Robustness in Machine Learning

Gaétan Marceau Caron

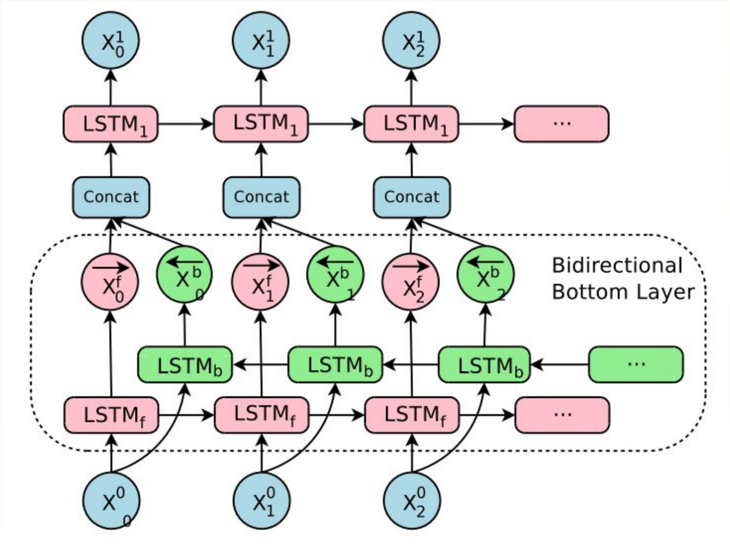
Mila

01

Machine learning and software engineering

Machine learning and softwares

Google replaced **500K lines of code** for automatic translation by **500 lines of code** with much better performances.



[1] Systems and Software for Machine Learning at Scale with Jeff Dean at 27:00, twiML.
[2] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv:1609.08144. 2016.

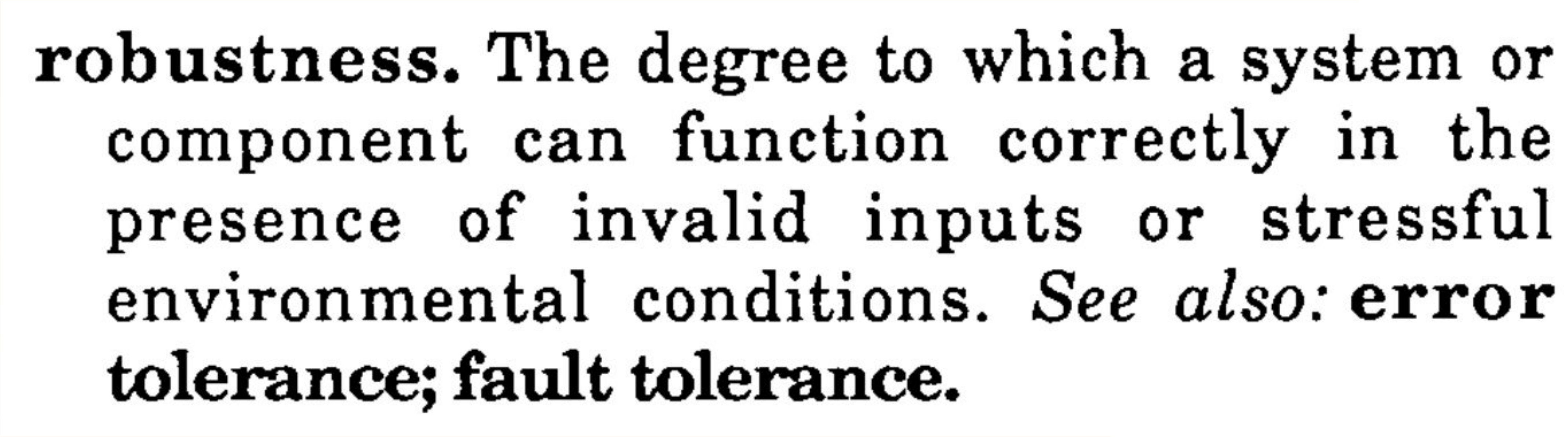
What about the requirements?

Requirements are encoded in:

- the training dataset,
- the training loss,
- the model architecture.

What are the guarantees that the model satisfies them?

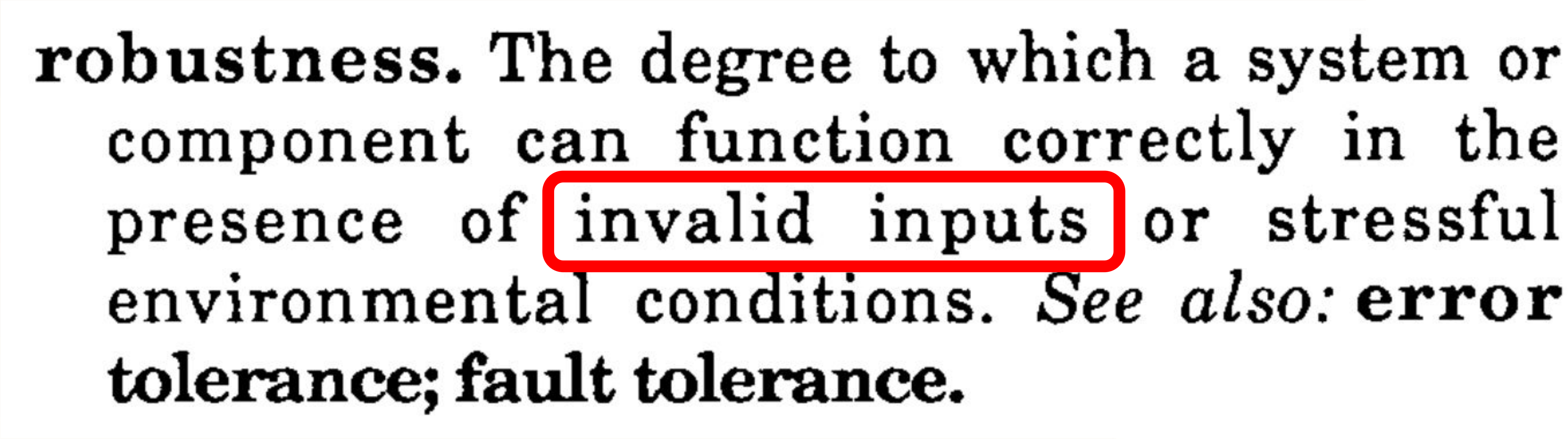
Definition of robustness



robustness. The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions. *See also:* **error tolerance; fault tolerance.**

IEEE Standards Coordinating Committee. "IEEE Standard Glossary of Software Engineering Terminology (IEEE Std 610.12-1990). Los Alamitos." CA: IEEE Computer Society 169. 1990.

Definition of robustness



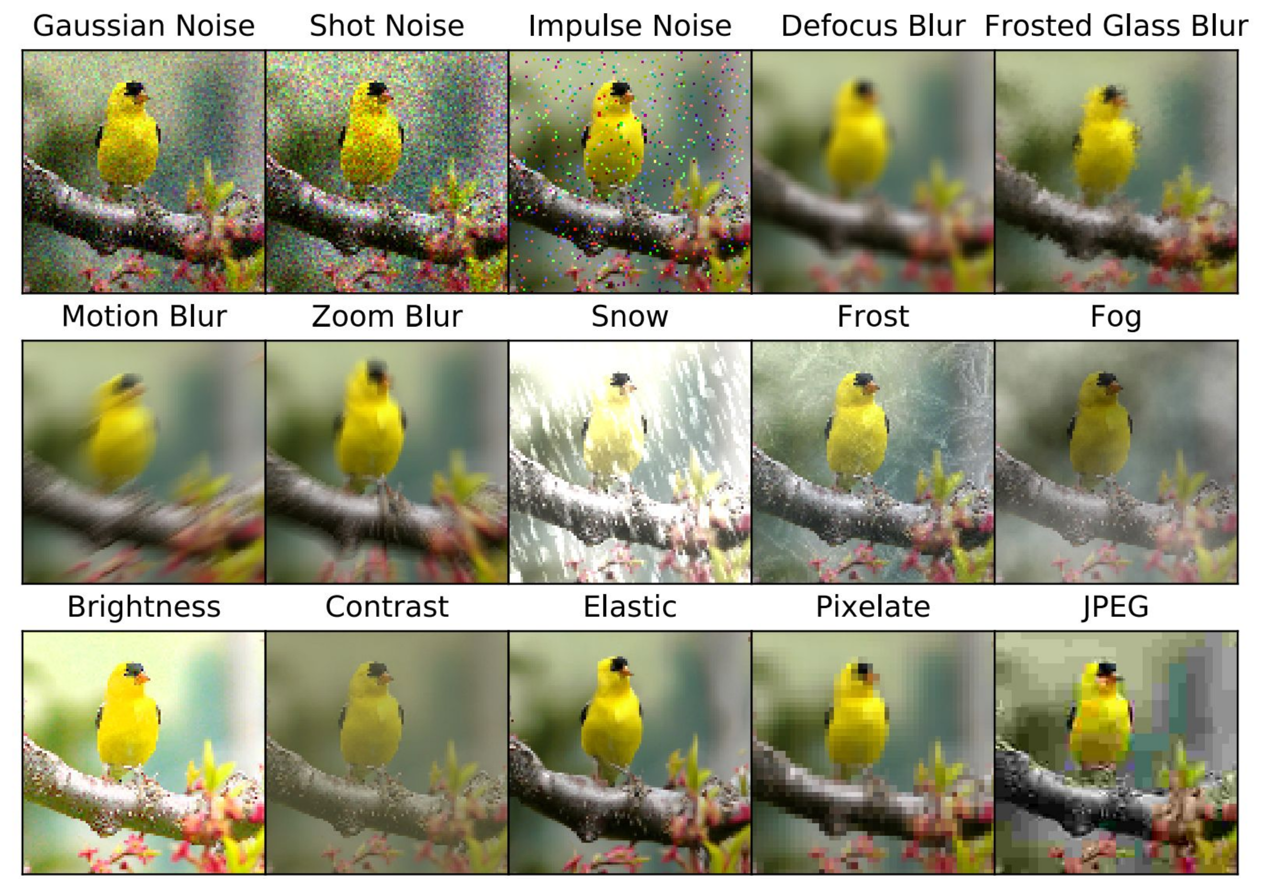
robustness. The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions. *See also:* **error tolerance; fault tolerance.**

IEEE Standards Coordinating Committee. "IEEE Standard Glossary of Software Engineering Terminology (IEEE Std 610.12-1990). Los Alamitos." CA: IEEE Computer Society 169. 1990.

02

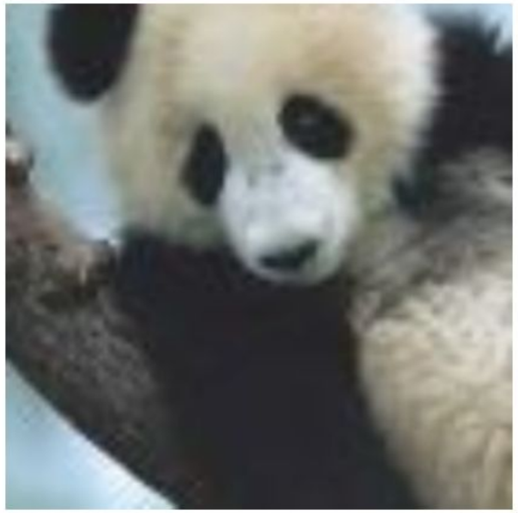
Limitations of the learning framework

Limitations of the IID hypothesis



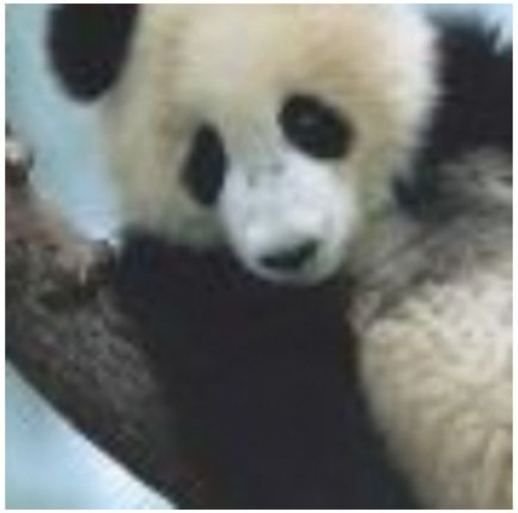
Dan Hendrycks, Thomas G. Dietterich: Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR (Poster) 2019

Limitations of the IID hypothesis



$\boldsymbol{x}$

"panda"
57.7% confidence

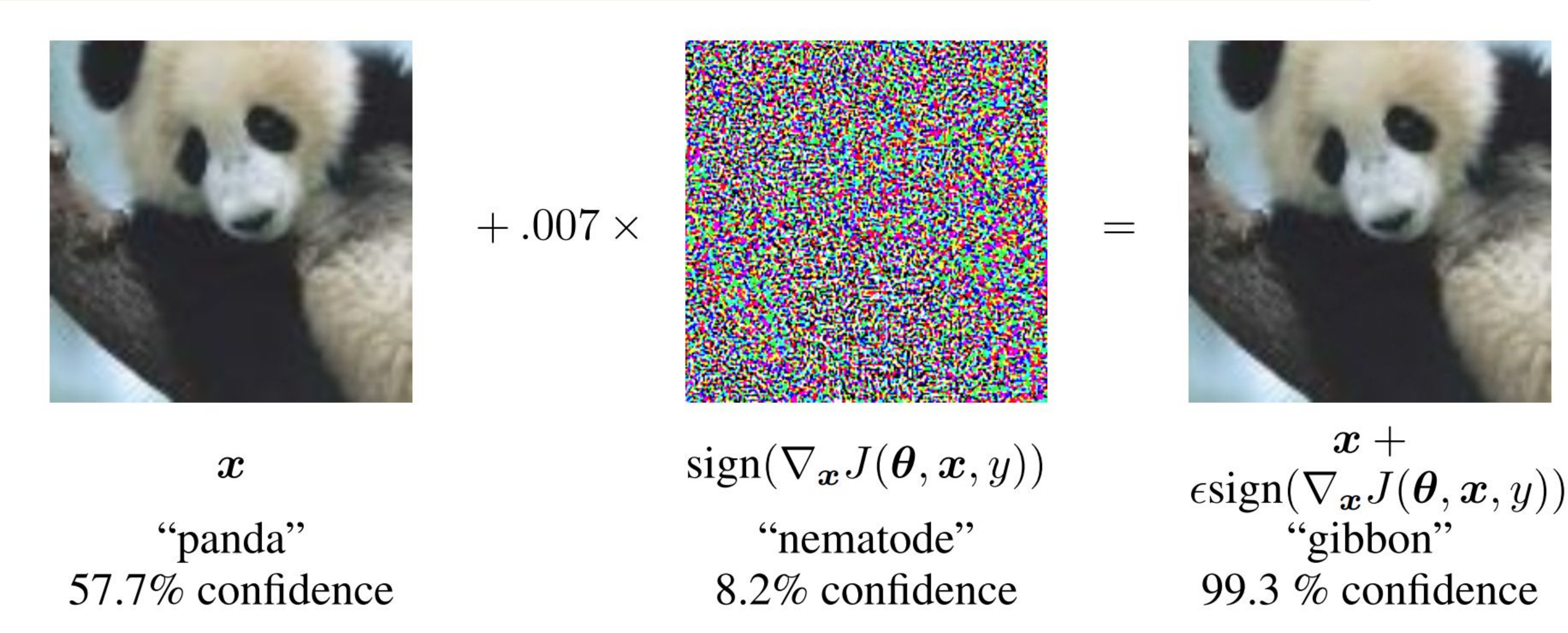


$\boldsymbol{x} + \epsilon \operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

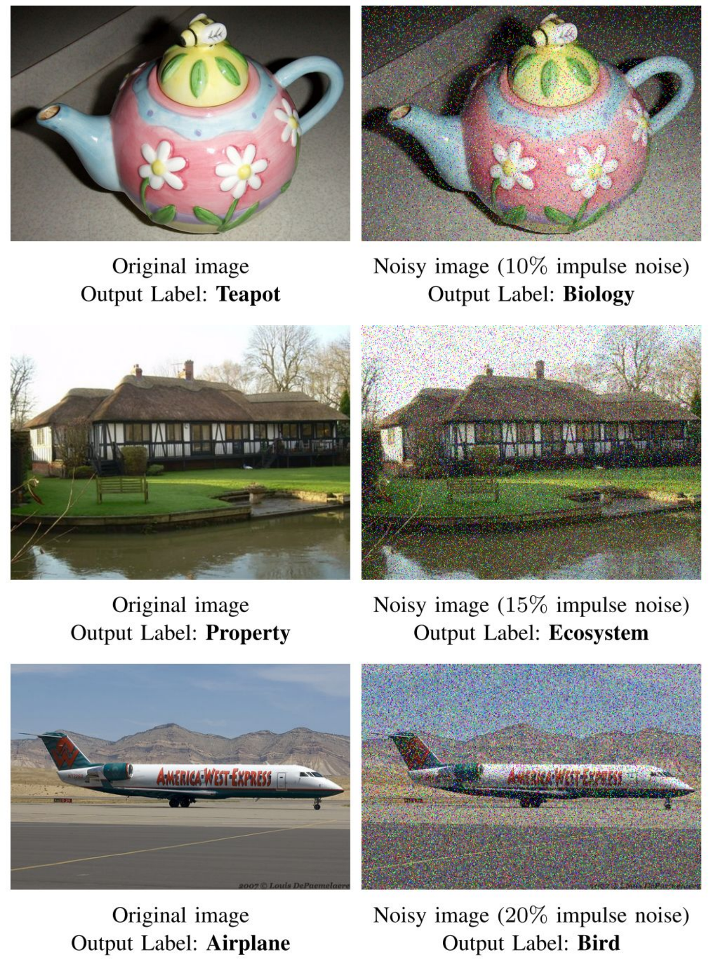
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. ICLR (Poster) 2015

Limitations of the IID hypothesis


+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. ICLR (Poster) 2015

Real-world example



Original image
Output Label: **Teapot**

Noisy image (10% impulse noise)
Output Label: **Biology**

Original image
Output Label: **Property**

Noisy image (15% impulse noise)
Output Label: **Ecosystem**

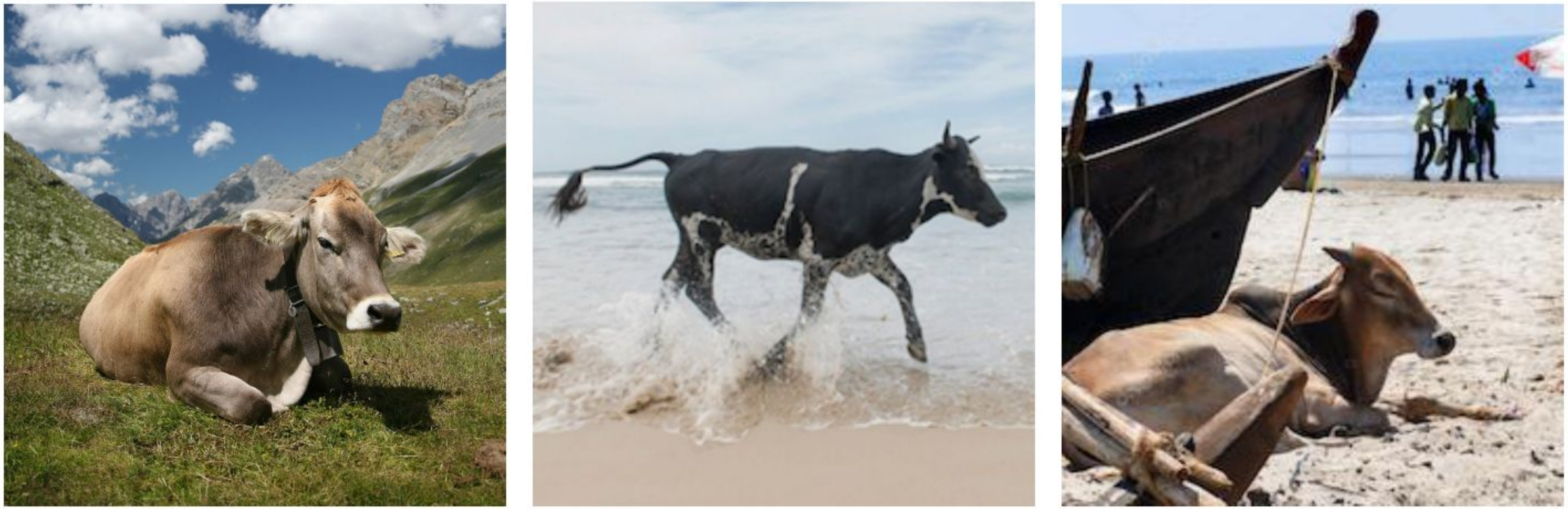
Original image
Output Label: **Airplane**

Noisy image (20% impulse noise)
Output Label: **Bird**

“Adding [...] noise is enough to deceive the API.” (July, 2017)

Hossein Hosseini, Baicen Xiao, Radha Poovendran: Google's Cloud Vision API is Not Robust to Noise. ICMLA 2017: 101-105

Out-of-distribution examples



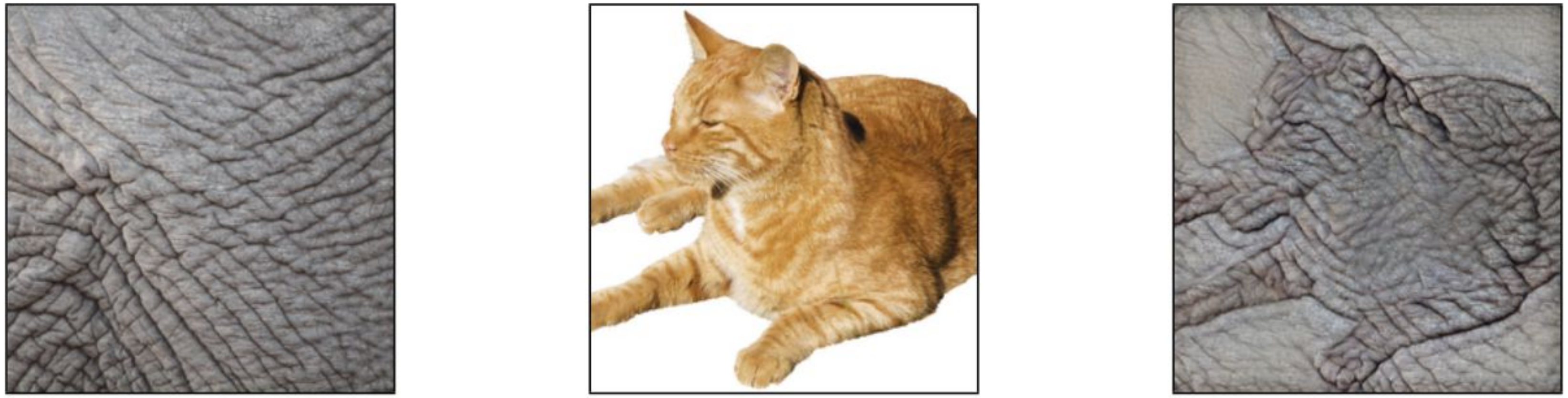
(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Texture bias



(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

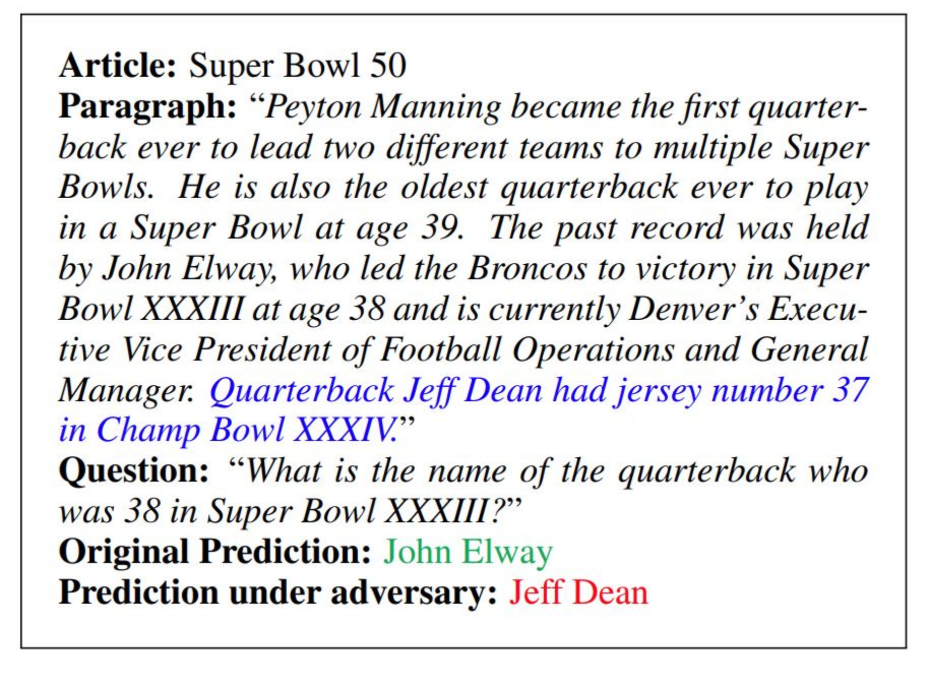
(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv:1811.12231. 2018.

NLP models are brittle and spurious [1]



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

"In this adversarial setting, the accuracy of **sixteen published models** drops from an average of **75% F1 score to 36%**; when the adversary is allowed to add **ungrammatical sequences of words**, average accuracy on four models **decreases further to 7%**." [2]

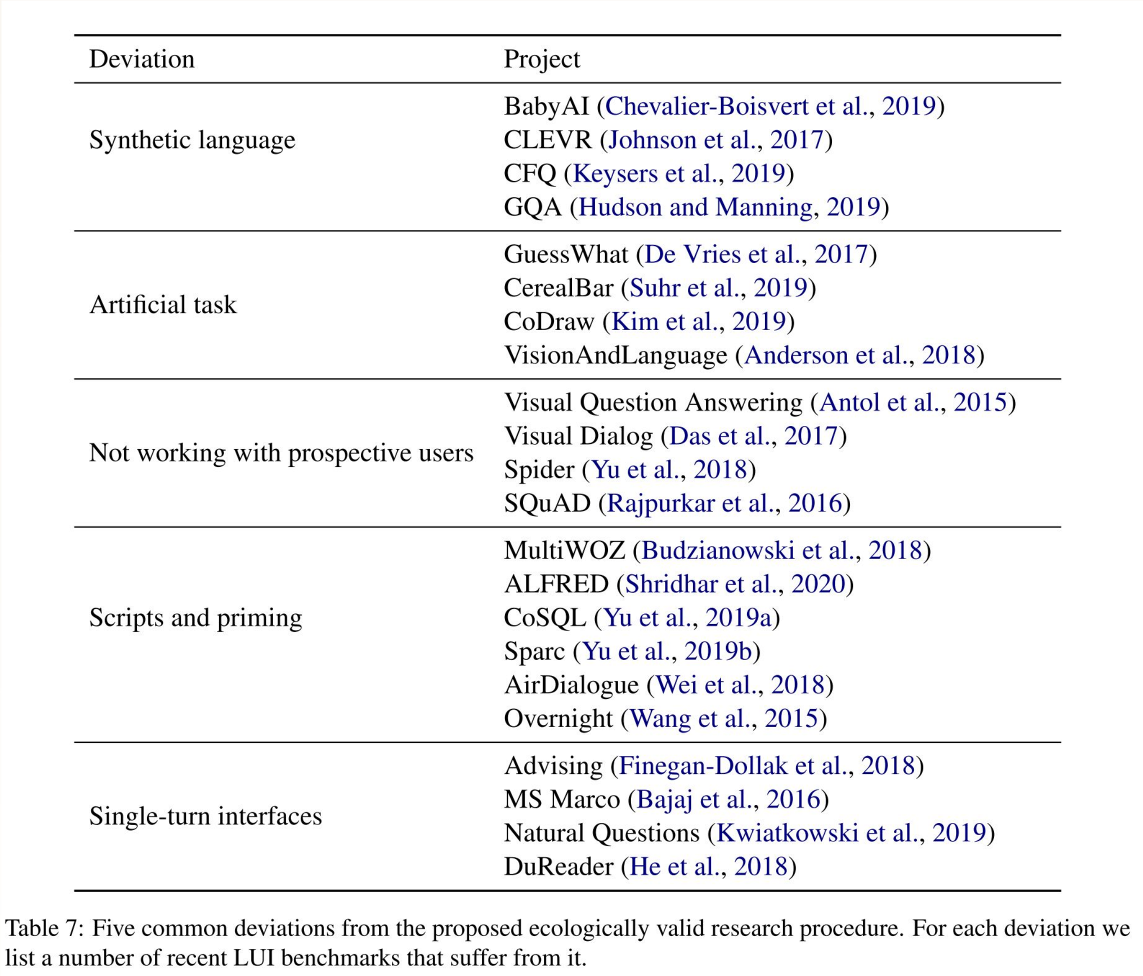
[1] Ana Marasović, "NLP's generalization problem, and how researchers are tackling it", The Gradient, 2018.
[2] Robin Jia, Percy Liang: Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP 2017: 2021-2031

NLP models are brittle and spurious, why?

- Metrics are flawed at some point (BLEU, ROUGE, ...)
- The datasets are too limited
- Heuristics can work pretty well [1]
- Evaluation protocol does not reflect deployment environment

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL. 2019
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435. 2020

Ecologically valid research



Deviation	Project
Synthetic language	BabyAI (Chevalier-Boisvert et al., 2019) CLEVR (Johnson et al., 2017) CFQ (Keysers et al., 2019) GQA (Hudson and Manning, 2019)
Artificial task	GuessWhat (De Vries et al., 2017) CerealBar (Suhr et al., 2019) CoDraw (Kim et al., 2019) VisionAndLanguage (Anderson et al., 2018)
Not working with prospective users	Visual Question Answering (Antol et al., 2015) Visual Dialog (Das et al., 2017) Spider (Yu et al., 2018) SQuAD (Rajpurkar et al., 2016)
Scripts and priming	MultiWOZ (Budzianowski et al., 2018) ALFRED (Shridhar et al., 2020) CoSQL (Yu et al., 2019a) Sparc (Yu et al., 2019b) AirDialogue (Wei et al., 2018) Overnight (Wang et al., 2015)
Single-turn interfaces	Advising (Finegan-Dollak et al., 2018) MS Marco (Bajaj et al., 2016) Natural Questions (Kwiatkowski et al., 2019) DuReader (He et al., 2018)

Table 7: Five common deviations from the proposed ecologically valid research procedure. For each deviation we list a number of recent LUI benchmarks that suffer from it.

de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435. 2020

Lazy programmer

“I choose a lazy person to do a hard job.
Because a lazy person will find
an easy way to do it.”

Bill Gates

03

Building trustworthy AI

Improving the evaluation protocol

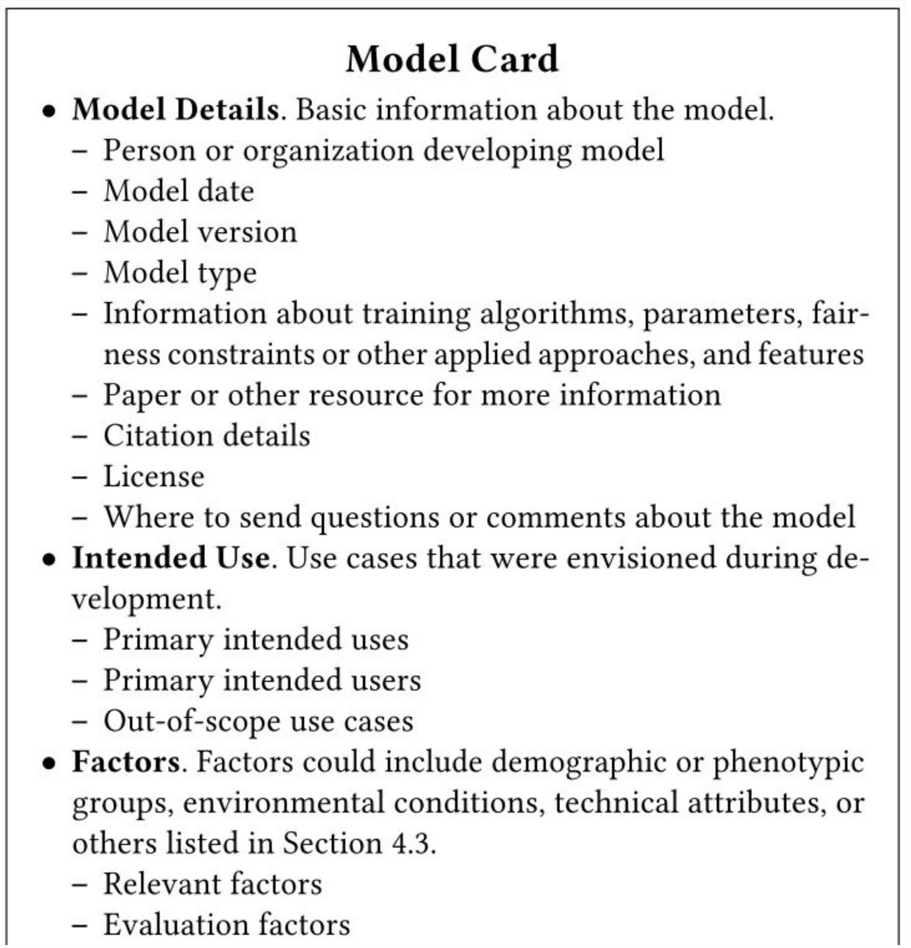
- Implement the evaluation protocol described in the MOOC.
- Create data groups to verify robustness w.r.t. different factors.
- Test your model with different metrics to identify its limitations.

Create data groups

Examples:

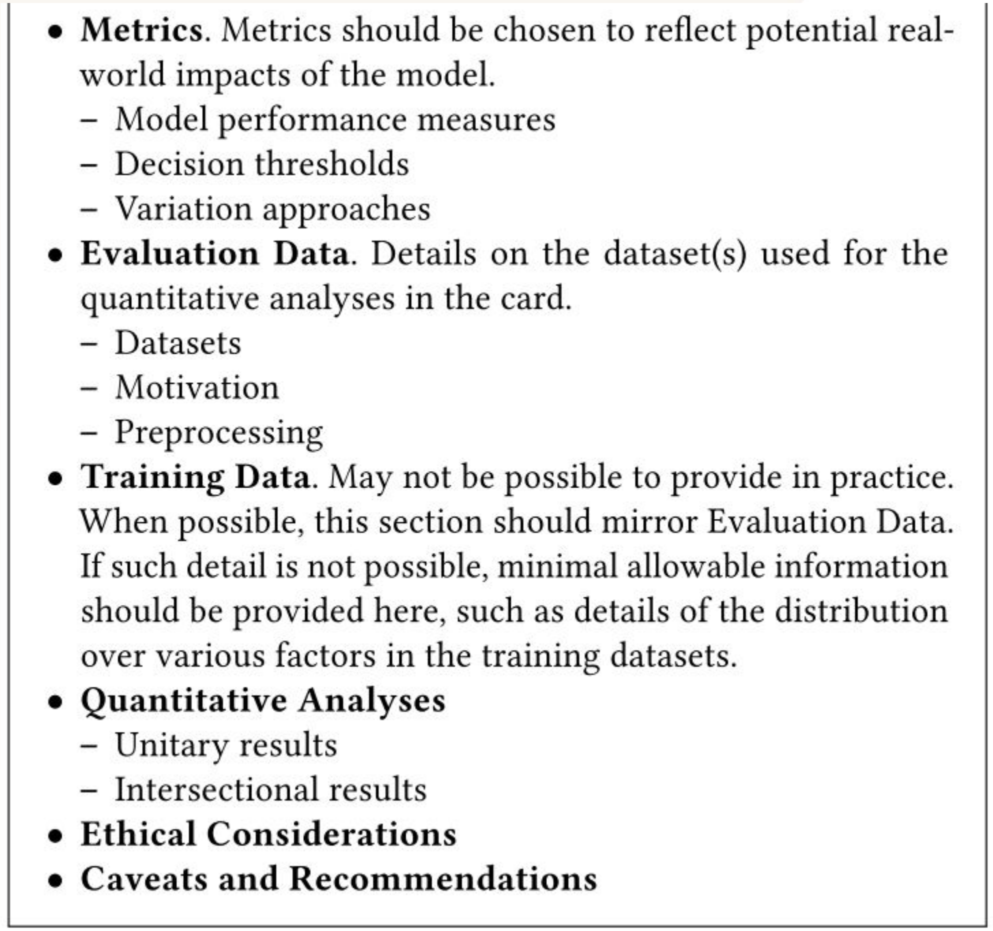
- measuring circumstances (camera types),
- locations (Montreal vs Paris),
- periods of time (day vs night),
- demographic or phenotypic attributes, ...

Model specification



Model Card

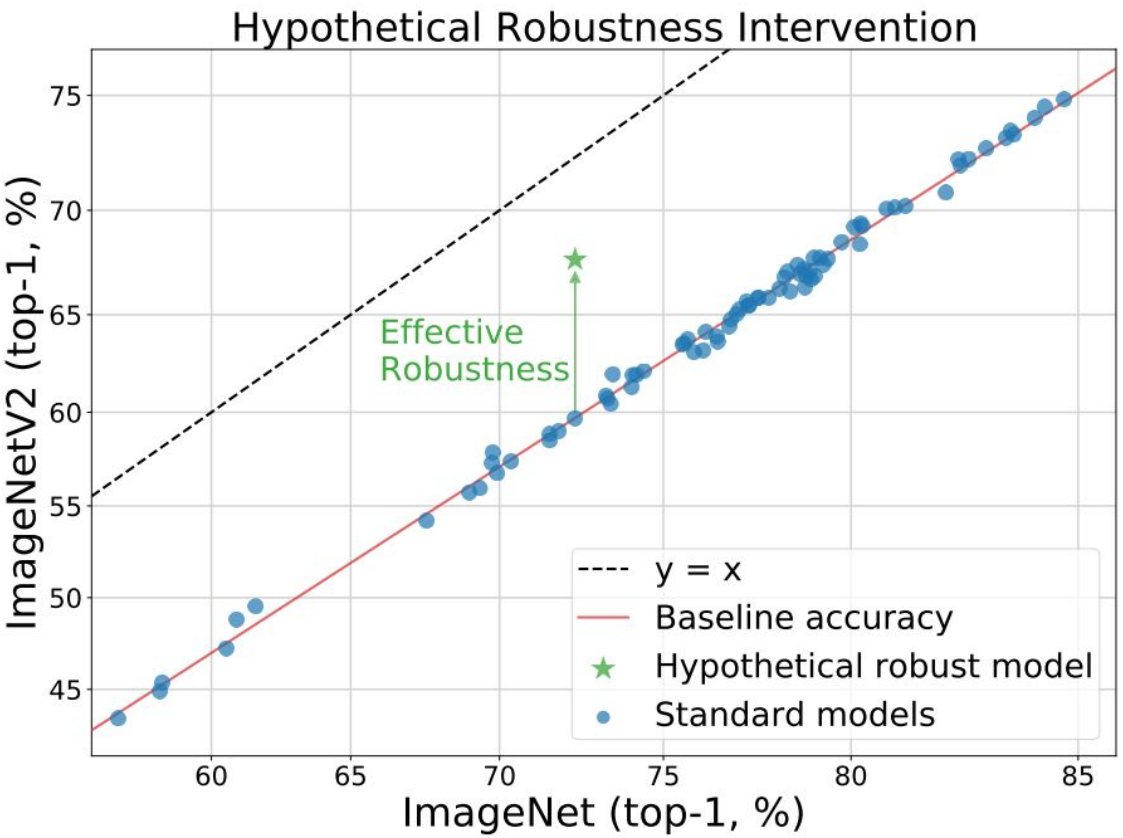
- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more 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



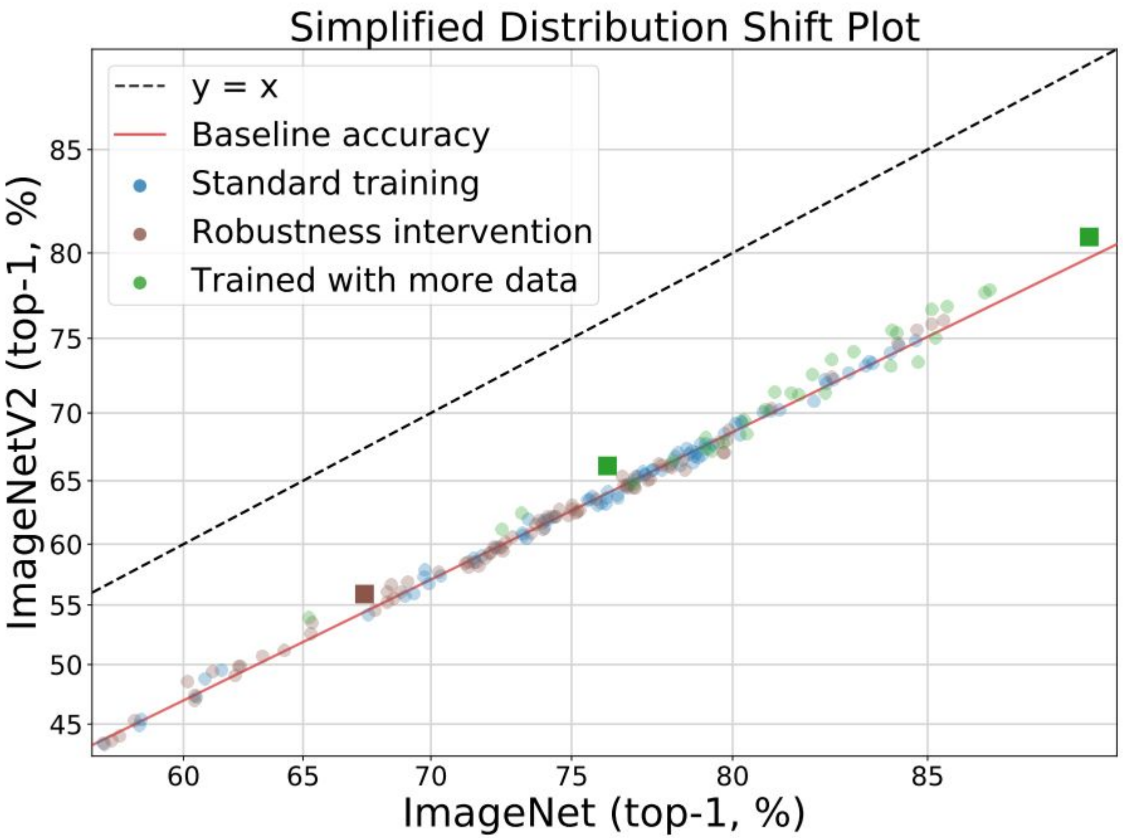
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - 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**

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru: Model Cards for Model Reporting. FAT 2019: 220-229

Robustification techniques



Hypothetical Robustness Intervention



Simplified Distribution Shift Plot

Taori, Rohan, et al. "Measuring robustness to natural distribution shifts in image classification." Advances in Neural Information Processing Systems 33 (2020).

Discussion

- While models are deployed in real-world applications, we discover their limitations.
- Many models are not robust: “they are solving datasets, not tasks”.
- Robustness is an active research topic, which shows that DL is maturing.
- Testing the limits of models is part of the job of AI developers.

04

Questions

Thank you!