

CS 360 Machine Learning

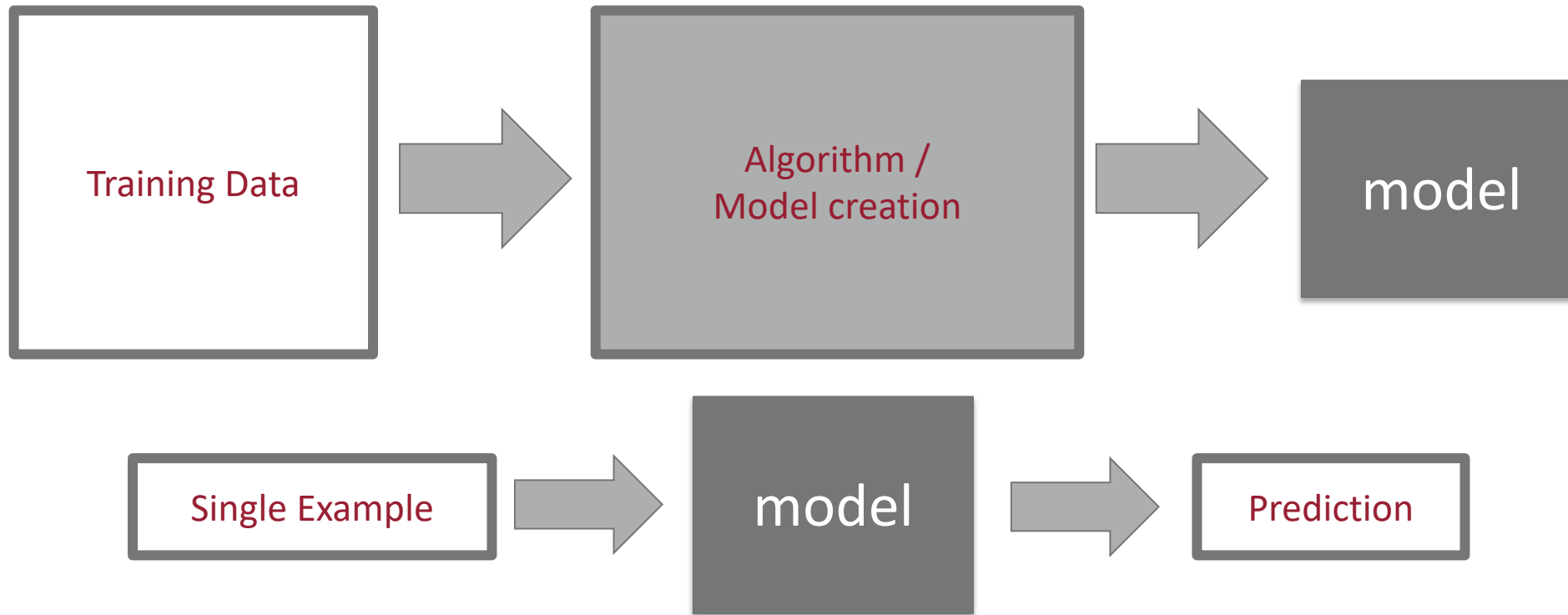
Sources of Error in an ML Pipeline



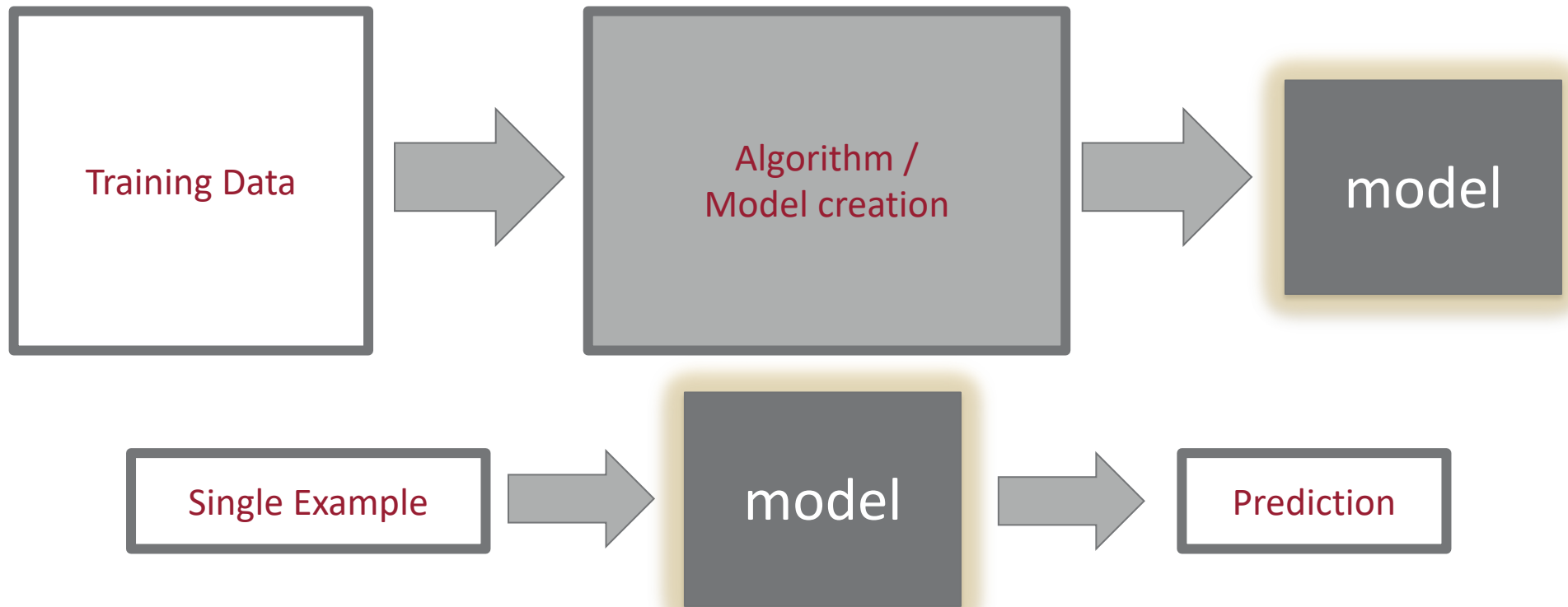
Haverford
COLLEGE

DEPARTMENT OF COMPUTER SCIENCE

Machine Learning Pipeline



Error Measures



Accuracy and other traditional error measures focus on evaluating the model against the test data.



Error Measures

Our standard confusion matrix:

	Predicted False	Predicted True
Test Label False	TN	FP
Test Label True	FN	TP

Accuracy and other traditional error measures focus on evaluating the model against the test data.



Fairness

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Proceedings of Machine Learning Research 81:1–15, 2018

Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini

MIT Media Lab 75 Amherst St. Cambridge, MA 02139

JOYAB@MIT.EDU

Timnit Gebru

Microsoft Research 641 Avenue of the Americas, New York, NY 10011

TIMNIT.GEBRU@MICROSOFT.COM

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	TPR(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	PPV (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	2.6	10.7	12.9	0.7	6.0	20.8	0.0	1.7
Face++	TPR(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	90.2	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	9.8	0.8
	PPV (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	0.7	21.3	16.5	4.7	0.7	34.5	0.8	9.8
IBM	TPR(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	34.7	0.3	7.1

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).



Fairness



U.S. Equal Employment Opportunity Commission

Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964

Employers now have a wide variety of algorithmic decision-making tools available to assist them in making employment decisions, including recruitment, hiring, retention, promotion, transfer, performance monitoring, demotion, dismissal, and referral. Employers increasingly utilize these tools in an attempt to save time and effort, increase objectivity, optimize employee performance, or decrease bias.

Many employers routinely monitor their more traditional decision-making procedures to determine whether these procedures cause disproportionately large negative effects on the basis of race, color, religion, sex, or national origin under Title VII of the Civil Rights Act of 1964 ("Title VII").^[1] Employers may have questions about whether and how to monitor the newer algorithmic decision-making tools. The Questions and Answers in this document address this and several closely related issues.

Title VII applies to all employment practices of covered employers, including recruitment,

This technical assistance document was issued upon approval of the Chair of the U.S. Equal Employment Opportunity Commission.

OLC Control Number: EEOC-NVTA-2023-2
Concise Display Name: Title VII and AI: Assessing Adverse Impact
Issue Date: 05-18-2023

4. What is a "selection rate"?

"Selection rate" refers to the proportion of applicants or candidates who are hired, promoted, or otherwise selected.^[12] The selection rate for a group of applicants or candidates is calculated by dividing the number of persons hired, promoted, or otherwise selected from the group by the total number of candidates in that group.^[13] For example, suppose that 80 White individuals and 40 Black individuals take a personality test that is scored using an algorithm as part of a job application, and 48 of the White applicants and 12 of the Black applicants advance to the next round of the selection process. Based on these results, the selection rate for Whites is 48/80 (equivalent to 60%), and the selection rate for Blacks is 12/40 (equivalent to 30%).

5. What is the "four-fifths rule"?

The four-fifths rule, referenced in the *Guidelines*, is a general rule of thumb for determining whether the selection rate for one group is "substantially" different than the selection rate of another group. The rule states that one rate is substantially different than another if their ratio is less than four-fifths (or 80%).^[14]

In the example above involving a personality test scored by an algorithm, the selection rate for Black applicants was 30% and the selection rate for White applicants was 60%. The ratio of the two rates is thus 30/60 (or 50%). Because 30/60 (or 50%) is lower than 4/5 (or 80%), the four-fifths rule says that the selection rate for Black applicants is substantially different than the selection rate for White applicants in this example, which could be evidence of discrimination against Black applicants.



Fairness Measures

Example with fairness-focused confusion matrices:

Confusion matrix for everyone

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	TN_{all}	FP_{all}
Test Label WAS hired	FN_{all}	TP_{all}

Confusion matrix for men

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	TN_{men}	FP_{men}
Test Label WAS hired	FN_{men}	TP_{men}

Confusion matrix for non-men

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	TN_{non}	FP_{non}
Test Label WAS hired	FN_{non}	TP_{non}

Fairness measures focus on evaluating the model against the test data **per demographic group**.



Fairness Measures

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	TN_{men}	FP_{men}
Test Label WAS hired	FN_{men}	TP_{men}

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	TN_{non}	FP_{non}
Test Label WAS hired	FN_{non}	TP_{non}

“disparate impact” measure ignores “mis”classification:

$$\frac{\text{rate of hiring for non-men}}{\text{rate of hiring for men}} \geq \frac{4}{5}$$

$$\frac{\left(\frac{FP_{non} + TP_{non}}{\text{total non-men}}\right)}{\left(\frac{FP_{men} + TP_{men}}{\text{total men}}\right)}$$

“error rate balance”, “equal odds”, or “equality of opportunity” measures are demographic-conditioned error measures:

$$\frac{\text{true positive rate for non-men}}{\text{true positive rate for men}} = 1$$

$$\frac{\left(\frac{TP_{non}}{TP_{non} + FN_{non}}\right)}{\left(\frac{TP_{men}}{TP_{men} + FN_{men}}\right)} = 1$$

$$\frac{\text{false positive rate for non-men}}{\text{false positive rate for men}} = 1$$



Fairness Measures: discussion

Non-men

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	542	170
Test Label WAS hired	23	56

Men

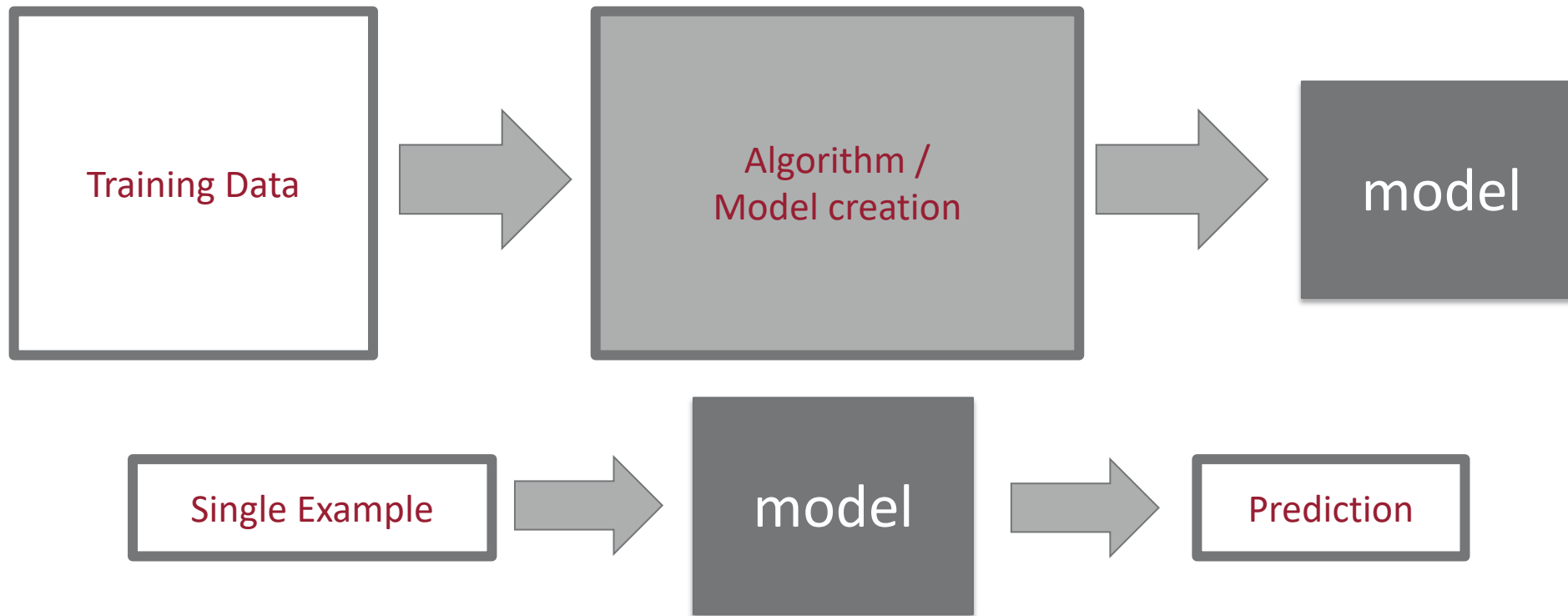
	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	1598	430
Test Label WAS hired	340	190

Consider these confusion matrices for a resume screening model:

- Calculate at least two different fairness measures
- What do you notice?
- Does this model appear “fair” to you?



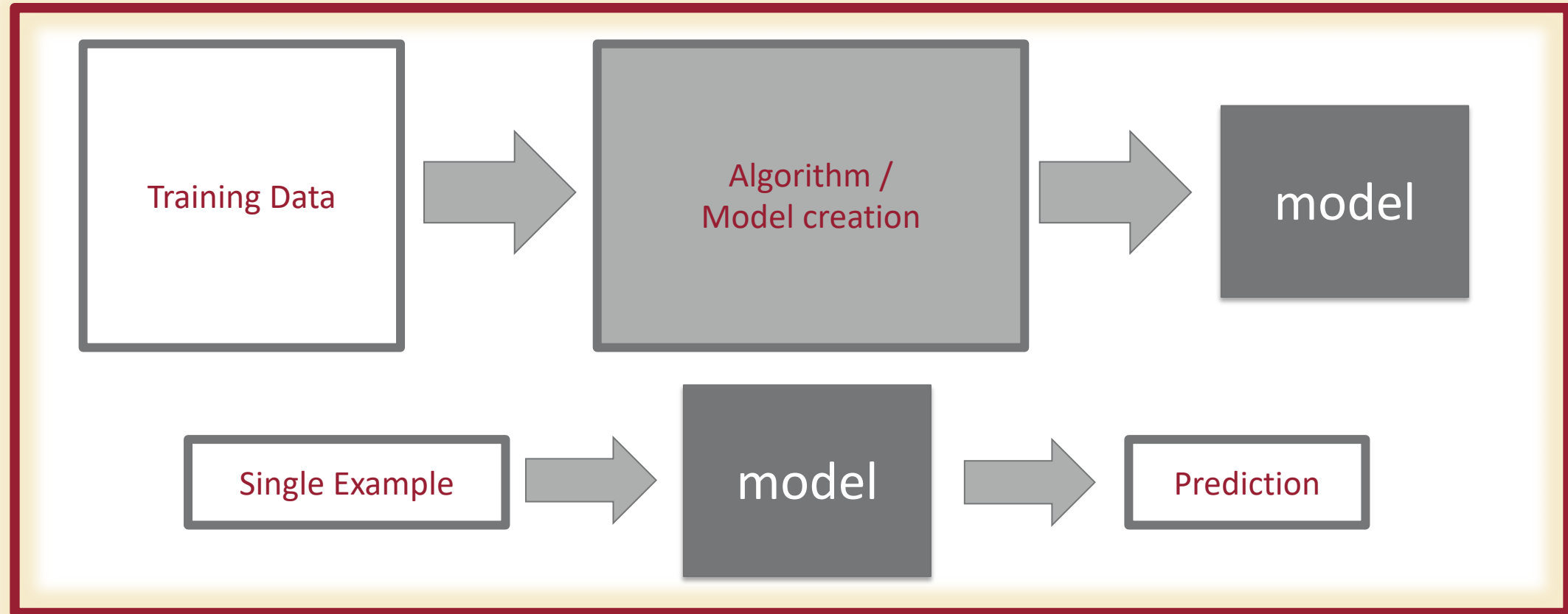
Sources of Error



In a real world problem, you've made assumptions throughout this pipeline – what if they're wrong?



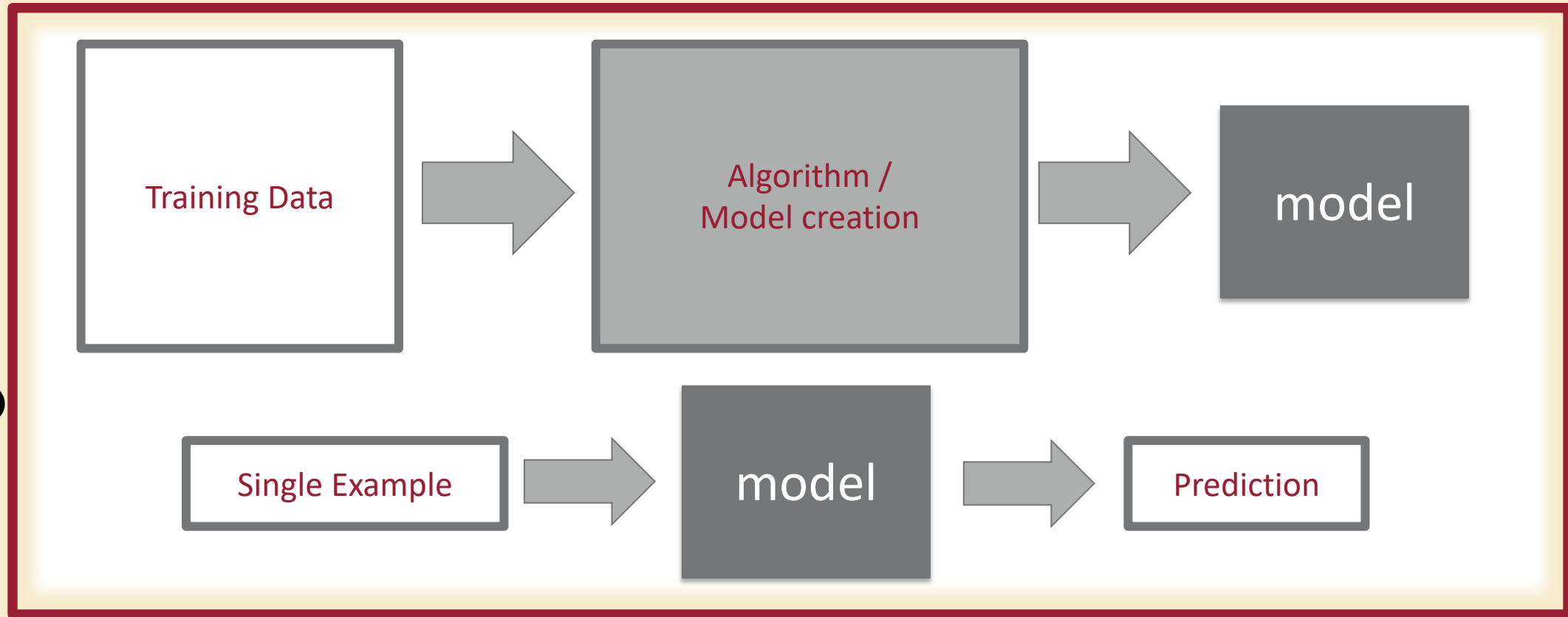
Sources of Error



Assumption 0: the problem is appropriate to solve with ML



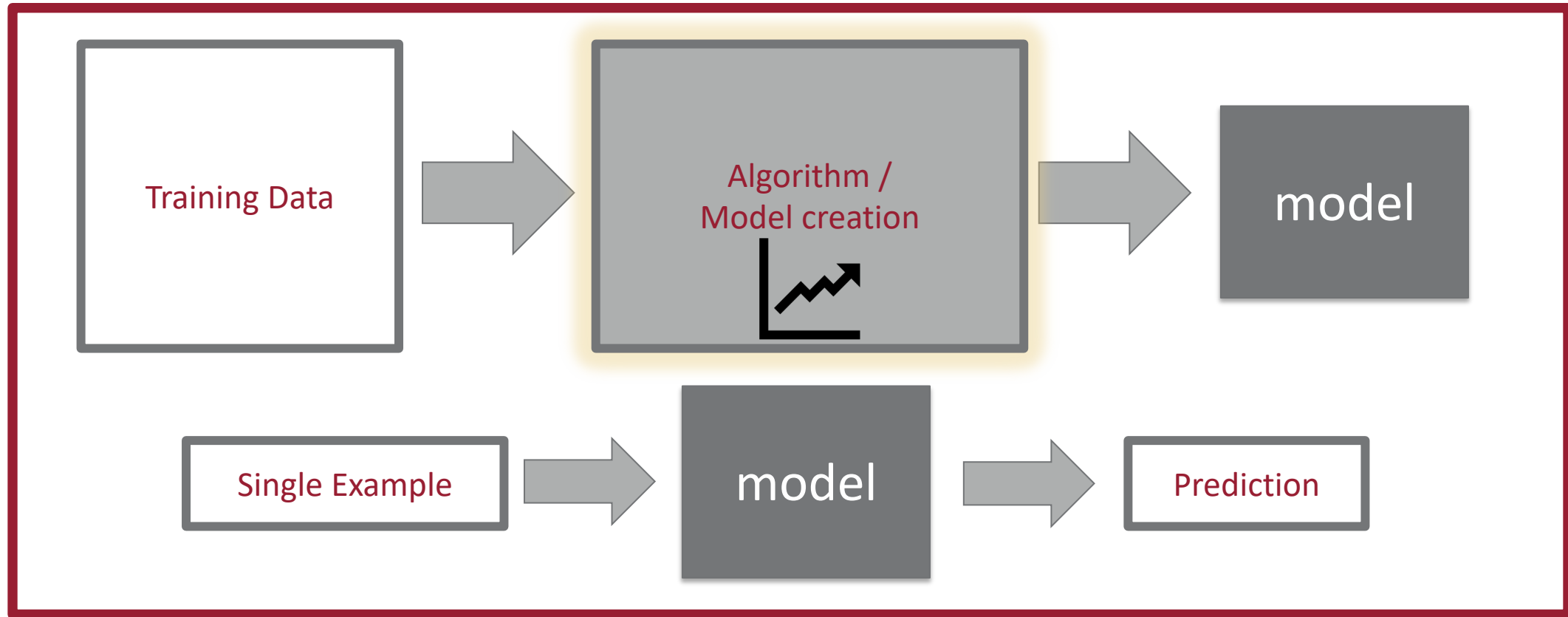
Sources of Error



Assumption 1: the real world won't change or impact the ML pipeline



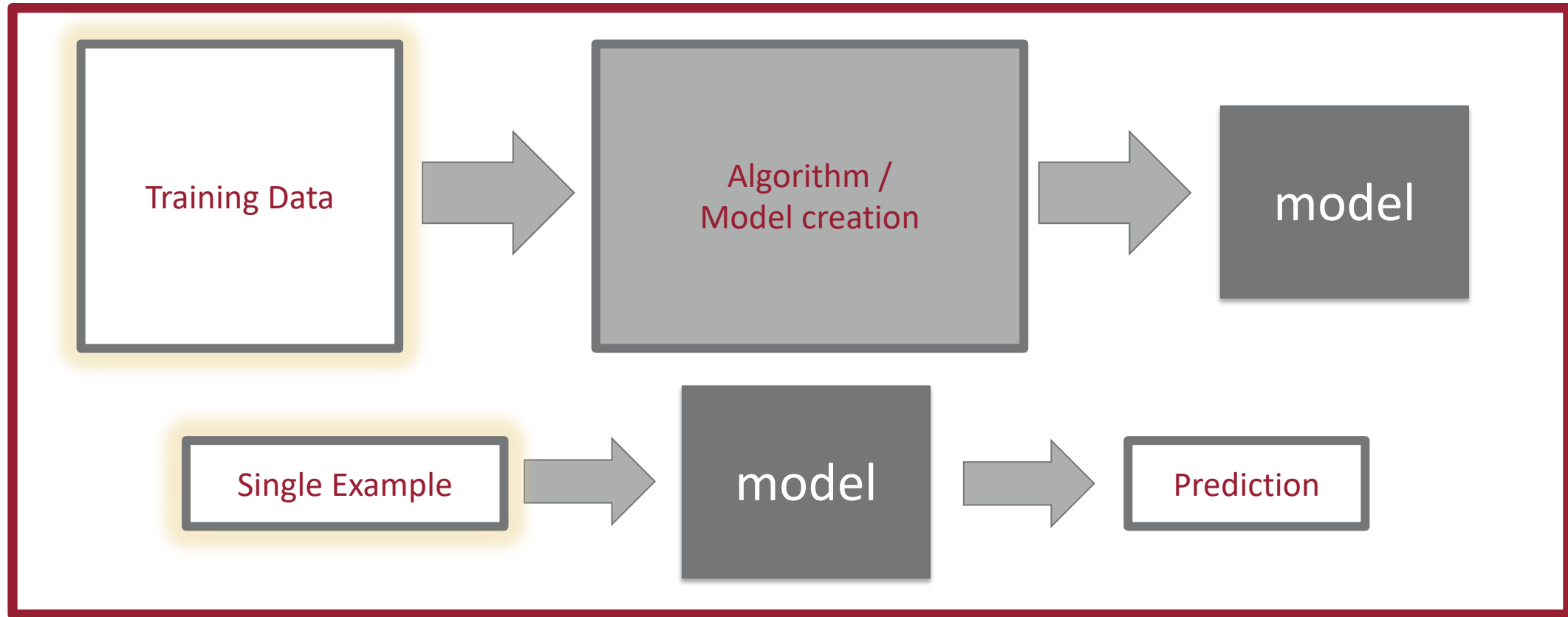
Sources of Error



Assumption 2: the chosen ML algorithm is appropriate to the real world context – does your model match the underlying phenomena and real-world societal understandings?



Sources of Error

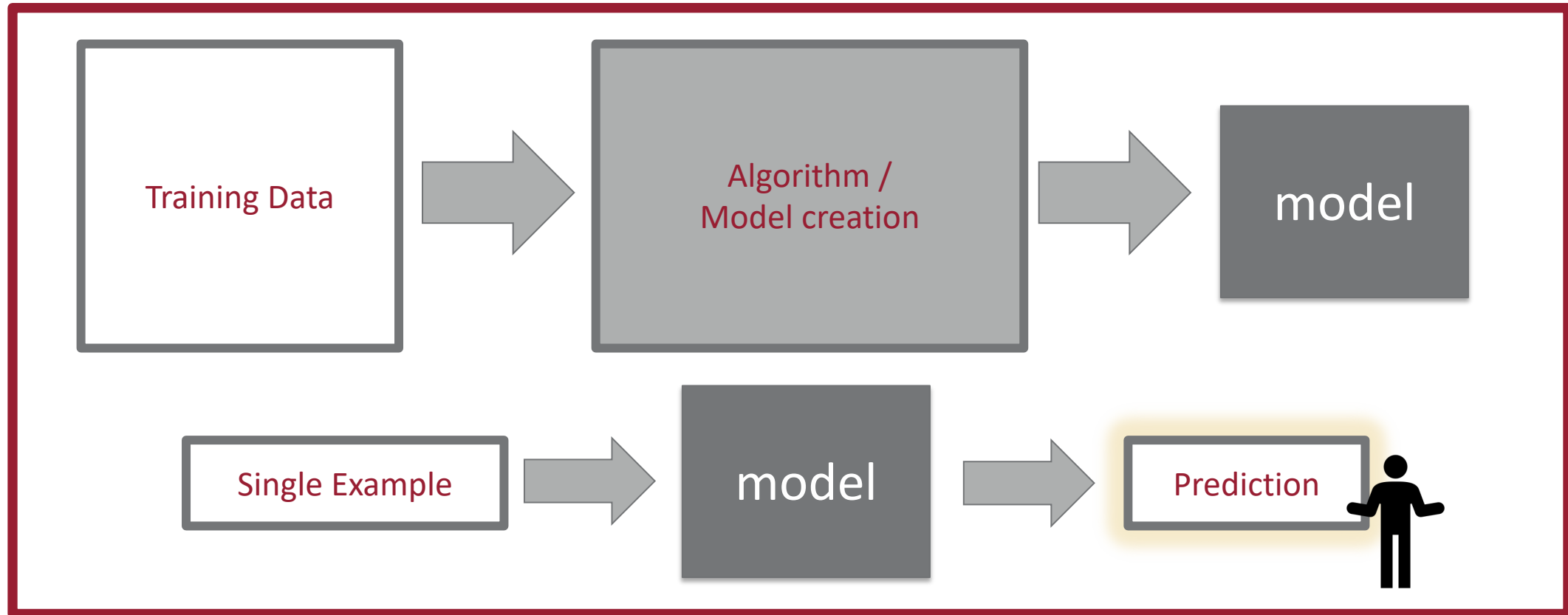


Assumption 3: the developed pipeline and/or model can be applied in a new context

- Assumptions about the training data and/or example distributions may not hold!



Sources of Error



Assumption 4: the resulting prediction will be applied correctly and in the appropriate context – what real-world considerations might you have forgotten?



What can we do about it?

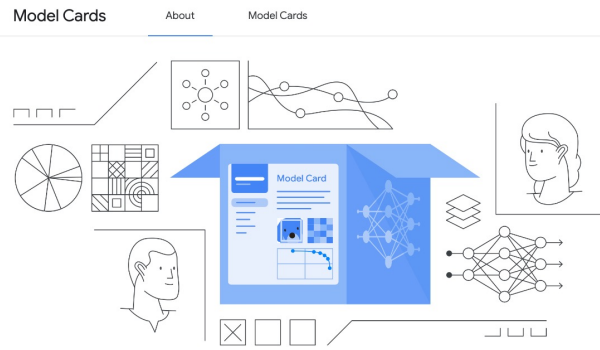
- Key idea: interrogate your assumptions, make them explicit
- One concrete version of this: model cards



Model Cards

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru



The value of a shared understanding of AI models

Whether it's knowing the nutritional content in our food, the conditions of our roads, or a medication's interaction warnings, we rely on information to make responsible decisions. But what about AI? Despite its potential to transform so much of the way we work and live, machine learning models are often distributed without a clear understanding of how they function. For example, under what conditions does the model perform best and most consistently? Does it have blind spots? If so, where? Traditionally, such questions have been surprisingly difficult to answer.



[huggingface_hub / src / huggingface_hub / templates / modelcard_template.md](https://huggingface_hub/src/huggingface_hub/templates/modelcard_template.md)

Uses

Direct Use

```
{{ direct_use | default("[More Information Needed]", true)}}
```

Downstream Use [optional]

```
{{ downstream_use | default("[More Information Needed]", true)}}
```

Out-of-Scope Use

```
{{ out_of_scope_use | default("[More Information Needed]", true)}}
```

Bias, Risks, and Limitations

```
{{ bias_risks_limitations | default("[More Information Needed]", true)}}
```

Recommendations

```
{{ bias_recommendations | default("Users (both direct and downstream) should be made aware of the risks, biases and limitations of the model. More information needed for further recommendations.", true)}}
```

How to Get Started with the Model

Use the code below to get started with the model.

```
{{ get_started_code | default("[More Information Needed]", true)}}
```

Training Details

Training Data

```
{{ training_data | defau
```

Amazon SageMaker

Developer Guide

Amazon SageMaker Model Cards

Use Amazon SageMaker Model Cards to document critical details about your machine learning (ML) models in a single place for streamlined governance and reporting.

Catalog details such as the intended use and risk rating of a model, training details and metrics, evaluation results and observations, and additional call-outs such as considerations, recommendations, and custom information. By creating model cards, you can do the following:

- Provide guidance on how a model should be used.
- Support audit activities with detailed descriptions of model training and performance.
- Communicate how a model is intended to support business goals.

Model cards provide prescriptive guidance on what information to document and include fields for custom information. After creating a model card, you can export it to a PDF or download it to share with relevant stakeholders. Any edits other than an approval status update made to a model card result in additional model card versions in order to have an immutable record of model changes.

Topics

- [Prerequisites](#)
- [Intended uses of a model](#)
- [Risk ratings](#)
- [Model card JSON schema](#)
- [Create a model card](#)
- [Manage model cards](#)
- [Cross-account support for Amazon SageMaker Model Cards](#)
- [Use model cards through the low-level APIs](#)
- [Model card FAQs](#)



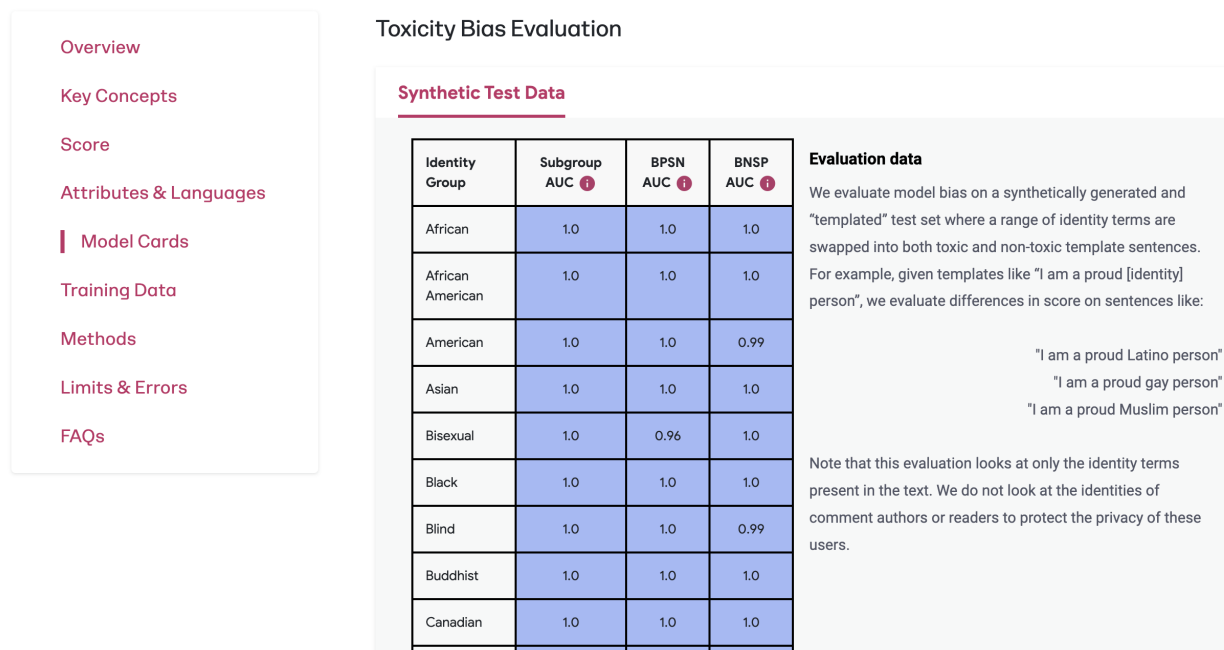
Model Cards

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru

<https://developers.perspectiveapi.com/s/about-the-api-model-cards>

 **Perspective** | Developers [About the API](#) [Docs](#) [Contact Us](#) Language ▾



Overview

Key Concepts

Score

Attributes & Languages

Model Cards

Training Data




Methods

Limits & Errors

FAQs

Toxicity Bias Evaluation

Synthetic Test Data

Identity Group	Subgroup AUC 	BPSN AUC 	BNSP AUC 
African	1.0	1.0	1.0
African American	1.0	1.0	1.0
American	1.0	1.0	0.99
Asian	1.0	1.0	1.0
Bisexual	1.0	0.96	1.0
Black	1.0	1.0	1.0
Blind	1.0	1.0	0.99
Buddhist	1.0	1.0	1.0
Canadian	1.0	1.0	1.0

Evaluation data

We evaluate model bias on a synthetically generated and "templated" test set where a range of identity terms are swapped into both toxic and non-toxic template sentences. For example, given templates like "I am a proud [identity] person", we evaluate differences in score on sentences like:

"I am a proud Latino person"
"I am a proud gay person"
"I am a proud Muslim person"

Note that this evaluation looks at only the identity terms present in the text. We do not look at the identities of comment authors or readers to protect the privacy of these users.

Model Card

- **Model Details.** Basic information about the model.
 - Person or organization developing model
 - Model date
 - Model version
 - Model type
 - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
 - Paper or other resource for more information
 - Citation details
 - License
 - Where to send questions or comments about the model
- **Intended Use.** Use cases that were envisioned during development.
 - Primary intended uses
 - Primary intended users
 - Out-of-scope use cases
- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
 - Relevant factors
 - Evaluation factors
- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
 - Model performance measures
 - Decision thresholds
 - Variation approaches
- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
 - Datasets
 - Motivation
 - Preprocessing
- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
 - Unitary results
 - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.



Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets



Datasheets

» key insights

- There are currently no industry standards for documenting machine learning datasets.
- Datasheets address this gap by documenting the contexts and contents of datasets: from their motivation, composition, collection process, and recommended uses.
- Datasheets for datasets can increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to choose the right dataset.
- Datasheets enable dataset creators to be intentional throughout the dataset creation process.
- Iterating on the design of datasheets with practitioners and legal experts helped improve the questions.
- Datasheets and other forms of data documentation are increasingly commonly released along with datasets.

REPRESENTATIVENESS

12. How representative is this dataset? What population(s), contexts (e.g., scripted vs. conversational speech), conditions (e.g., lighting for images) is it representative of?

How was representativeness ensured or validated?

What are known limits to this dataset's representativeness?

13. What demographic groups (e.g., gender, race, age, etc.) are identified in the dataset, if any?

How were these demographic groups identified (e.g., self-identified, inferred)?

What is the breakdown of the dataset across demographic groups? Consider also reporting intersectional groups (e.g., race x gender) and including proportions, counts, means or other relevant summary statistics.

Note: This information can help a user of this dataset understand what groups are represented in the dataset. This has implications for the performance of models trained on the dataset and on its appropriateness for fairness evaluations – e.g., comparisons of performance across groups.

DATA QUALITY

14. Is there any missing information in the dataset? If yes, please explain what information is missing and why (e.g., some people did not report their gender).

Note: Consider the impact of missing information on appropriate uses of this dataset.

15. What errors, sources of noise, or redundancies are important for dataset users to be aware of?

Note: Consider how errors, noise, redundancies might impact appropriate uses of this dataset.

16. What data might be out of date or no longer available (e.g., broken links in old tweets)?

17. How was the data validated/verified?

18. What are potential validity issues a user of this dataset needs to be aware of (e.g., survey answers might not be truthful, age was guessed by a model and might be incorrect, GPA was used to quantify intelligence)?

19. What are other potential data quality issues a user of this dataset needs to be aware of?



Research



System Cards

AI SYSTEM CARDS

Empowering people to learn more
about the technology powering
Facebook and Instagram

 Meta AI

There are four sections in every AI system card:

- An overview of the AI system
- A section explaining how the AI system works, which includes a summary of the steps involved in creating experiences on Facebook and Instagram
- A section describing how to customize the content that is shown. This includes descriptions of system controls and instructions for how each person can control and customize their experience.
- A section explaining how the AI delivers content, which includes an explanation of how some of the significant prediction models inform the overall AI system and produce product experiences

Research

DALL·E 3 system card

Abstract

DALL·E 3 is an artificial intelligence system that takes a text prompt as an input and generates a new image as an output. DALL·E 3 builds on DALL·E 2 by improving caption fidelity and image quality. In this system card, we share the work done to prepare DALL·E 3 for deployment, including our work on external expert red teaming, evaluations of key risks, and mitigations to reduce the risks posed by the model and reduce unwanted behaviors.

GPT-4 System Card

OpenAI

March 23, 2023

Abstract

Large language models (LLMs) are being deployed in many domains of our lives ranging from browsing, to voice assistants, to coding assistance tools, and have potential for vast societal impacts.[1, 2, 3, 4, 5, 6, 7] This system card analyzes GPT-4, the latest LLM in the GPT family of models.[8, 9, 10] First, we highlight safety challenges presented by the model's limitations (e.g., producing convincing text that is subtly false) and capabilities (e.g., increased adeptness at providing illicit advice, performance in dual-use capabilities, and risky emergent behaviors). Second, we give a high-level overview of the safety processes OpenAI adopted to prepare GPT-4 for deployment. This spans our work across measurements, model-level changes, product- and system-level interventions (such as monitoring and policies), and external expert engagement. Finally, we demonstrate that while our mitigations and processes alter GPT-4's behavior and prevent certain kinds of misuses, they are limited and remain brittle in some cases. This points to the need for anticipatory planning and governance.[11]



Model Cards: discussion

Non-men

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	542	170
Test Label WAS hired	23	56

Men

	Predicted don't hire	Predicted DO hire
Test Label wasn't hired	1598	430
Test Label WAS hired	340	190

Consider these same confusion matrices for a resume screening model:

- What do you wish you knew about the data used to train the model?
- If this model were going to be deployed, what information would you want to make sure any hiring manager had to help them make a final decision?

