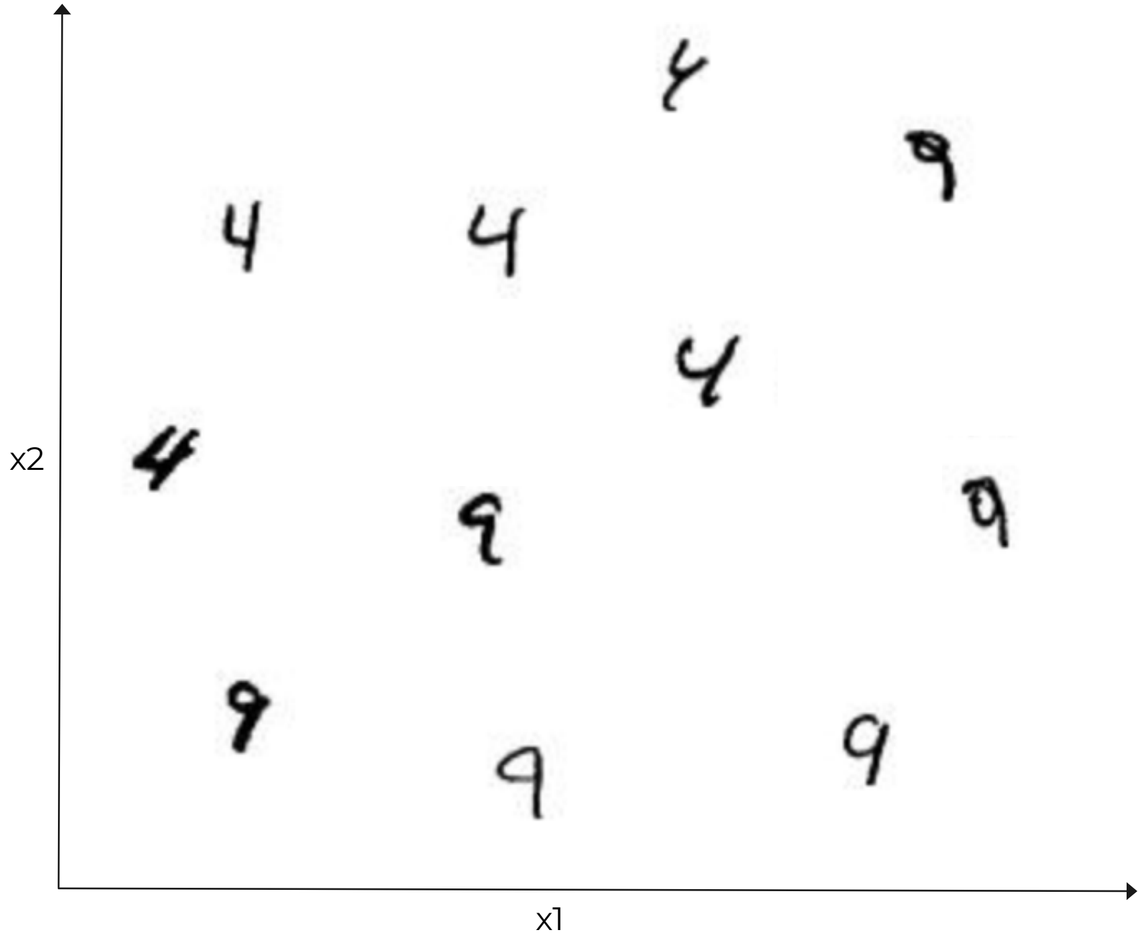
Stress Testing Models to Understand their Decisions

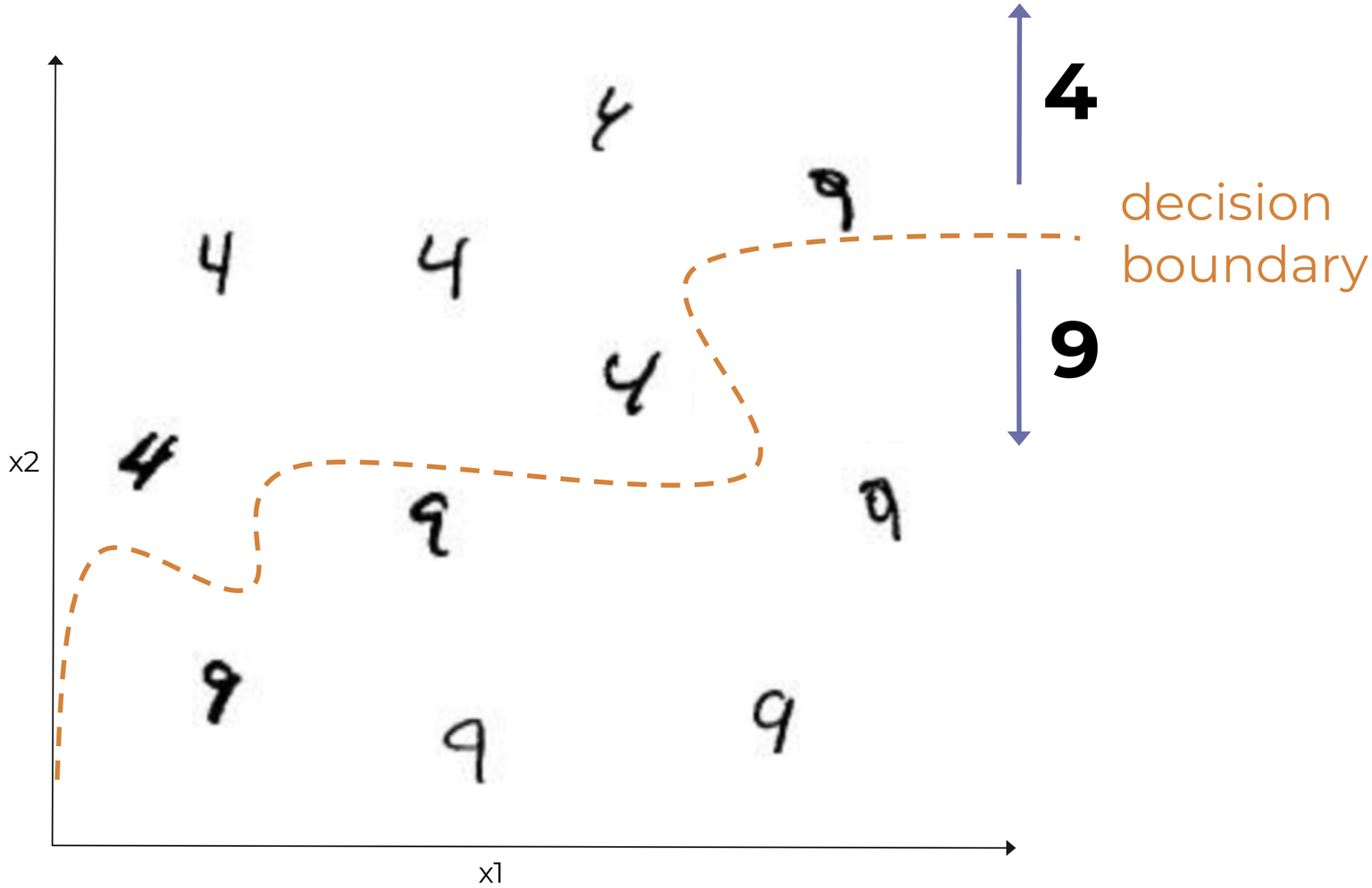
Inês Gomes

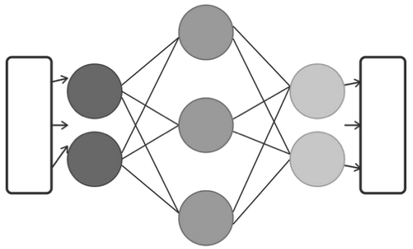
Data Makers Fest, 5 May 2026

Example: distinguish handwritten digits



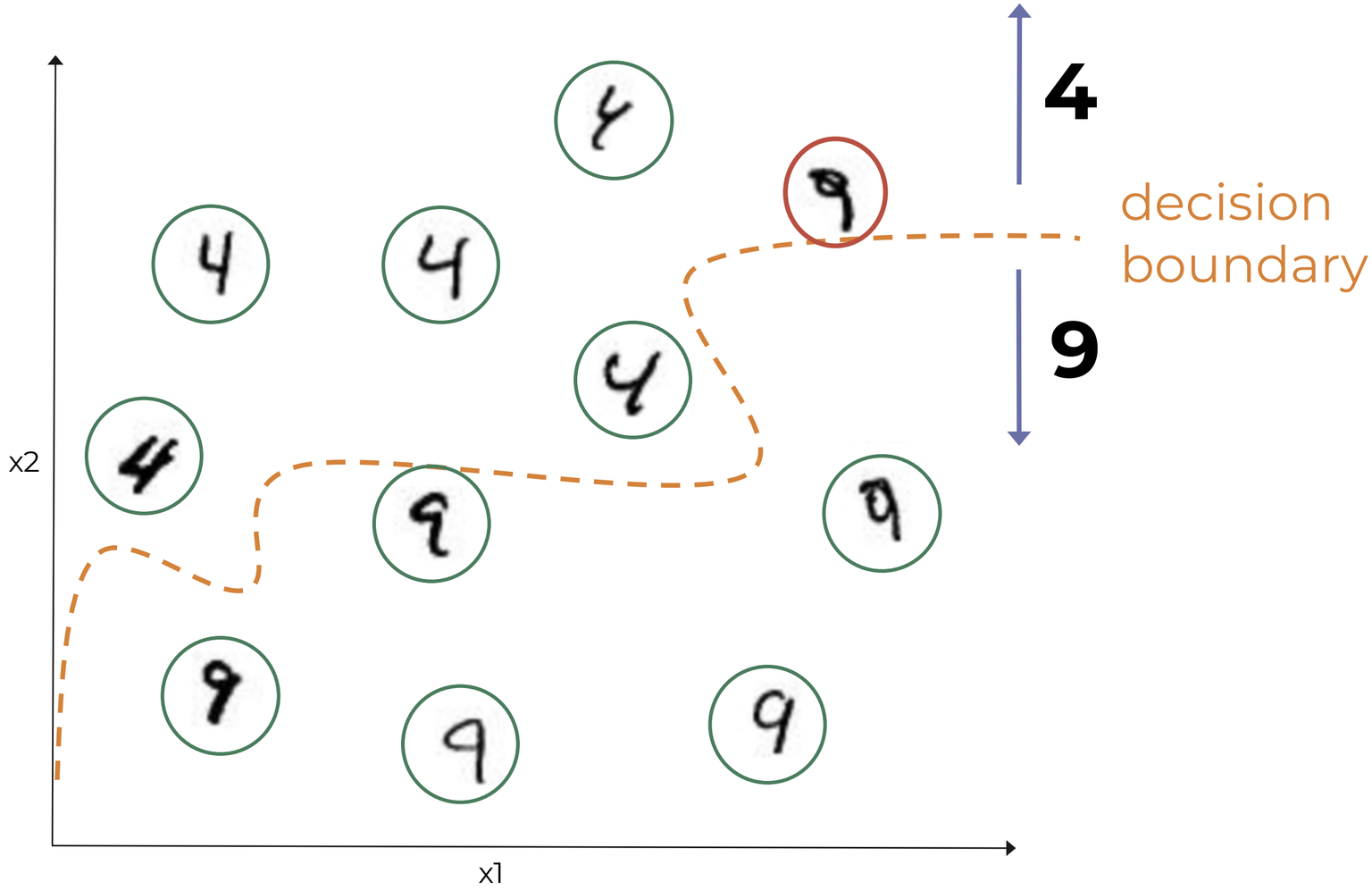
Example: distinguish handwritten digits

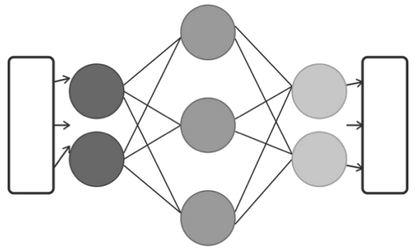




CNN

Example: distinguish handwritten digits





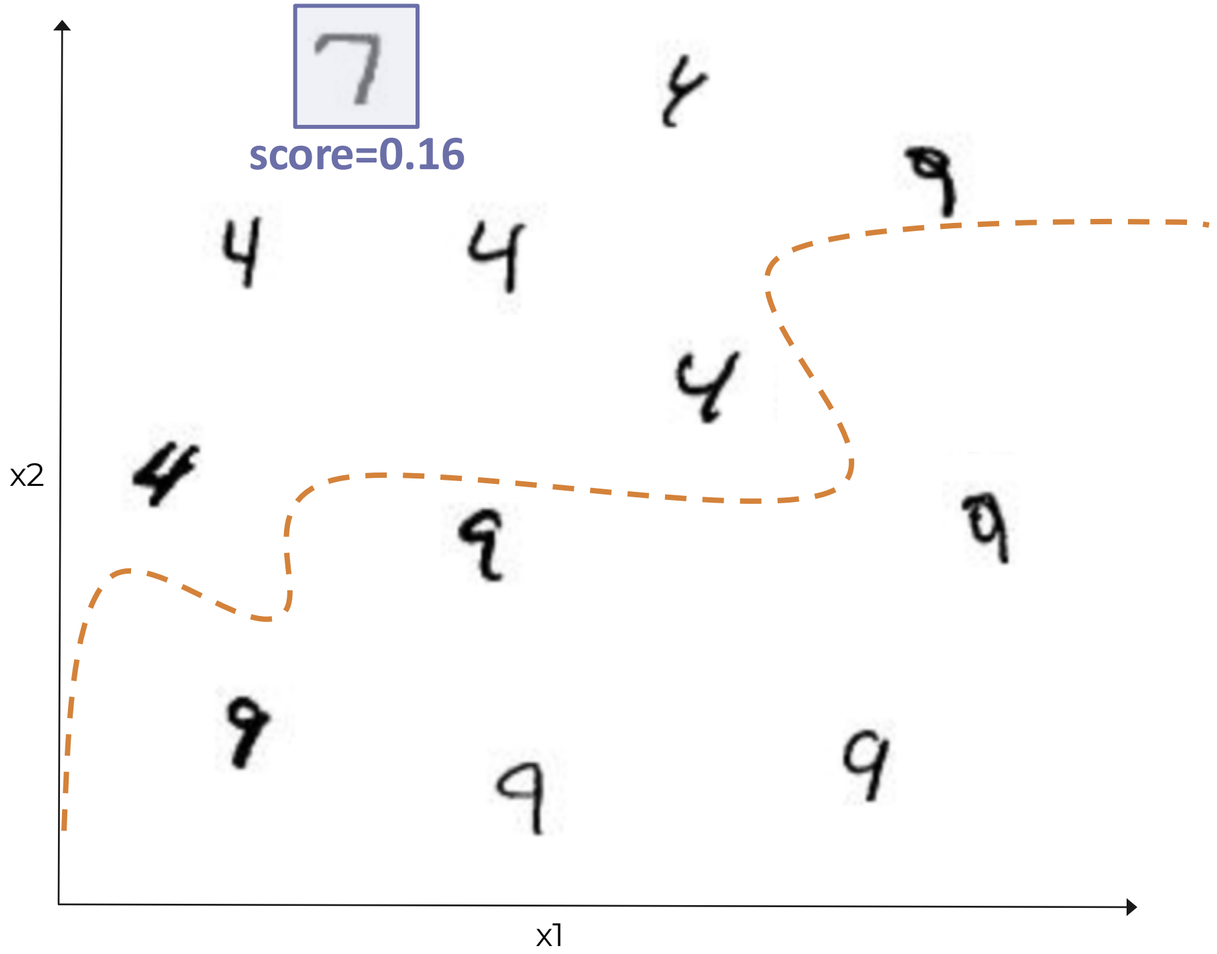
CNN

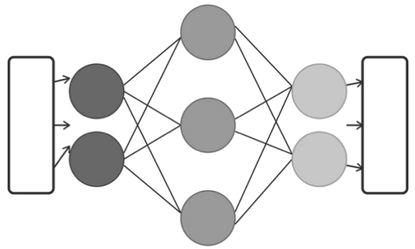
accuracy = 98%

BUT...

does 98% accuracy tell the whole story?

Out-of-distribution: visually unlike any training sample

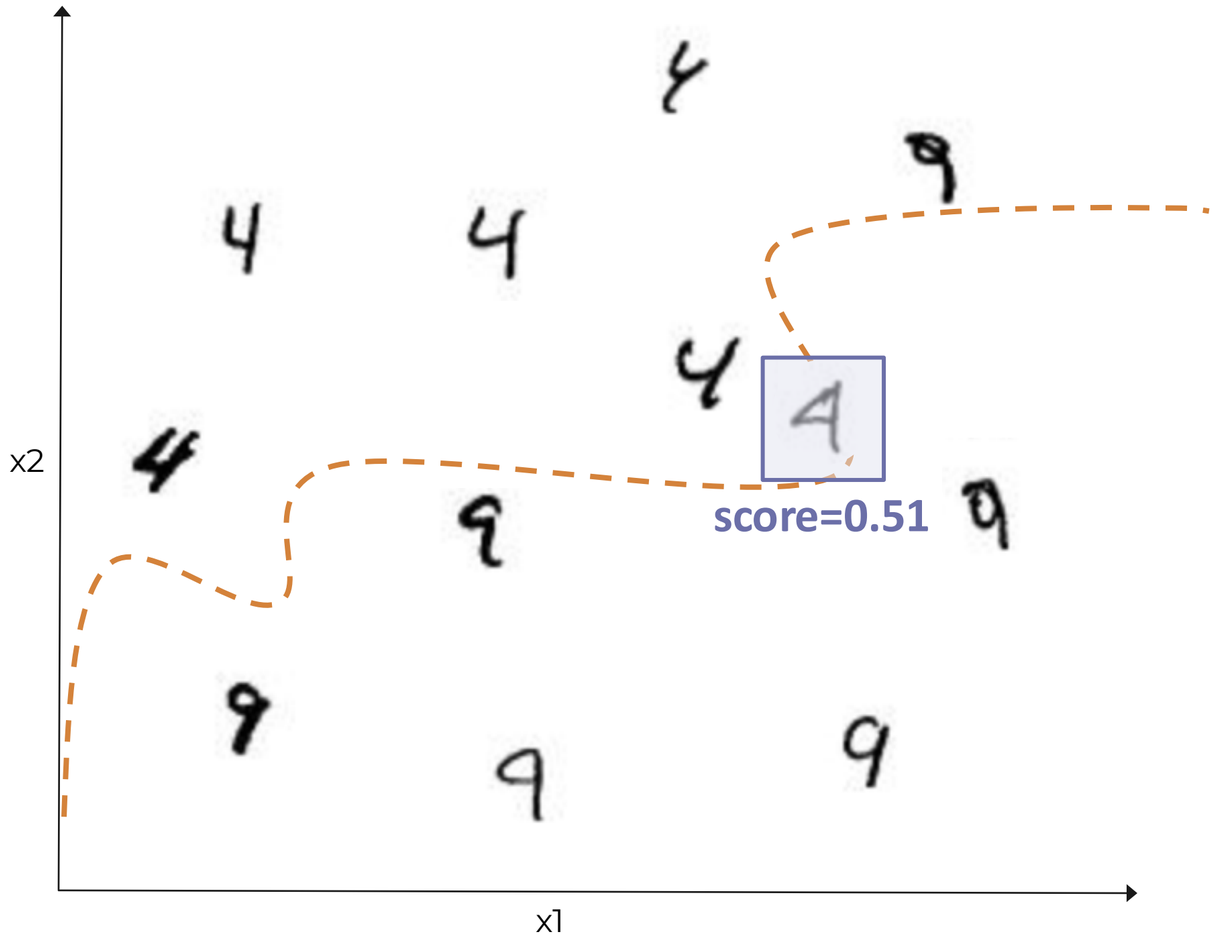


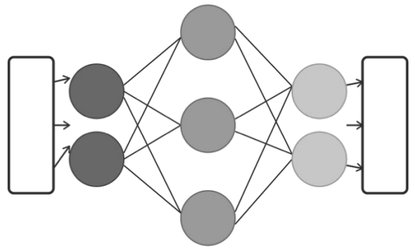


CNN

accuracy = 98%

Visual ambiguity: image sits between two or more classes

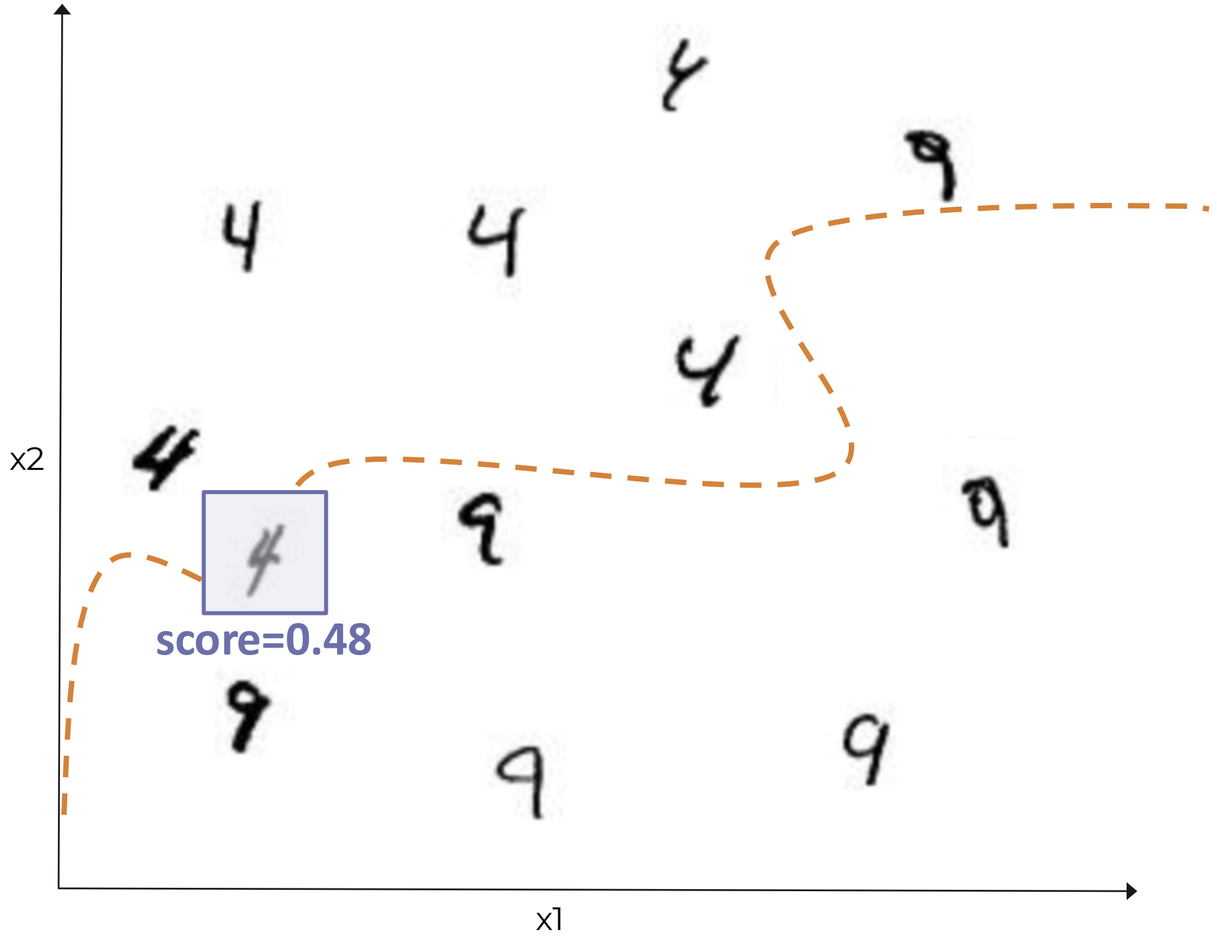


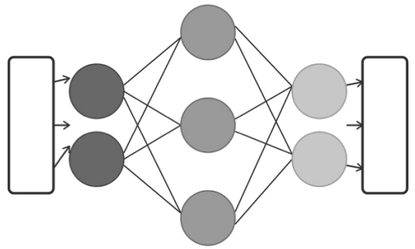


CNN

accuracy = 98%

Underrepresented subgroup: rare style that leads to not confident predictions

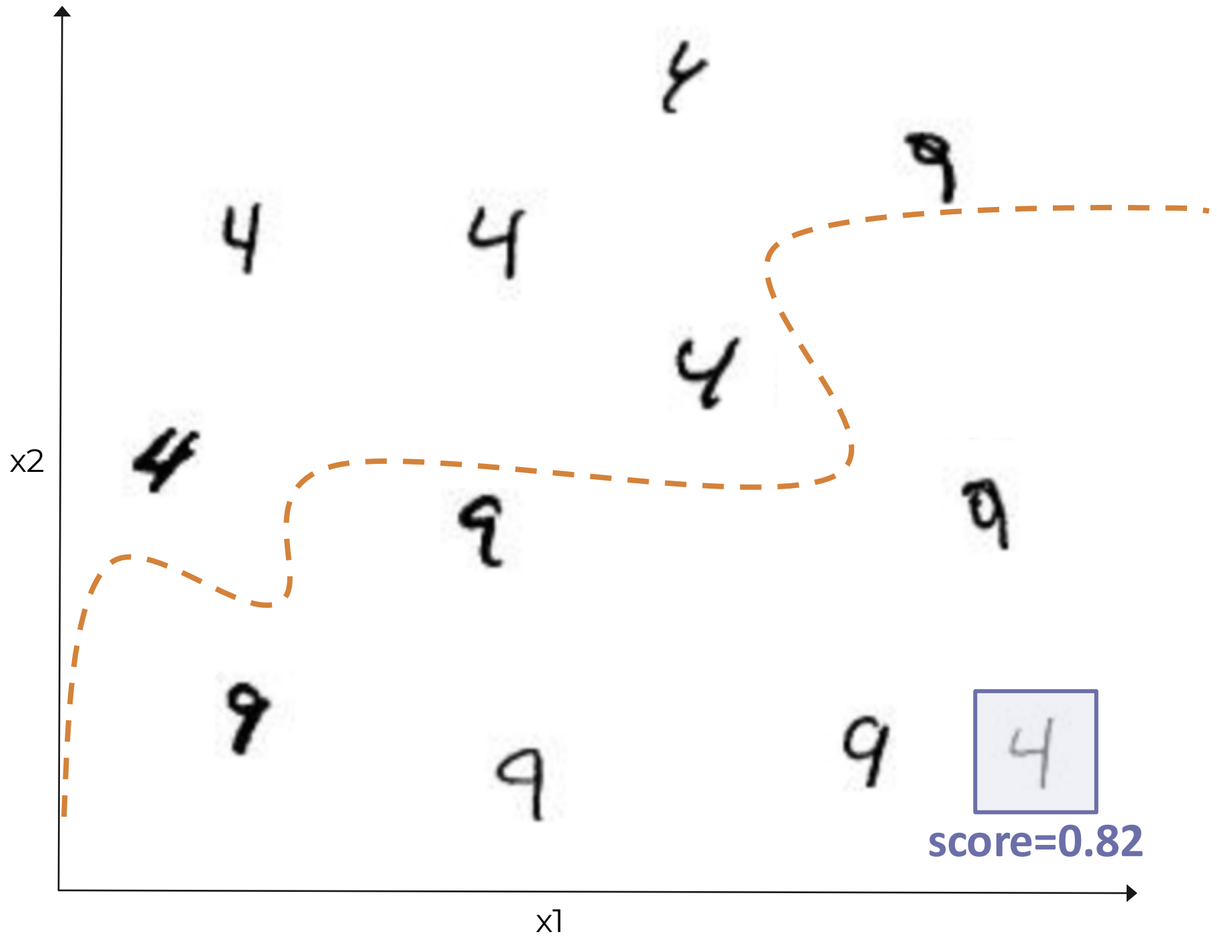


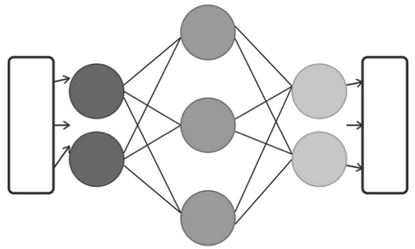


CNN

accuracy = 98%

Adversarial perturbation: imperceptible noise that flips the prediction

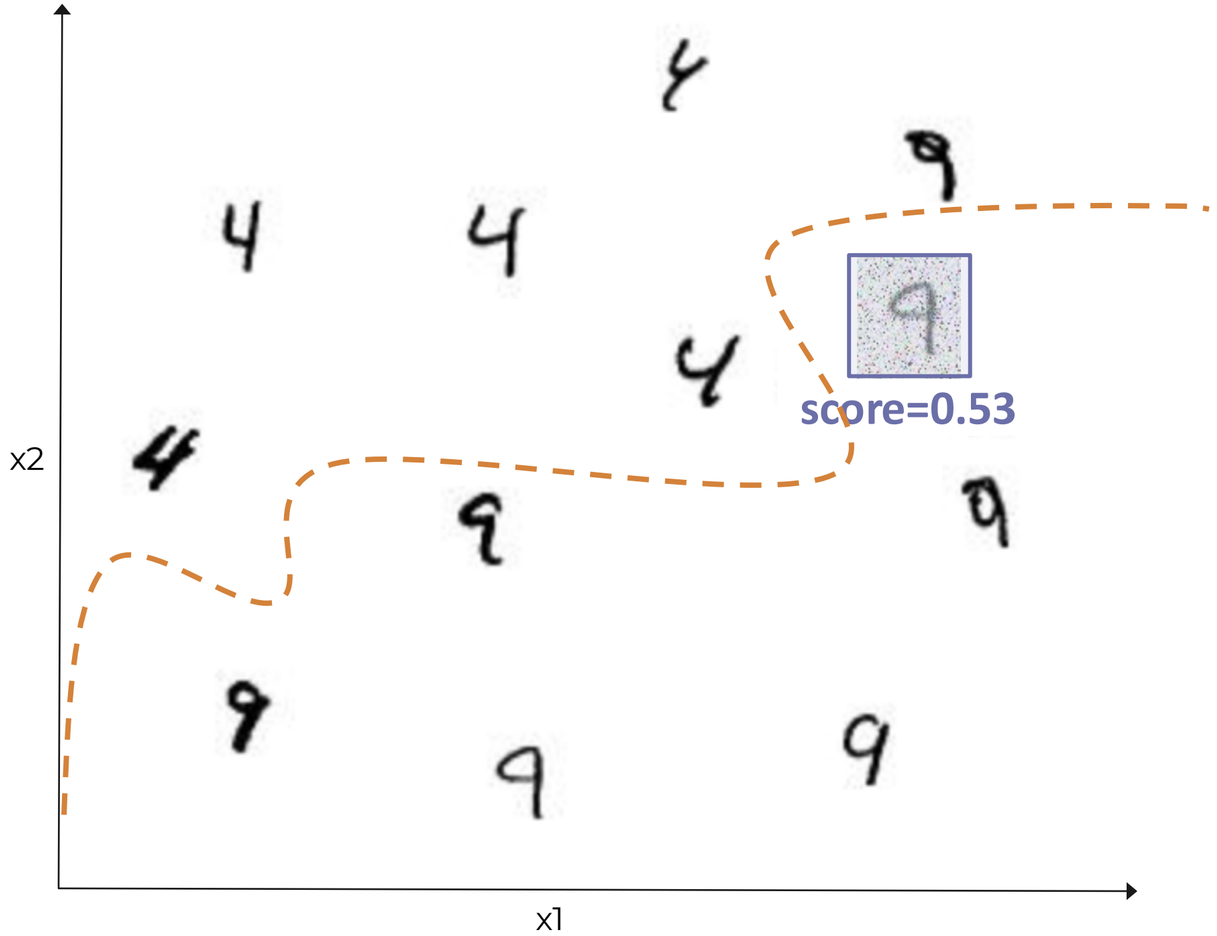


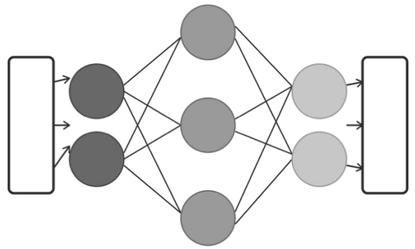


CNN

accuracy = 98%

Corruption & noise: blur or artifacts that degrade input quality





CNN

accuracy = 98%

high performance
on test set

≠

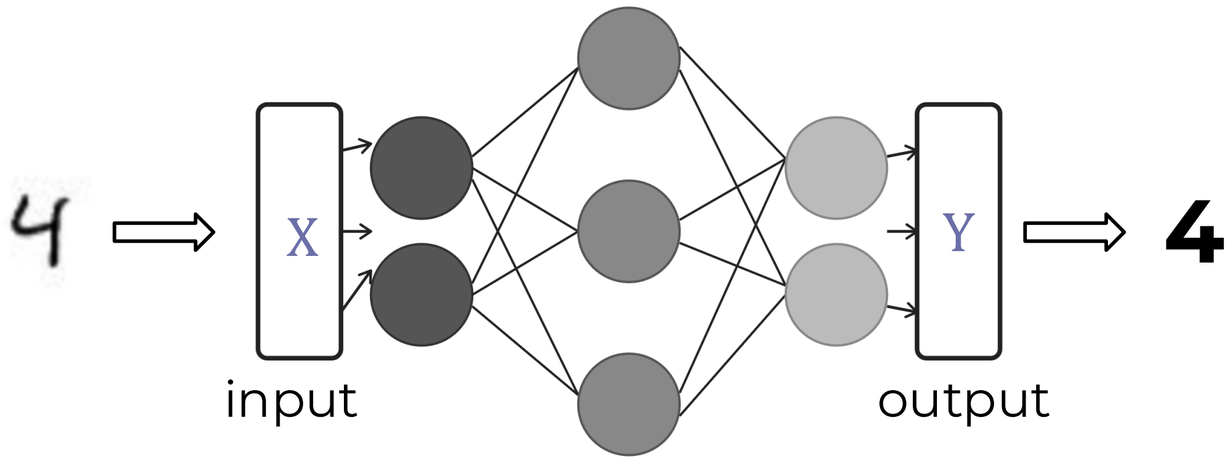
trustworthy
behavior

... because limitations may go undetected!

Why do models fail?

How we formalize it...

A model (f) is a mapping between input and output



$$f: X \rightarrow Y$$

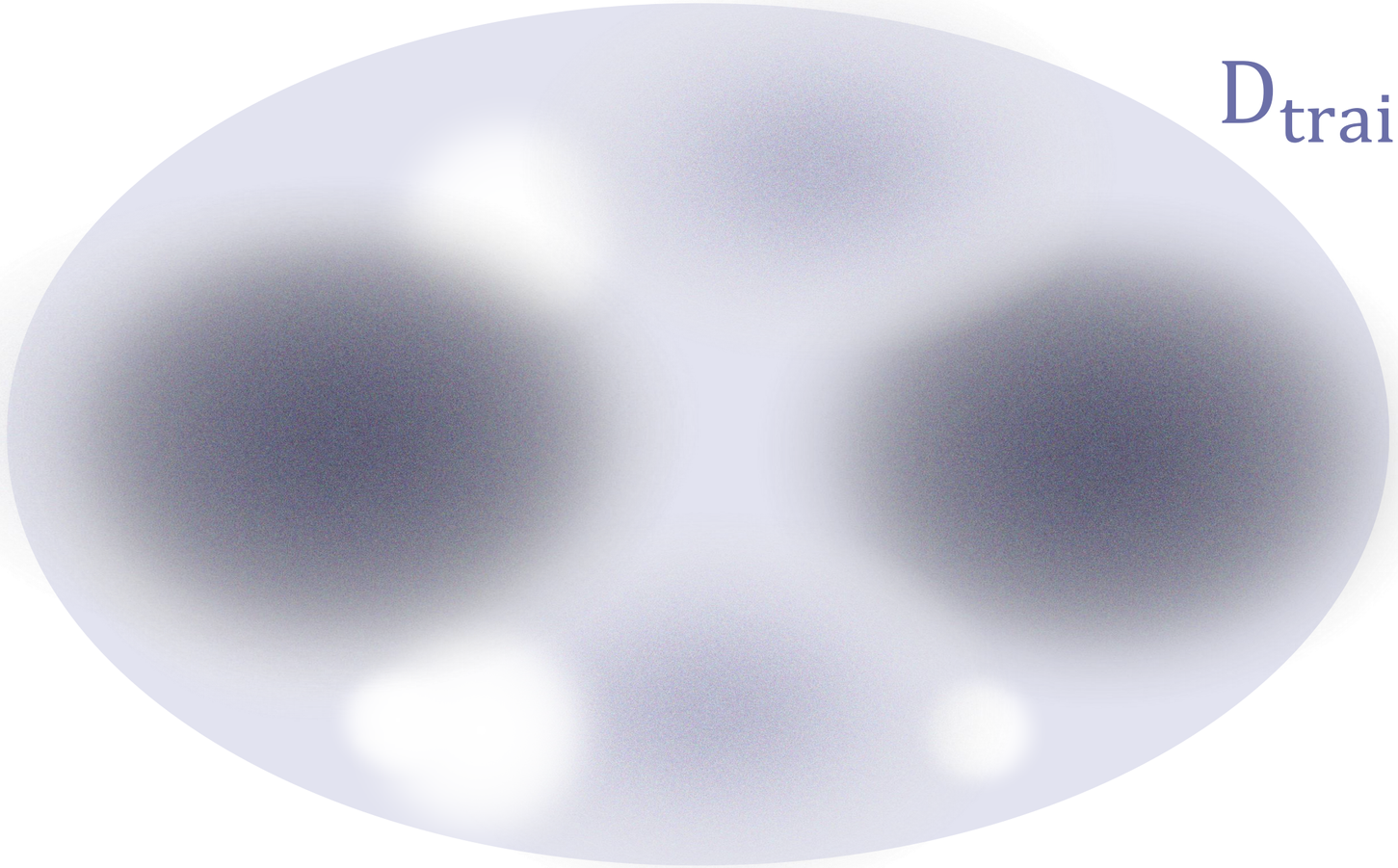
$$(x, y) \sim D_{train}$$

In a perfect world...

An ideal model f^* would map perfectly the inputs and outputs from the perfect distribution D^*

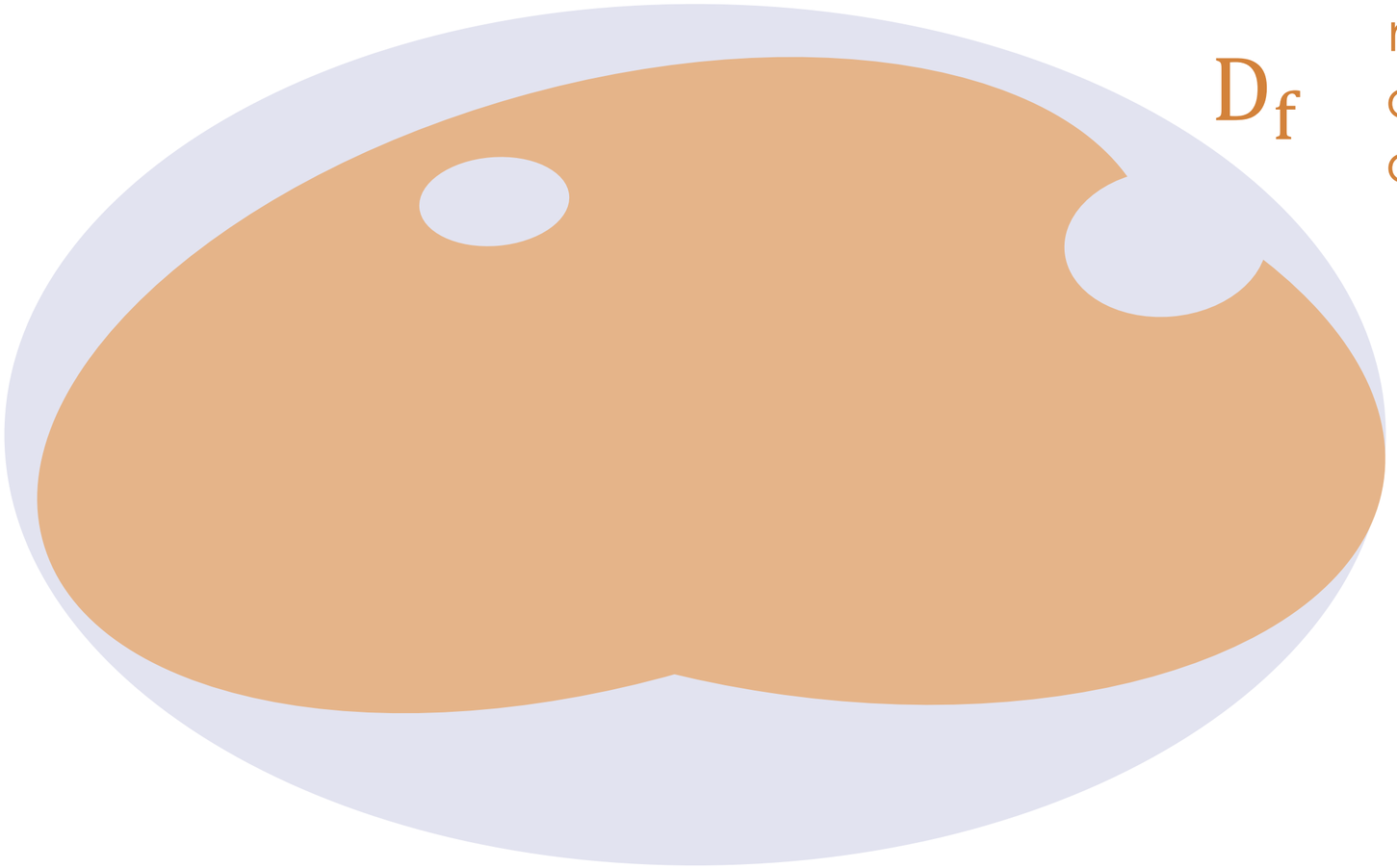
Real world: training/test distributions

Finite samples drawn from the perfect distribution D^*



$D_{train}, D_{test} \sim D^*$

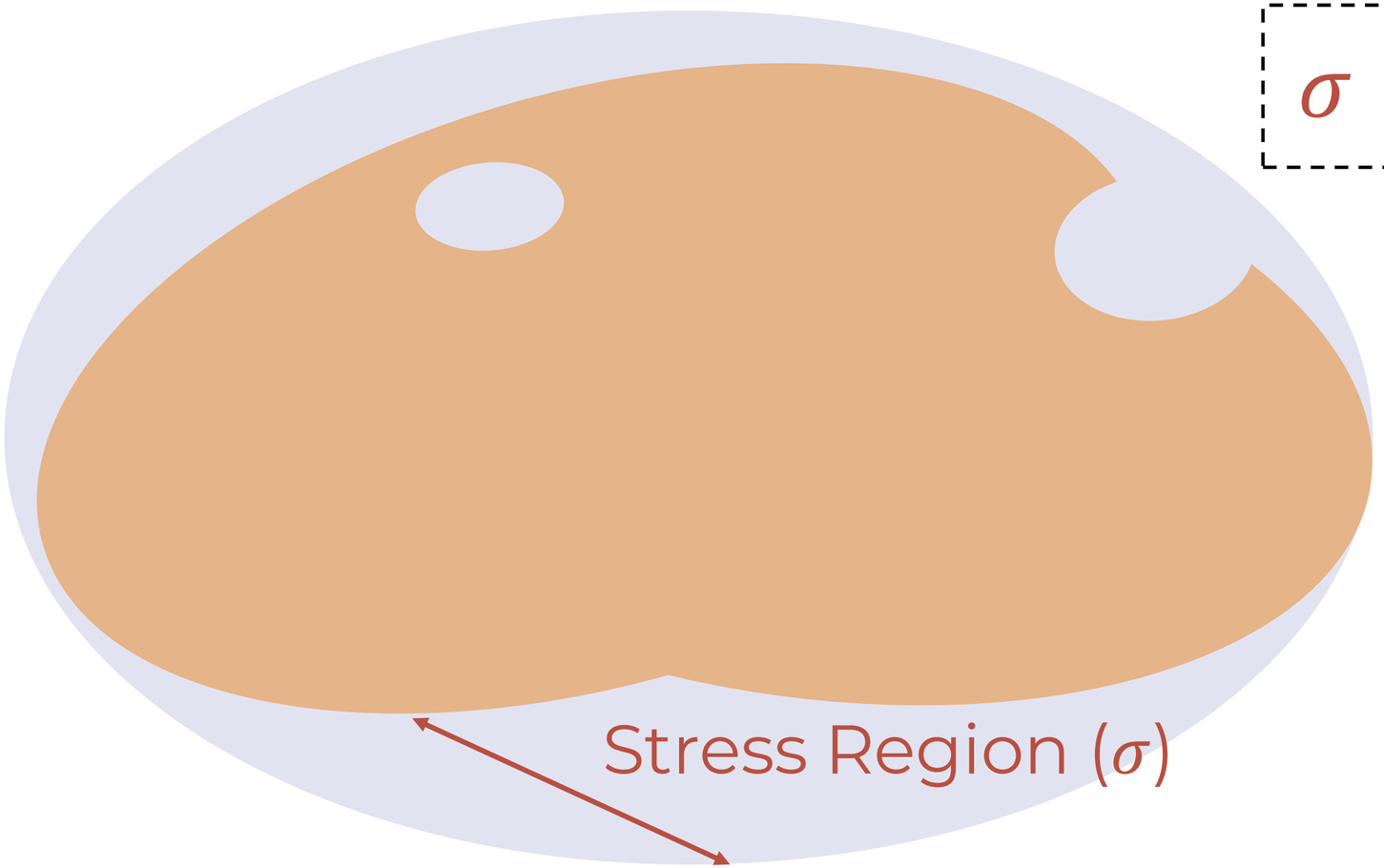
What the algorithm learned correctly from training data



D_f region of competence of model f

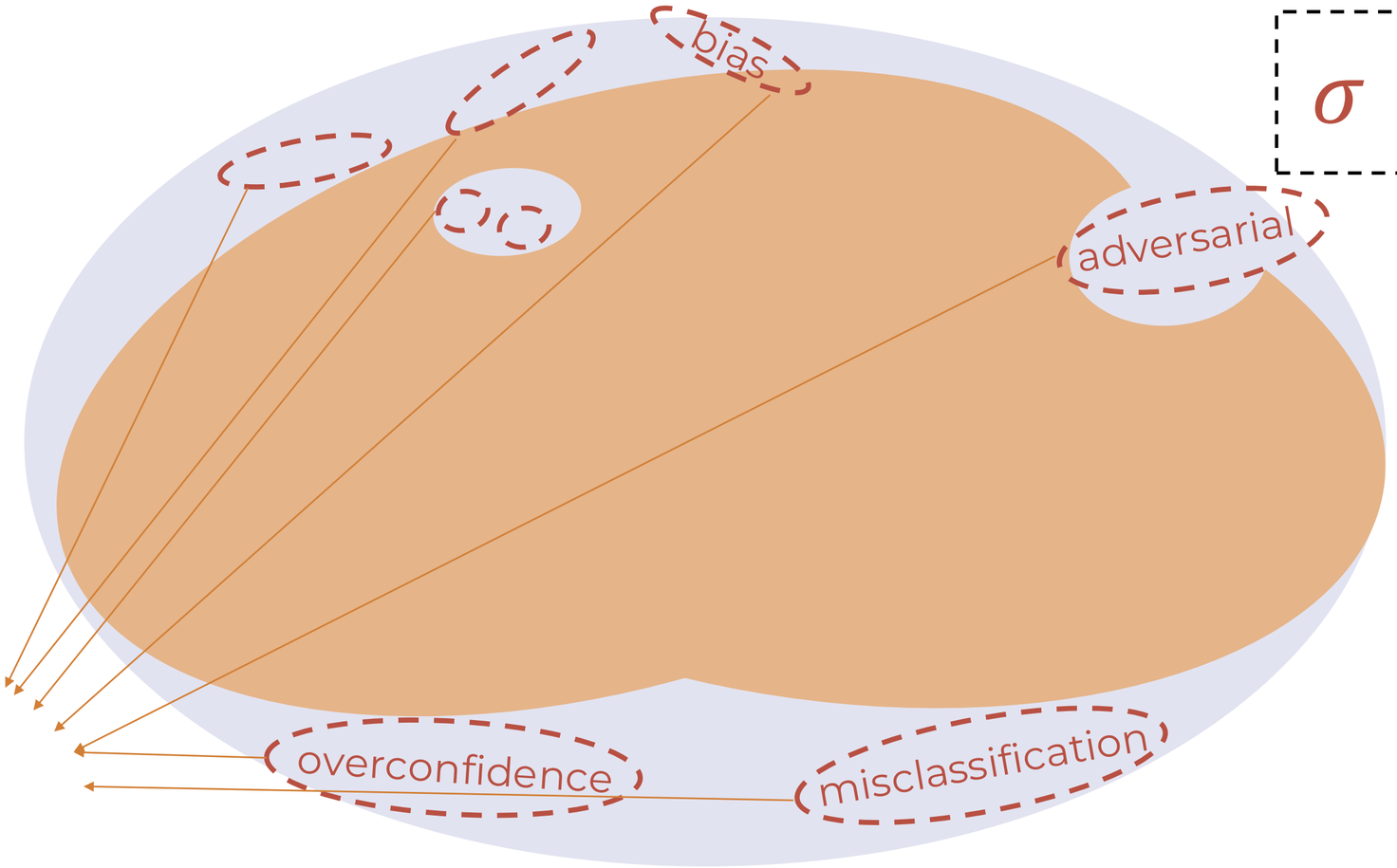
Stress region: where limitations emerge

$$\sigma = D^* - D_f$$



Stress region: where limitations emerge

Set of distributions under which the model does not behave as expected



$$\sigma = D^* - D_f$$

$$D_{f,stress} \sim \sigma$$

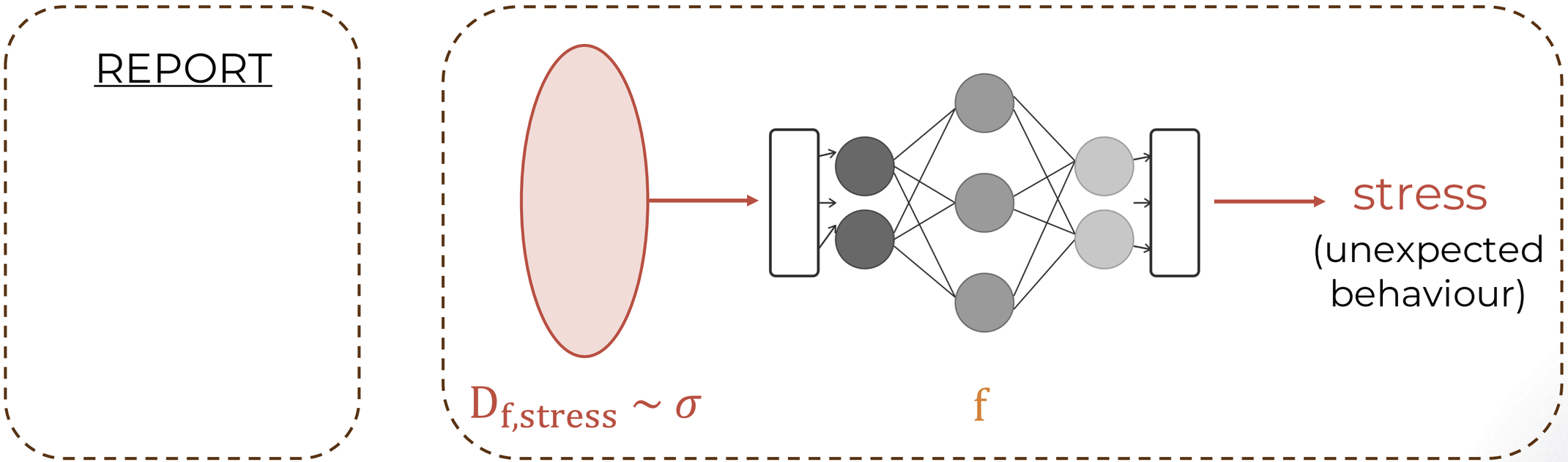
How can we expose model limitations

in the stress region?

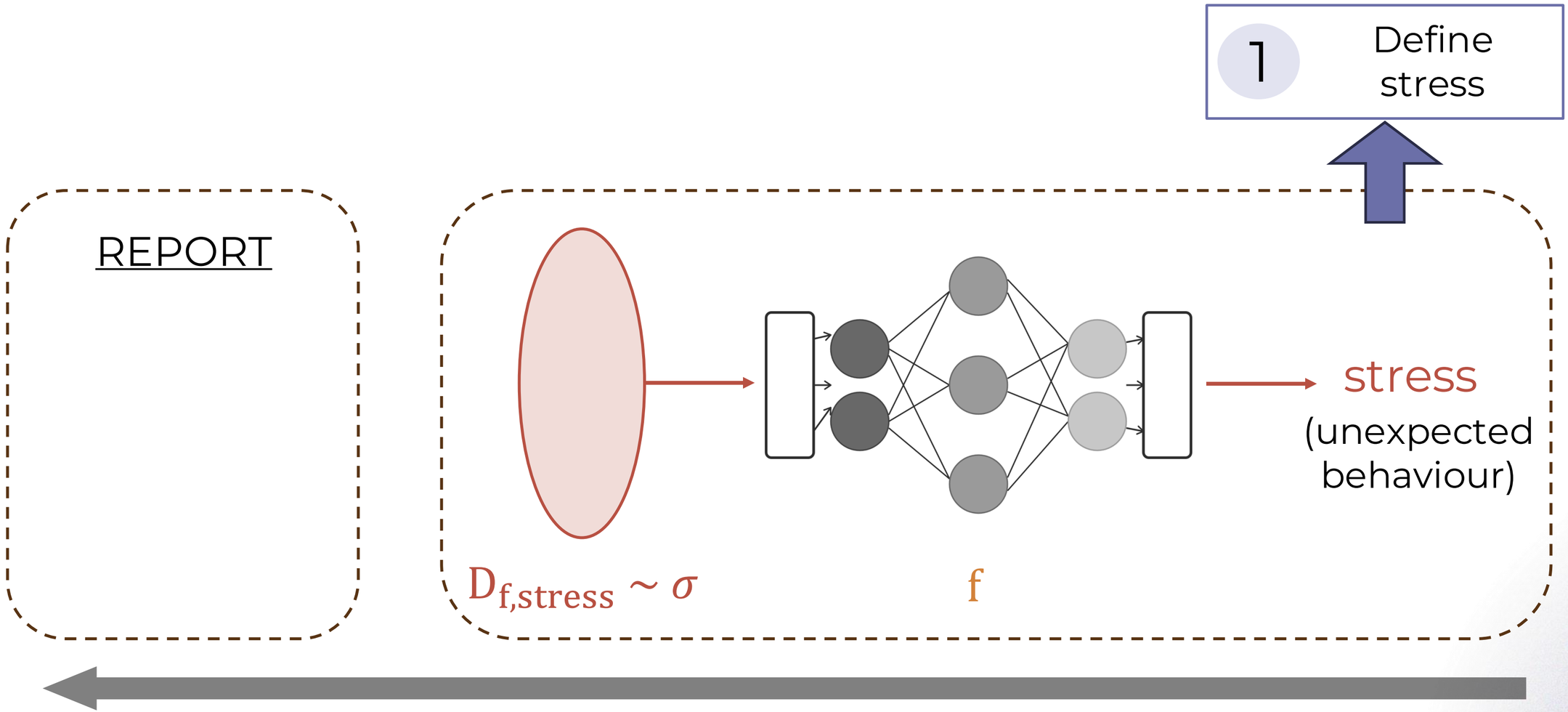
Stress Testing: a framework

Stress Testing: map the limits of the model competence

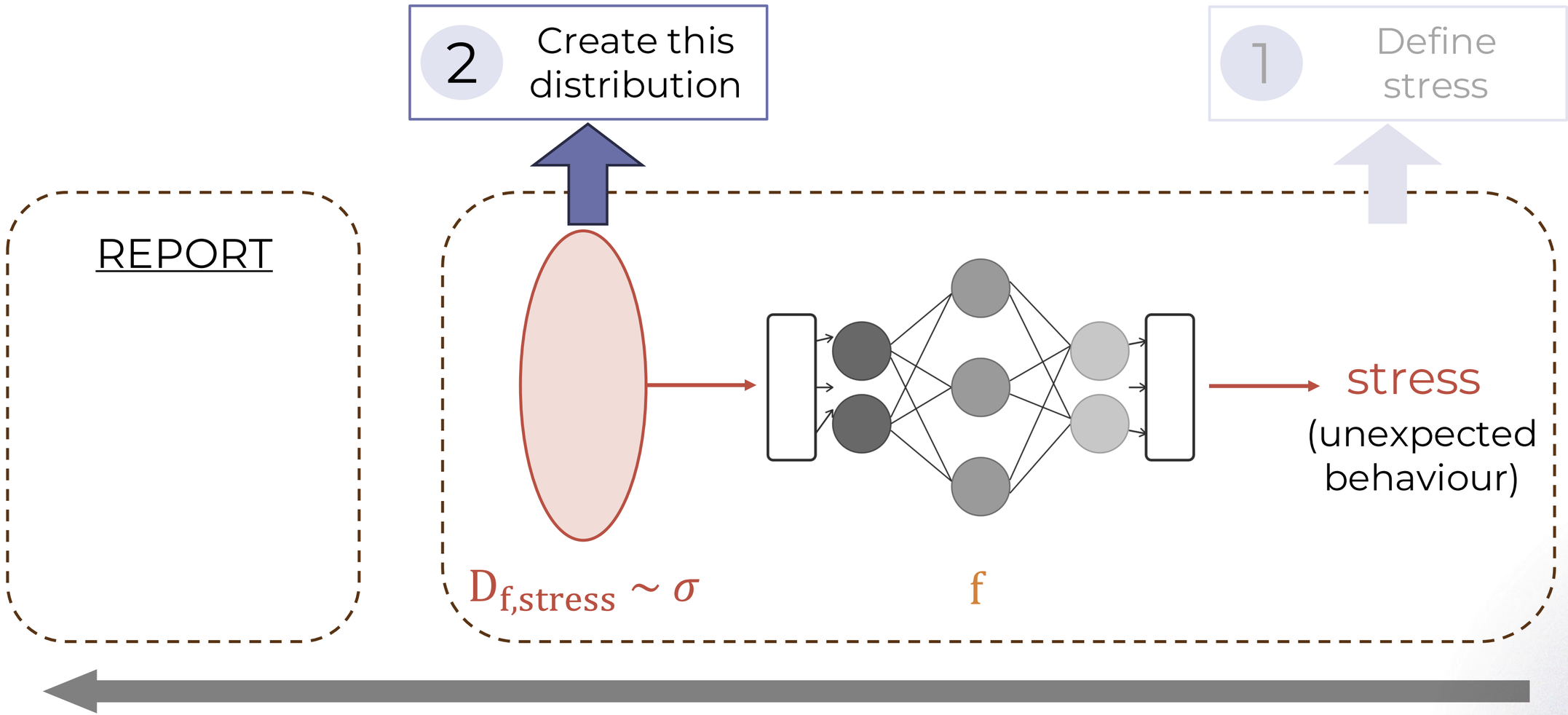
Data-driven way to report model limitations



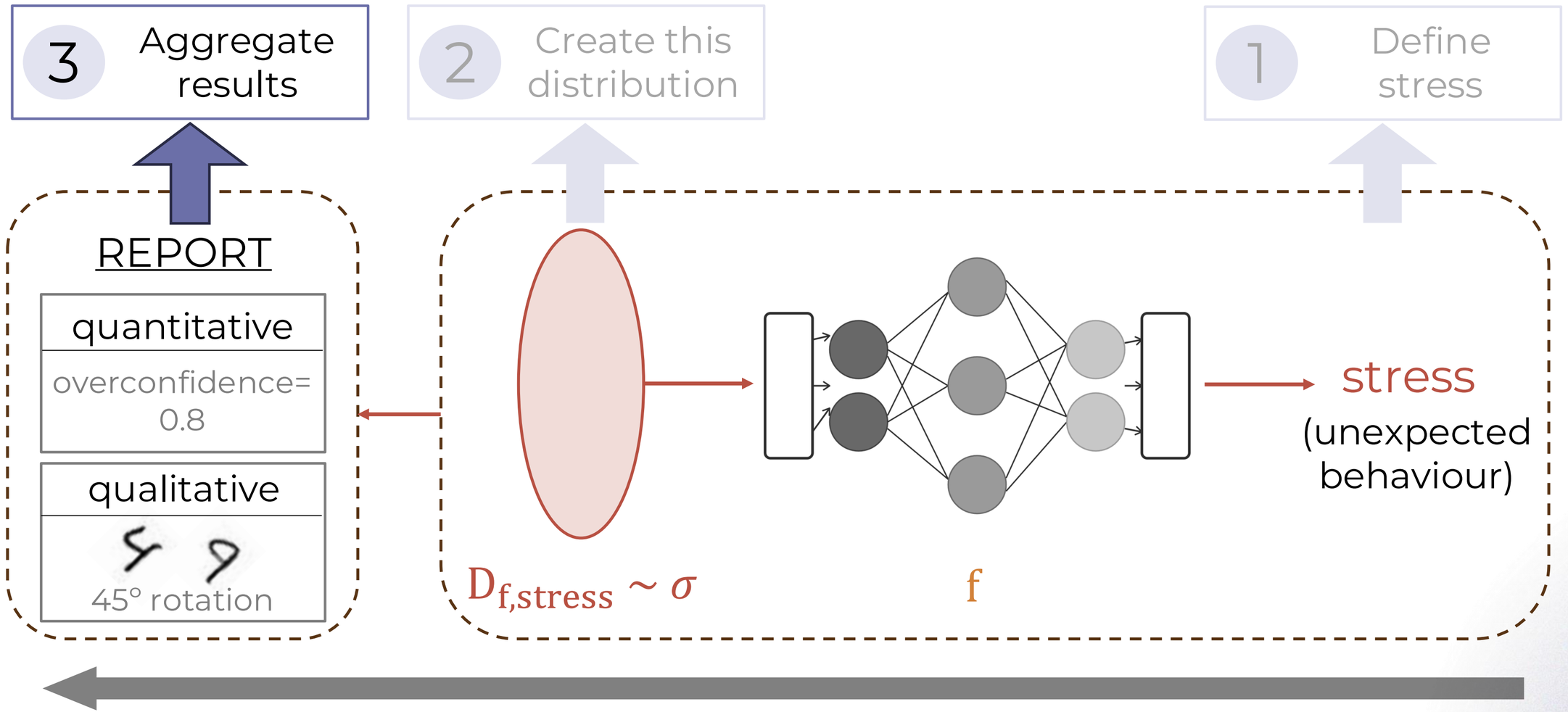
Stress Testing



Stress Testing

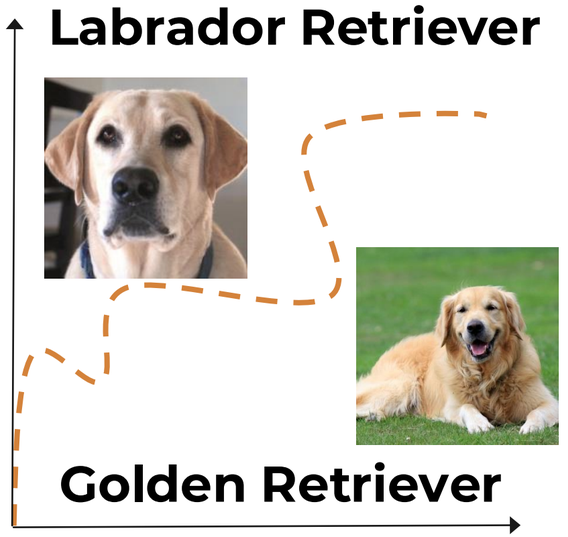


Stress Testing: understand model decisions

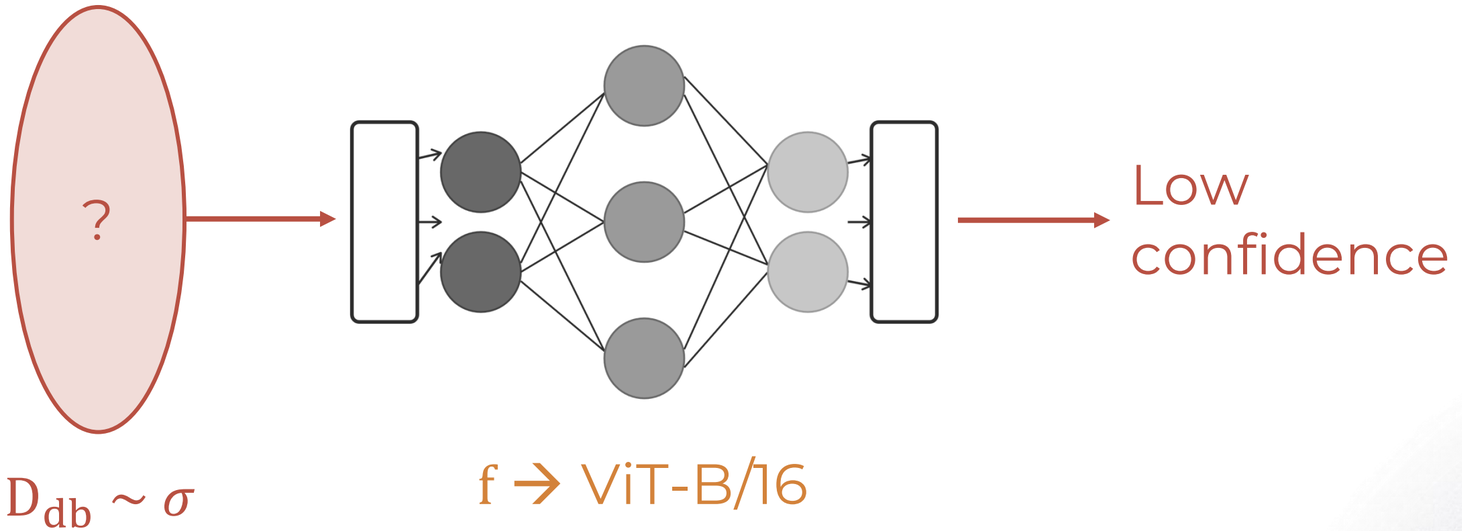


Application:
Stress Testing Decision Boundaries of Image Classifiers

What happens at decision boundaries?



REPORT



$D_{db} \sim \sigma$

f → ViT-B/16

D_{train}: ImageNet D^*: visual world

Stress Testing Decision Boundaries

KLDB: measures if an instance is in a decision boundary of N classes

REPORT

$D_{db} \sim \sigma$

f

Low confidence

KLDB( , )~0

Stress Testing Decision Boundaries

Synthetic Data Generation
Stable Diffusion + classifier guidance

REPORT

$D_{db} \sim \sigma$

f

Low confidence

KLDB(,)~0

Stress Testing: understand model decisions

clustering + xAI
(WIP)

REPORT

puppies

Synthetic Data Generation

$D_{db} \sim \sigma$

f

Low confidence

KLDB(,)~0

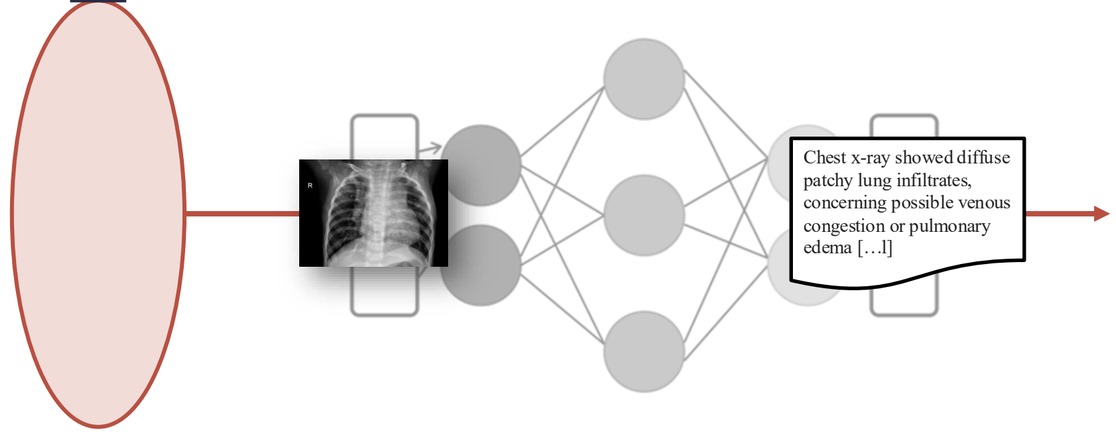
Other applications: medical report generation

synthetic data generation + image perturbations

REPORT

entropy=0.4

“patients with pacemakers cause a drop in the identification of lung masses of 30%”



stress: Incorrect or low-quality text

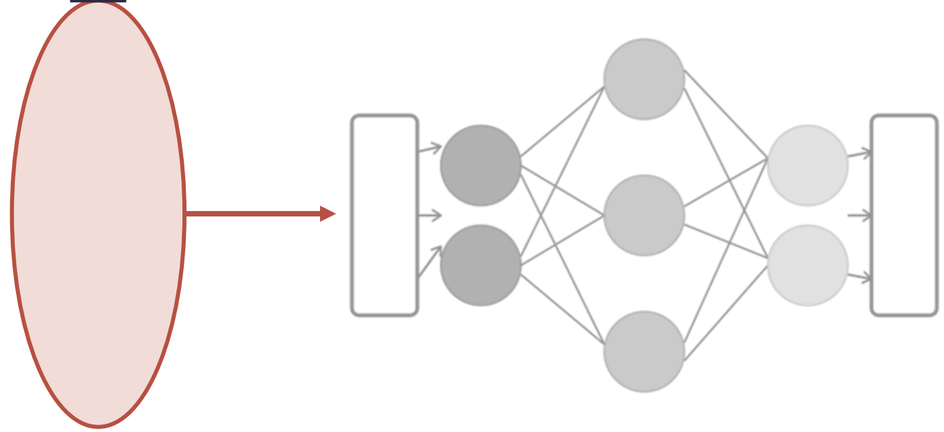
Other applications: time series forecasting

metaclassifier for stress prediction

data augmentation (meta features, time series)

REPORT

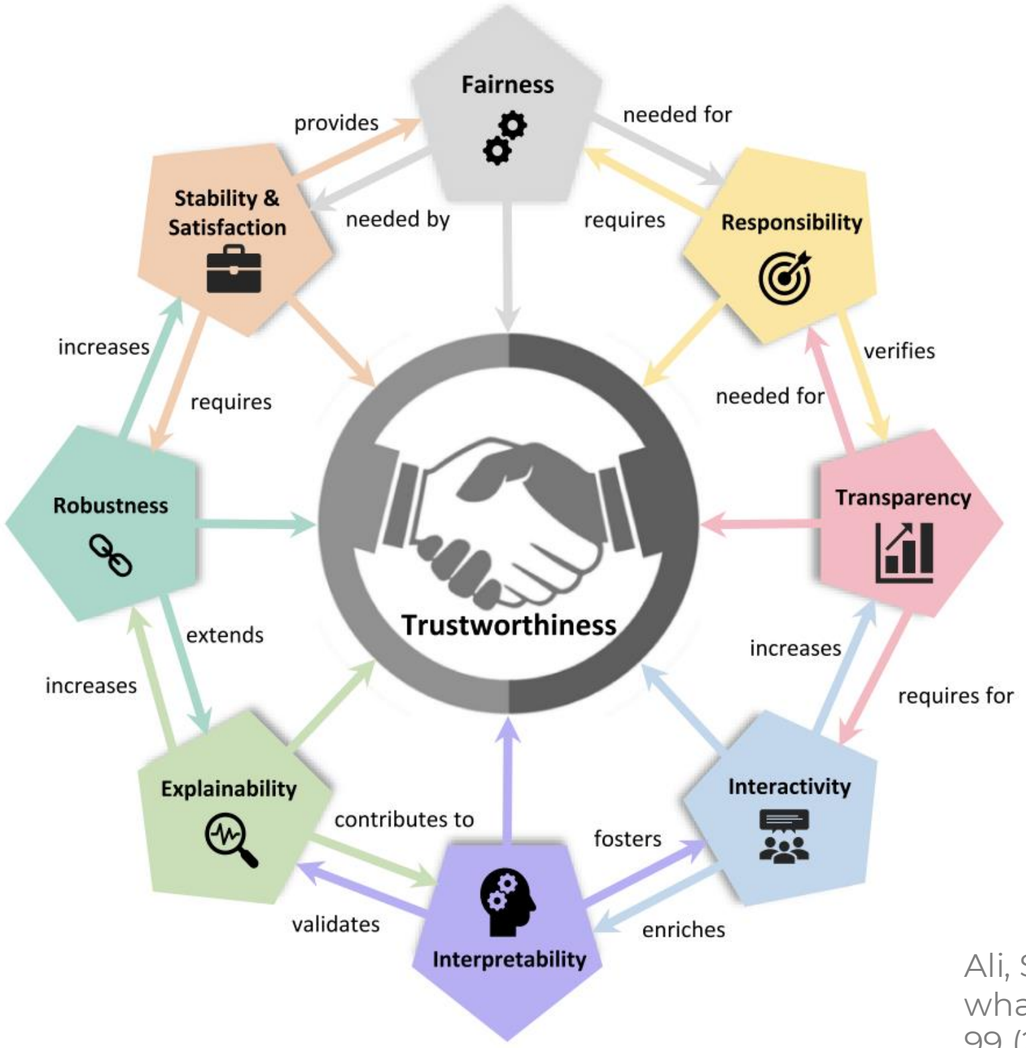
stress=70%



stress:
error and/or uncertainty

The bigger picture: positioning stress testing

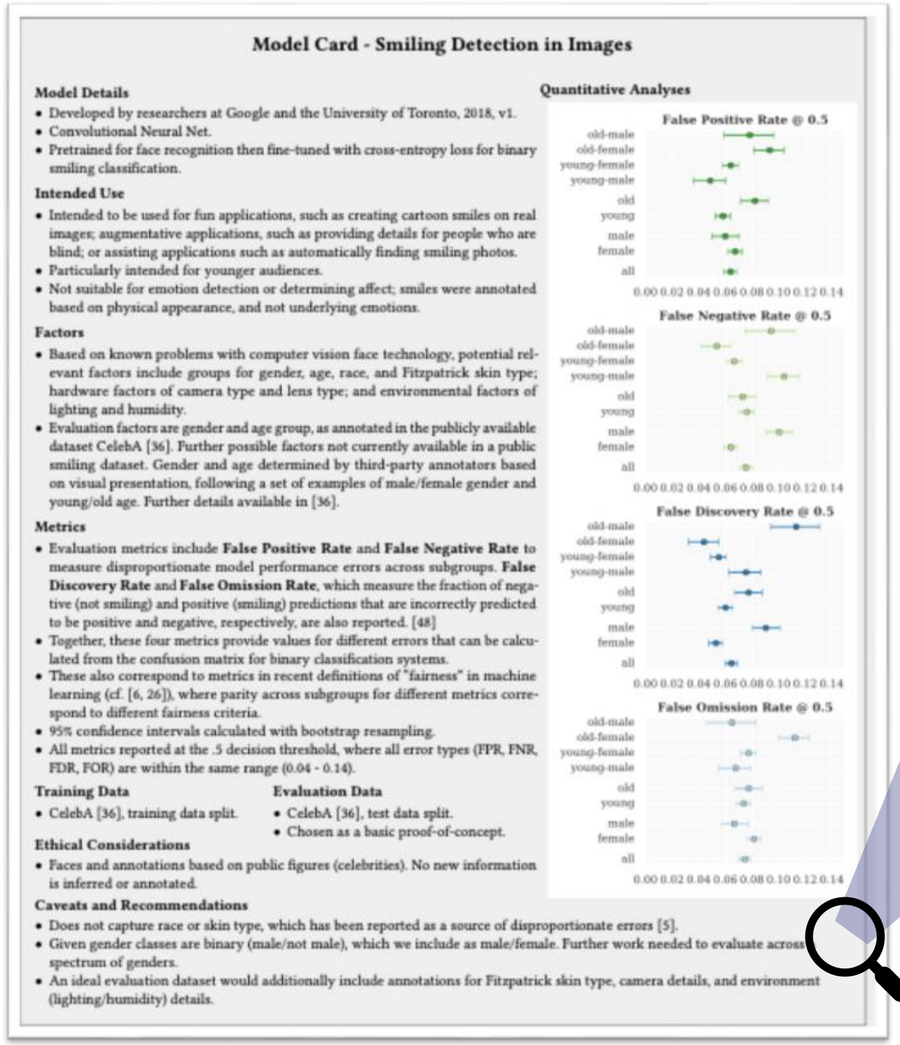
Stress Testing & Trustworthiness



stress testing complements standard evaluation

Ali, Sajid, et al. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." *Information fusion* 99 (2023)

Stress Testing & Responsible AI



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

stress testing generates the evidence
to fill model cards

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Mitchell, Margaret, et al. "Model cards for model reporting." *Proceedings of the conference on fairness, accountability, and transparency*. 2019

Stress Testing: framework limitations

- Stress definition shapes the framework
  - poor specification may lead to wrong finding or missed failures
- Finding/creating the stress distribution is non-trivial
  - options such as synthetic data generation may introduce new biases
  - the probe may be imperfect
- General concept with costly implementation

Key takeaways

- High performance on a test set does not guarantee trustworthy behaviour
  - limitations may go undetected…
- Stress testing probes the stress region to map the limits of a model’s competence
  - supporting responsible AI
- General framework: applicable to any model, any domain, any definition of stress

Inês Gomes
(image)

Flávia Carvalhido
(text-image)

Ricardo Inácio
(time series)

Ons Zammel
(text/text-image)

Moisés Santos
(time series)

André Restivo

Carlos Soares

Cátia Teixeira

Henrique L. Cardoso

Jan van Rijn

Luís F. Teixeira

Marília Barandas

Thomas Bäck

Gomes, I., Teixeira, L. F., Van Rijn, J. N., Soares, C., Restivo, A., Cunha, L., & Santos, M. (2024). Finding patterns in ambiguity: interpretable stress testing in the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Woskshops* (pp. 8316-8321).

Brito, A., Santos, M., Folgado, D., & Soares, C. (2025). tsMIST: Model Sensitivity Analysis with Time Series Morphing. In International Conference on Discovery Science (pp. 270-285).

Carvalhido, F., Cardoso, H. L., Cerqueira, V., & Soares, C. (2025). XAI in Medical Image Report Generation: Unlocking Stress Testing as a Responsible AI Practice for Multimodal Models. Proceedings of the Second Multimodal, Affective and Interactive eXplainable AI Workshop (MAI-XAI 2025) co-located with the 28th European Conference on Artificial.

Inácio, R., Cerqueira, V., Barandas, M., & Soares, C. (2025). Mast: interpretable stress testing via meta-learning for forecasting model robustness evaluation. *Machine Learning*, *114*(11), 251.

Teixeira, C., Gomes, I., Soares, C., & van Rijn, J. N. (2025). Hubris Benchmarking with AmbiGANs: Assessing Model Overconfidence with Synthetic Ambiguous Data. In *International Conference on Discovery Science* (pp. 476-491).

Gomes, I., Santos, M., Restivo, A., Soares, C., Teixeira, L. F., Van Rijn, J. N., Bäck, T. (2026). Stress Testing the Decision Boundaries of Image Classifiers via Latent DIffusion. Submitted to *Machine Learning*.

Thank you!

Do you have any questions?

ines.gomes@fe.up.pt

contacts

feedback

Stress Testing Models to Understand Their Decisions

Universiteit Leiden The Netherlands
liacs Leiden Institute of Advanced Computer Science
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
LIACC
Center for Responsible AI
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU
UKRI UK Research and Innovation
AISYM4MED
Project funded by
Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Funded by the European Union under the **Horizon Europe Framework** Programme Grant Agreement Nº: 101095387. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.