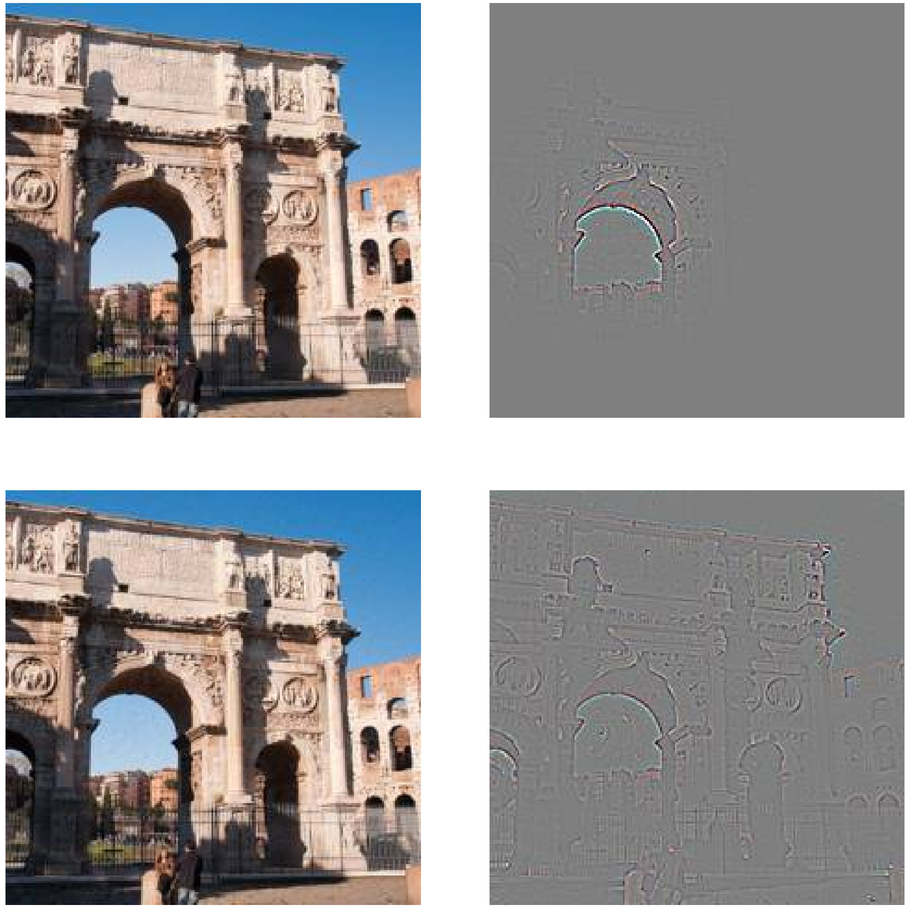
Zürcher Hochschule
für Angewandte Wissenschaften

Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps

Deep Learning Day 2018, Sep 14th, Winterthur, Switzerland

Mohammadreza Amirian



Swiss Alliance for
Data-Intensive Services

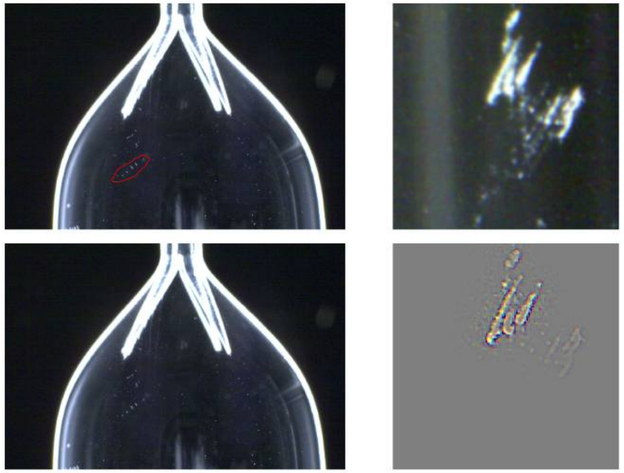
datalab
www.zhaw.ch/datalab

Motivation

Zürcher Hochschule
für Angewandte Wissenschaften

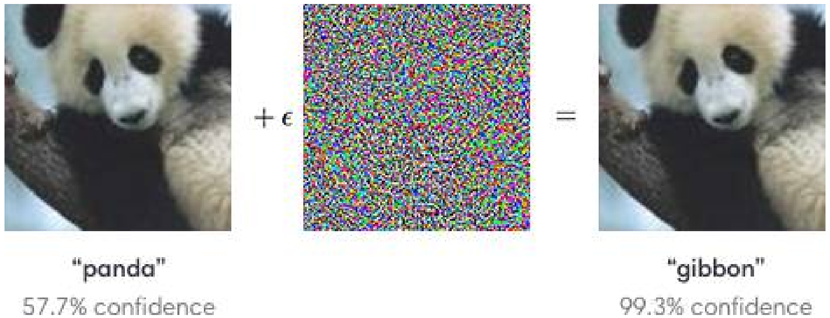
Explainable AI:

- How does the networks learn?
- How does the networks decide?



Reliability in AI:

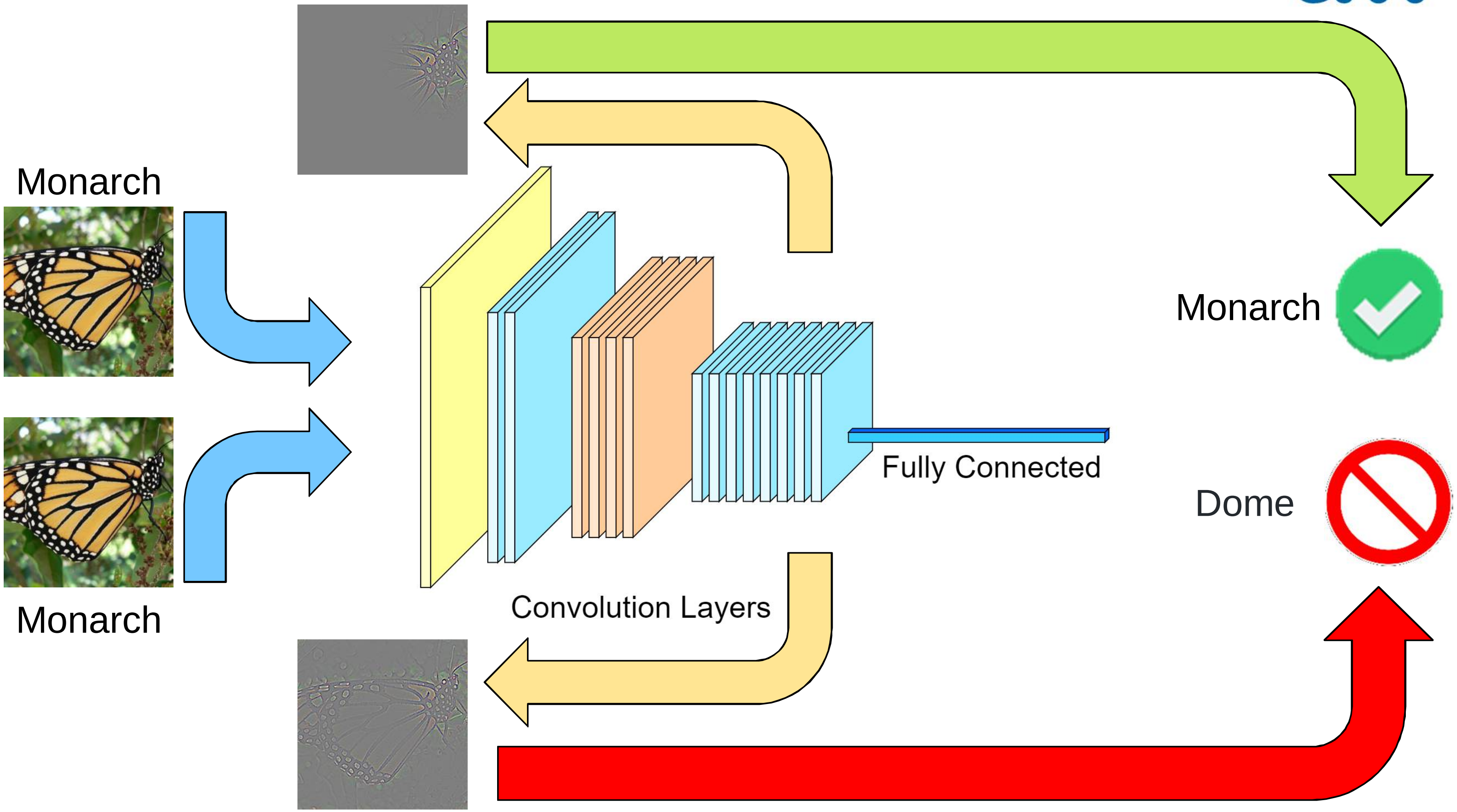
- Adversarial perturbations.



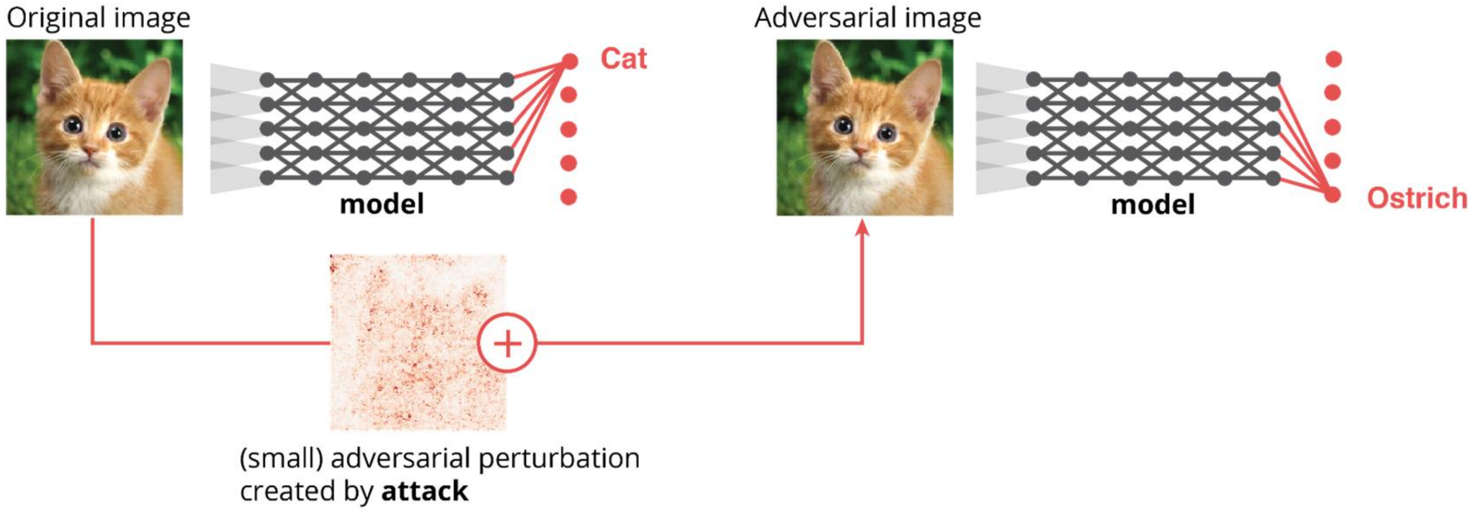
https://blog.openai.com/adversarial-example-research/

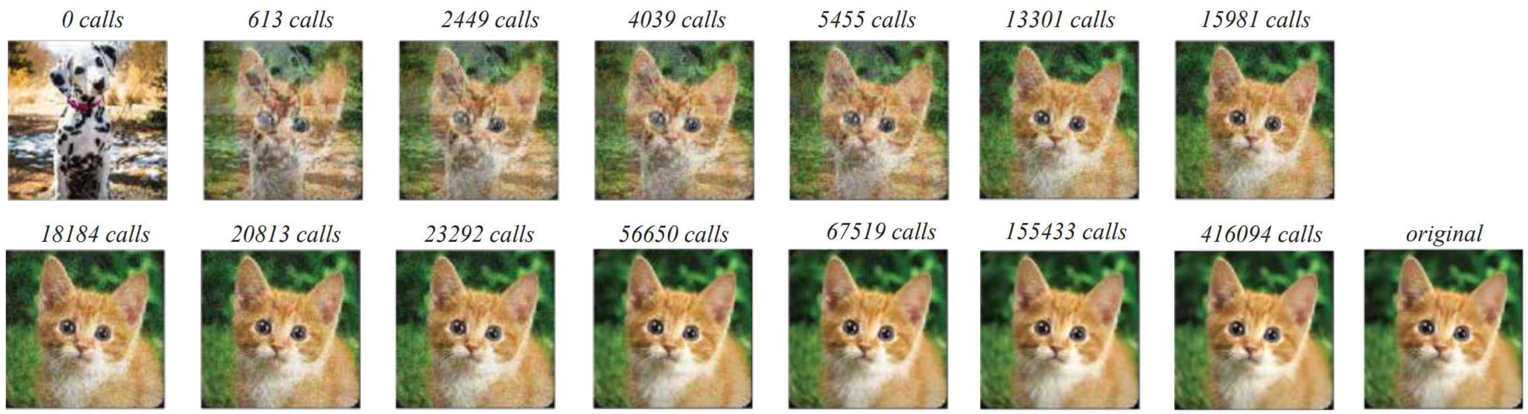
AM(1

Underlying Idea



1. Adversarial Perturbations

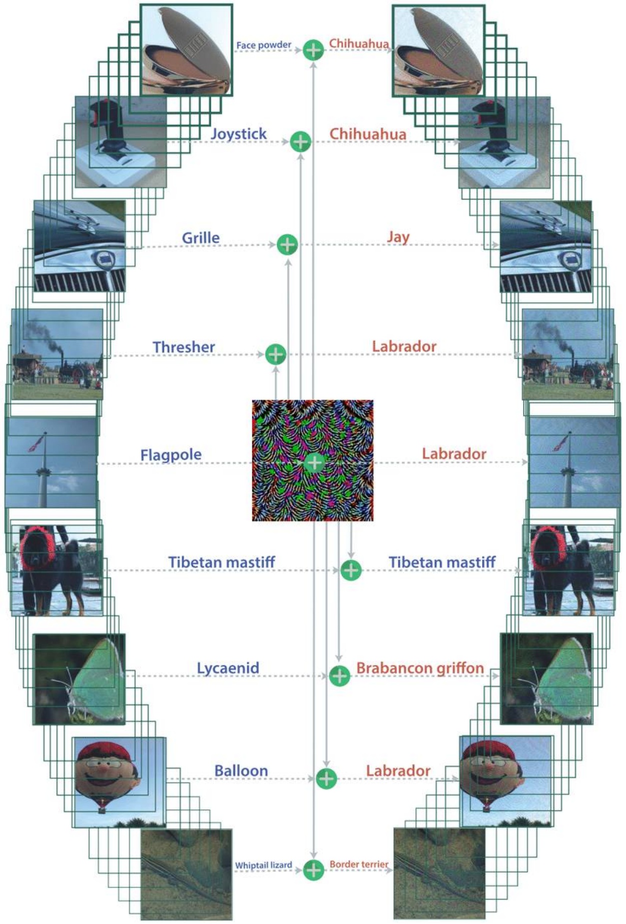




https://www.crowdai.org/challenges/nips-2018-adversarial-vision-challenge
https://arxiv.org/pdf/1712.04248.pdf

Common Perturbation Scenarios

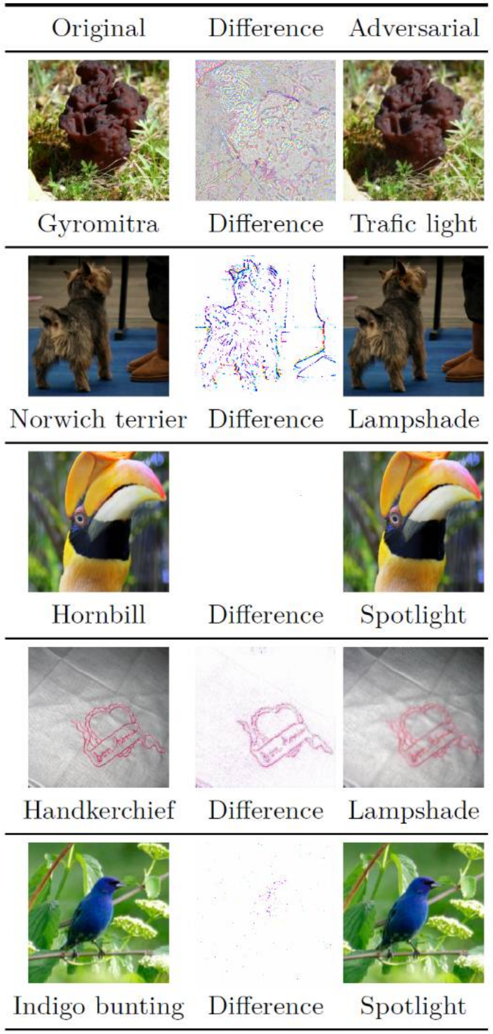
- Non-targeted attack
- Targeted attack
- White box
- Black box with probing
- Black box without probing
- Digital attack
- Physical attack
- Universal attack



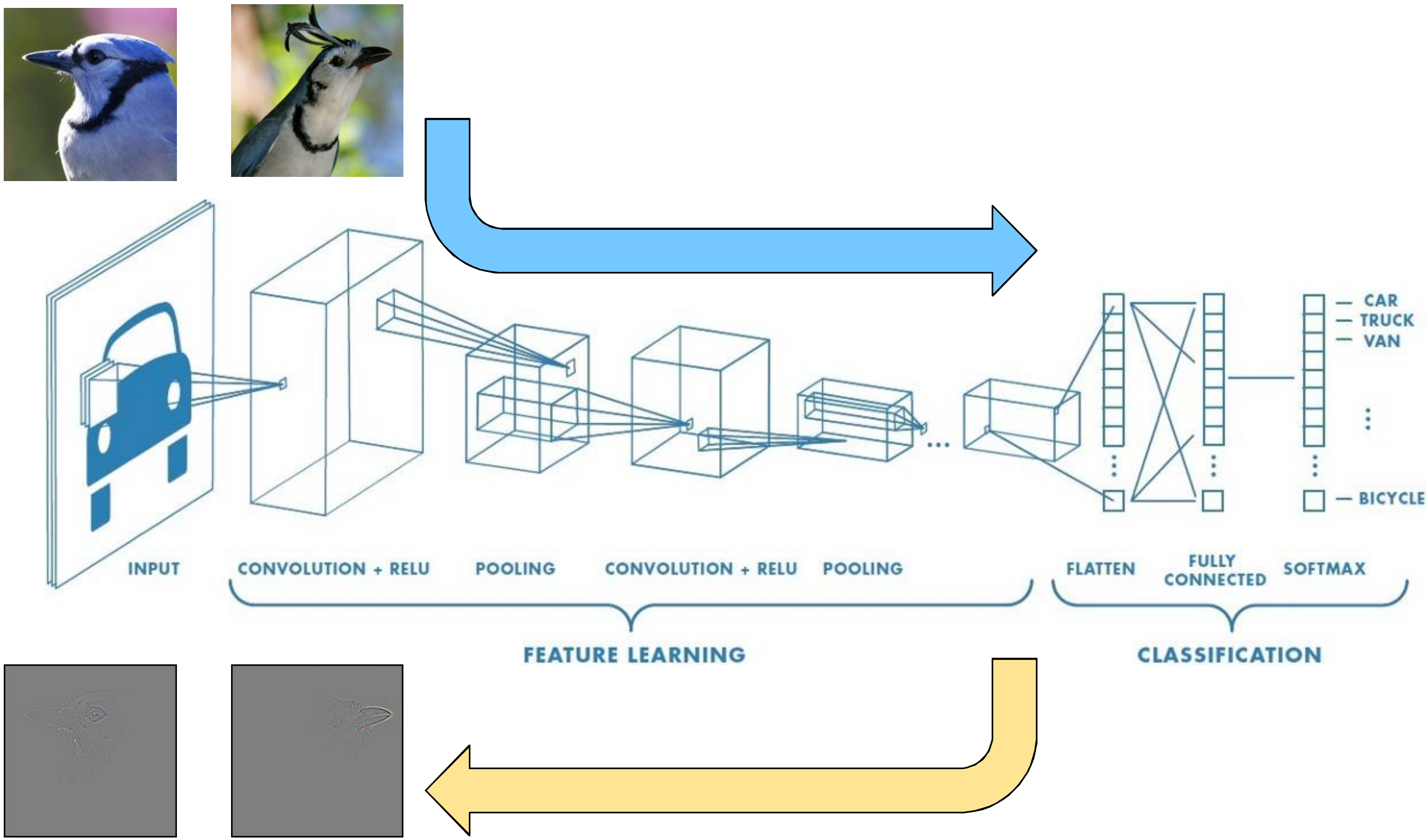
https://www.youtube.com/watch?v=piYnd_wYlT8
http://openaccess.thecvf.com/content_cvpr_2017/papers/Moosavi-Dezfooli_Universal_Adversarial_Perturbations_CVPR_2017_paper.pdf

Computing Adversarial Perturbations

- Fast Gradient Sign Method (FGSM)
- Gradient attack
- One-pixel attack
- DeepFool
- Boundary attack

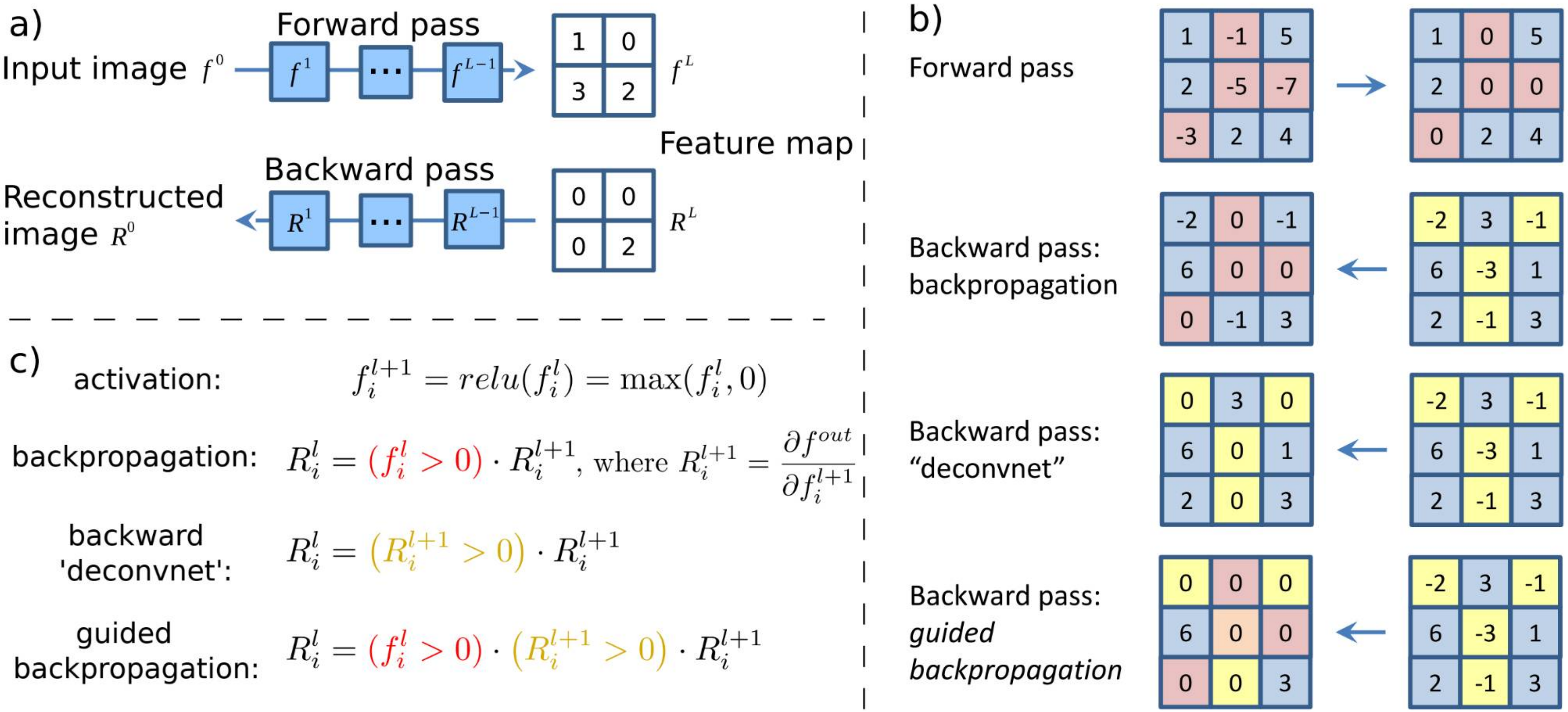


2. Feature Response Visualization



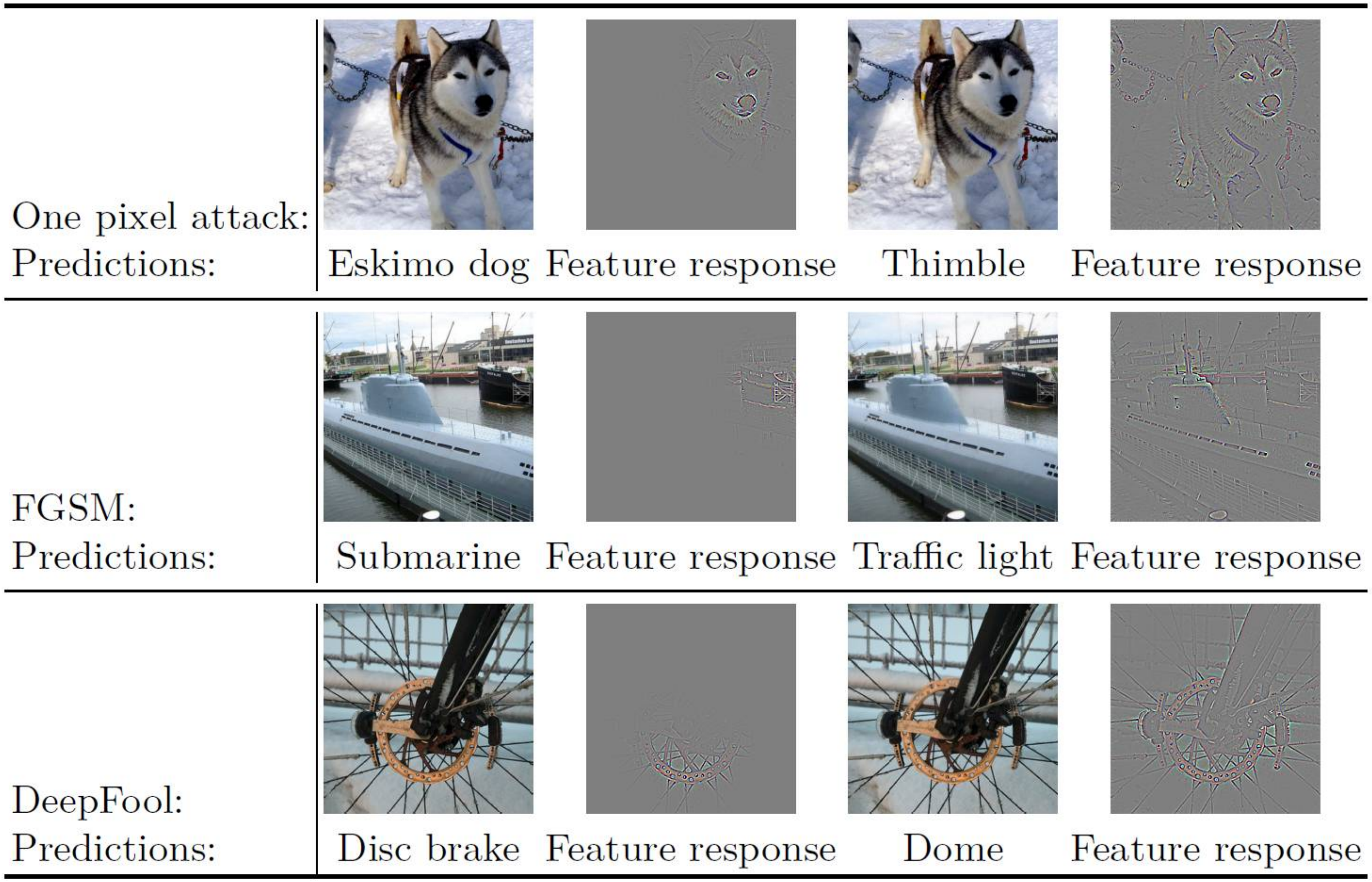
https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050

Computing Feature Responses

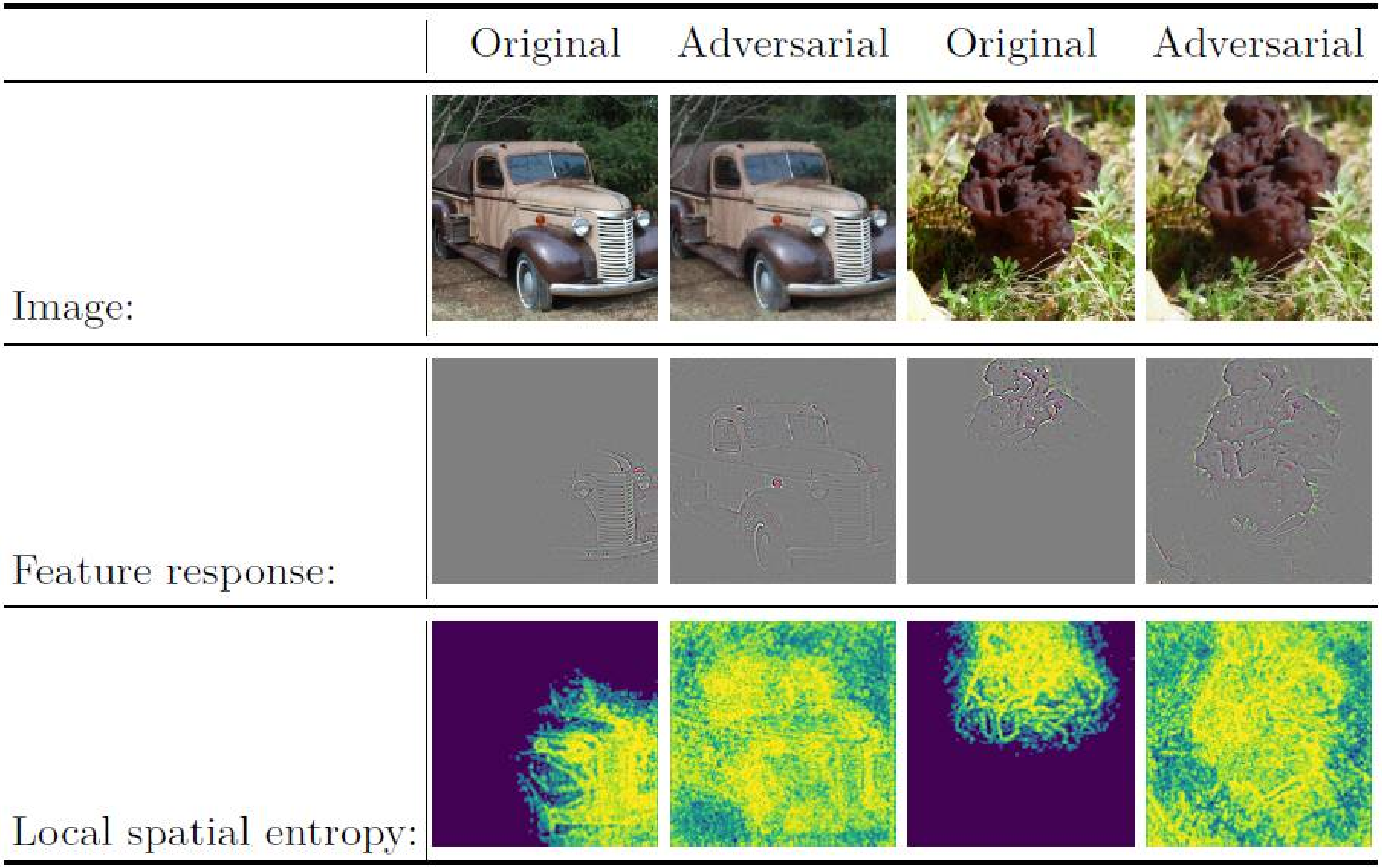


https://arxiv.org/pdf/1412.6806.pdf

3. Tracing Adversarial Perturbations



4. Feature-Based Adversarial Image Detection

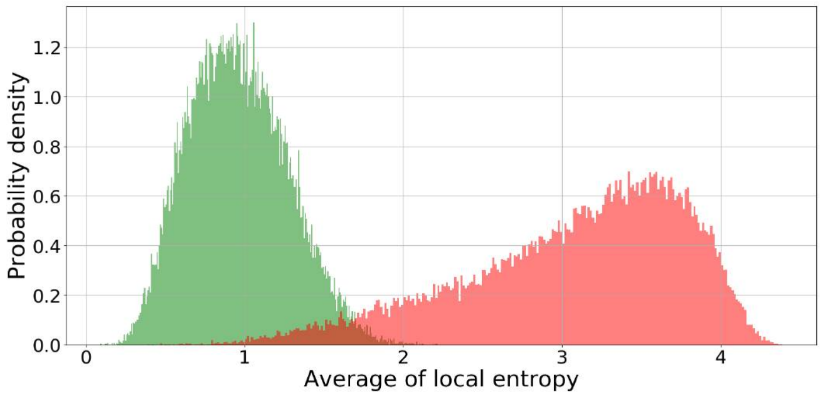


Decision Metric and Method Performance

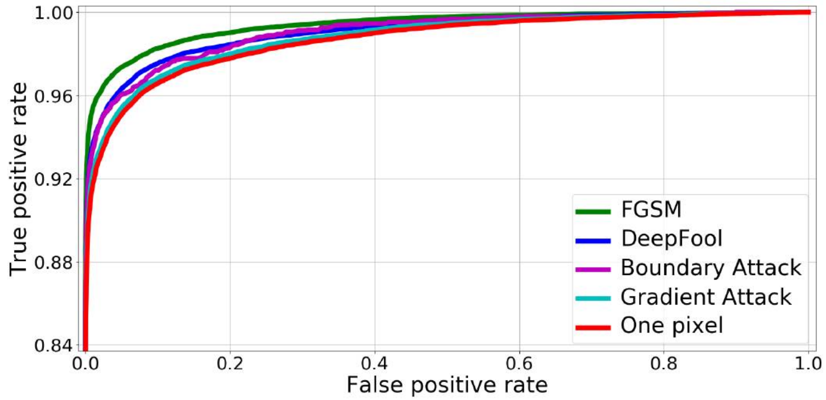
- Average local spatial entropy:

$$S_k = -\sum_i \sum_j \boldsymbol{h}_k(i,j)\log_2(h_k(i,j))$$

- Results:



Histogram



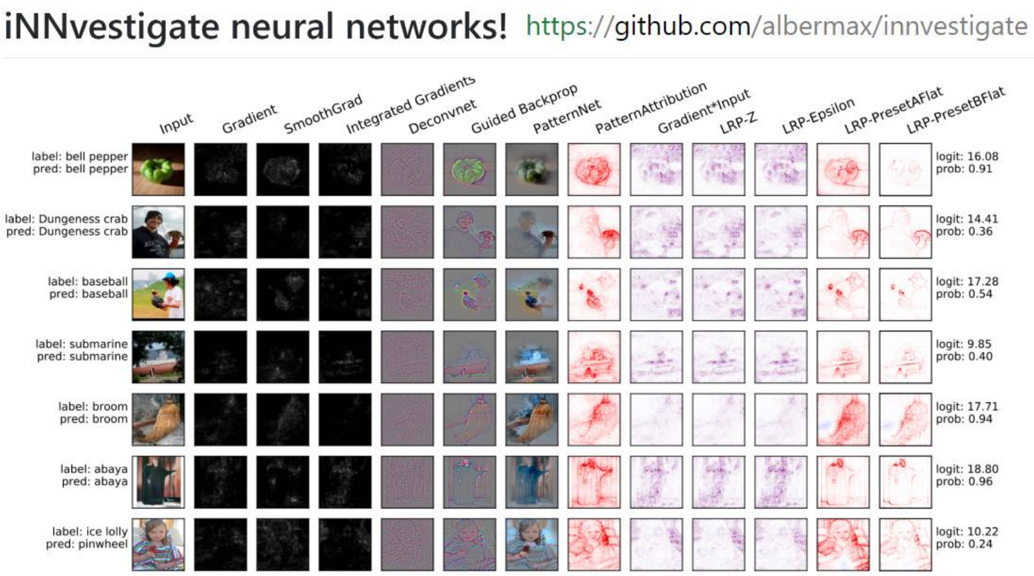
ROC curves

Numerical Evaluation and Comparison

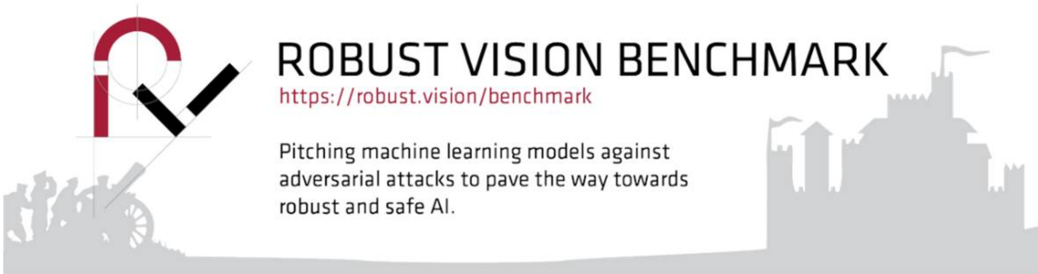
Adversarial attack	#Images (run time [days])	Success rate	Ground truth confidence	Target class confidence	False positive rate 1%	5%	10%
FGSM	50,014 (3)	0.925	0.022	0.588	0.954	0.974	0.983
Gradient attack	50,014 (15)	0.499	0.052	0.371	0.922	0.954	0.969
One pixel attack	50,014 (32)	0.620	0.037	0.463	0.917	0.951	0.966
DeepFool	47,858 (42)	0.606	0.041	0.446	0.936	0.963	0.976
Boundary attack	4,013 (17)	0.940	0.023	0.583	0.934	0.960	0.972

Method	Dataset	Network	Attack	Performance Recall	Precision	AUC
Uncertainty density estimation	SVHN	LeNet	FGSM	-	-	0.890
Adaptive noise reduction	ImageNet (4 classes)	CaffeNet	DeepFool	0.956	0.911	-
Feature squeezing	ImageNet-1000	VGG19	Several attacks	0.859	0.917	0.942
Statistical analysis	MNIST	Self-designed	FGSM ($\epsilon = 0.3$)	0.999	0.940	-
Feature response (our approach)	ImageNet validation	VGG19	Several attacks	0.979	0.920	0.990

References:



https://www.crowdai.org/challenges/nips-2018-adversarial-vision-challenge



Conclusions and Future Works

- Feature response visualizations help to debug & understand
- Improving the adversarial perturbation detection algorithm based on the feature responses
- Design a defense algorithm against adversarial perturbations
- Using the feature responses to design and enhance the network architectures

datalab
www.zhaw.ch/datalab

Swiss Alliance for
Data-Intensive Services

On me:

- Research Assistant and Ph.D. student in AI/ML
- mohammadreza.amirian@zhaw.ch
- Collaboration: datalab@zhaw.ch
- 058 934 47 62

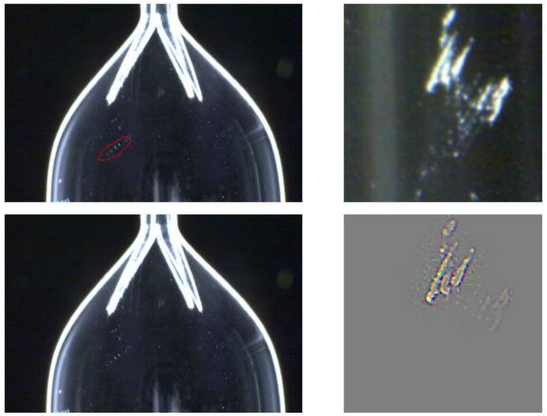
Research interests:

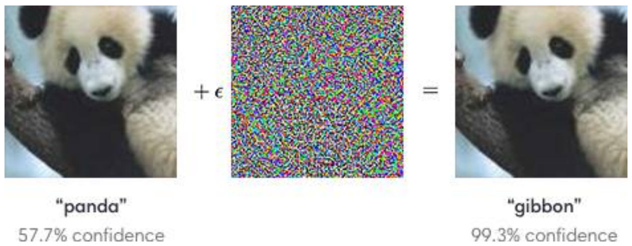
- Deep learning
- Explainable AI
- Medical image processing

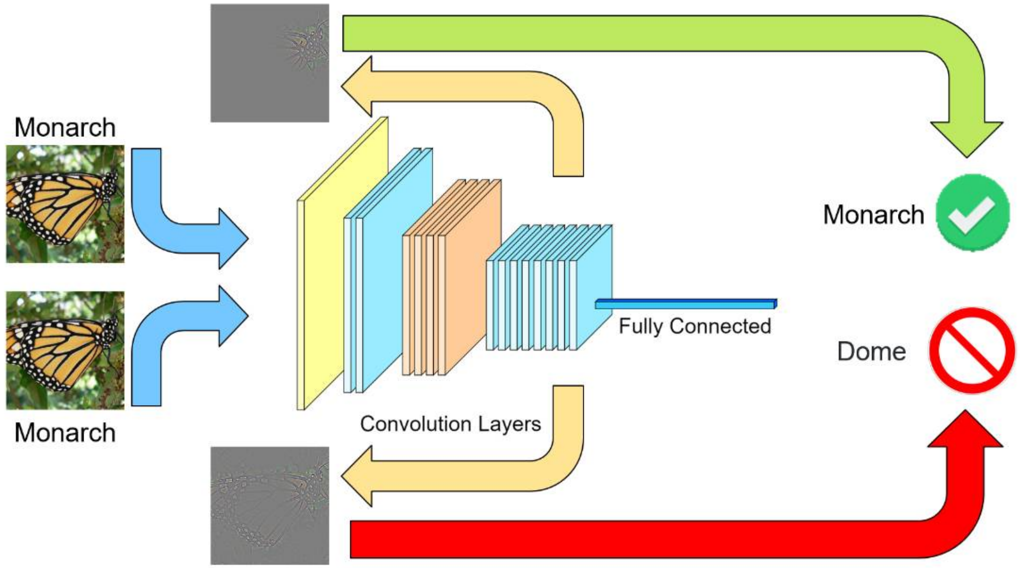
➔ Happy to answer questions & requests.

APPENDIX

Any Question?







datalab
www.zhaw.ch/datalab

BW-TEC
ENGINEERING AUTOMATION CONSULTING

data service
Swiss Alliance for
Data-Intensive Services

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency