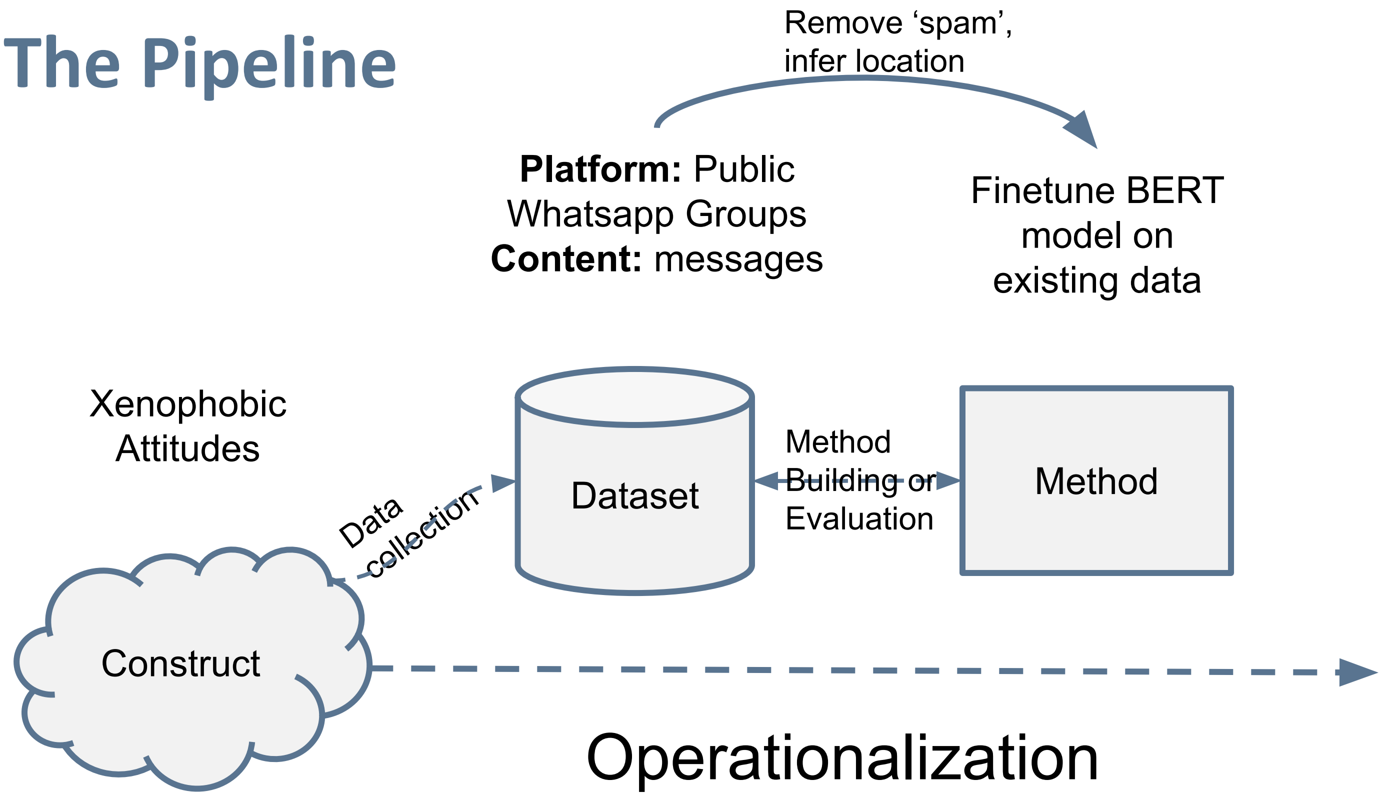
gesis Leibniz Institute for the Social Sciences

The Pipeline

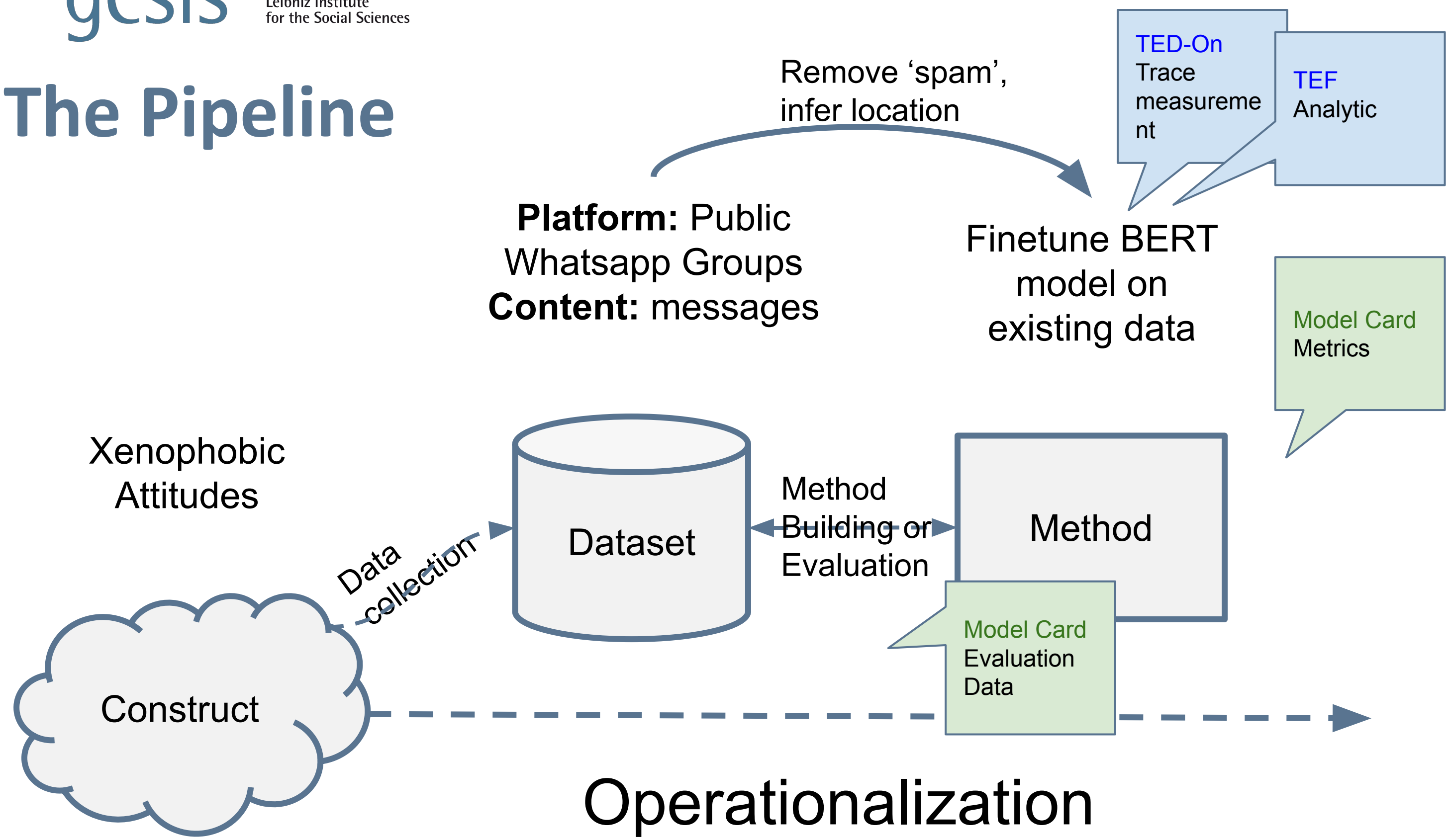


gesis Leibniz Institute for the Social Sciences

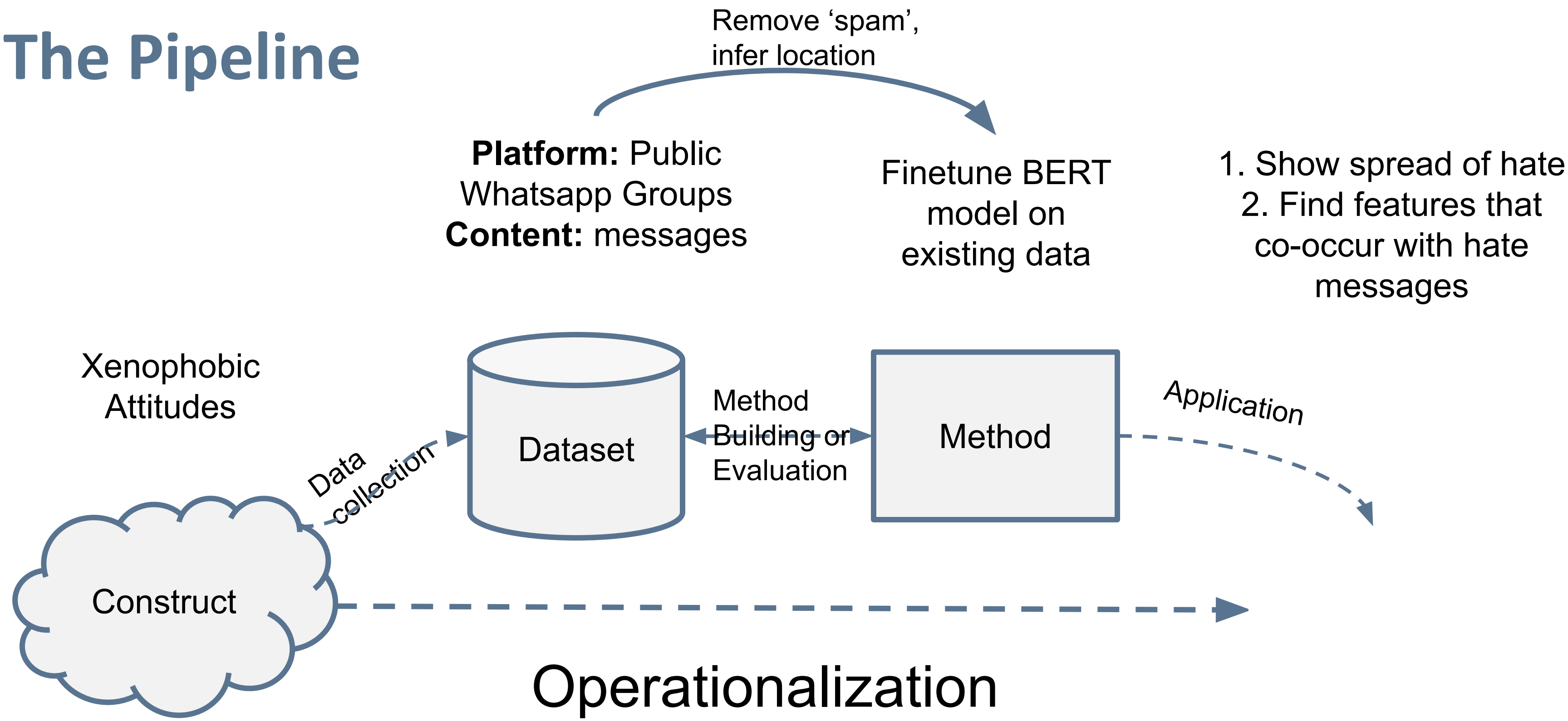
The Pipeline



The Pipeline

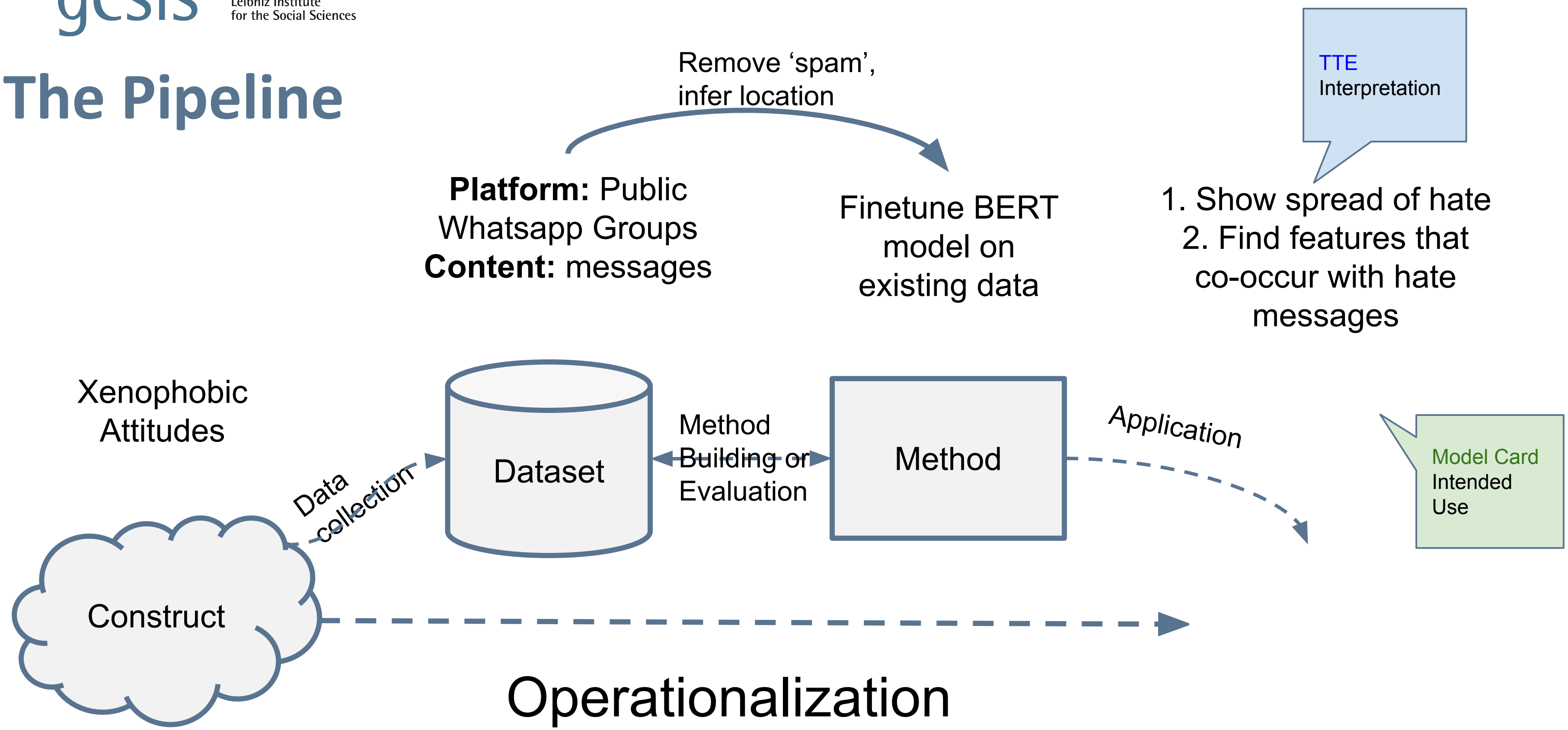


The Pipeline

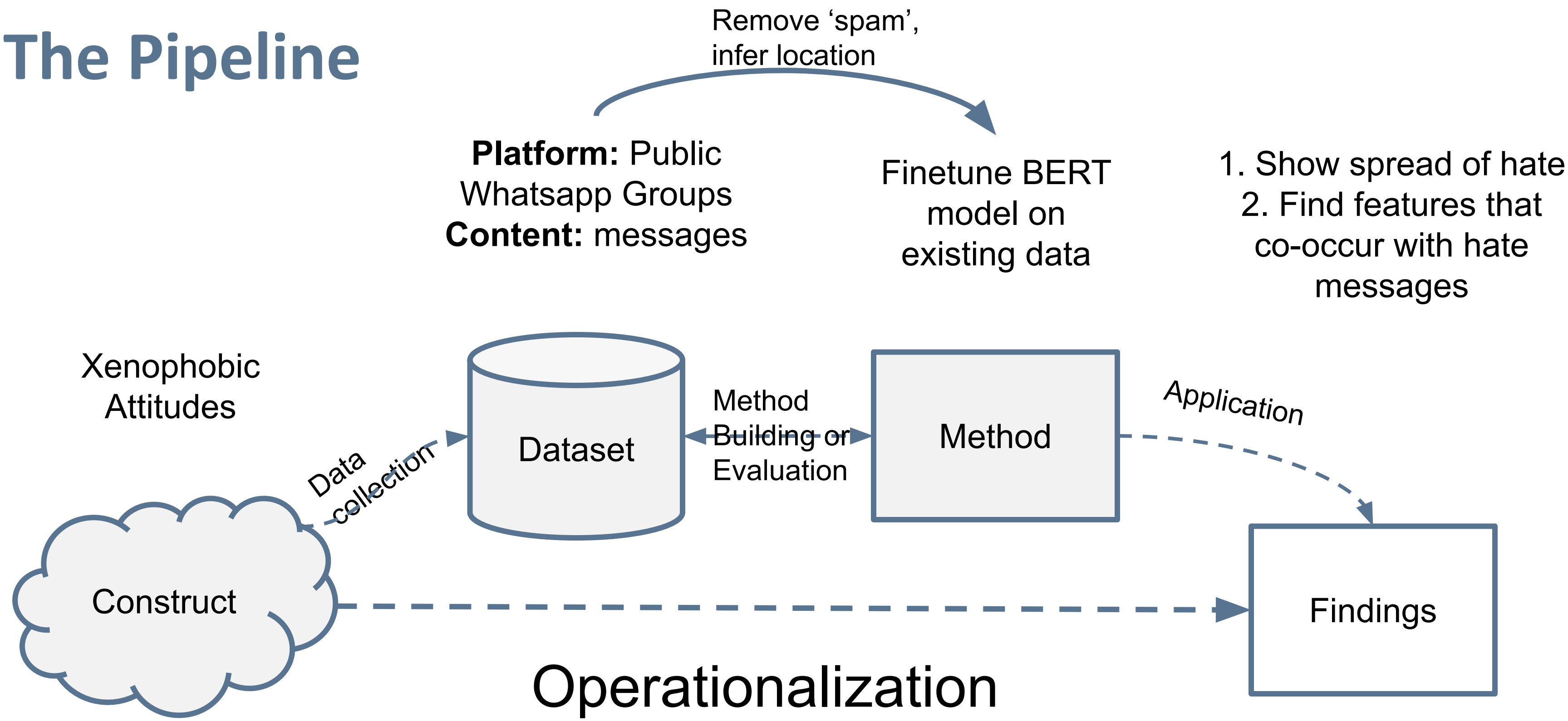


gesis Leibniz Institute for the Social Sciences

The Pipeline

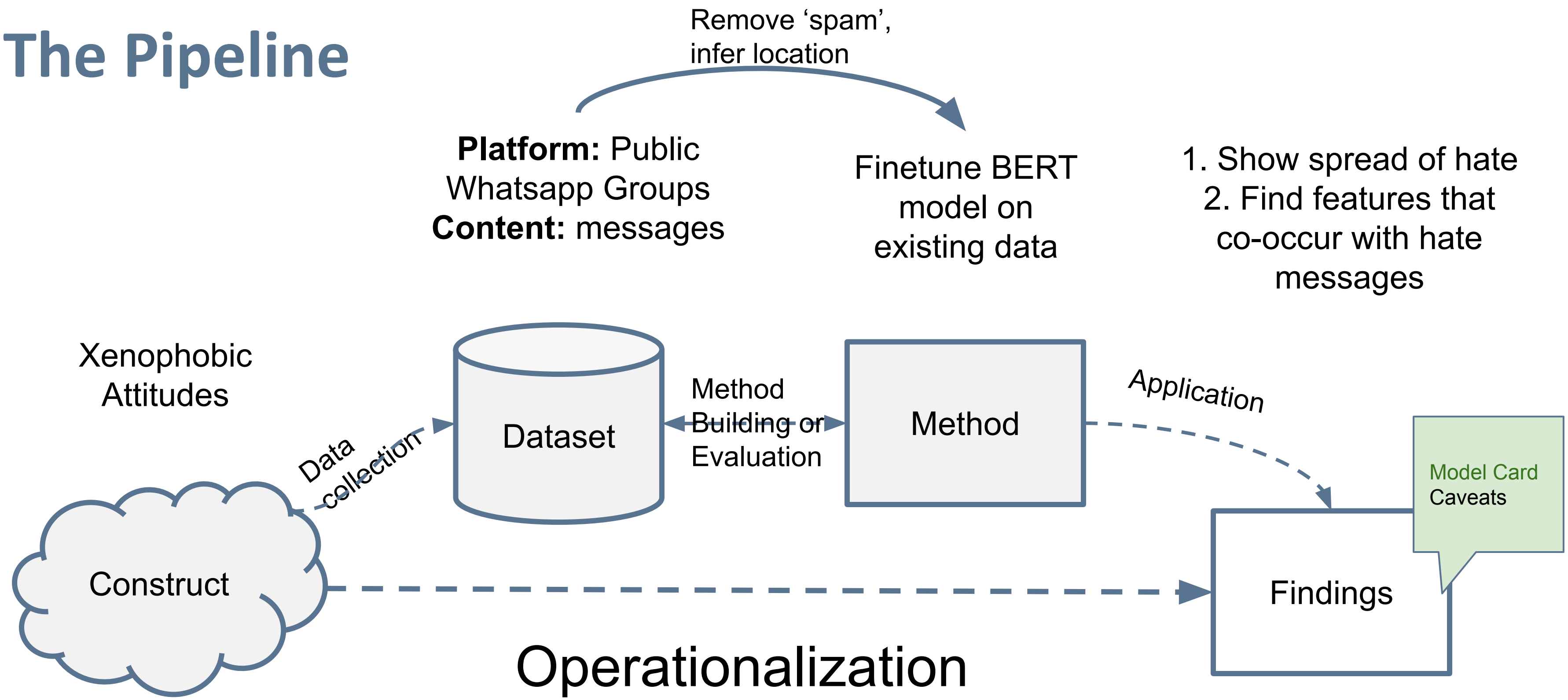


The Pipeline



Findings could help
design interventions
1. Early detection
2. Nudges to spread
awareness

The Pipeline

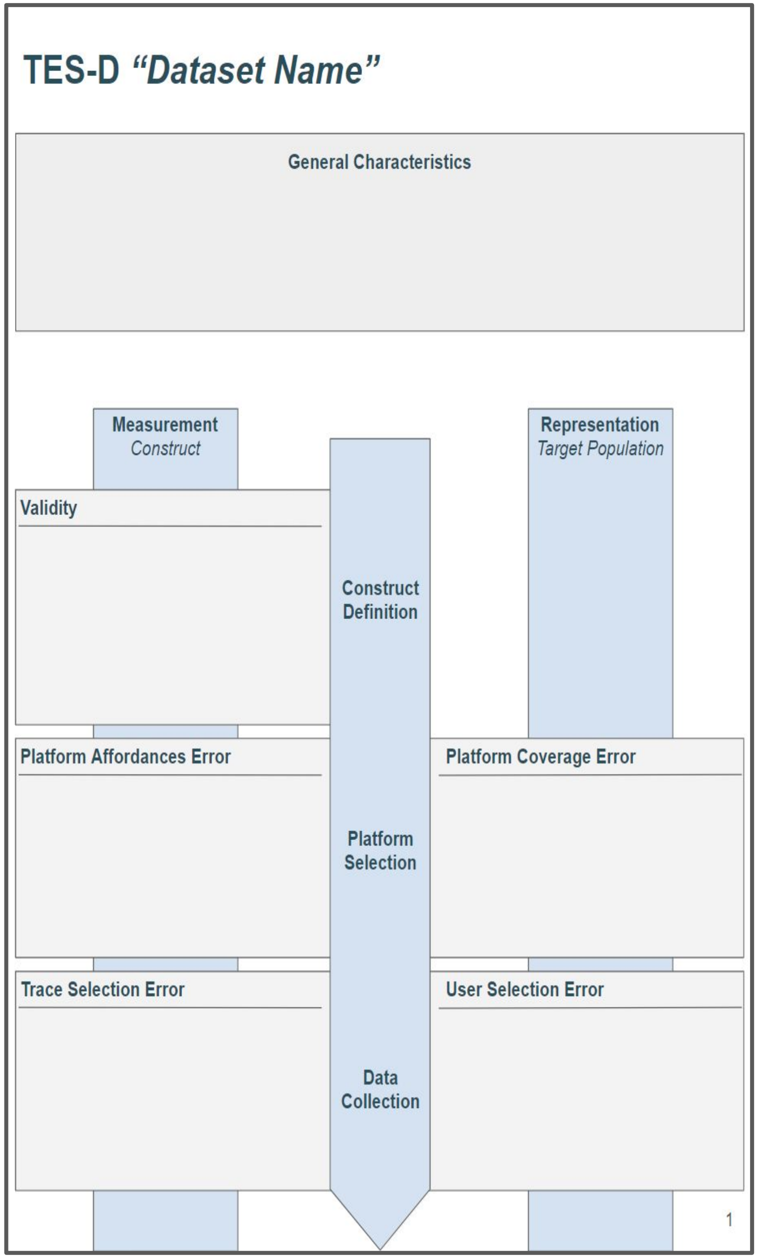


Findings could help design interventions
1. Early detection
2. Nudges to spread awareness

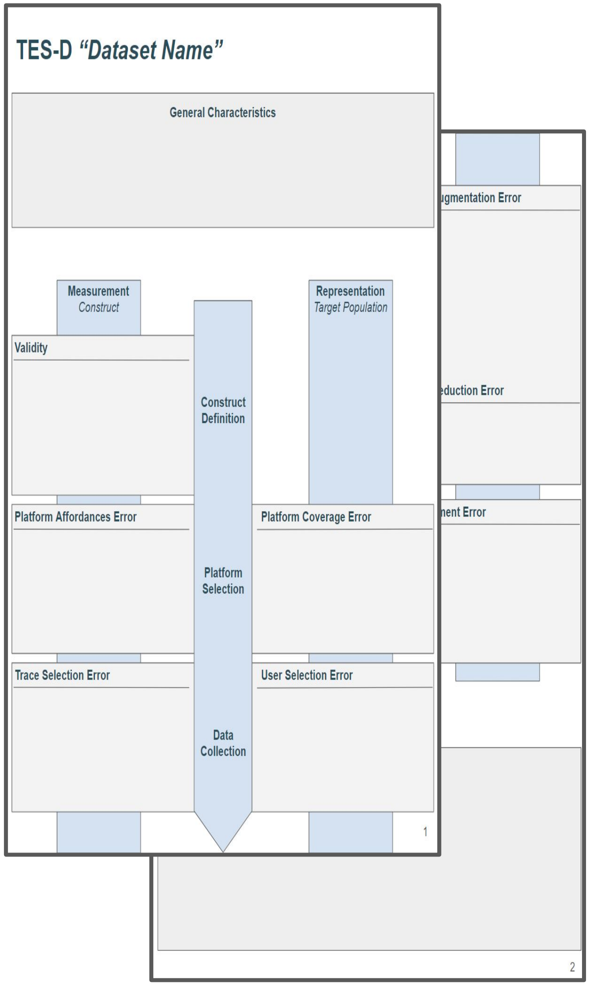
Towards Documentation Templates for Social Media Datasets

Applying the Documentation

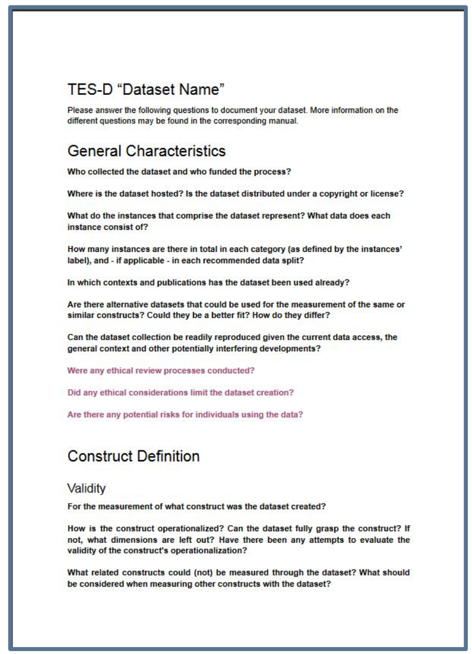
- ❖ TES-D as a documentation template informed by an underlying error framework
  - ➢ Structured along the abstracted research process and the associated errors identified in TED-On
  - ➢ Set of questions per type of error to guide the documentation process
- ❖ Questions as well as brief explanations available from the TES-D manual



Overview of TES-D materials



Template



TES-D "Dataset Name"

Please answer the following questions to document your dataset. More information on the different questions may be found in the corresponding manual.

General Characteristics

Who collected the dataset and who funded the process?

Where is the dataset hosted? Is the dataset distributed under a copyright or license?

What do the instances that comprise the dataset represent? What data does each instance consist of?

How many instances are there in total in each category (as defined by the instances' label), and - if applicable - in each recommended data split?

In which contexts and publications has the dataset been used already?

Are there alternative datasets that could be used for the measurement of the same or similar constructs? Could they be a better fit? How do they differ?

Can the dataset collection be readily reproduced given the current data access, the general context and other potentially interfering developments?

Were any ethical review processes conducted?

Did any ethical considerations limit the dataset creation?

Are there any potential risks for individuals using the data?

Construct Definition

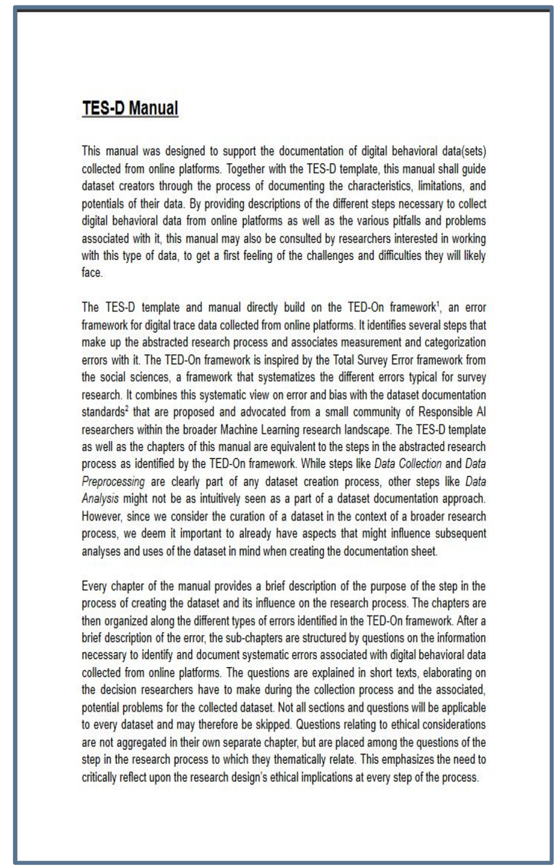
Validity

For the measurement of what construct was the dataset created?

How is the construct operationalized? Can the dataset fully grasp the construct? If not, what dimensions are left out? Have there been any attempts to evaluate the validity of the construct's operationalization?

What related constructs could (not) be measured through the dataset? What should be considered when measuring other constructs with the dataset?

Questions (in detail)



TES-D Manual

This manual was designed to support the documentation of digital behavioral data(sets) collected from online platforms. Together with the TES-D template, this manual shall guide dataset creators through the process of documenting the characteristics, limitations, and potentials of their data. By providing descriptions of the different steps necessary to collect digital behavioral data from online platforms as well as the various pitfalls and problems associated with it, this manual may also be consulted by researchers interested in working with this type of data, to get a first feeling of the challenges and difficulties they will likely face.

The TES-D template and manual directly build on the TED-On framework[1], an error framework for digital trace data collected from online platforms. It identifies several steps that make up the abstracted research process and associates measurement and categorization errors with it. The TED-On framework is inspired by the Total Survey Error framework from the social sciences, a framework that systematizes the different errors typical for survey research. It combines this systematic view on error and bias with the dataset documentation standards[2] that are proposed and advocated from a small community of Responsible AI researchers within the broader Machine Learning research landscape. The TES-D template as well as the chapters of this manual are equivalent to the steps in the abstracted research process as identified by the TED-On framework. While steps like *Data Collection* and *Data Preprocessing* are clearly part of any dataset creation process, other steps like *Data Analysis* might not be as intuitively seen as a part of a dataset documentation approach. However, since we consider the curation of a dataset in the context of a broader research process, we deem it important to already have aspects that might influence subsequent analyses and uses of the dataset in mind when creating the documentation sheet.

Every chapter of the manual provides a brief description of the purpose of the step in the process of creating the dataset and its influence on the research process. The chapters are then organized along the different types of errors identified in the TED-On framework. After a brief description of the error, the sub-chapters are structured by questions on the information necessary to identify and document systematic errors associated with digital behavioral data collected from online platforms. The questions are explained in short texts, elaborating on the decision researchers have to make during the collection process and the associated, potential problems for the collected dataset. Not all sections and questions will be applicable to every dataset and may therefore be skipped. Questions relating to ethical considerations are not aggregated in their own separate chapter, but are placed among the questions of the step in the research process to which they thematically relate. This emphasizes the need to critically reflect upon the research design's ethical implications at every step of the process.

Manual with detailed examples