FROM AWARENESS TO ACTION

Defining, Assessing, & Improving the Quality of Digital Trace Data

Dr. Valerie Hase, LMU Munich

orcid.org/0000-0001-6656-4894

valeriehase

www.valerie-hase.com

Can I use data donations to understand how citizens engage with news online?

(Hase & Haim, 2024)

?

Can I use APIs to understand which news is shared across platforms?

(Hase et al., 2023)

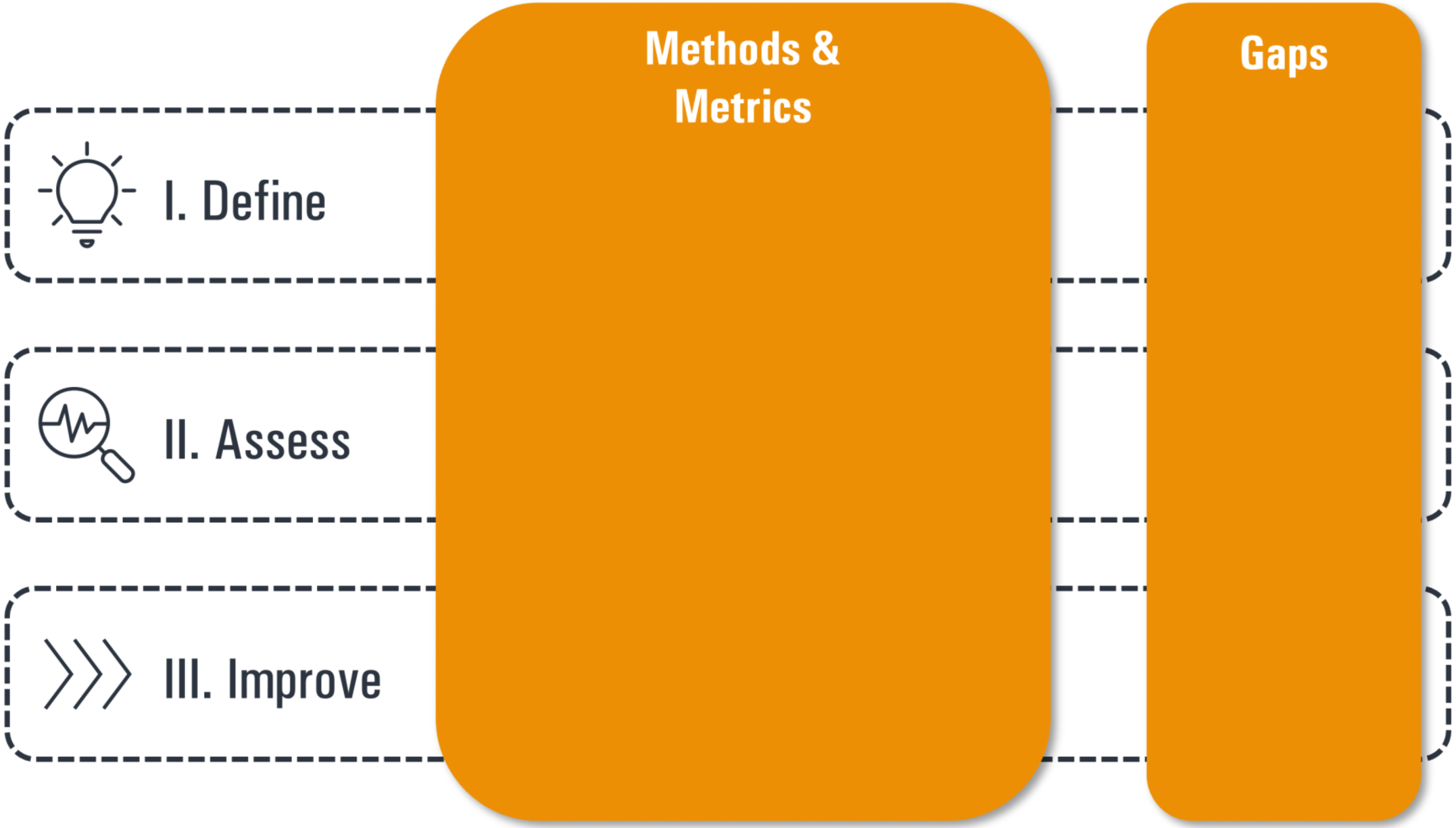
QUALITY FRAMEWORK

I. Define: What are criteria for evaluating quality?

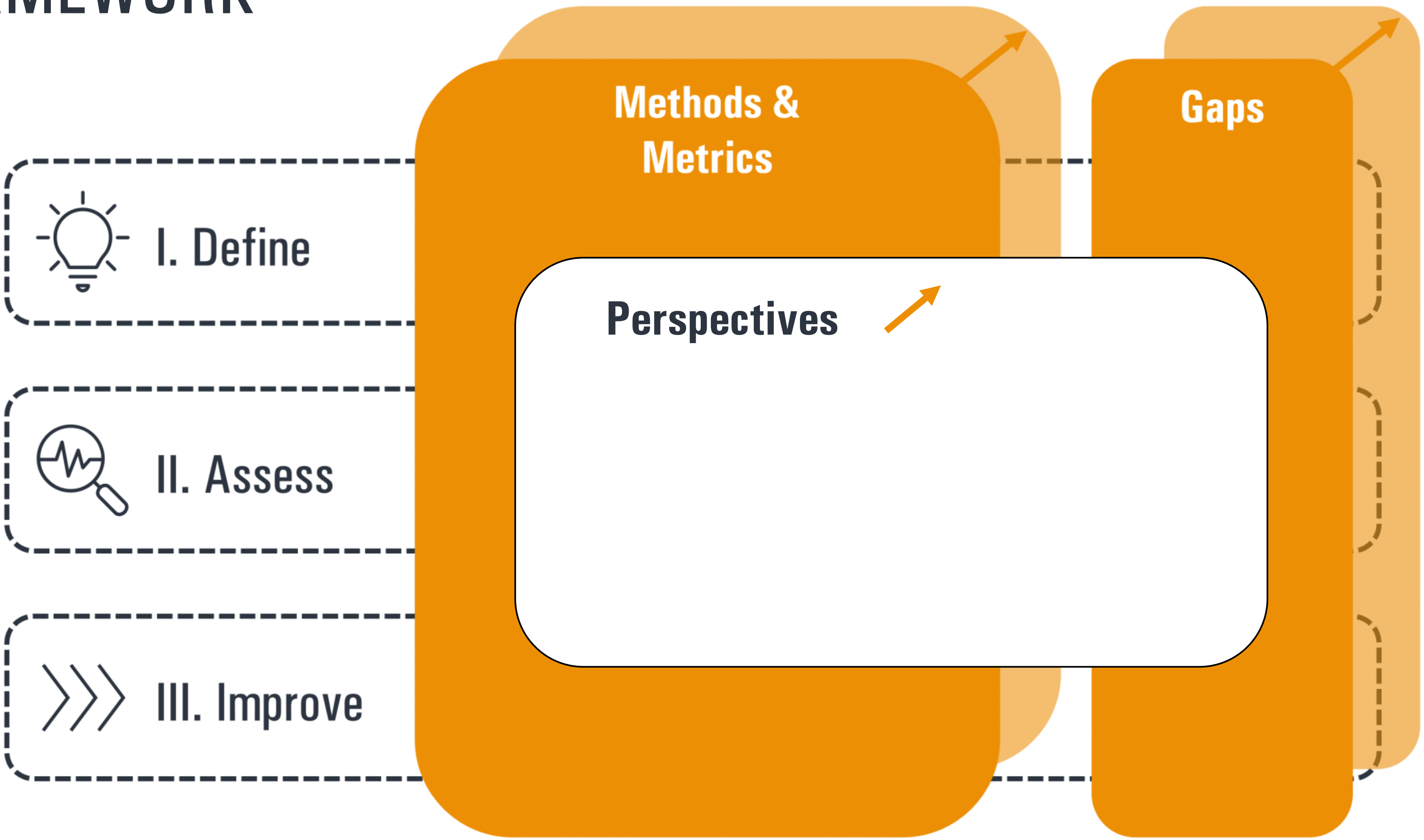
II. Assess: How do I assess quality?

III. Improve: How do I improve quality?

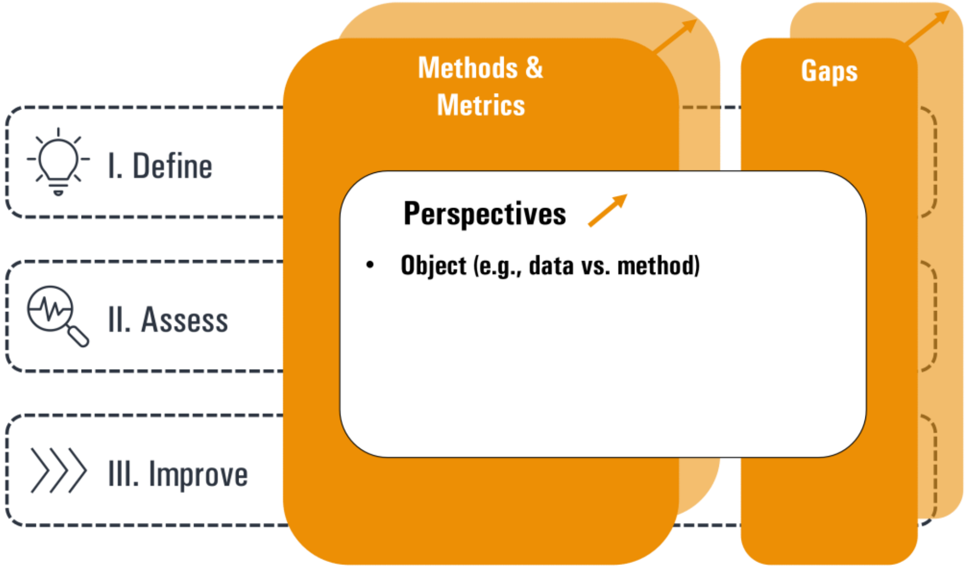
QUALITY FRAMEWORK



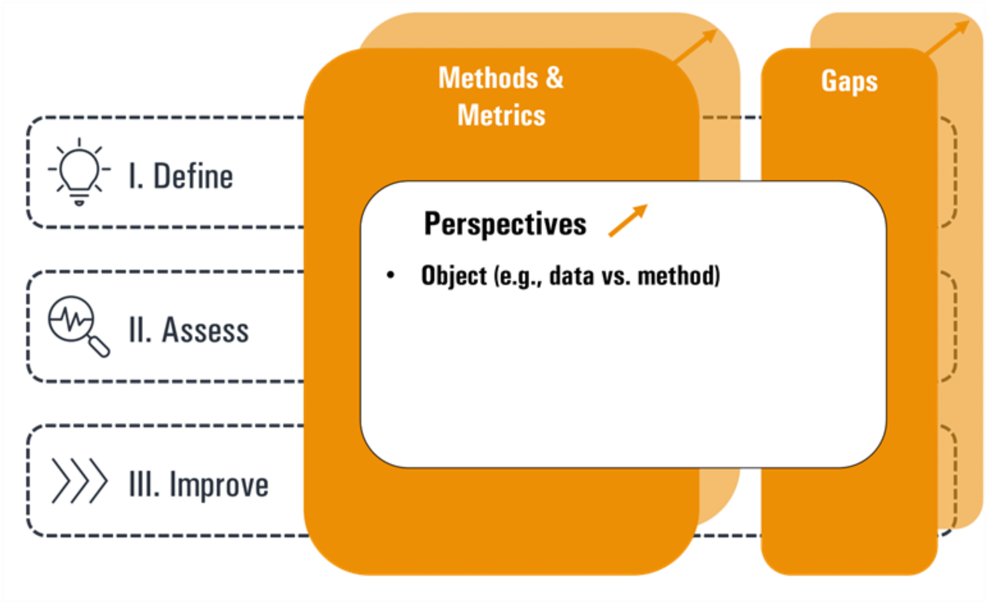
QUALITY FRAMEWORK



QUALITY FRAMEWORK



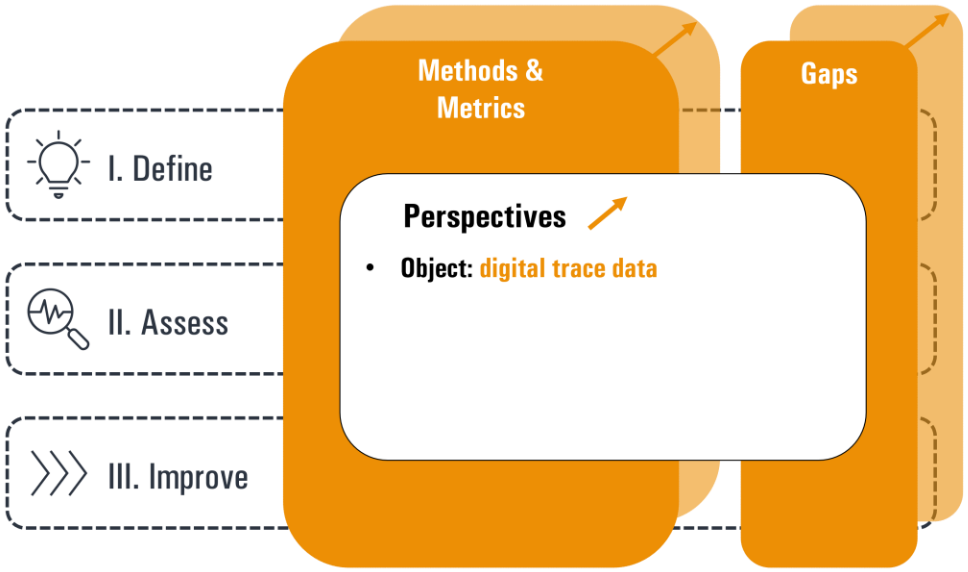
QUALITY FRAMEWORK



How „good“ is my data set?
(or meta-data, variable)

How „good“ is my analysis method?

QUALITY FRAMEWORK



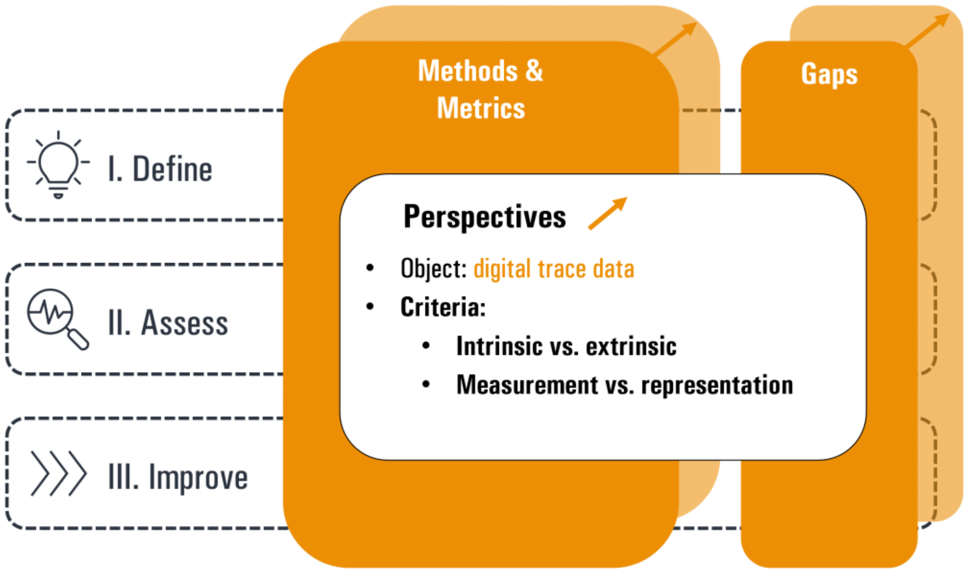
How „good“ is my data set?
(or meta-data, variable)

Focus: „found“ digital trace data

- Platform-centric approaches
  (e.g., APIs, industry collaborations)
- User-centric approaches
  (e.g., data donation, tracking, sensors)

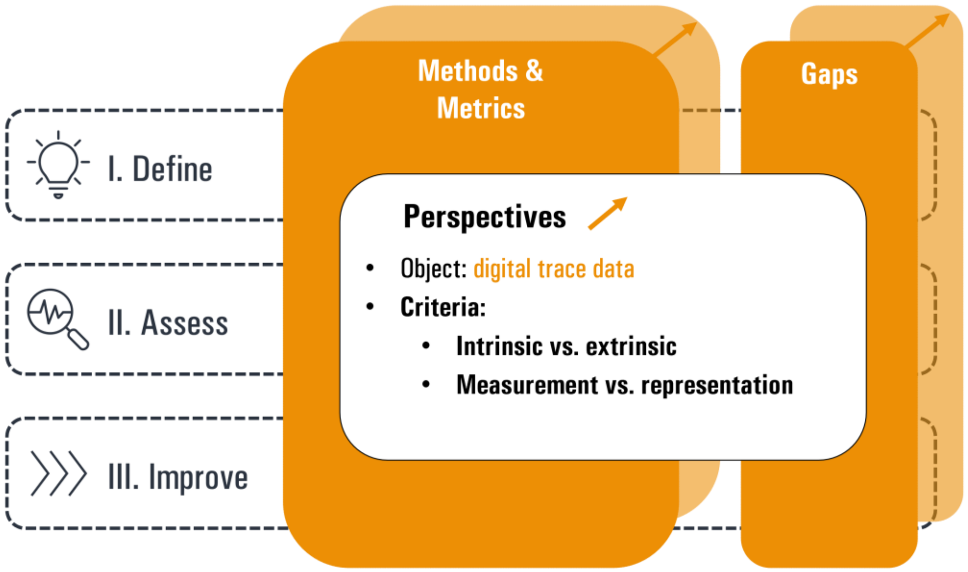
QUALITY FRAMEWORK

see similarly Birkenmaier et al., 2024; Daikeler et al., 2024



QUALITY FRAMEWORK

see similarly Birkenmaier et al., 2024; Daikeler et al., 2024

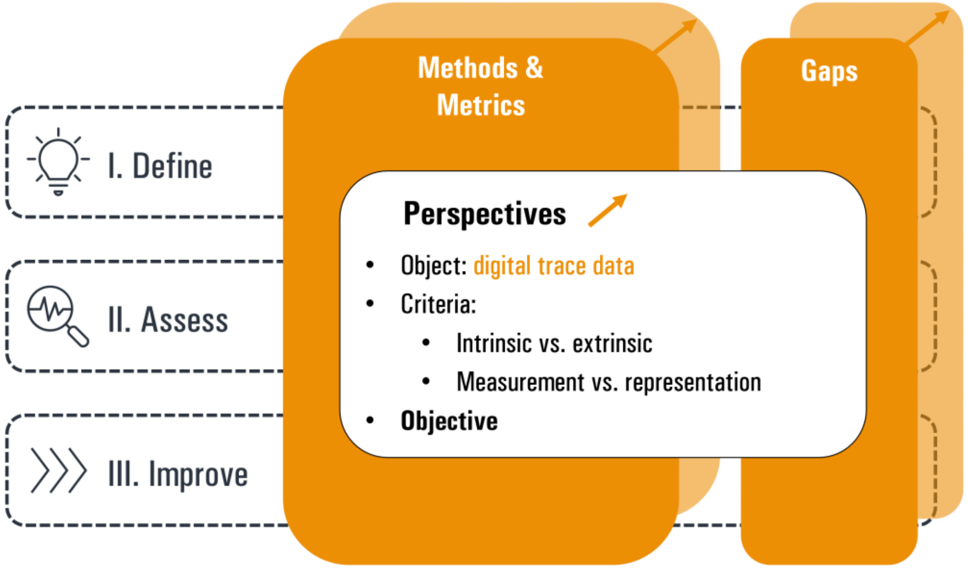


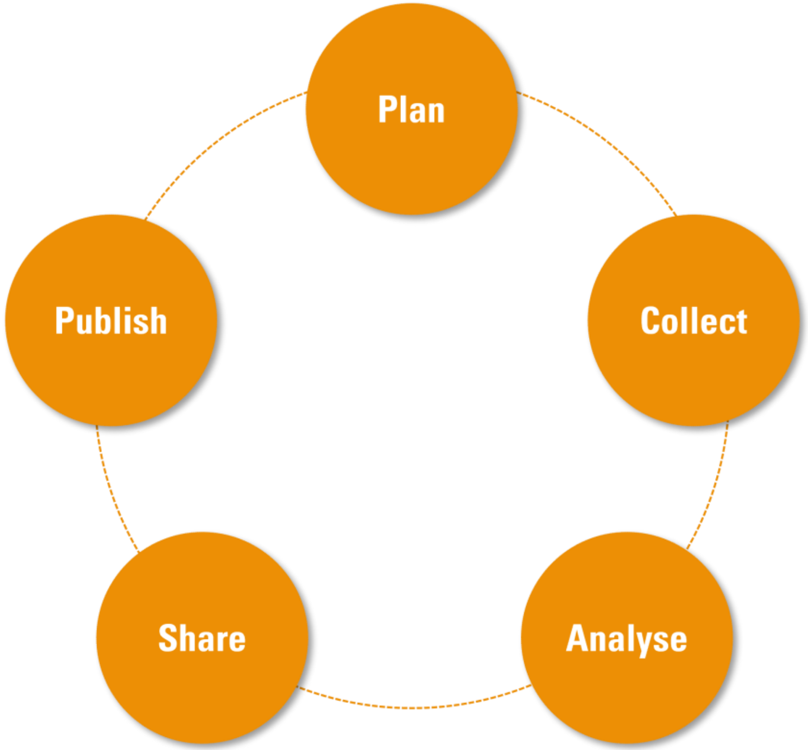
Intrinsic: How „correct“ is my data?
(e.g., measurement, representation)

Extrinsic: How „usable“ is my data?
(e.g., FAIR, CARE principles)

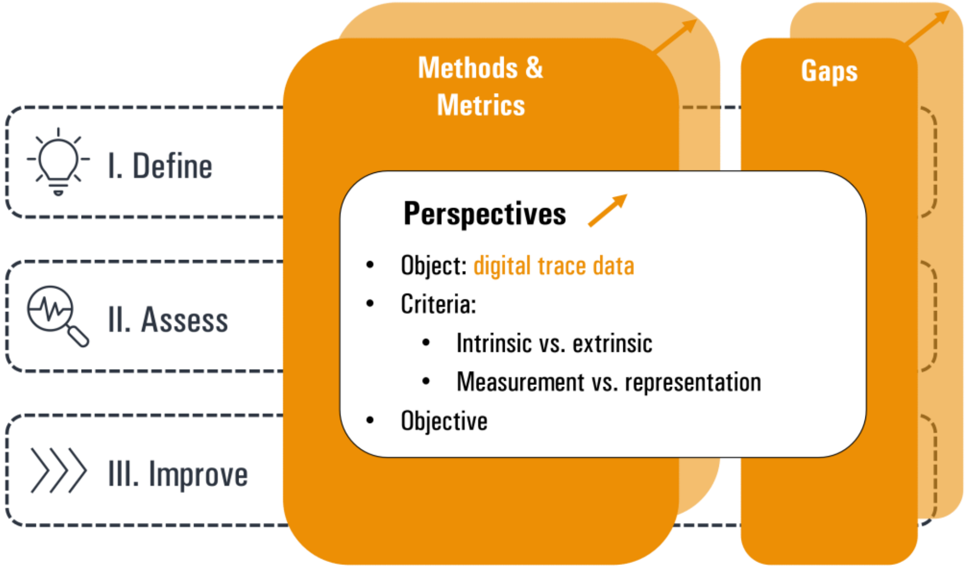
QUALITY FRAMEWORK

see similarly Rfll, 2020





MAIN QUESTION



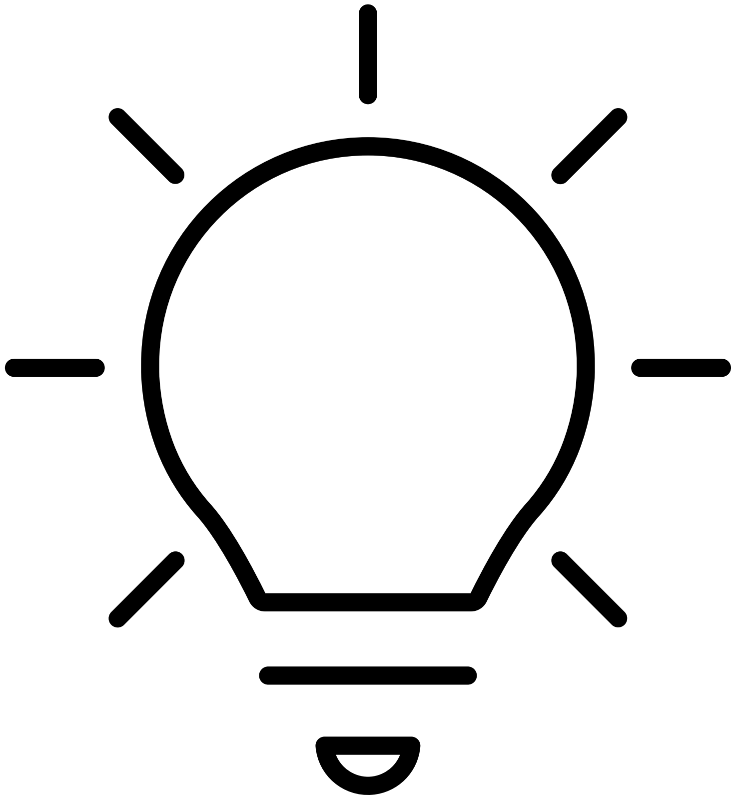
?

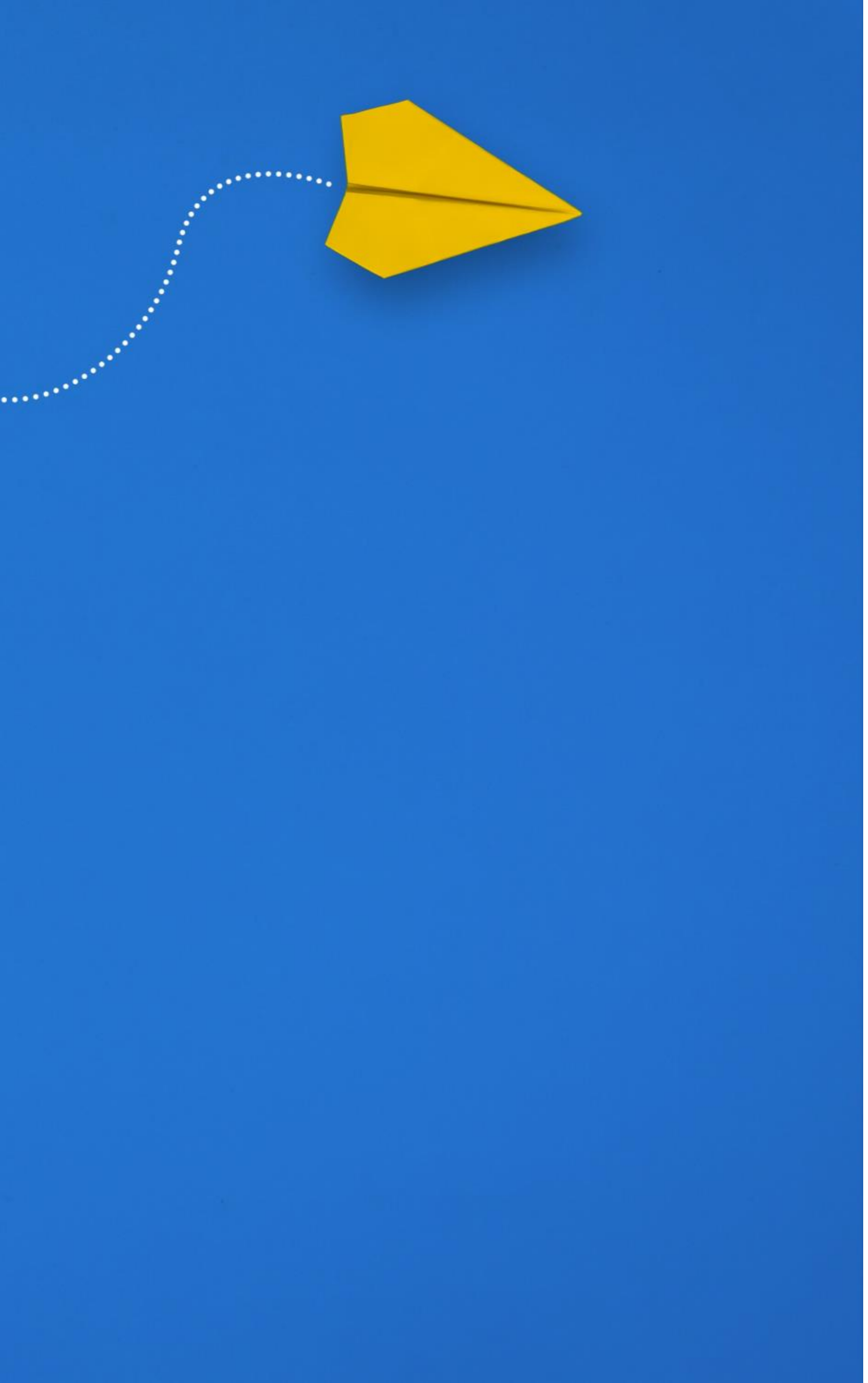
How can we define, assess, & improve the quality of digital trace data for research?

I. DEFINE QUALITY

In CSS (and beyond), data quality is a problem we have **ignored for too long**.

With increasing awareness, we have started to adapt & develop quality criteria – which also led to a **lack of conceptual agreement**.

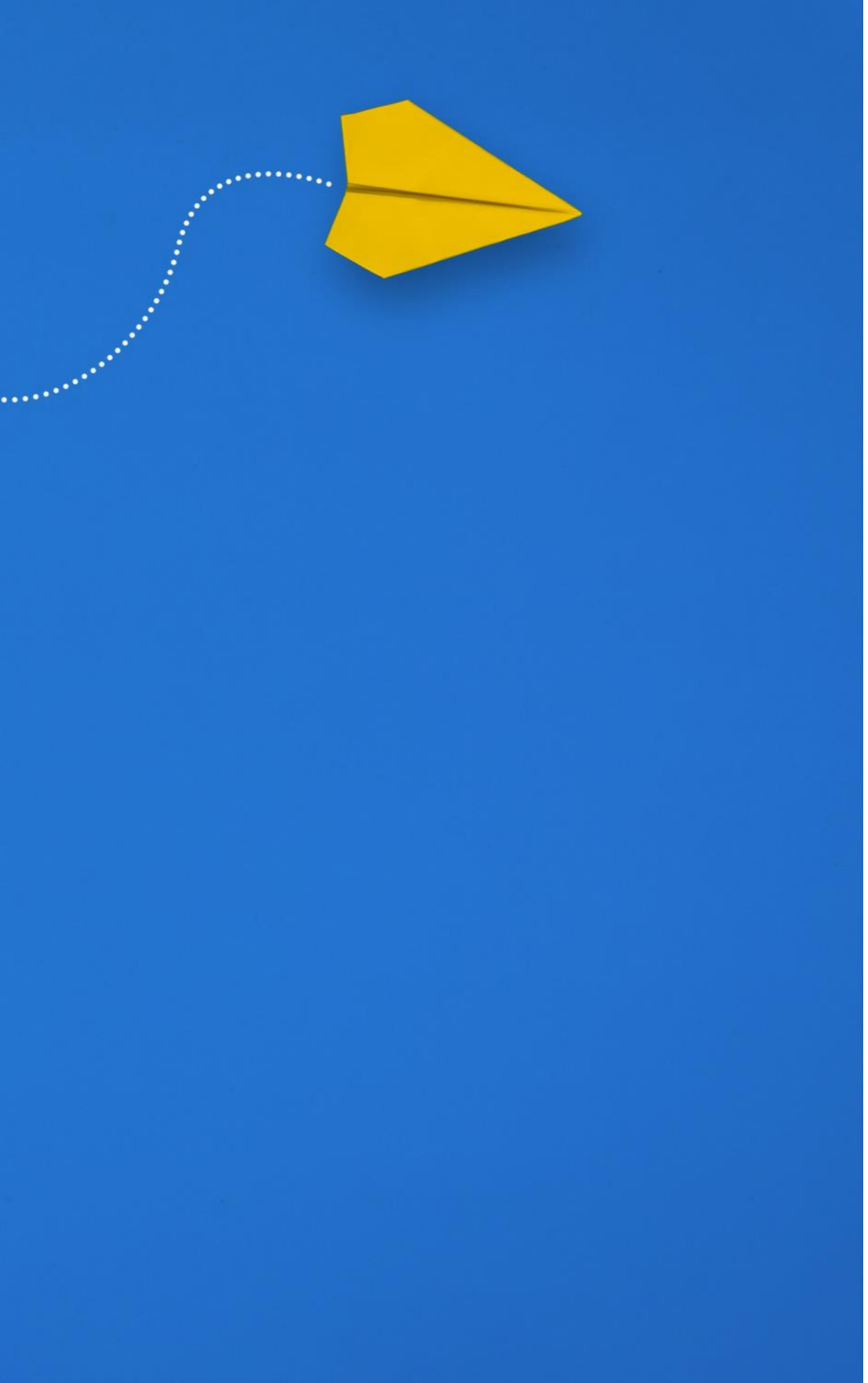




DEFINE QUALITY: METHODS & METRICS

- Frameworks
  - Error frameworks (Daikeler et al., 2024)
  - Data quality frameworks: FAIR (Wilkinson et al., 2016), CARE (Carroll et al., 2021)

e.g., Batini et al., 2009; Daikeler et al., 2024; Ijab et al., 2019; Theh et al., 2020

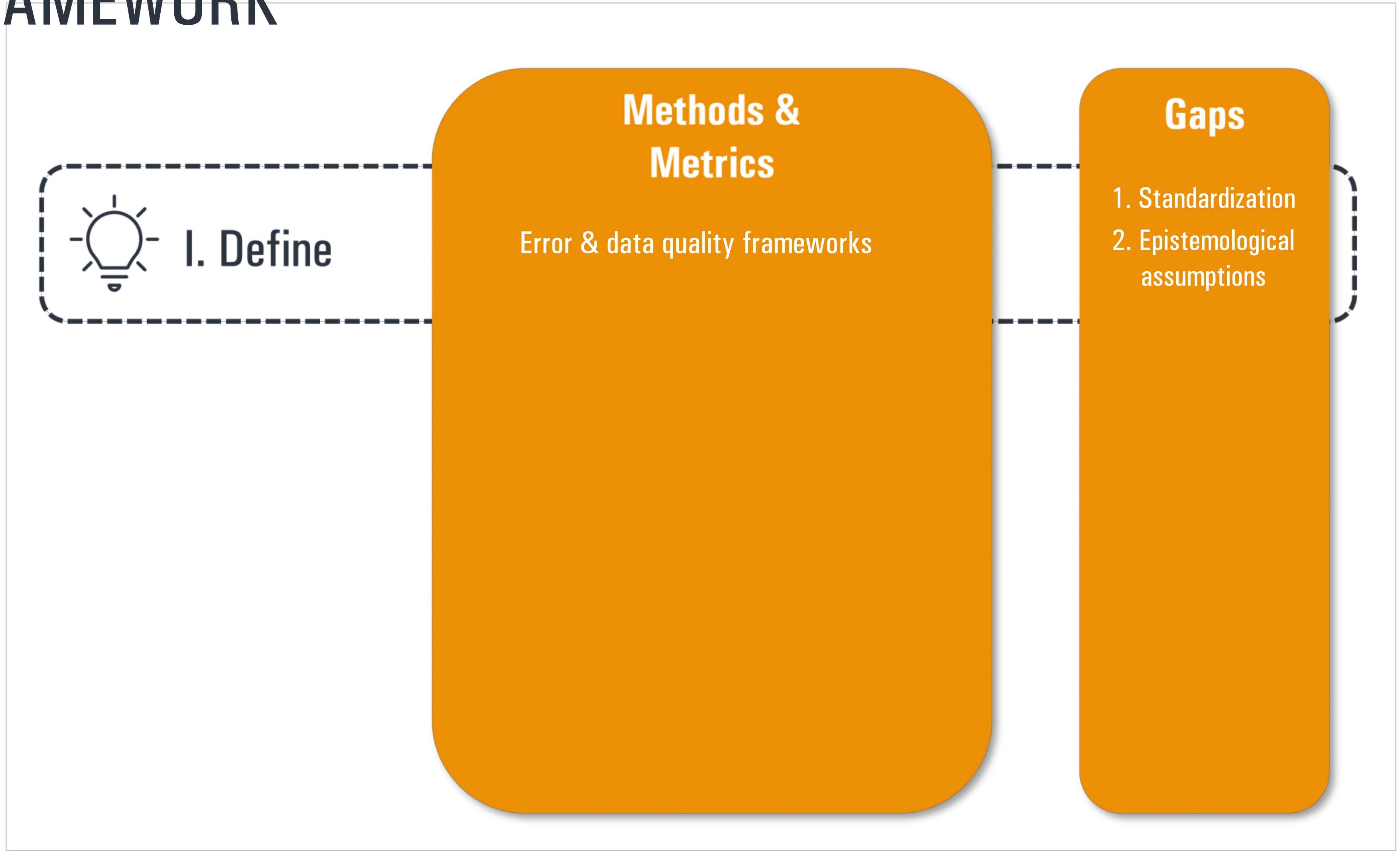


DEFINE QUALITY: GAPS

(Birkenmaier et al., 2024; Hammersley, 1997; Kitching, 2014; Shugars, 2024)

- **Balance** between unification & specialization across methods/disciplines
- **Integrating epistemologies**: Can we use “bad data” (e.g., “bias”) constructively?

QUALITY FRAMEWORK



QUALITY FRAMEWORK

I. Define

II. Assess

Methods & Metrics

Error & data quality frameworks

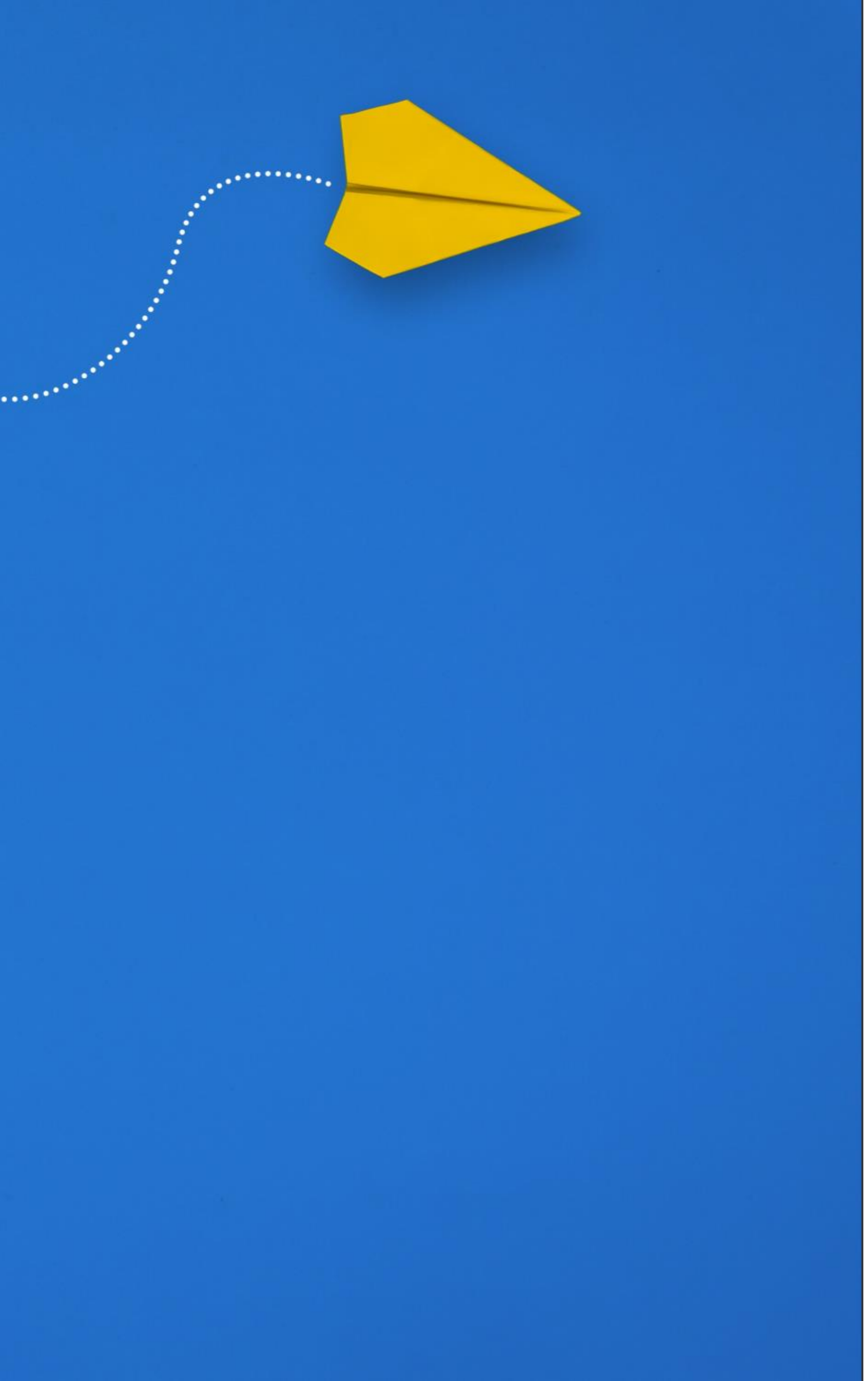
Gaps

1. Standardization
2. Epistemological assumptions

II. ASSESS QUALITY

In CSS, there is a **"critical" turn** dedicated to assessing data quality.

Given the **lack of standardized methods & metrics**, we still ask: "how good is good enough?"



ASSESS QUALITY

- Not yet a standard
  - Only 55% of psychological studies assess internal quality (Gottfried et al., 2024)
  - External quality sometimes tested (Batzdorfer et al., 2024; Eder & Jedinger, 2019)

EXAMPLE: DATA DONATION STUDY

Can I use data donations to understand how citizens engage with news online?
(Hase & Haim, 2024)

?

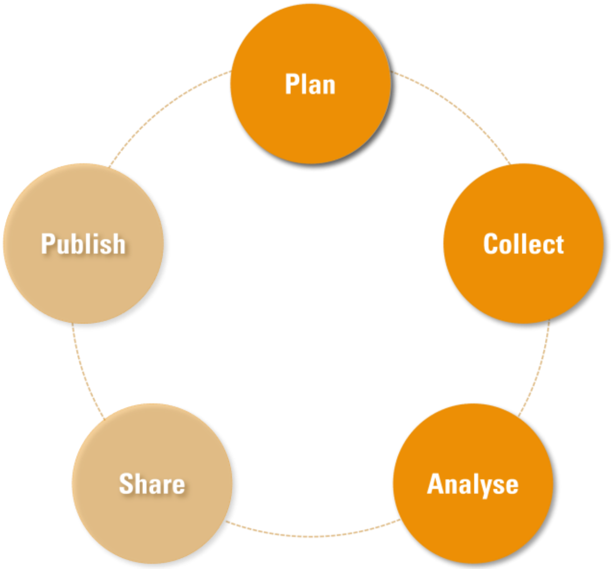
EXAMPLE: DATA DONATION STUDY

How prevalent are errors of representation in data donation studies?

2 survey experiments:
online panel (*N* = 2,309) &
student sample (*N* = 345)
(see Haim et al., 2023 for tool)

N = 423 data donation packages
(Facebook, Instagram, X/Twitter, YouTube)

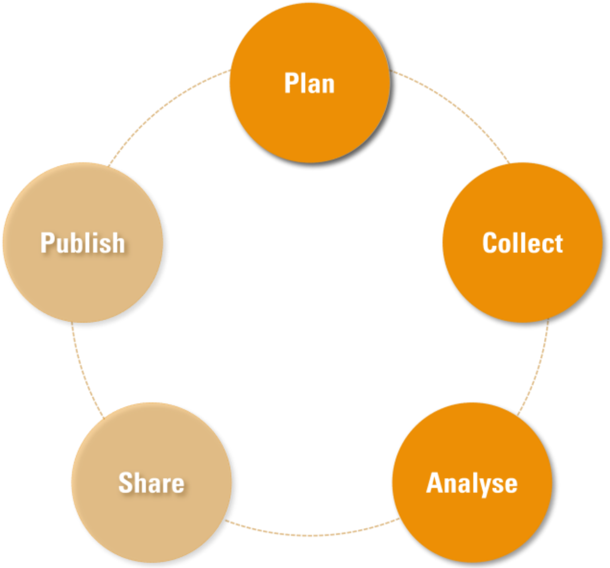
Can we use *also* data to study digital news engagement?



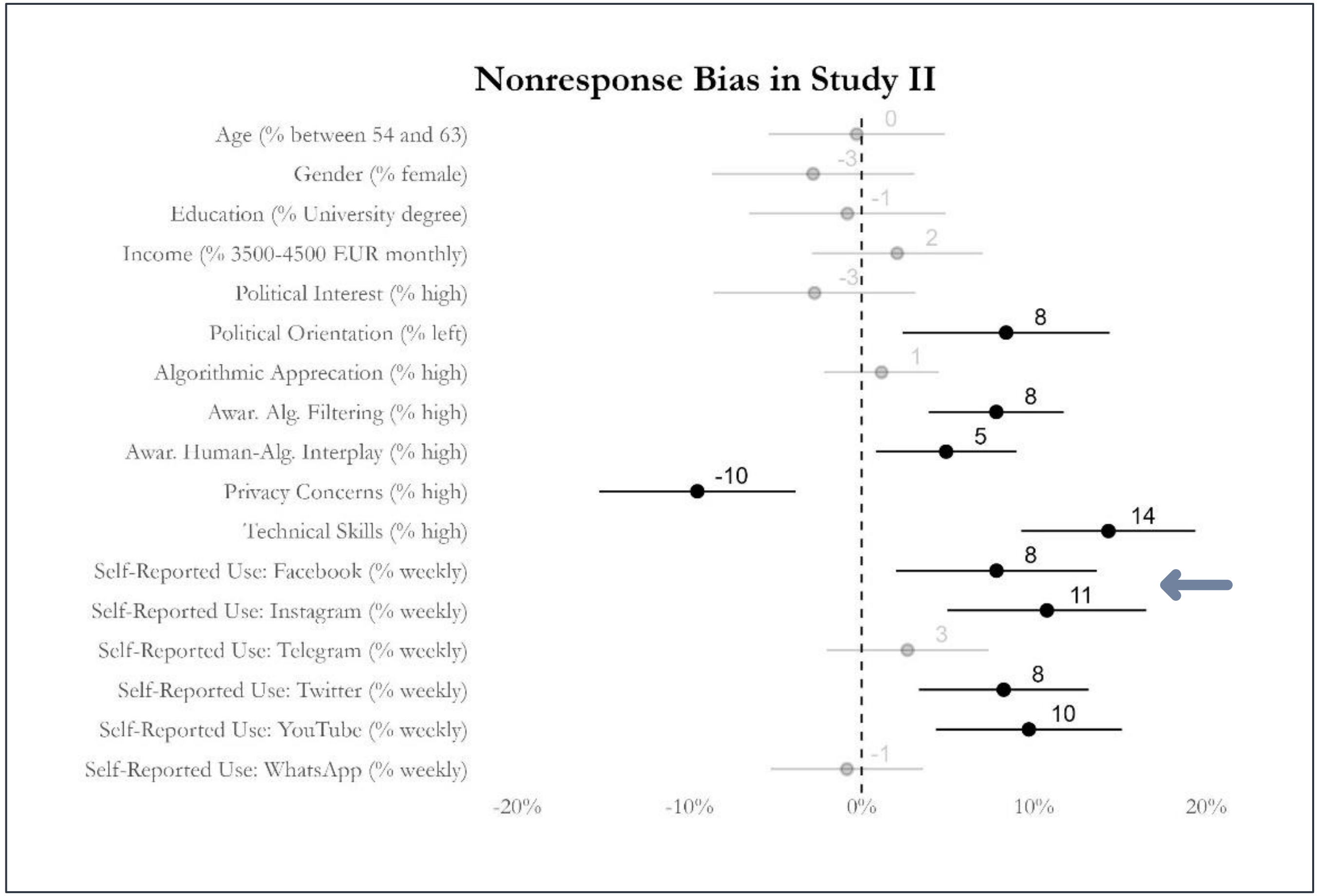
EXAMPLE: DATA DONATION STUDY

Intrinsic (error of representation):

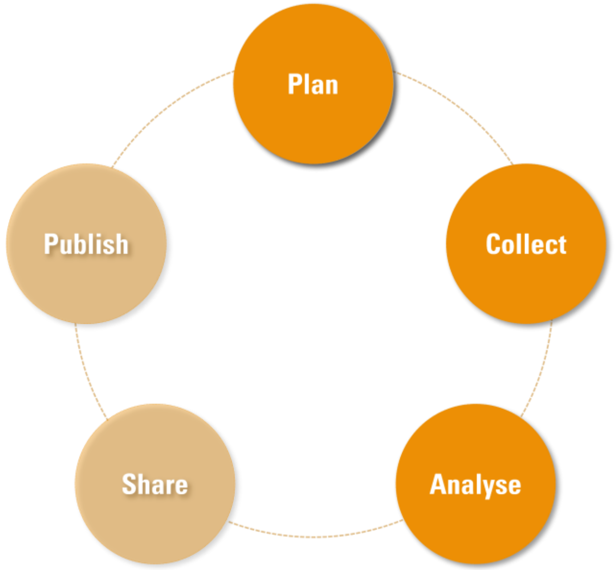
- ✓ Track drop-out with para data
  - e.g., 63% response rate survey vs. 12% response rate data donation
- ✓ Capture predictors of drop-out with survey data
  - e.g., average non-response bias of 6-7%



EXAMPLE: DATA DONATION STUDY



EXAMPLE: DATA DONATION STUDY



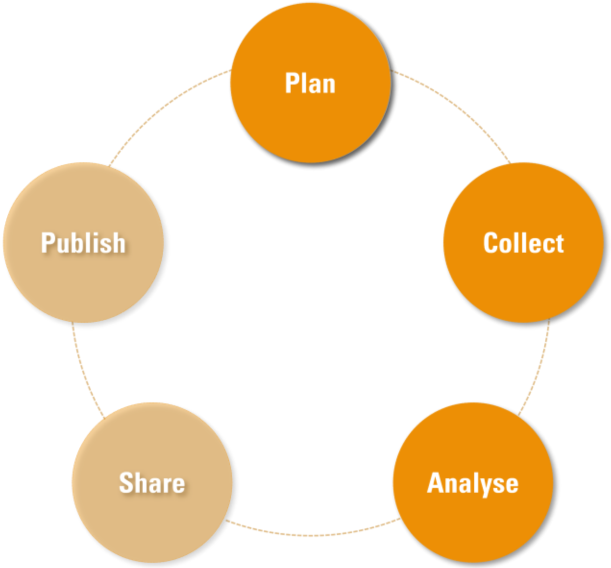
Intrinsic (error of representation):

✓ Track drop-out via para data

✓ Capture predictors of drop-out with survey data

× Disentangle different errors (coverage, non-response)

EXAMPLE: DATA DONATION STUDY

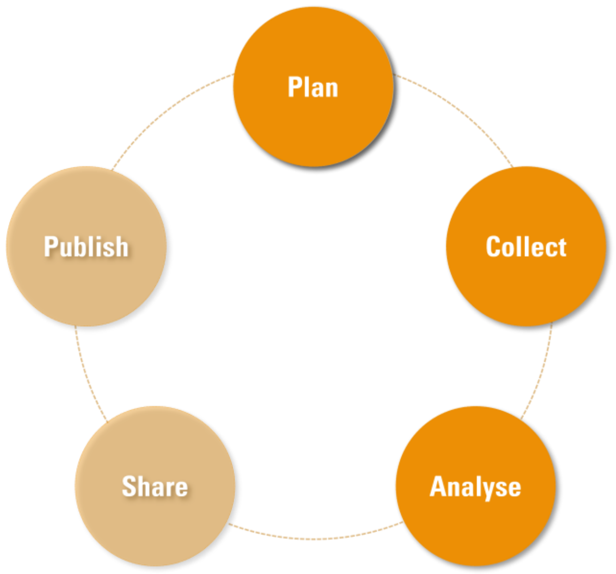


Intrinsic (measurement error):

✓ Track missing data via error logging

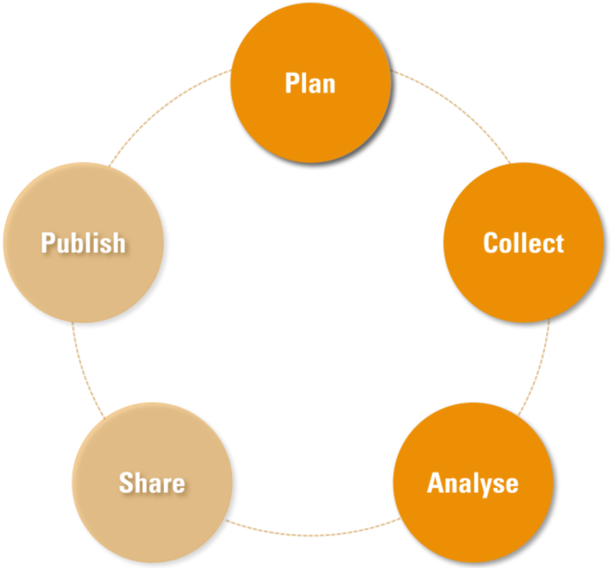
- e.g., tool failed to upload DDPs from 2 participants
- e.g., 9% of participants deleted data

EXAMPLE: DATA DONATION STUDY



Intrinsic (measurement error):

- ✓ Track missing data via error logging
- ✓ Compare different data sources
  - e.g., low correlation self-reported & observed engagement

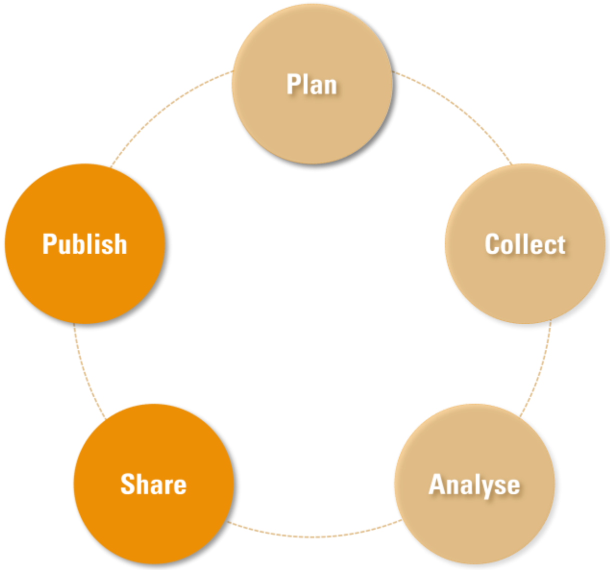


EXAMPLE: DATA DONATION STUDY

Intrinsic (measurement error):

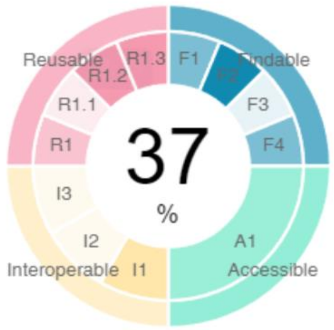
- ✓ Track missing data via error logging
- ✓ Compare different data sources
- × Variation across preprocessing pipelines
  - e.g., classifying news engagement with dictionary vs. ML
  - e.g., classifying news engagement using different metrics/time thresholds

EXAMPLE: DATA DONATION STUDY

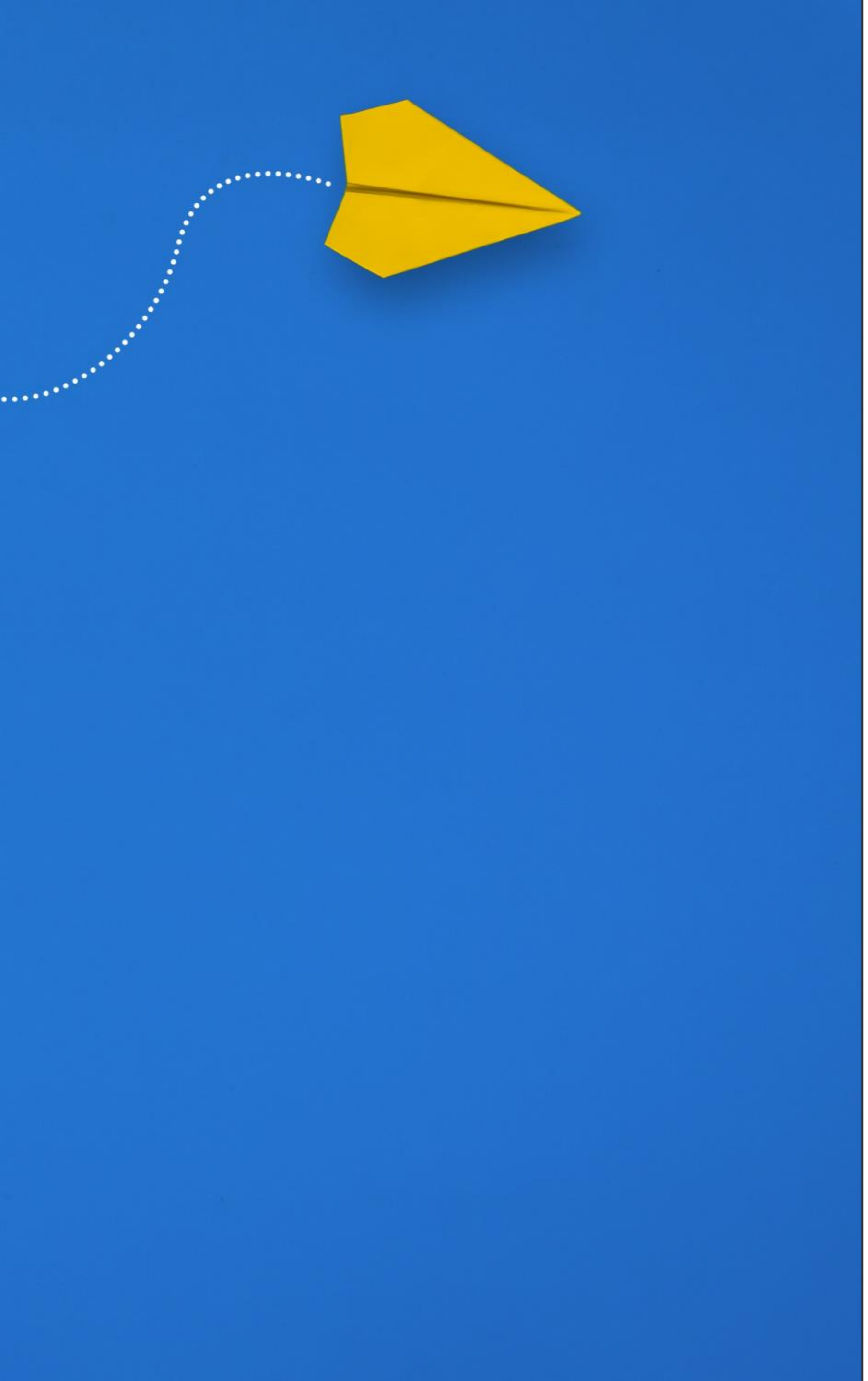


Extrinsic (e.g., FAIR, CARE):

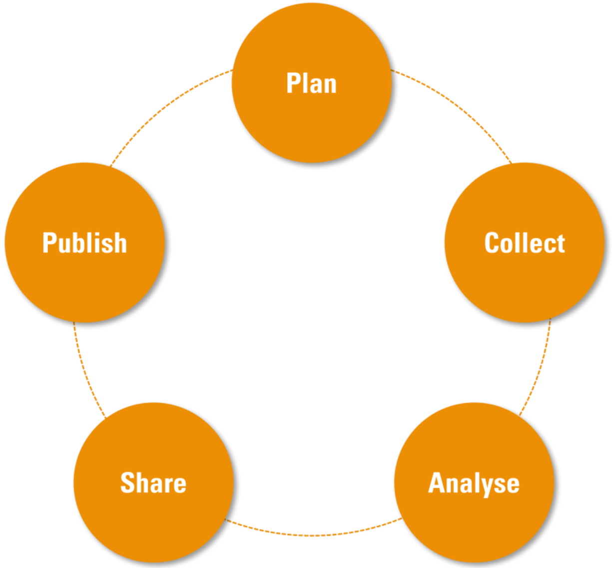
- ✓ Shared preregistration, code, data, data documentation
- × Adhered to FAIR principles

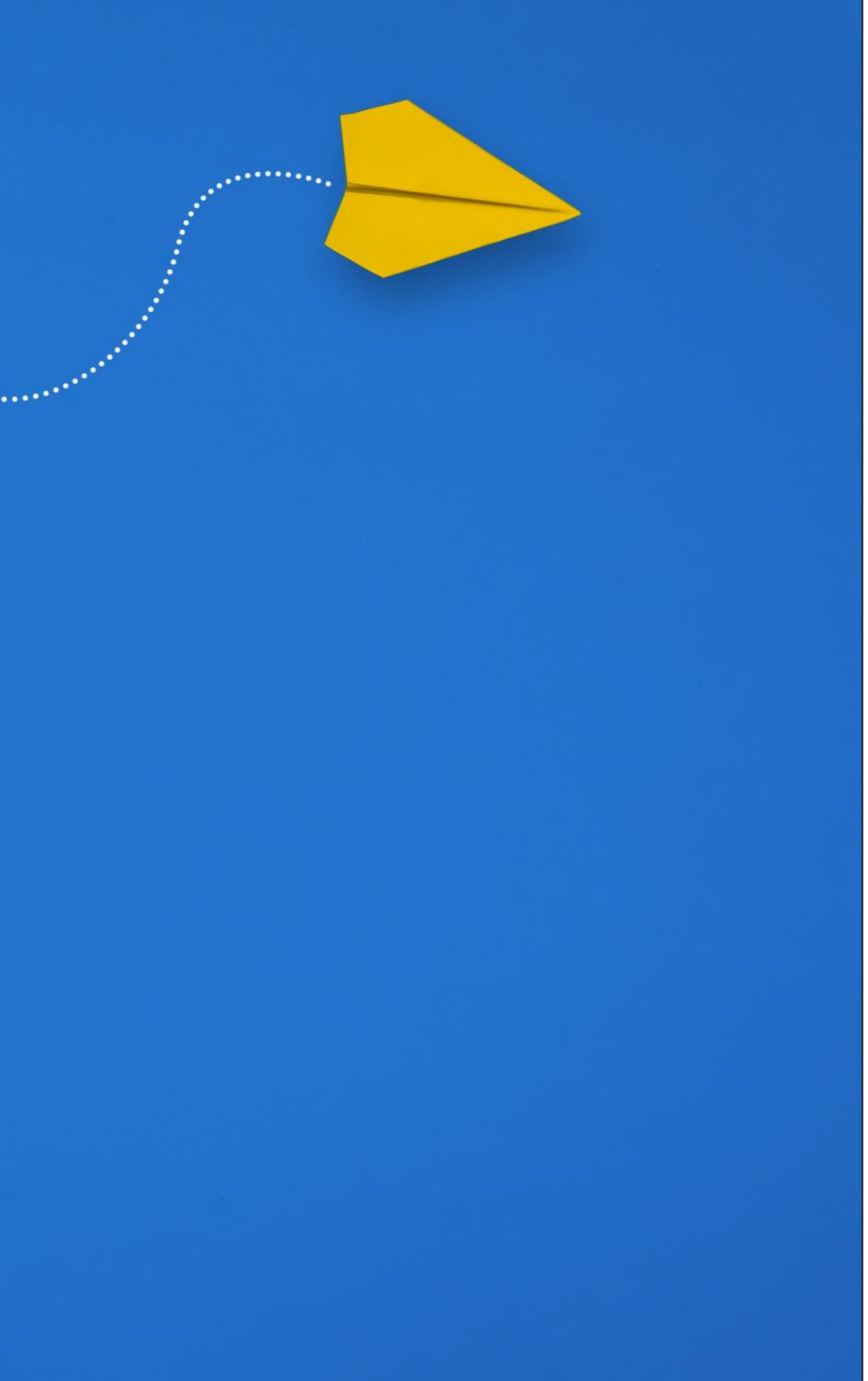


	Score earned:	Fair level:
Findable:	4 of 7	moderate
Accessible:	1 of 3	initial
Interoperable:	1 of 4	initial
Reusable:	3 of 10	initial



ASSESS QUALITY: METHODS & METRICS

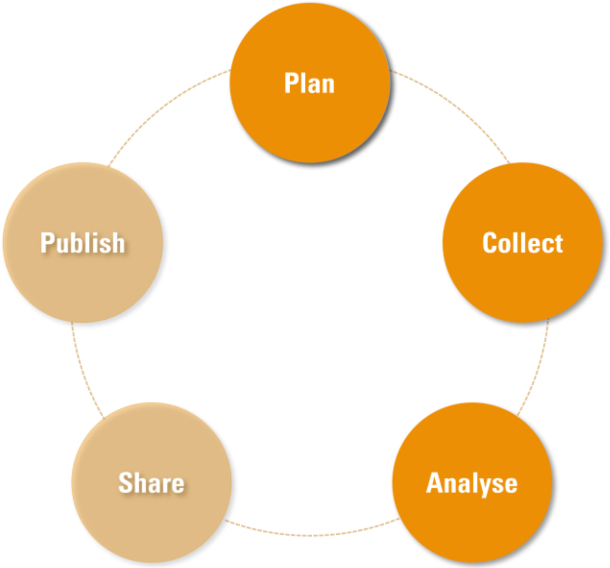


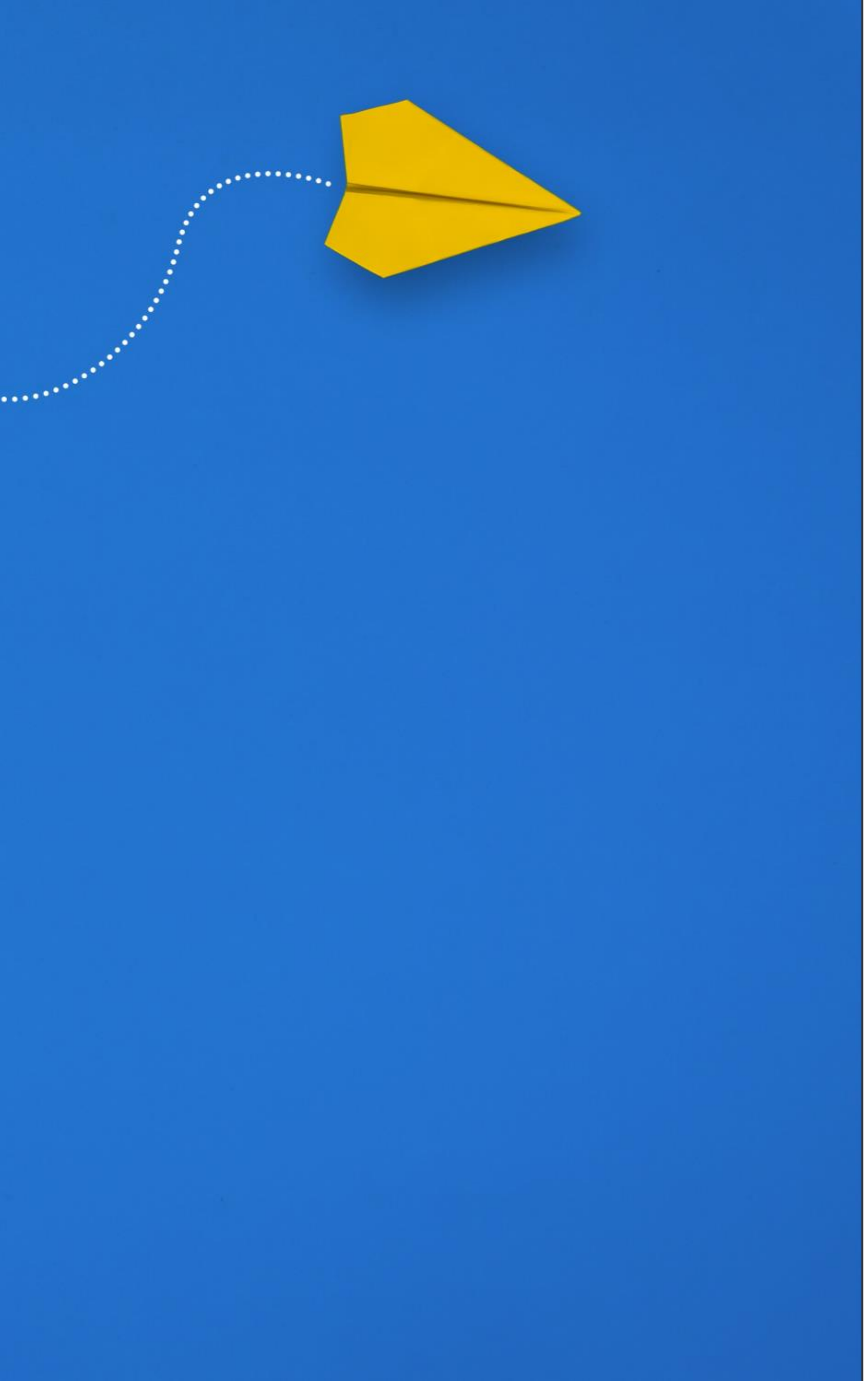


ASSESS QUALITY: METHODS & METRICS

1. "How to"- Guidelines

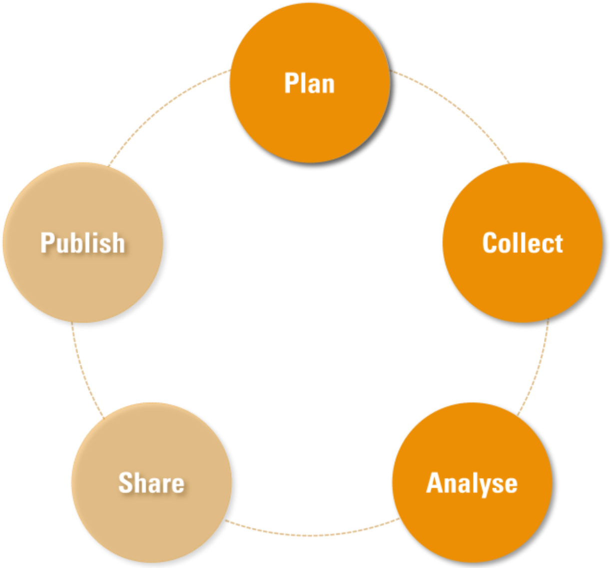
- Data donation (Carrière et al., 2024)
- Tracking (Clemm von Hohenberg et al., 2024)
- Scraping (Boegershausen et al., 2022)
- Machine learning (Kapoo et al., 2024)

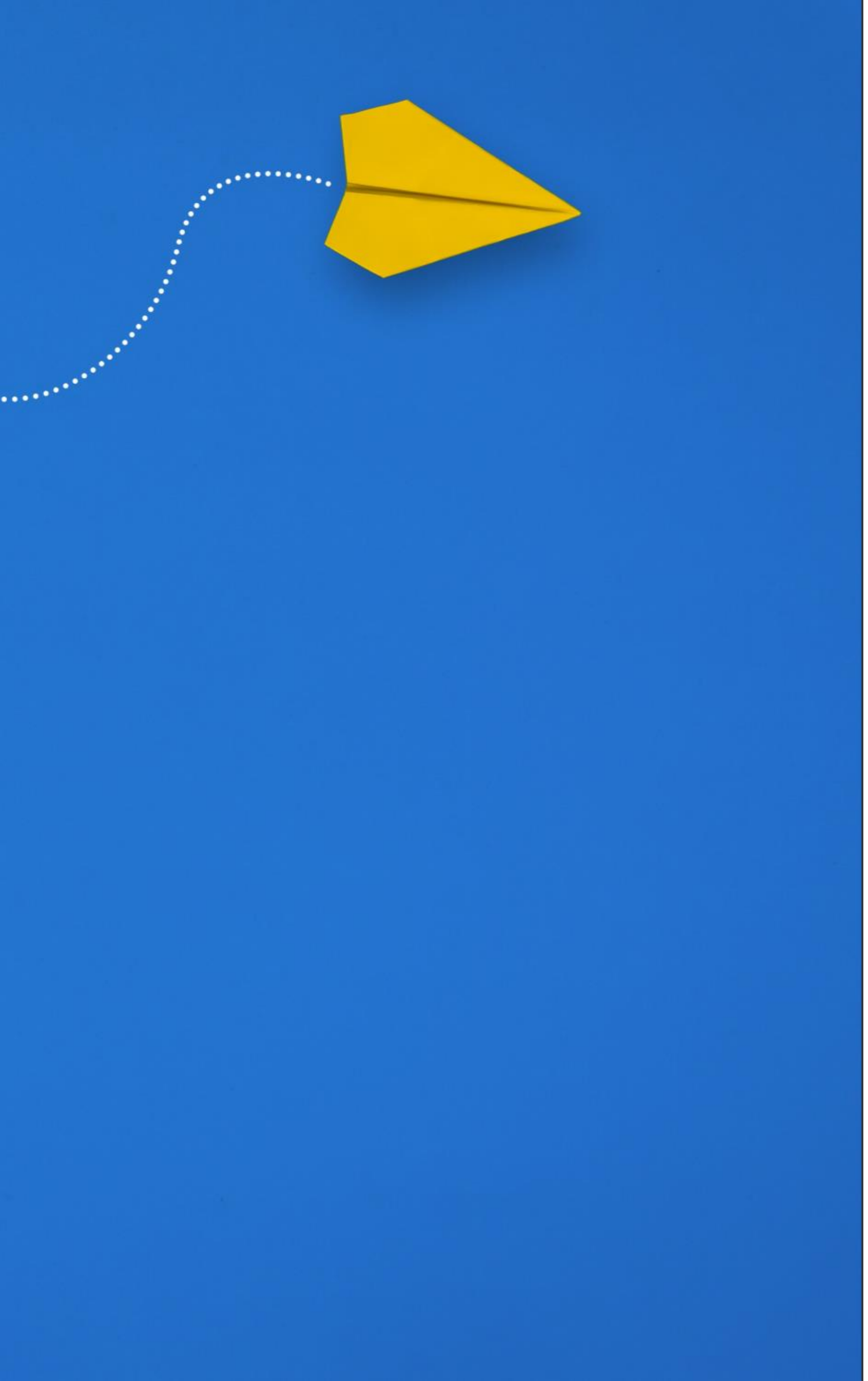




ASSESS QUALITY: METHODS & METRICS

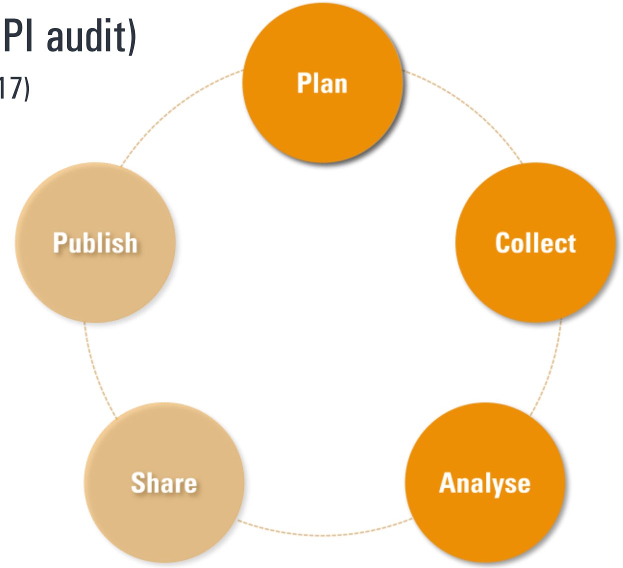
1. “How to”- Guidelines
2. **Para data from initial data collection**
   - log error (e.g., response latency, missing data)
   - qualitative data helpful!

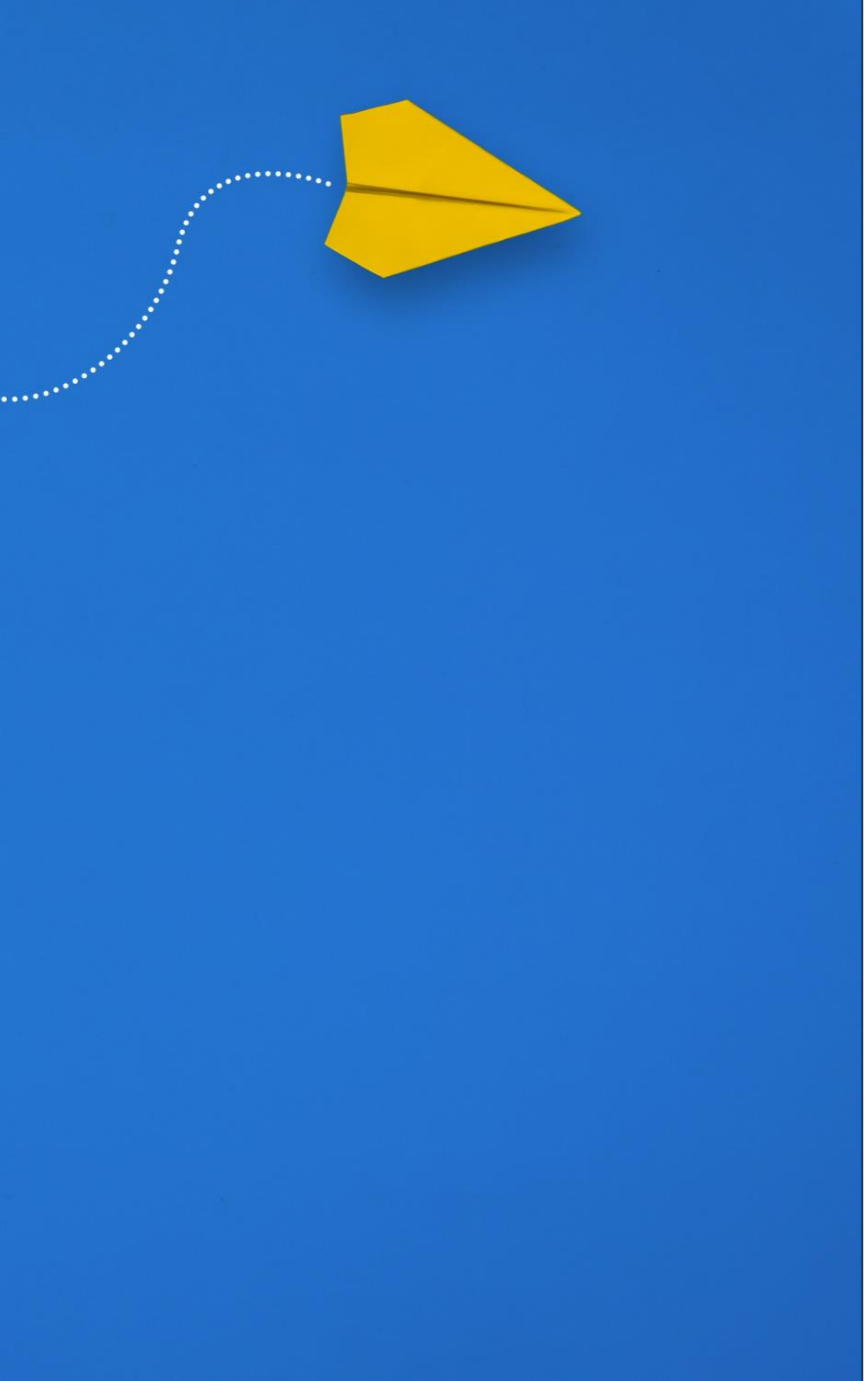




ASSESS QUALITY: METHODS & METRICS

1. "How to"- Guidelines
2. Para data from initial data collection
3. **Additional data collection/analysis methods**
   - API vs. scraping: understand NAs (e.g., API audit) (Pearson et al., 2024; Pfeffer et al., 2023; Tromble et al., 2017)
   - Multiverse approaches (Bosch et al., 2023)
   - MultiTrait Multi Method(MTMM) models (Cernat et al., 2024)

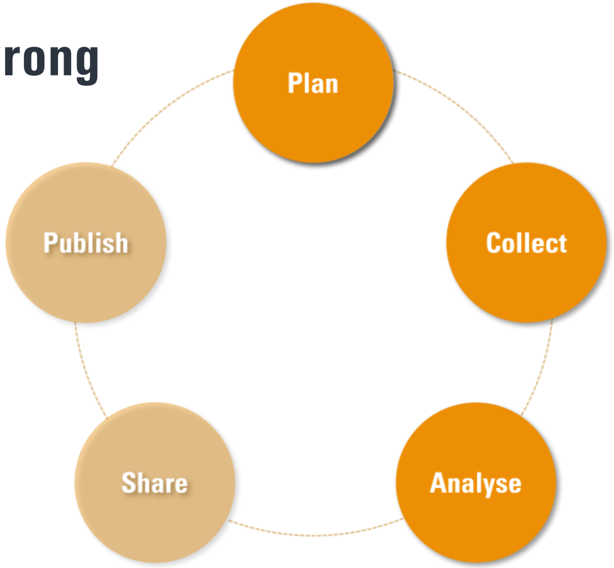


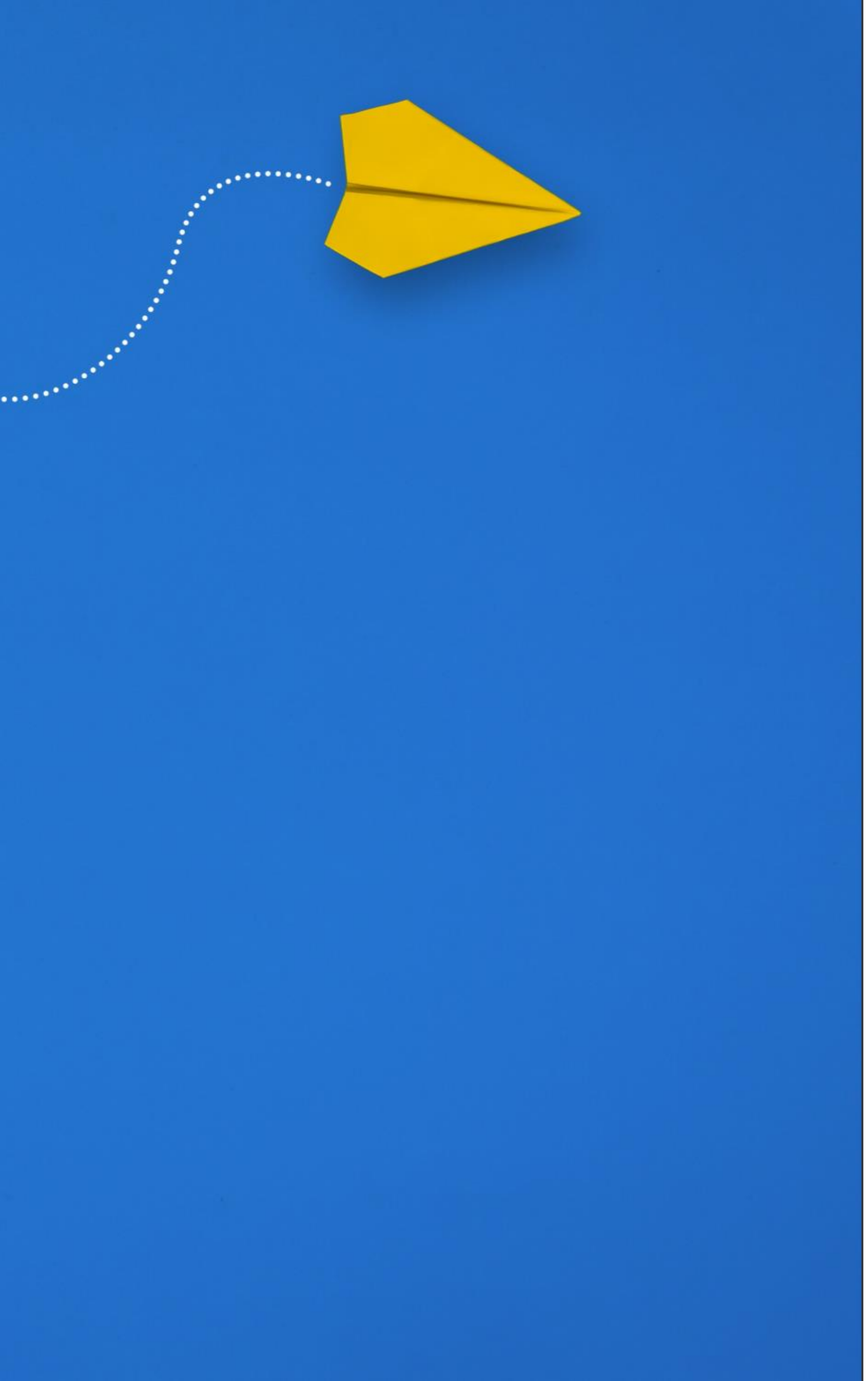


ASSESS QUALITY: METHODS & METRICS

1. "How to"- Guidelines
2. Para data from initial data collection
3. Additional data collection/analysis methods
4. **Simulate what could have gone wrong**
   - measurement error: bots (Schmitz et al., 2022)
   - representation error: device-specific tracking (Bosch et al., 2024)

Implications for direction, consistency, & size of effects





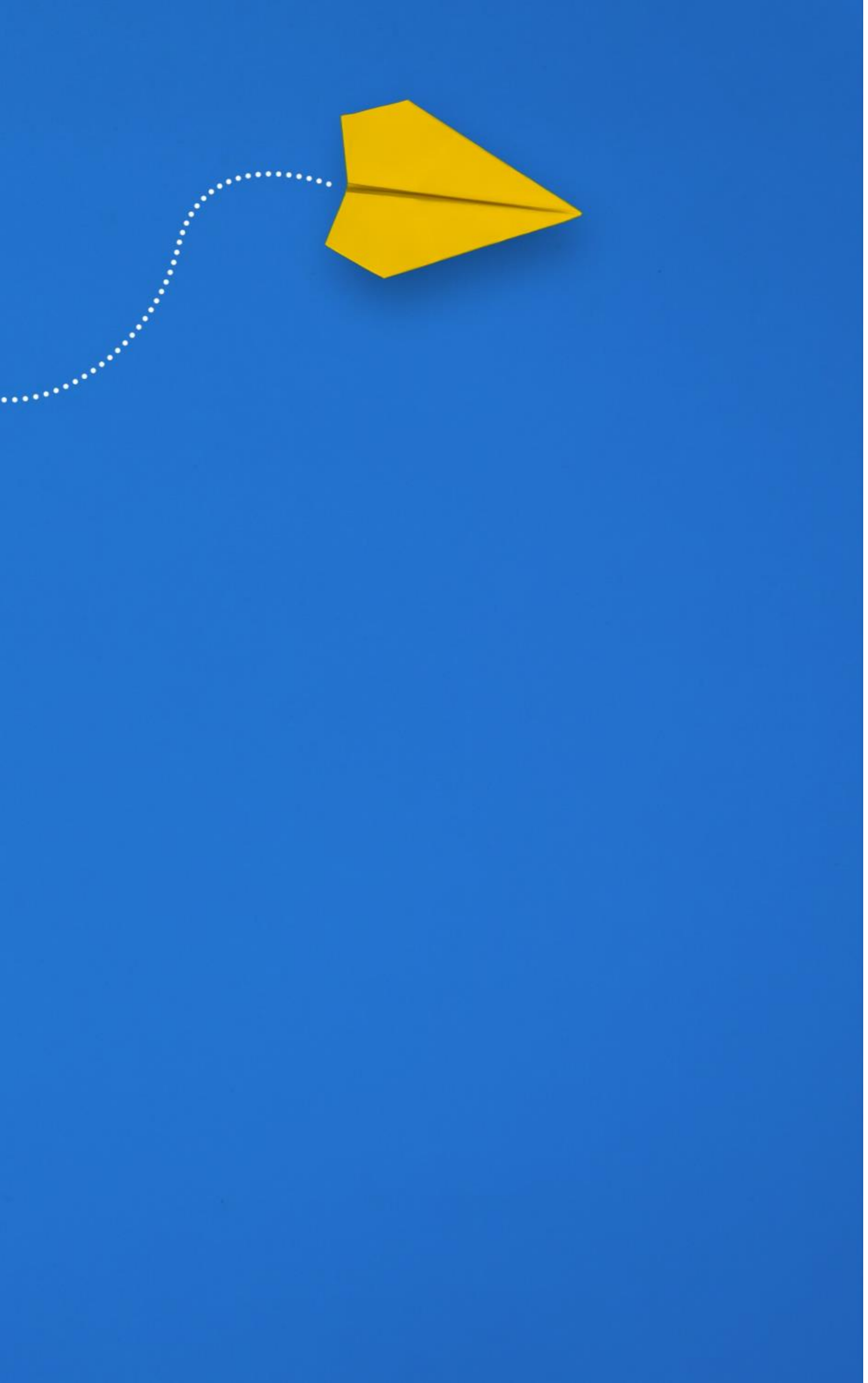
ASSESS QUALITY: METHODS & METRICS

5. "How to"- guidelines & assessment tools

- FAIR checklists (Bahim et al., 2020)
- Assessment tools like F-UJI (Devaraju & Huber, 2021; Devaraju et al., 2022)

Table 1: FAIR data maturity model indicators.

FAIR	ID	Indicator	Priority
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	□□□ Essential
F1	RDA-F1-01D	Data is identified by a persistent identifier	□□□ Essential
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	□□□ Essential
F1	RDA-F1-02D	Data is identified by a globally unique identifier	□□□ Essential
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	□□□ Essential



ASSESS QUALITY: GAPS

- **Missing agreement** upon… (Birkenmaier et al., 2024)
  - methods
  - metrics
  - thresholds for inacceptable quality
- **Unclear predictors** of quality issues (e.g., difference to surveys)

QUALITY FRAMEWORK

I. Define

II. Assess

Methods & Metrics

Error & data quality frameworks

Para data (e.g., response latency, drop-out)
Add. data (e.g., MAR, retrieval/recall precision)
Simulations (e.g., downstream effect sizes)
Tools like F-UJI (e.g., FAIR-score)

Gaps

1. Standardization
2. Epistemological assumptions

1. Methods, metrics, thresholds
2. Predictors

QUALITY FRAMEWORK

I. Define

II. Assess

III. Improve

Methods & Metrics

Error & data quality frameworks

Para data (e.g., response latency, drop-out)
Add. data (e.g., MAR, retrieval/recall precision)
Simulations (e.g., downstream effect sizes)
Tools like F-UJI (e.g., FAIR-score)

Gaps

1. Standardization
2. Epistemological assumptions

1. Methods, metrics, thresholds
2. Predictors

III. IMPROVE QUALITY

Criticizing our methods is great – but could (and should) we not **do more**?

Be **critical, but constructive**:

Adapting existing (or developing new) error correction approaches as the next step in CSS.

EXAMPLE: API STUDY

Can I use APIs to understand which news is shared across platforms?
(Hase et al., 2023)

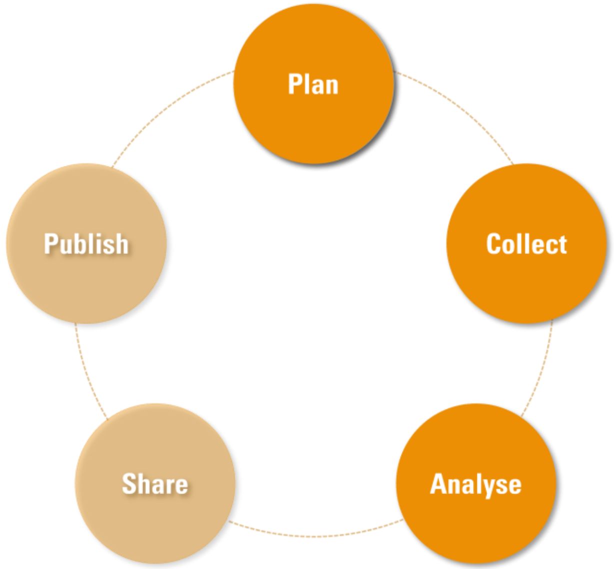
?

EXAMPLE: API STUDY

How diverse is news across digital platforms?

Content analysis German media:
N = 11,000 posts/images/videos

	WEB	FB	INST	TWI
Step 1. Data collection	Crawling & scraping	API	API	API
Step 2. Analysis	Automated text (e.g, BERT transformer) & video analysis (e.g., face detection)			

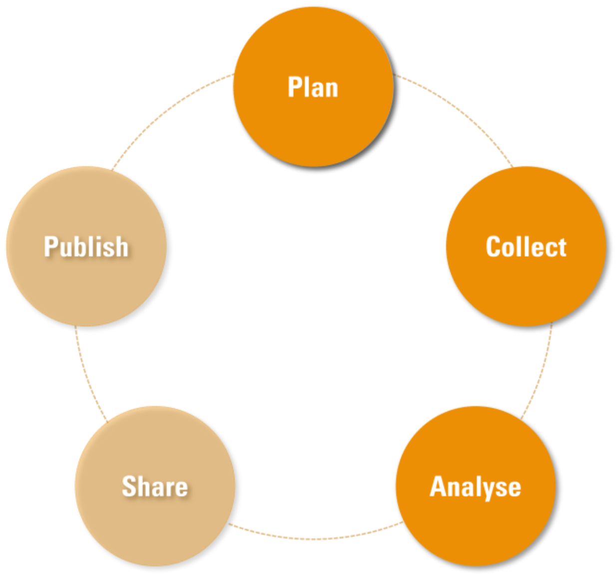


EXAMPLE: API STUDY

Intrinsic (error of representation & measurement error):

✓ Combine data collection methods

- e.g., (1) assess non-random missingness → (2) improve retrieval recall/precision via scraping, API, & manual collection



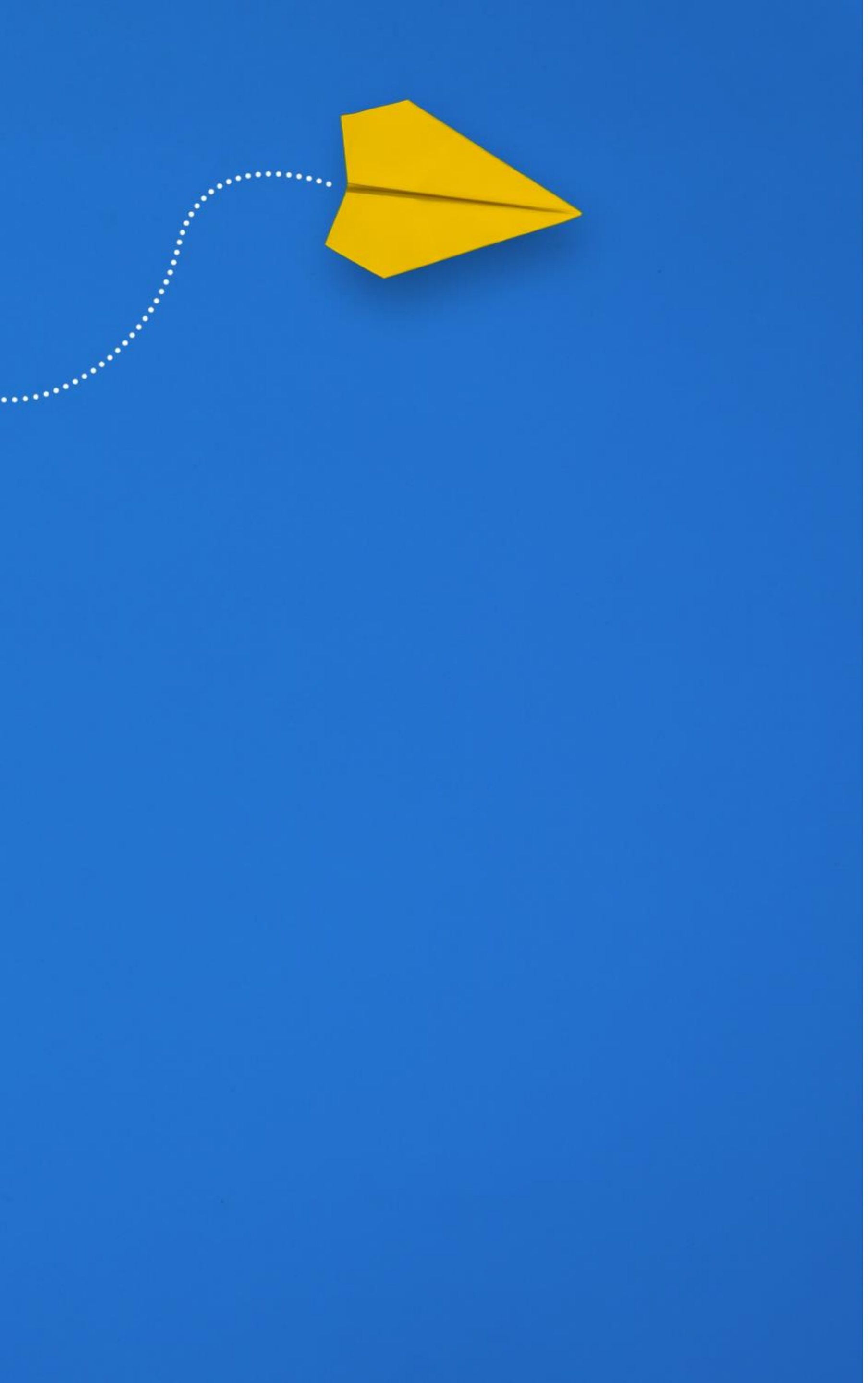
EXAMPLE: API STUDY

Intrinsic (error of representation & measurement error):

✓ Combine data collection methods

× Improve misclassification through error correction methods

- e.g., improve errors in statistical ML inference via packages like misclassificationmodels (TeBlunthuis et al., 2024) or predictionerror (Fong & Tyler, 2021)



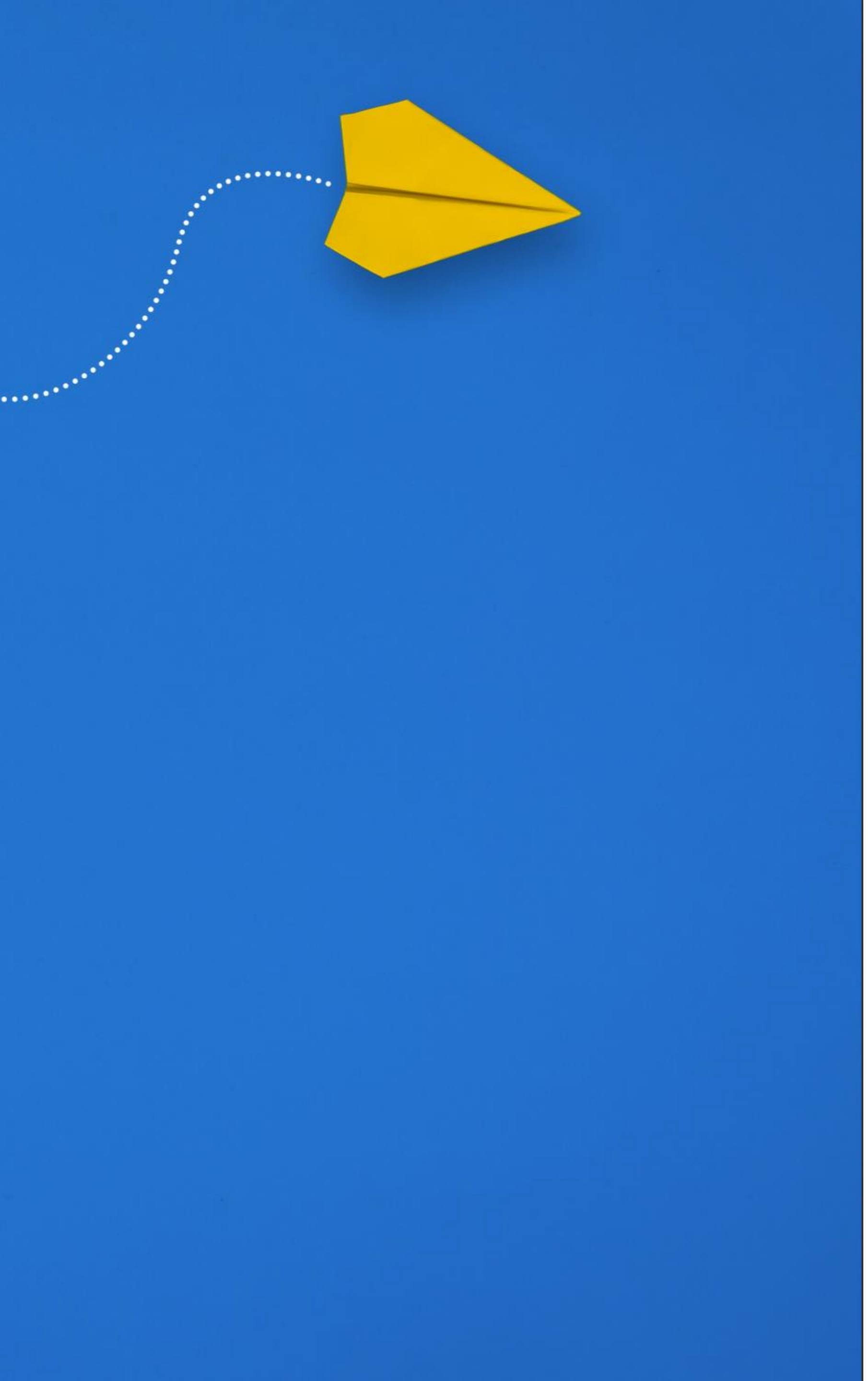
IMPROVE QUALITY

- Interdisciplinary “clash”:

  different definitions of quality + *different quality assessments* =

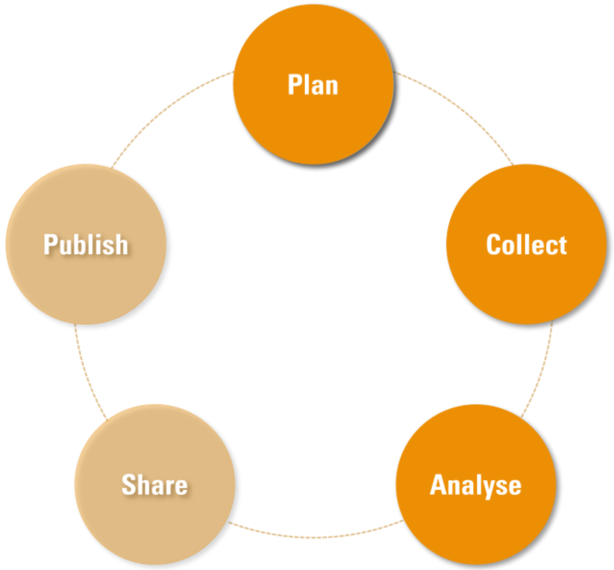
  very different error correction approaches

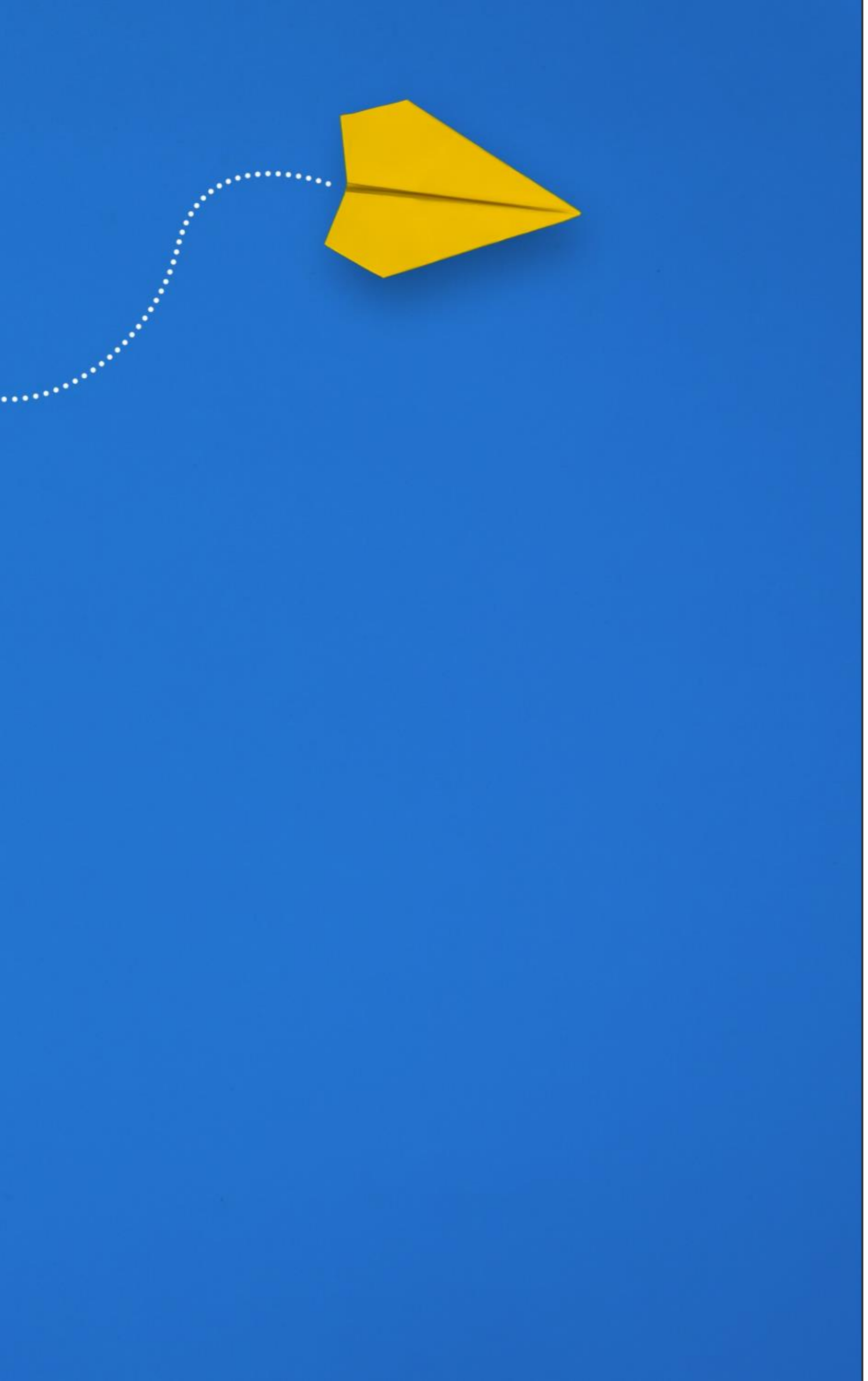
- Take **advantage** of this: Many ways to improve quality!



IMPROVE QUALITY: METHODS & METRICS

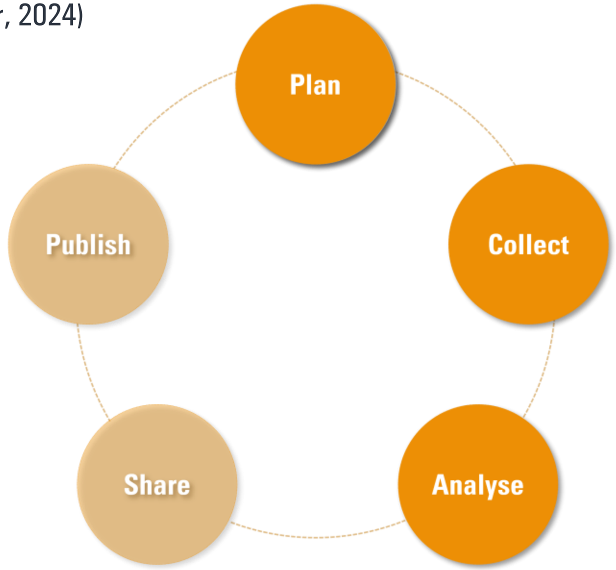
1. **Plan ahead**
   - Talk to IRB, data protection officer, data stewards, …
   - Data management plan (e.g., use files), preregistration
   - Consider non-proprietary methods

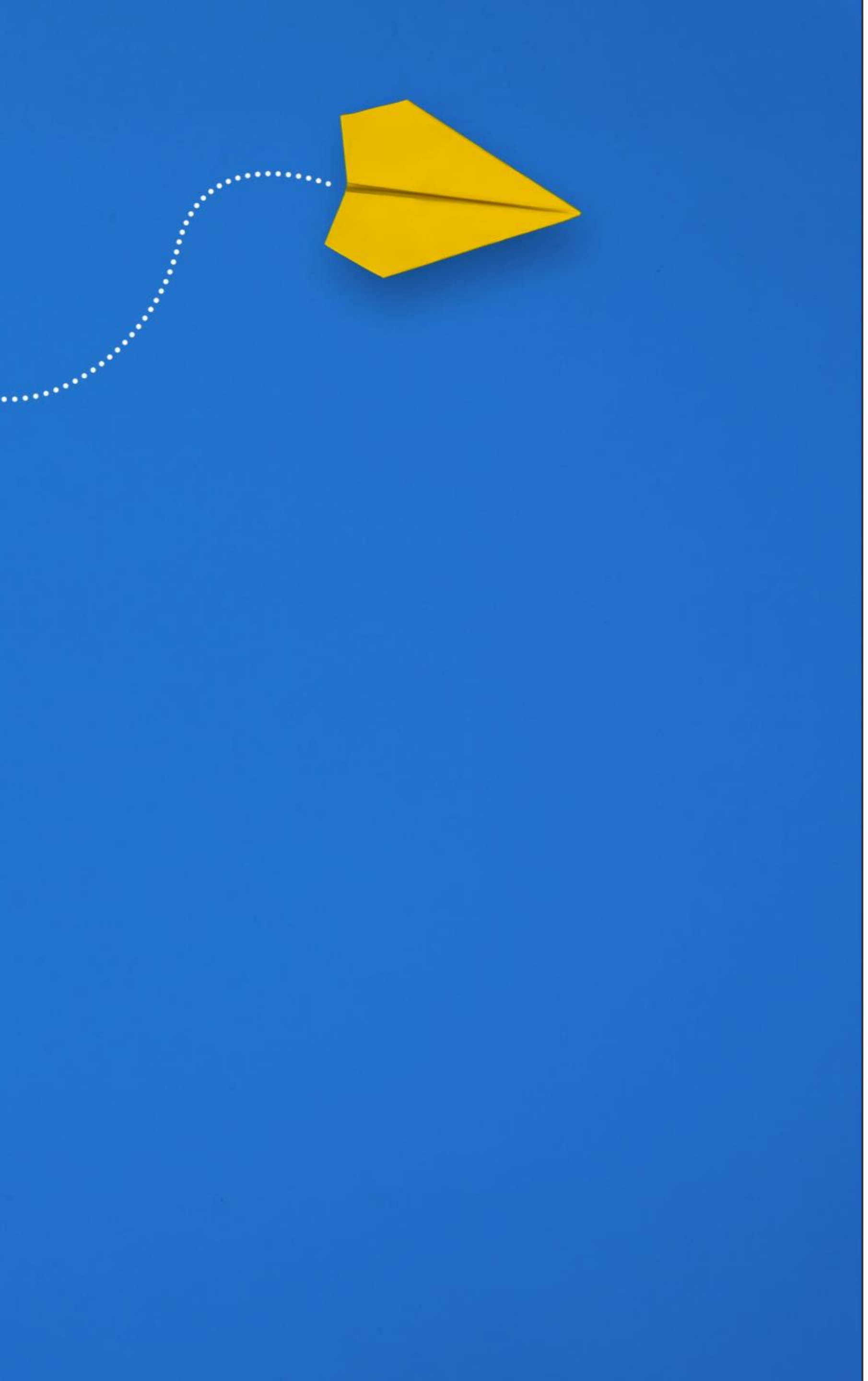




IMPROVE QUALITY: METHODS & METRICS

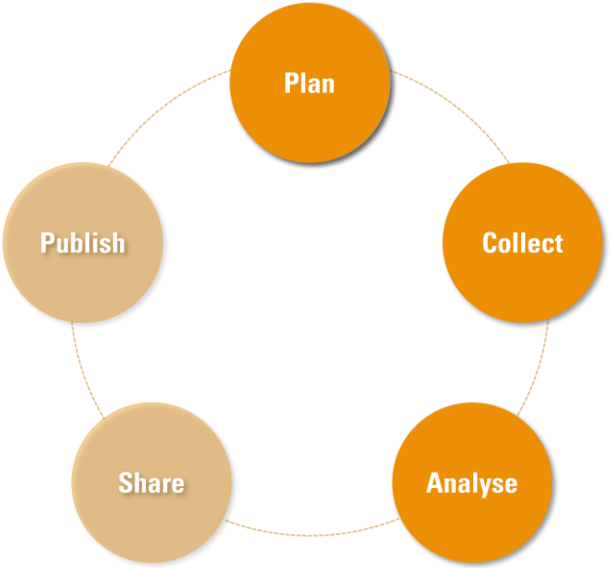
1. Plan ahead
2. **Combine methods for data collection**
   - Repeated/different data access
   - Rehydration (Knöpfle & Schatto-Eckrodt, 2024; Knüpfer, 2024)

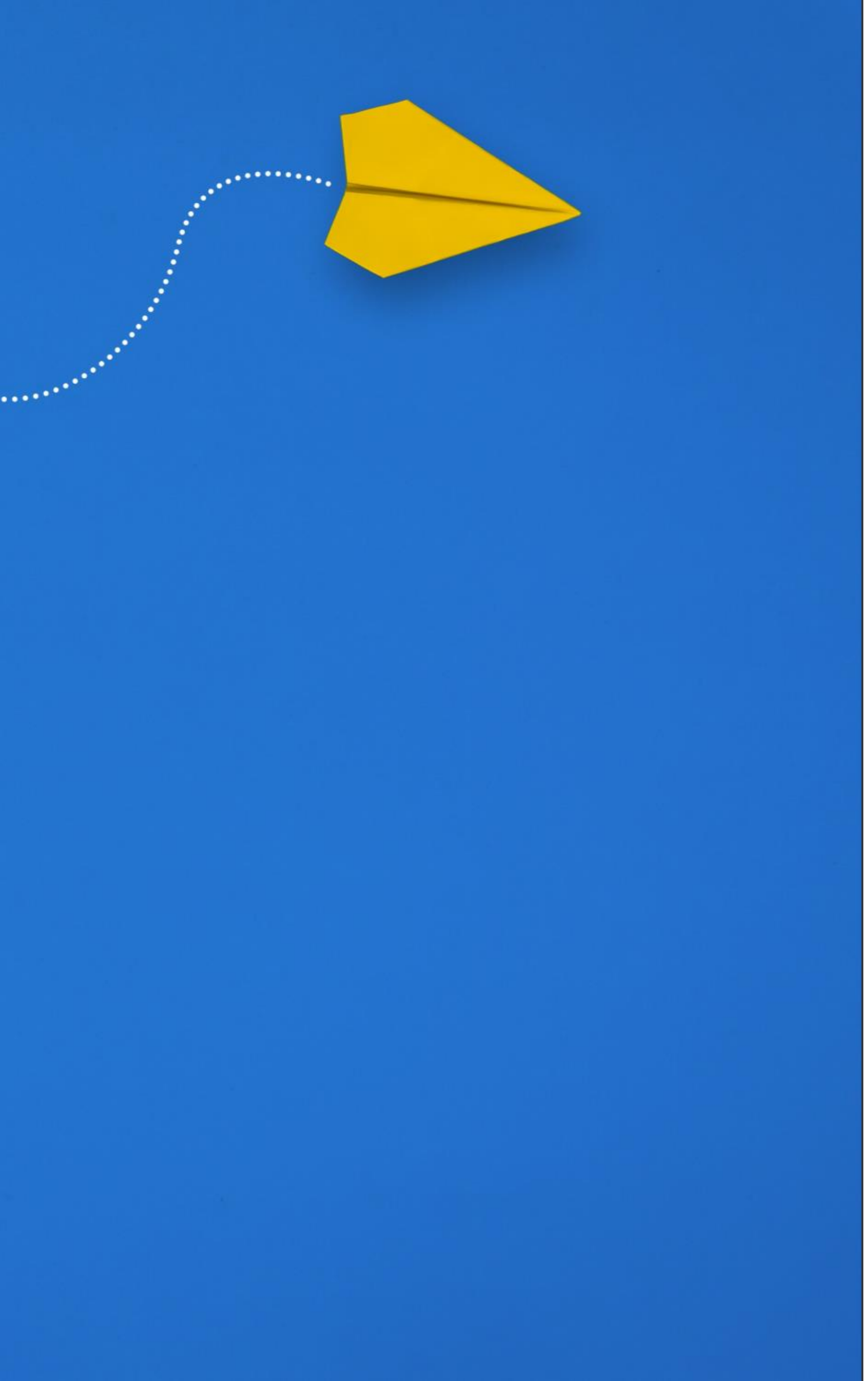




IMPROVE QUALITY: METHODS & METRICS

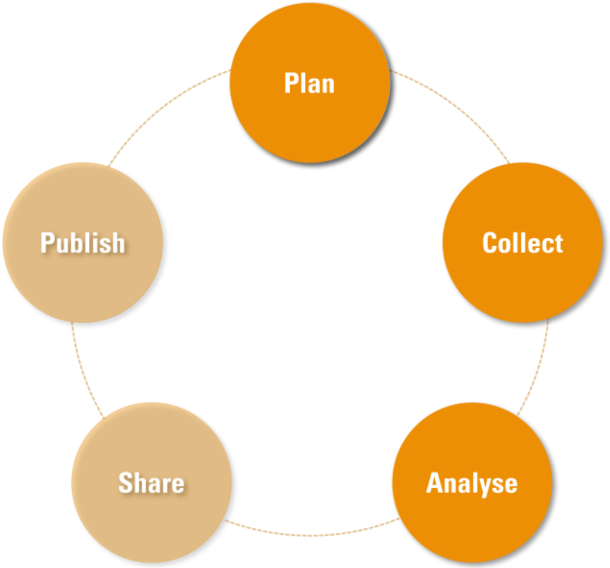
1. Plan ahead
2. Combine methods for data collection
3. **Turn “found” to “designed” data where possible**
   - Use survey design methods
     (Hase & Haim, 2024; Keusch et al., 2024)

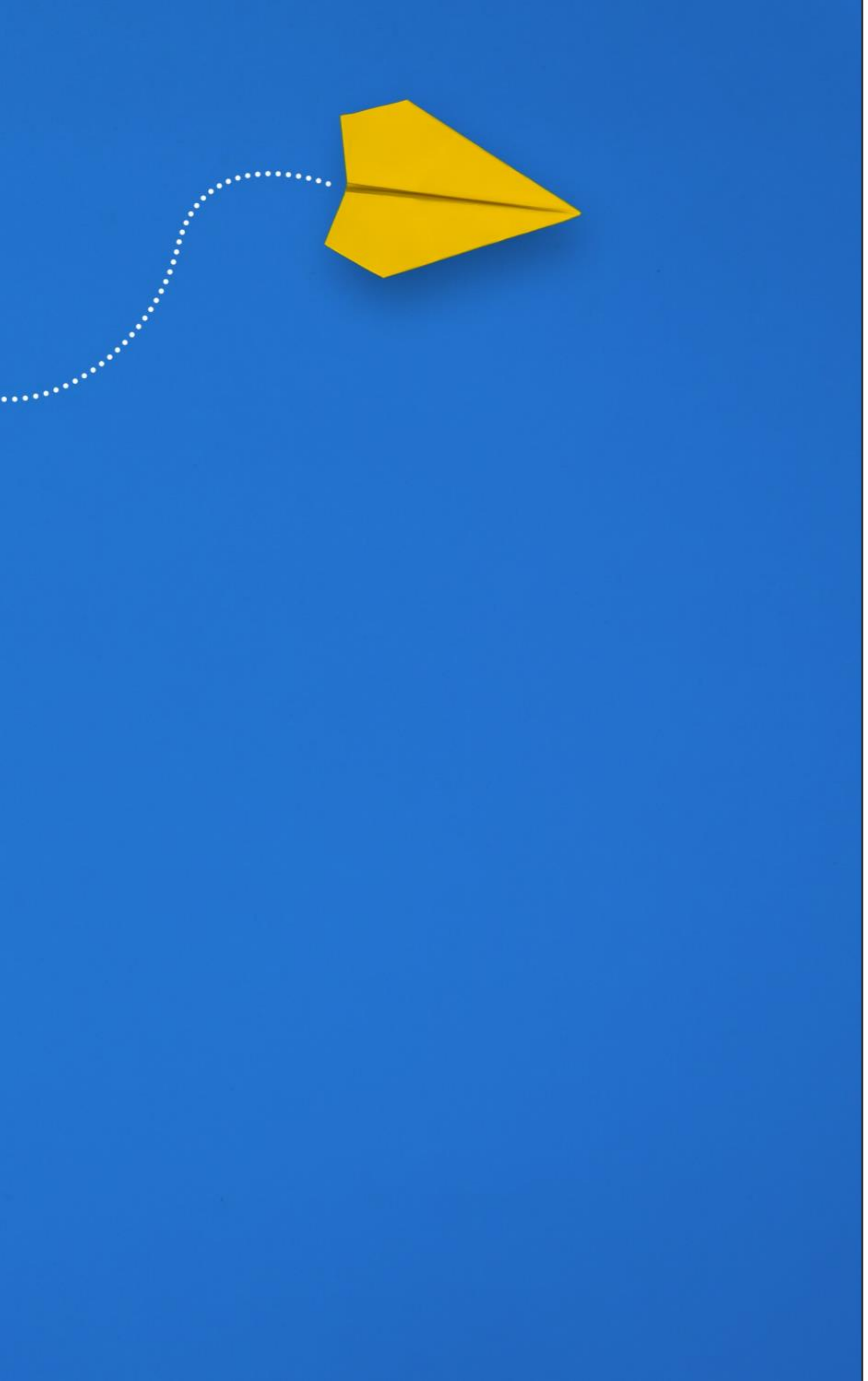




IMPROVE QUALITY: METHODS & METRICS

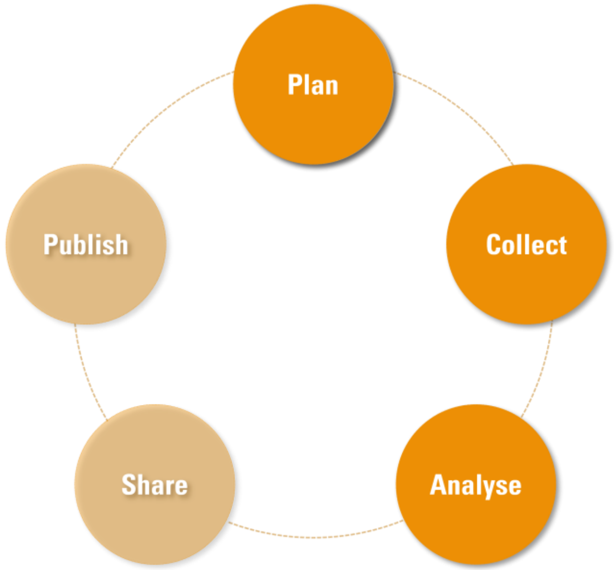
1. Plan ahead
2. Combine methods for data collection
3. Turn “found” to “designed” data where possible
4. **Statistically correct for errors**
   - e.g., weighting to correct for drop-out
     (Pak et al., 2022)
   - e.g., ML-classification for preprocessing
     (Fong & Tyler, 2021; TeBlunthuis et al., 2024)

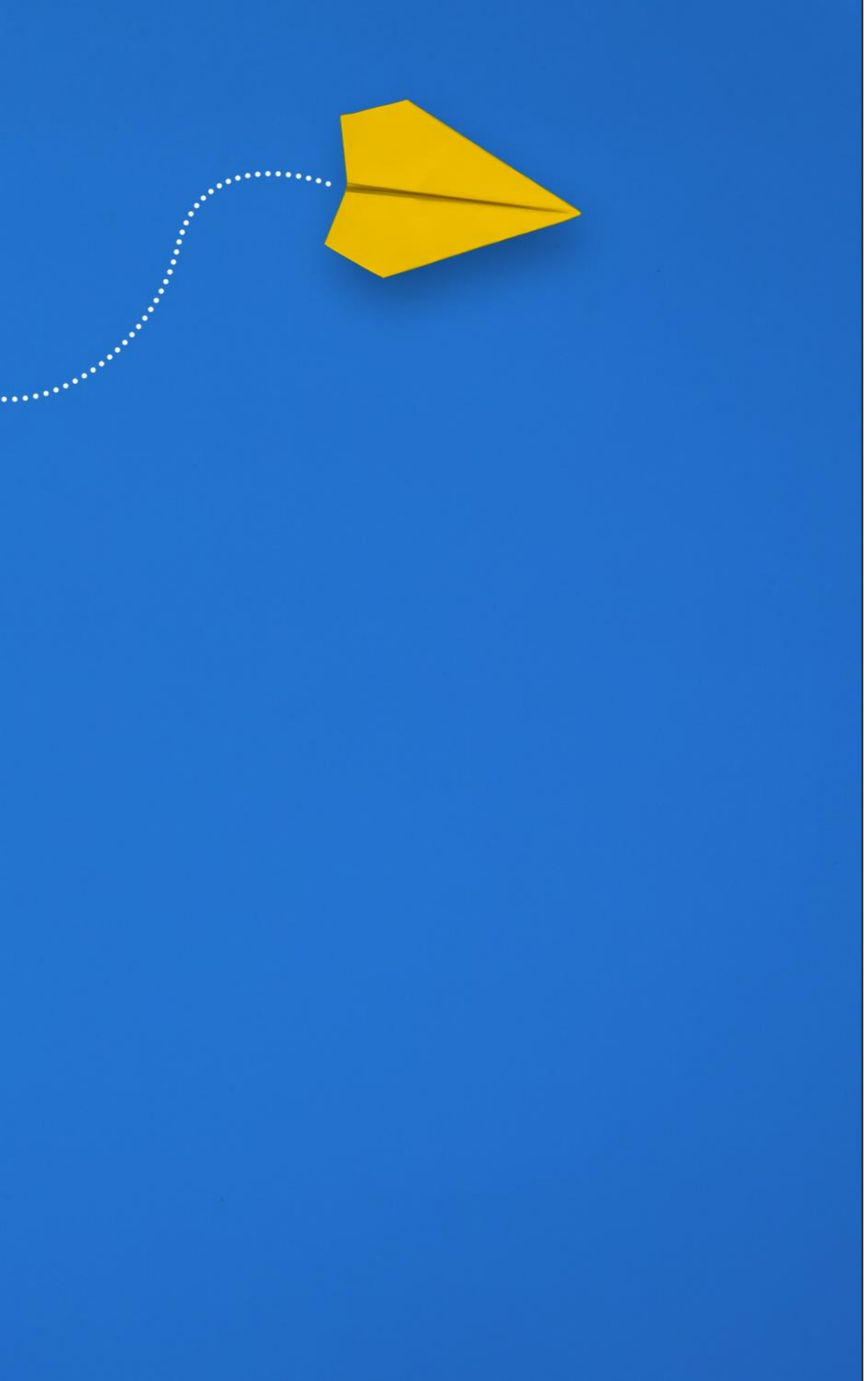




IMPROVE QUALITY: METHODS & METRICS

1. Plan ahead
2. Combine methods for data collection
3. Turn “found” to “designed” data where possible
4. Statistically correct for errors
5. **Ask different questions**
   - e.g., test effects of interventions on rather than describe individual behavior (Straub et al., 2024; Yu et al., 2024)

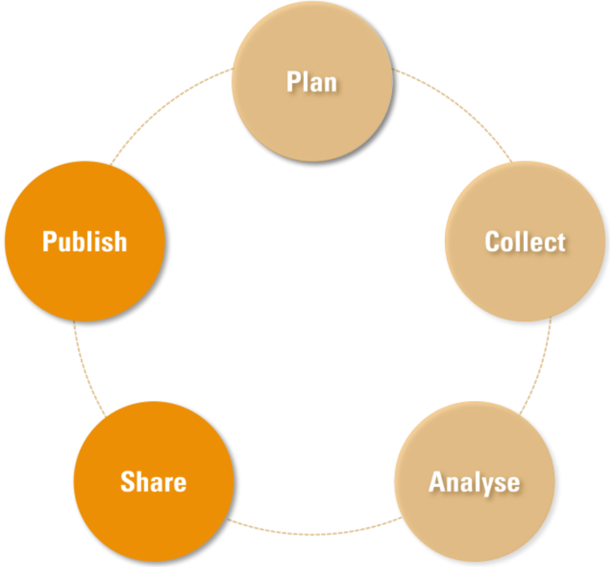


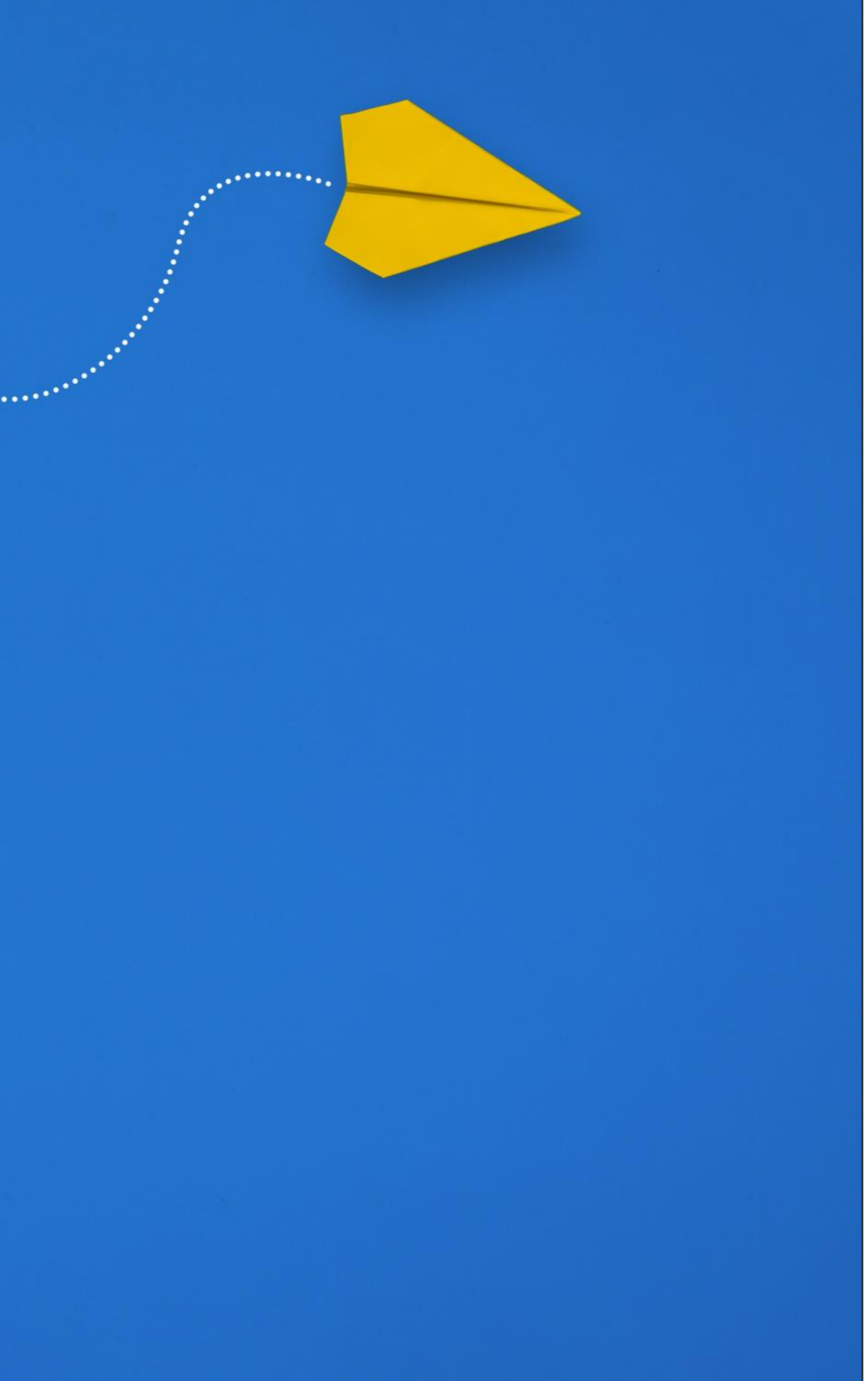


IMPROVE QUALITY: METHODS & METRICS

6. Document everything, including errors

- Datasheets for Datasets (Gebru et al., 2021)
- Data Statements for NLP (Bender & Friedman, 2018)
- Total Error Sheets for Datasets (Fröhling et al., 2023)

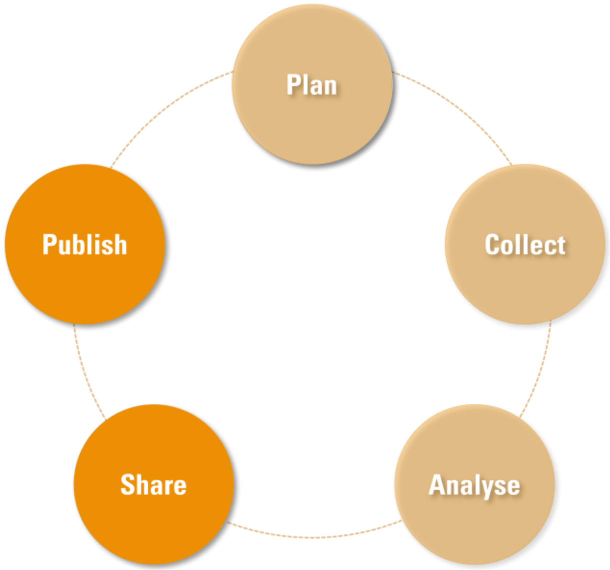


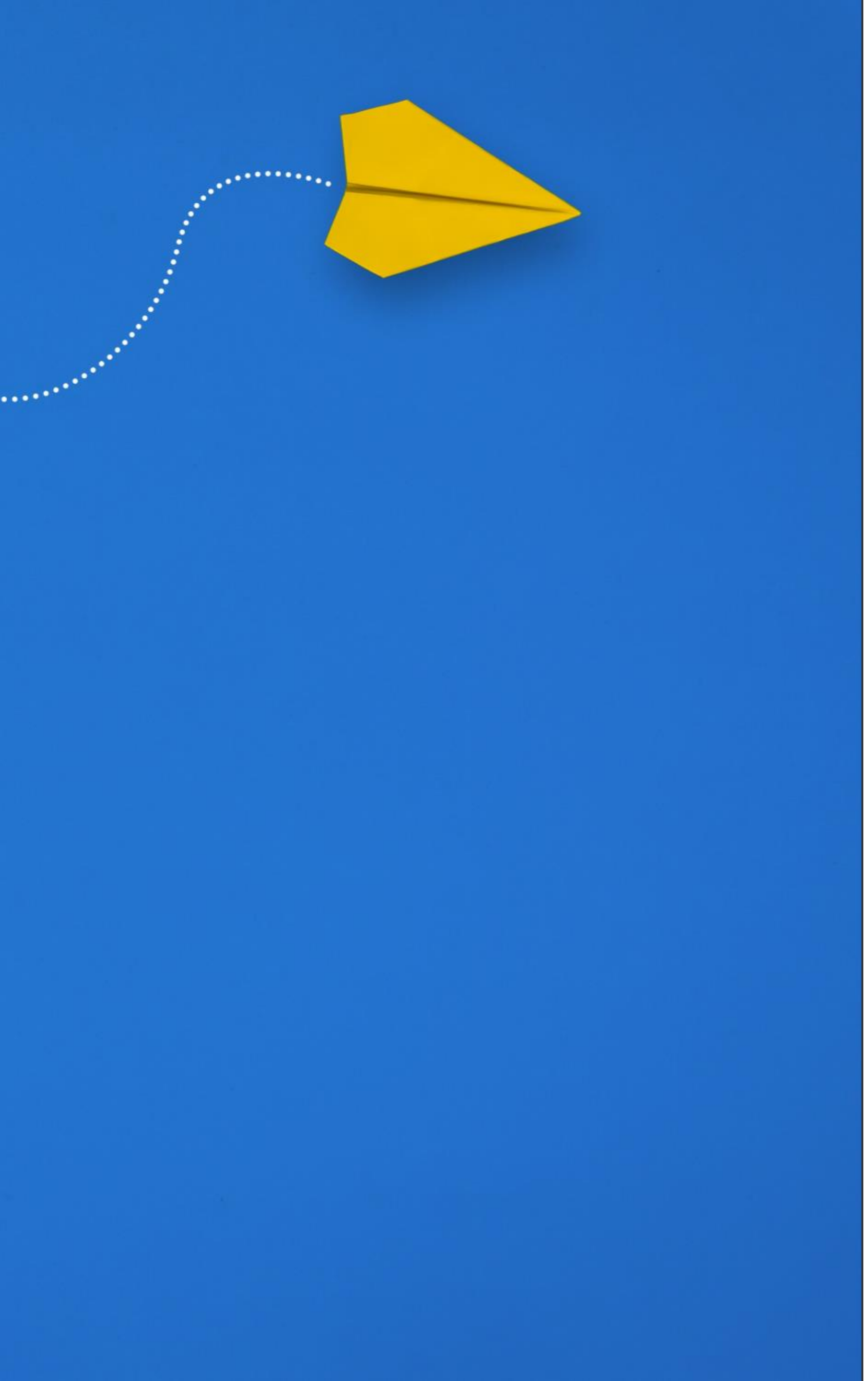


IMPROVE QUALITY: METHODS & METRICS

6. Document everything, including errors

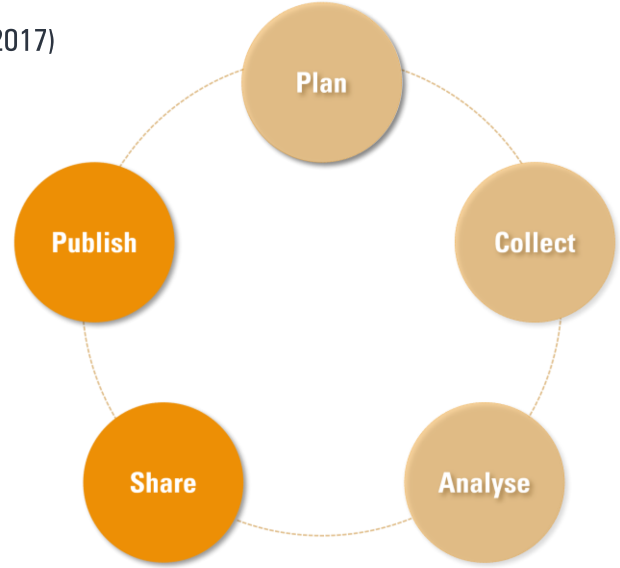
7. **Engage in community-based initiatives**
   - Collective data collection (Pfeffer et al., 2023)
   - Policy efforts, e.g. around DSA (Hase et al., 2024; Jaursch et al., 2024)

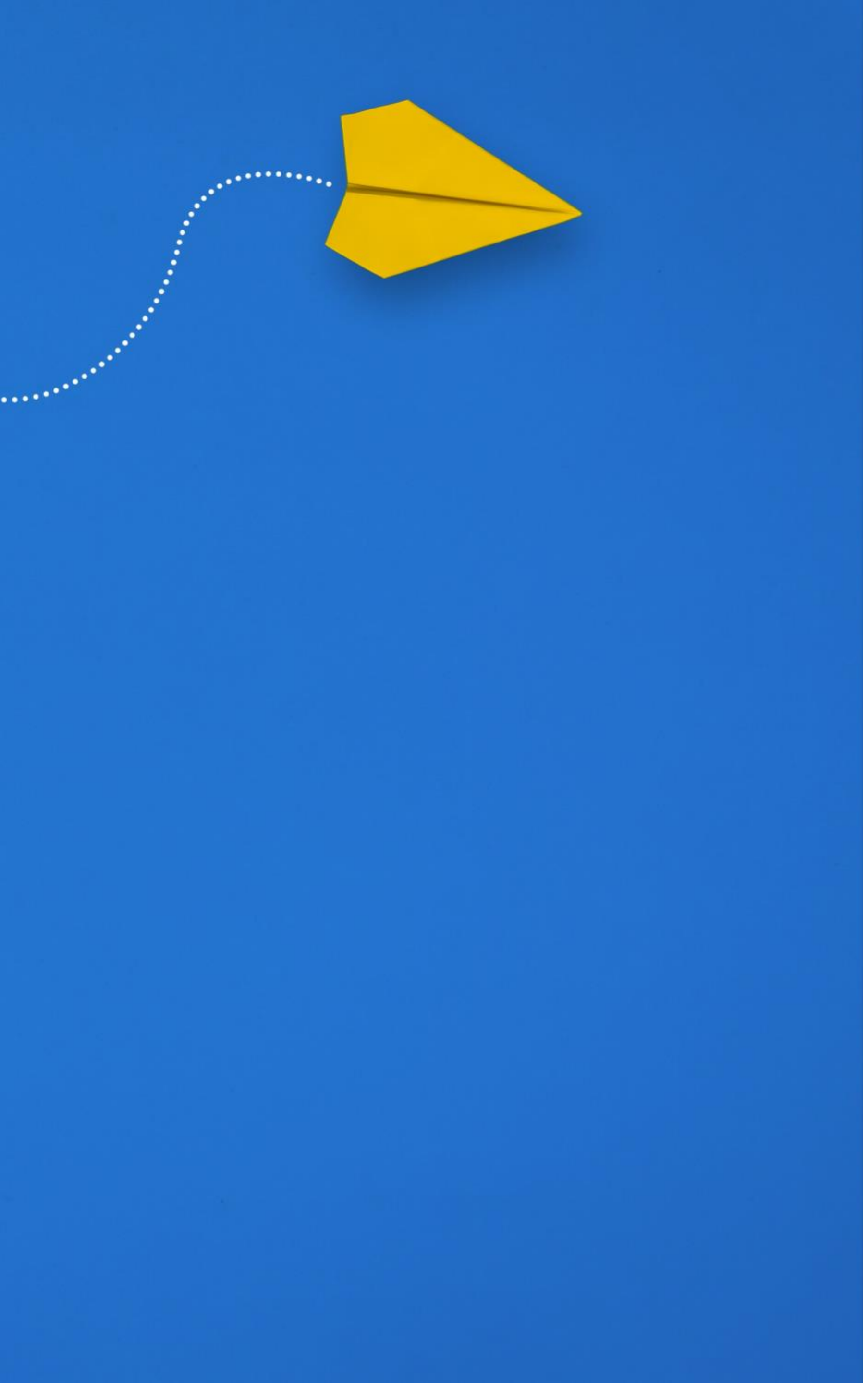




IMPROVE QUALITY: METHODS & METRICS

6. Document everything, including errors
7. Engage in community-based initiatives
8. **Push for infrastructural changes**
   - Peer-reviewed data publications (Carpenter, 2017)
   - Quality check badges (Gottfried et al., 2024)
   - Funding of infrastructure initiatives





IMPROVE QUALITY: GAPS

- **Transferability** of existing error correction methods to CSS
- **Constructive perspective** on bias
  - Identify sub-populations by making “big data” small (Baek et al., 2022)
  - Explore power structures in society (Cabitza et al., 2023; Kathirgamalingam et al., 2024)

QUALITY FRAMEWORK

	Methods & Metrics	Gaps
I. Define	Error & data quality frameworks	1. Standardization 2. Epistemological assumptions
II. Assess	**Para data** (e.g., response latency, drop-out) **Add. data** (e.g., MAR, retrieval/recall precision) **Simulations** (e.g., downstream effect sizes) **Tools** like F-UJI (e.g., FAIR-score)	1. Methods, metrics, thresholds 2. Predictors
III. Improve	**Plan ahead** (e.g., data management plan) **Make „found“ data „designed“** **Use statistical correction** (e.g., ML correction) **Push for infrastructural change** (e.g., badges)	1. Transferability 2. Constructive view on „bias“

Dr. Valerie Hase, LMU Munich

orcid.org/0000-0001-6656-4894

valeriehase

www.valerie-hase.com