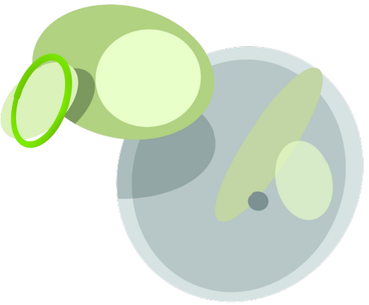
The Data Cards Playbook

A toolkit for purposeful and people-centric dataset documentation for transparency in AI systems.

https://pair-code.github.io/datacardsplaybook/

#datacardsplaybook

Introduction
01 Ask
02 Inspect
03 Answer
04 Audit



Evaluation Gaps in ML Practice

IN THIS SECTION

Using a framework of assumptions, identify gaps in dataset documentation that can adversely affect a reader's evaluation of the dataset.

INSTRUCTIONS

Though originally designed for ML models, this worksheet lists six common assumptions and their corresponding evaluation gaps. Use this table to audit a Data Card for possible gaps and remediative actions.

OUTCOMES

Evaluation and recommendations for completed Data Cards to account for
a wide range of factors and reader needs.

This worksheet was adapted from Evaluation Gaps in Machine Learning Practice, by Hutchinson, et al.

ACTIVITY LEVEL

Advanced

Six Assumptions

1. CONSEQUENTIALISM

Assumes that changes to the ecosystem in which a dataset is used are out of scope when determining if a specific application of a dataset is good or bad.

To test, ask "Does the Data Card focus only on measurable future impacts or first order consequences?"

Gaps: Provenance, Social Responsibility

2. ABSTRACTABILITY FROM CONTEXT

Assumes that the inputs and ground truth in the dataset do not need to capture socially important yet sensitive aspects of the environment.

To test, ask "Does the Data Card describe applicable system dynamics and creator positionalities?"

Gaps: System Considerations, Interpretative Epistemics

3. INPUT MYOPIA

Assumes that the utility of a given feature in the dataset to a task is limited and/or independent of the effect of other features.

To test, ask "Does the Data Card describe causal relationships between different features and the tasks that the dataset was intended for?"

Gaps: Disaggregated Analysis

4. QUANTIFIABILITY

Assumes that the impacts on individuals are reducible to numbers, trivializing the difficulty in comparing benefits and costs.

To test, ask "Does the Data Card provide explanations and information that adequately speaks to a range of downstream impacts?"

Gaps: Incommensurables

5. FAILURE CASES ARE EQUIVALENT

Assumes that all errors and error rates captured by a defined set of metrics are equivalent, even if error magnitudes vary.

To test, ask "Does the Data Card focus only on measurable future impacts or first order consequences?"

Gaps: Disparate Harms & Benefits

6. TEST DATA VALIDITY

Assumes that the methods used to estimate performance results model the behavior in the ecosystem that the dataset will be used in.

To test, ask "Does the Data Card focus only on measurable future impacts or first order consequences?"

Gaps: Data Drifts

10 min

Gap: Provenance

Assumption: Consequentialism

Gap occurs when: the Data Card minimizes or omits important value-oriented considerations that went into the construction of the dataset.

Examples of gaps in the Data Card include missing inter-rater policies related to collection, parameters related to the selection, inclusion and exclusion of data points from the dataset.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Social Responsibility

Assumption: Consequentialism

Gap occurs when: the Data Card does not report on the bounds and limitations related to the social contracts that guide the ecosystem in which the dataset was designed to be used.

Examples of gaps in the Data Card include missing assessments of human rights, social and ethical impact; privacy or principle-oriented assessments, if any.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: System Considerations

Assumption: Abstractability from Context

Gap occurs when: the Data Card describes the dataset as a benchmark for a set of industry standard models without paying attention to interpretability or explainability methods that describe results, or omits any human-in-the-loop or feedback loops in the ecosystems in which the dataset was created or will be used.

Examples of gaps in the Data Card include missing inter-rater policies related to collection, parameters related to the selection, confounding or control effects in evaluations.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Interpretive Epistemics

Assumption: Abstractability from Context

Gap occurs when: the Data Card is implicitly positivist in position and the information contained is seen as socially and culturally independent that is interpreted in a fixed set of ways and contexts.

Examples of gaps in the Data Card include missing decisions, rationales, as well as past research, experience, and personal insights that impacted the creation or curation of the dataset.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Disaggregated Analysis

Assumption: Input Myopia

Gap occurs when: the Data Card presents evaluation and robustness statistics in aggregate, and are not broken down by meaningful and potentially sensitive intersections in variables of interest.

Examples of gaps in the Data Card include missing fairness analysis, disaggregated statistics, correlations, risks and trade-offs.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Incommensurables

Assumption: Quantifiability

Gap occurs when: harms and benefits of using a dataset in the intended use cases are assumed to be comparable in the same scale, and may disproportionately impact underrepresented groups.

Examples of gaps in the Data Card include missing data points, sampling errors when pertinent, and any other qualitative impacts that cannot be inferred from the dataset itself, such as suitable and unsuitable use cases.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Disparate Harms & Benefits

Assumption: Failure Cases are Equivalent

Gap occurs when: the Data Card describes the impact of errors generally – without breaking down errors by how offensive or harmful they might be.

Examples of gaps in the Data Card include missing descriptions of the model(s) used to demonstrate performance, description of evaluation processes, expected performance and any known caveats that readers should be aware of.

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

Gap: Data Drifts

Assumption: Test Data Validity

Gap occurs when: the Data Card describes a distribution of the data that diverges from that of the ecosystem, evaluation samples, or don't account from system feedback effects.

Examples of gaps in the Data Card include missing sources, its features, shapes and any warnings about the datasets in use

GAP OBSERVED

✍️ Provide a description of this gap as observed in the Data Card or the assessment of the dataset.

READER IMPACT

✍️ Which reader or audience group will this gap most likely impact?

SOURCE + EVIDENCE

✍️ If at all, what is the source of the gap? If possible, point to evidence in the Data Card.

MITIGATION / ACTION

✍️ How might this gap be reduced? If at all, what strategies can help dataset owners take to complete the evaluation or explicitly make the reader aware of this gap?

10 min

GAP OBSERVED	SOURCE + EVIDENCE	READERS IMPACTED	MITIGATION / ACTION
1 . ✍️ Largest Gaps observed	✍️ Sources and evidence of gaps	1 . ✍️ Reader and audience groups most impacted	1 . ✍️ Actionable steps for Data Card Creator
2 . ✍️ Largest Gaps observed	✍️ Sources and evidence	2 . ✍️ Reader and audience groups moderately impacted	2 . ✍️ Actionable steps for Data Card Creator
3 . ✍️ Largest Gaps observed	✍️ Sources and evidence	3 . ✍️ Reader and audience groups somewhat impacted	3 . ✍️ Actionable steps for Data Card Creator

Checklist

YOU SHOULD NOW HAVE

- ✔ Audited a completed Data Card for six assumptions
- ✔ Identified specific gaps in documentation that can be corrected
- ✔ Feedback that the Data Card producers can use to improve the Data Card
- ✔ Identified aspects of a dataset that might not be remediated at this time, but can inform future dataset development and documentation practices

#datacardsplaybook

The Data Cards Playbook ↗ is an adaptable toolkit of participatory activities, conceptual frameworks, and guidance that support Responsible AI practices for transparency in dataset documentation.

If you've adapted, implemented, or have feedback for this guidance, we'd love to hear from you at https://github.com/pair-code/datacardsplaybook ↗.

Find the complete playbook at
https://pair-code.github.io/datacardsplaybook ↗