Disentangling Adversarial Robustness and Generalization

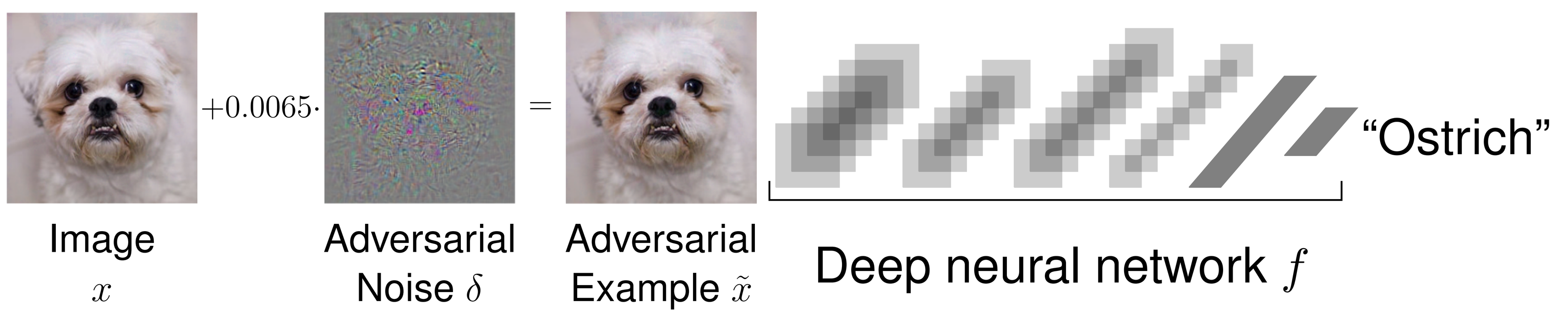
max planck institut informatik

David Stutz, Matthias Hein and Bernt Schiele

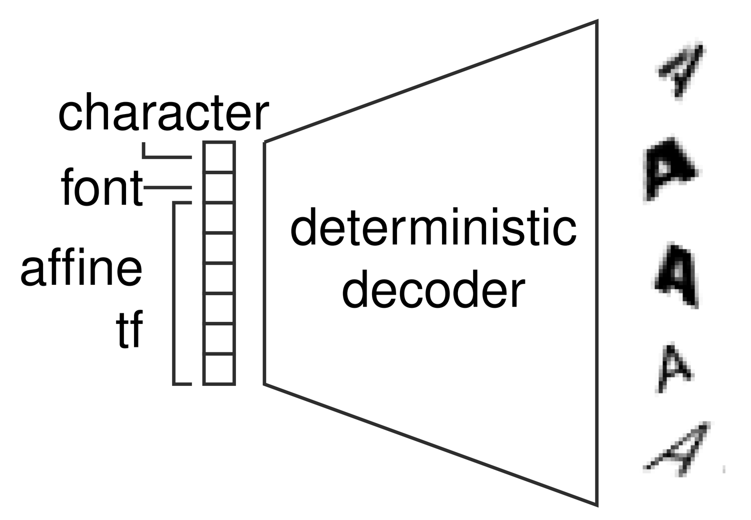
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Problem

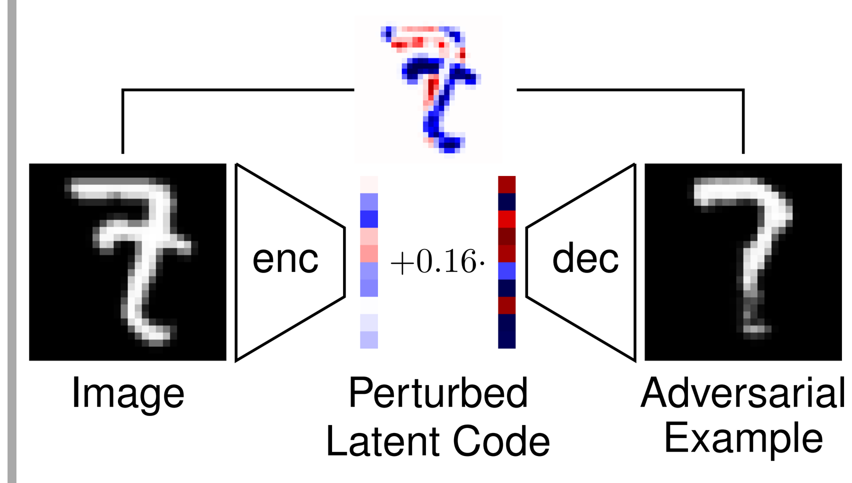
Adversarial examples: imperceptibly perturbed images causing mis-classification.



FONTS (Synthetic)



EMNIST



Background

Deep neural networks (simplified):

Element-wise activation function

$$f(x; w) = h(w_L^T h(w_{L-1}^T h(\dots h(w_1^T x))))$$

Set of weight matrices $\{w_l\}_{l=1}^L$

Training with dataset $\{(x_n, y_n)\}_{n=1}^N$:

Target label y_n for input x_n

$$w^\star = \underset{w}{\text{argmin}} \sum_n \mathcal{L}(f(x_n; w), y_n)$$

Cross-entropy loss $\mathcal{L}$

Adversarial examples:

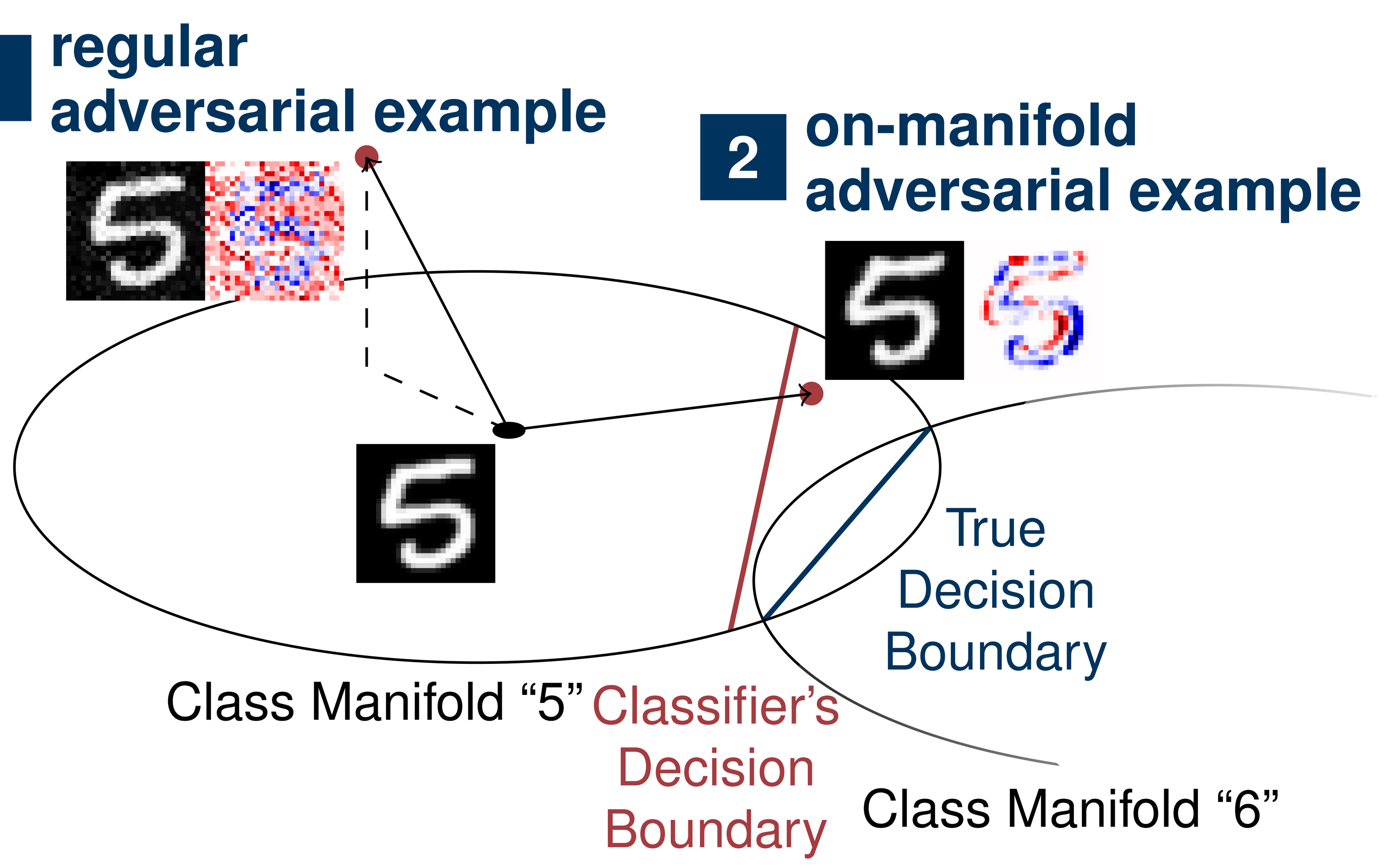
Maximize loss w.r.t. true label y

$$\tilde{x} = x + \delta \quad \text{with} \quad \delta = \underset{\|\delta\|_\infty \le \epsilon}{\text{argmax}} \mathcal{L}(f(x + \delta; w^\star), y)$$

"Imperceptible" adversarial noise

Contributions

Are accurate *and* robust models possible?

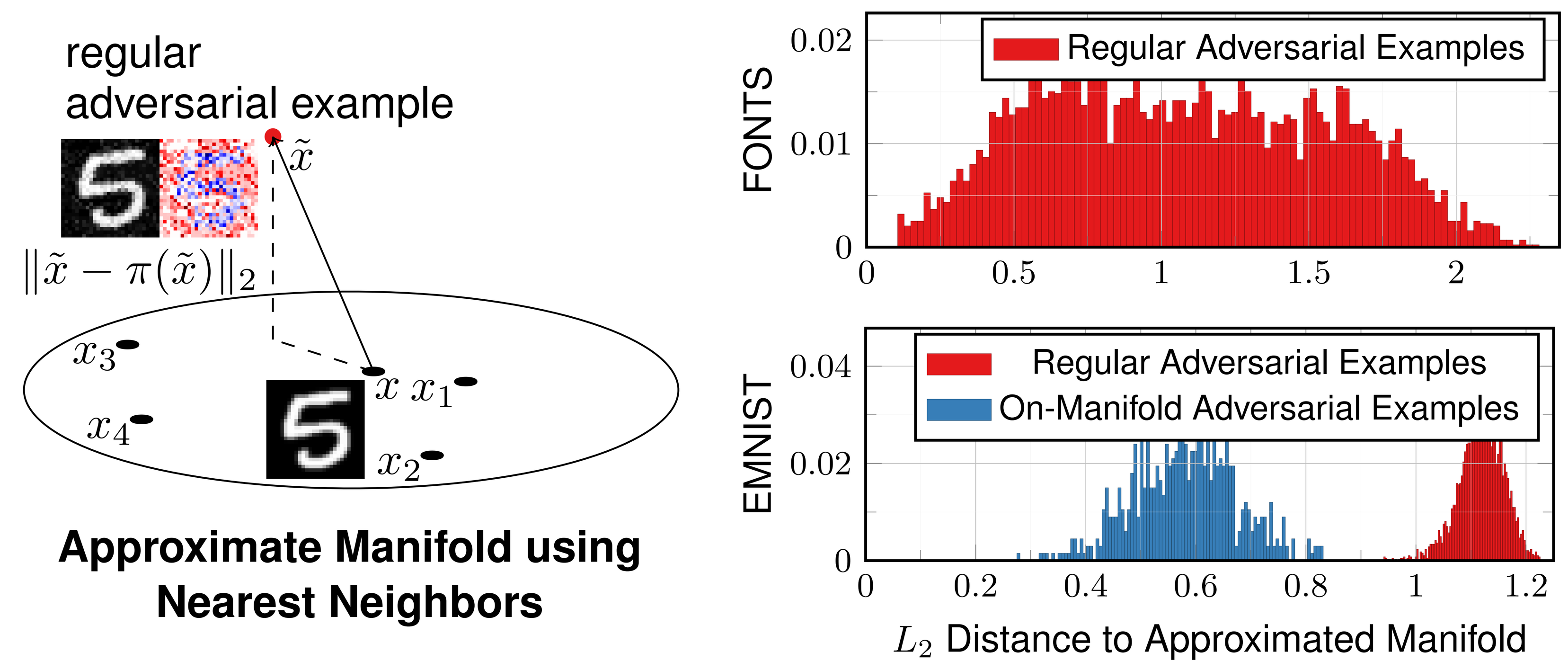


3 On-manifold robustness *is* generalization.

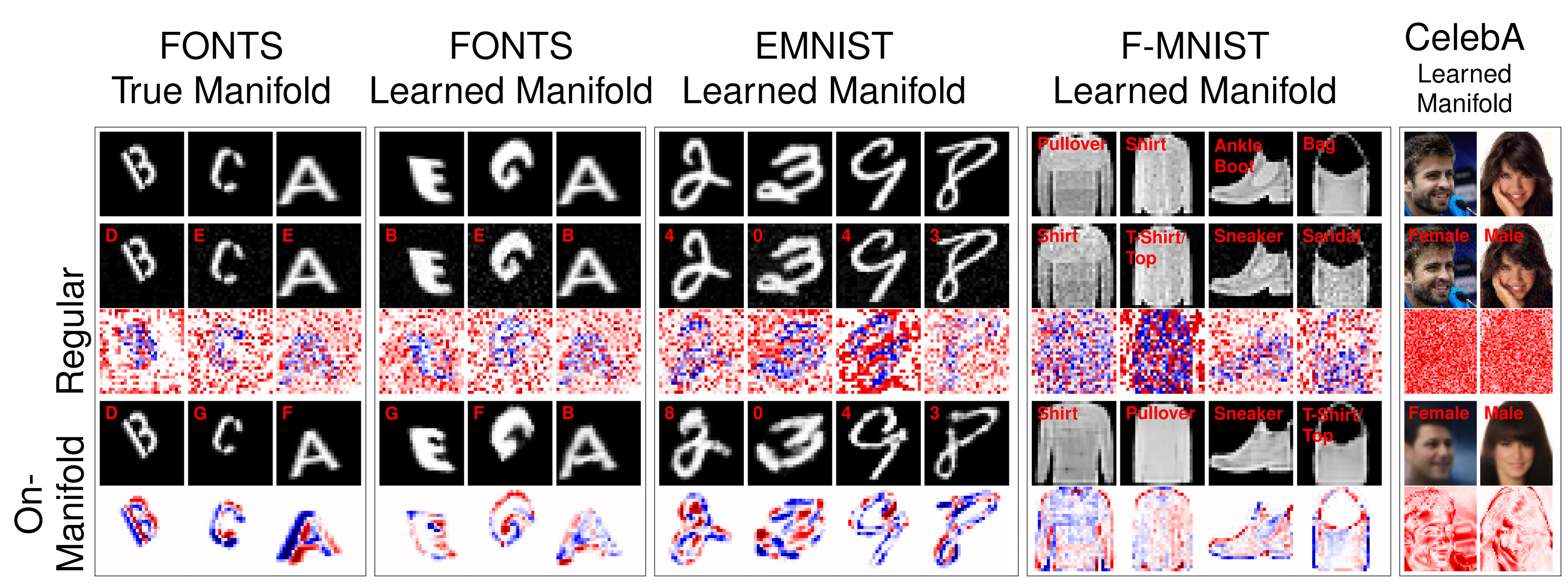
4 Robustness and generalization *not* contradicting.

▶ **Robustness has higher sample complexity.**

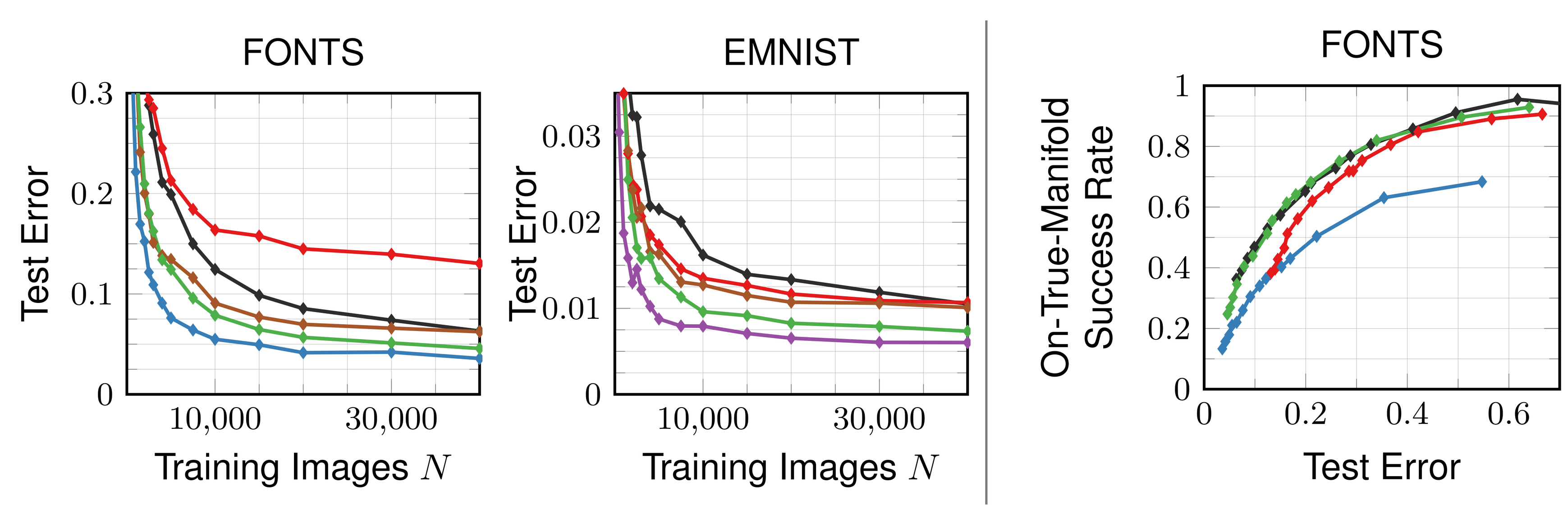
1 Adversarial Examples Leave Manifold



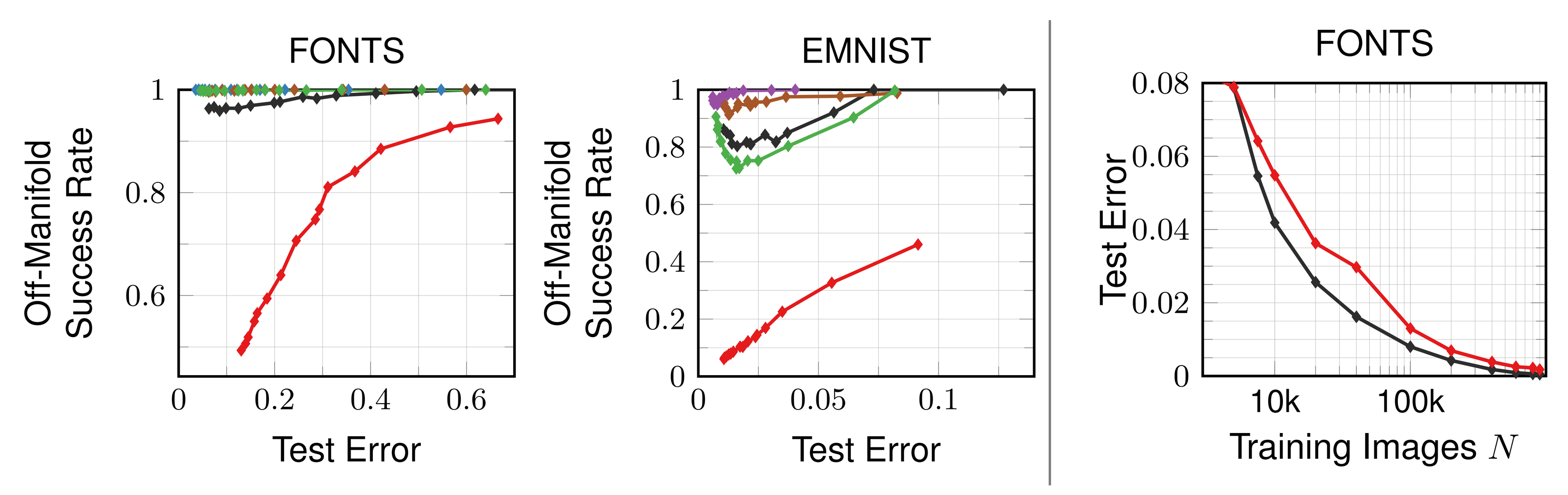
2 On-Manifold Adversarial Examples



3 On-Manifold Robustness *is* Generalization



4 Robustness Independent of Generalization



Paper, Code and Data:
davidstutz.de/hlf2019

- Normal Training
- Adversarial Training
- Adversarial Training w/ On-*True*-Manifold Adversarial Examples
- Adversarial Training w/ On-*Learned*-Manifold Adversarial Examples
- Adversarial Training w/ Adversarial Transformations