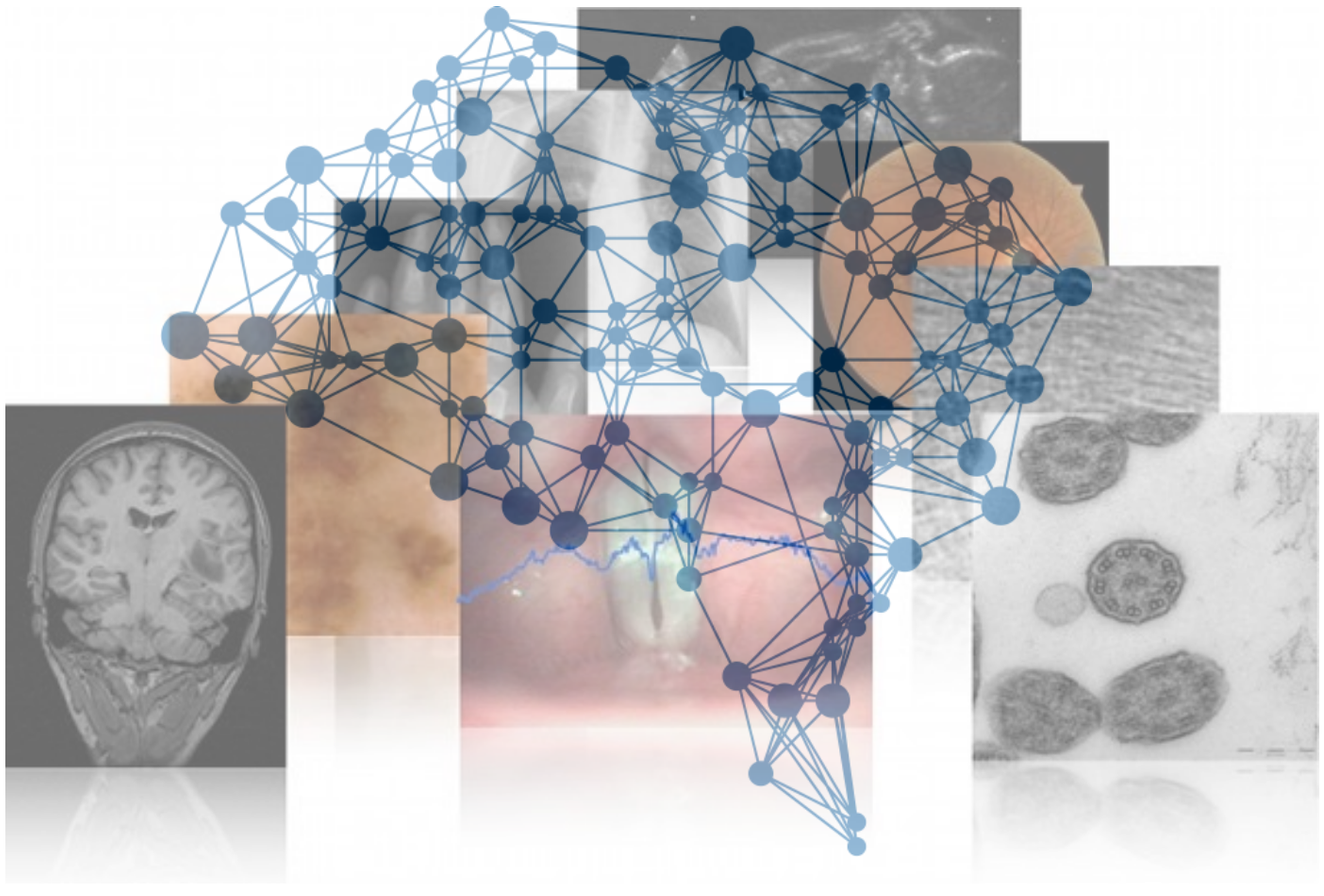
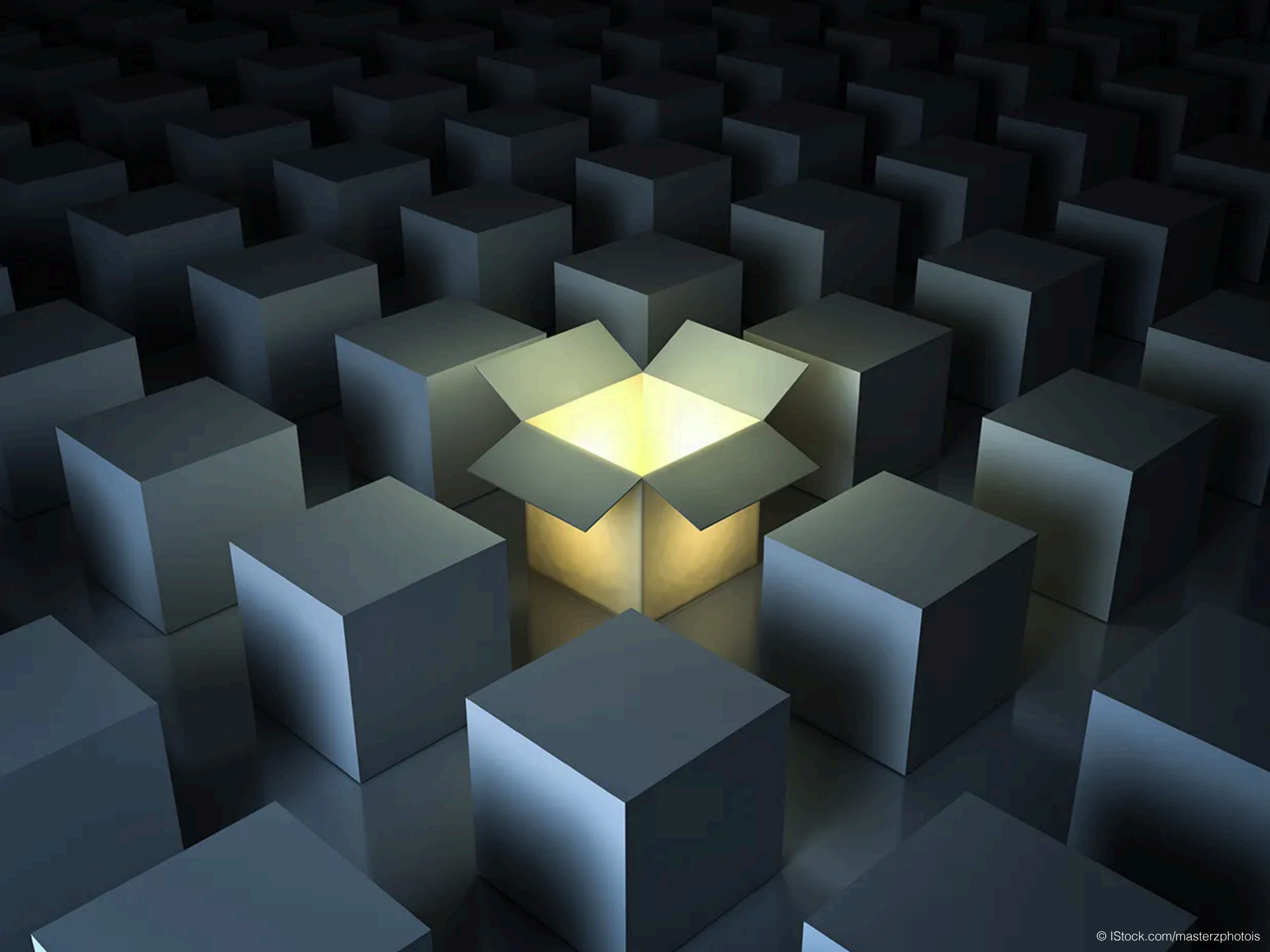


Explainable AI

1. **Einführung**
2. Modell-spezifische Methoden
3. Modell-unspezifische Methoden



Clever Hans

© Wikipedia

Interpretierbarkeit

Kann ein Mensch den Grund für eine Entscheidung verstehen?

Kann ein Mensch die Entscheidung eines Modells vorhersagen?

© Dr. Dataman

Erklärbarkeit

Ansätze, um interne Funktionen oder Abläufe des Modells zu verdeutlichen

Explainable AI

Warum ist Interpretierbarkeit wichtig?

↳ Akzeptanz erhöhen in public mind

↳ Debugging (Bias z.B. erkennen)

↳ Fairness gewährleisten z.B. KI entscheidet über Kredit
→ muss objektiv und fair sein

Explainable AI

sehr komplex → sehr genau → nicht einfach interpretierbar

↳ post-hoc-Explanation

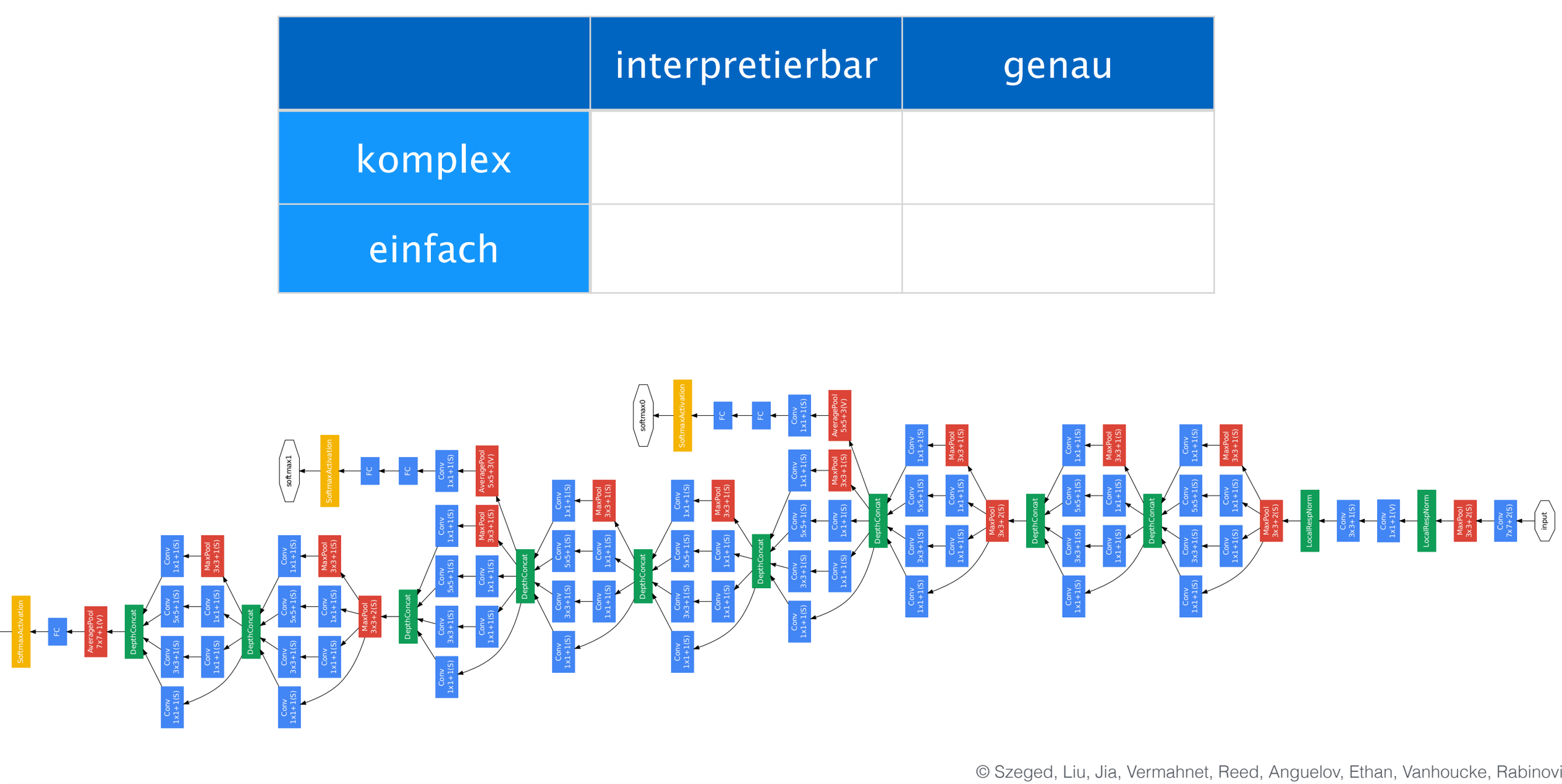
	interpretierbar	genau
komplex	✗	✓
einfach	✓	✗

Einfache Modelle genauer zu machen

oder komplexe Modelle versuchen zu interpretieren = post-hoc-explanation

Explainable AI

	interpretierbar	genau
komplex		
einfach		



© Szeged, Liu, Jia, Vermahnet, Reed, Anguelov, Ethan, Vanhoucke, Rabinovich

Explainable AI

Post Hoc Explanation

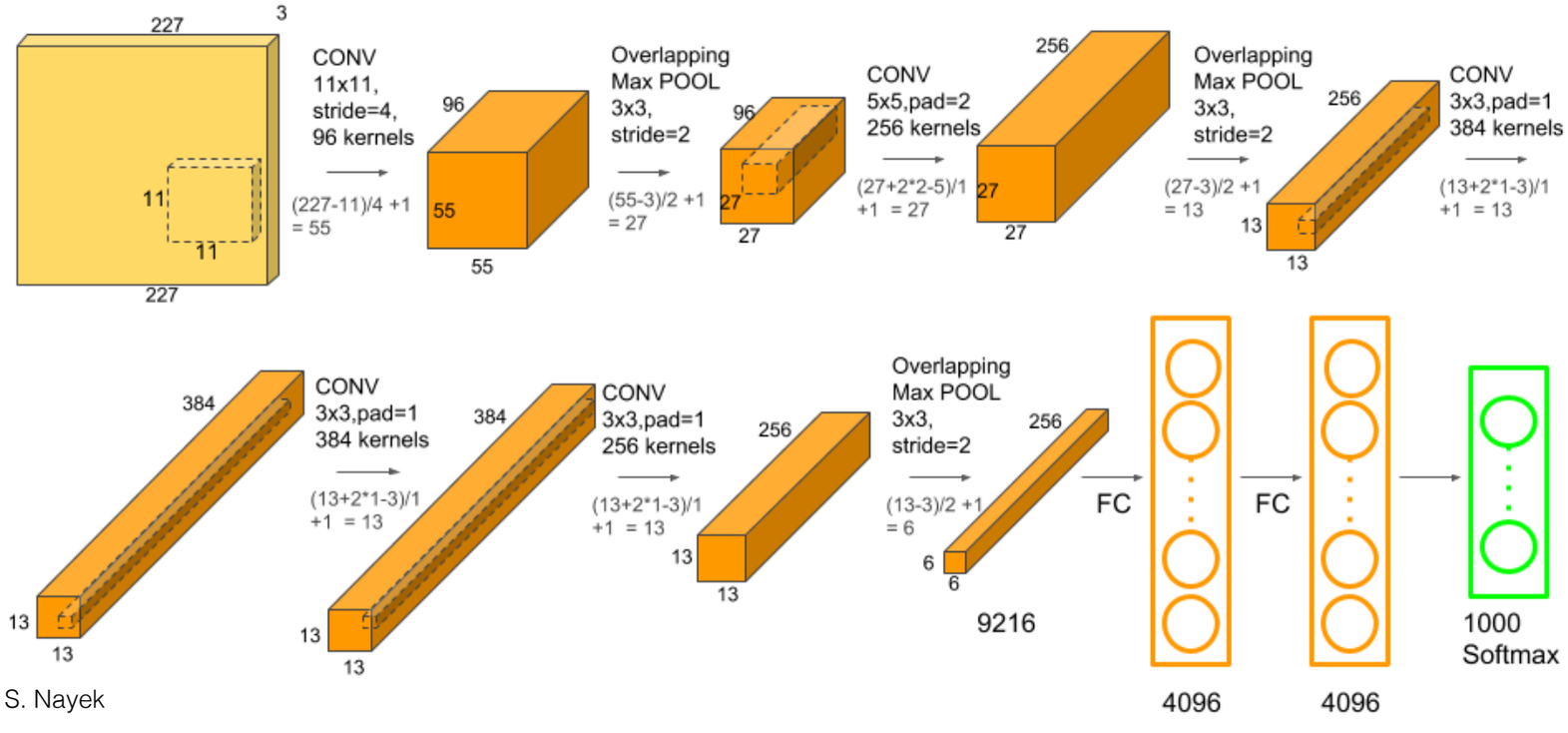
↳ je tiefer und komplexer ein NN, umso schwieriger ist die Interpretation

Explainable AI

1. Einführung
2. **Modell-spezifische Methoden**
3. Modell-unspezifische Methoden

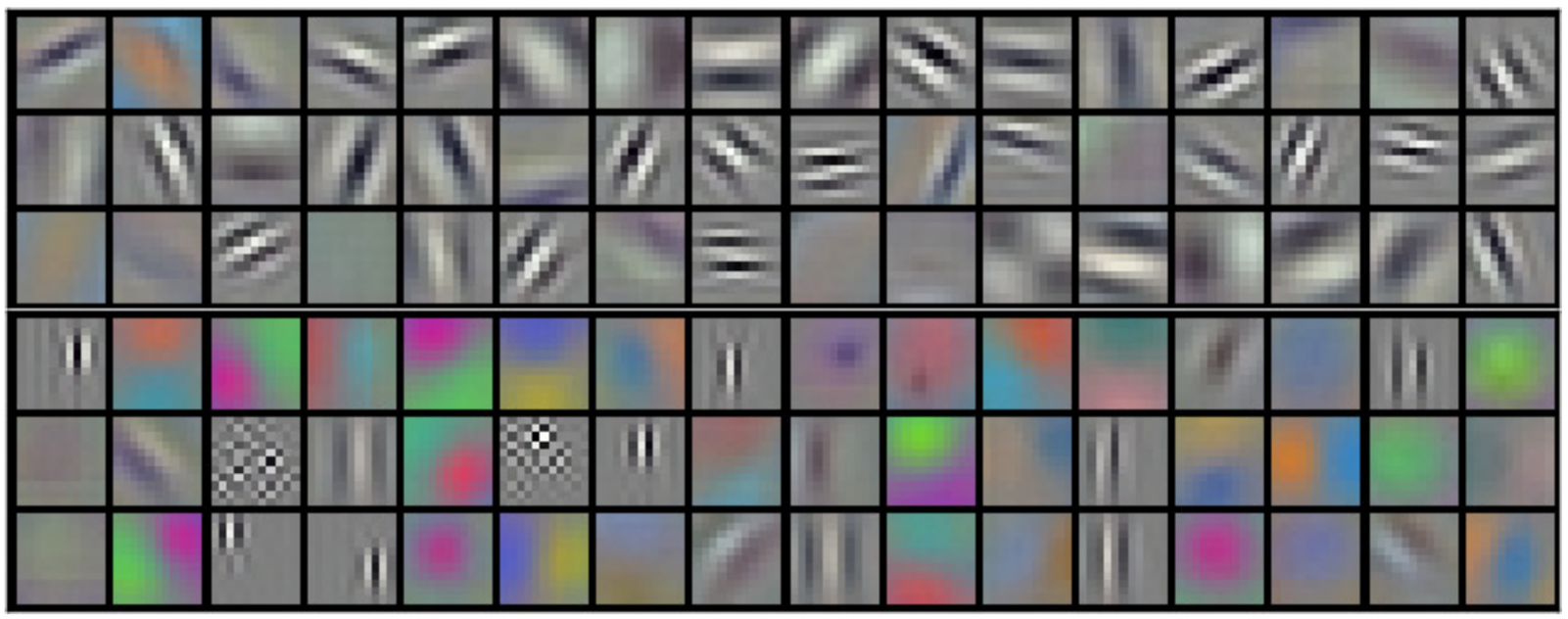
vs.

Visualisierung der Gewichte



Alex Net

Gewicht 1. Schicht, Faltungskerne visualisiert

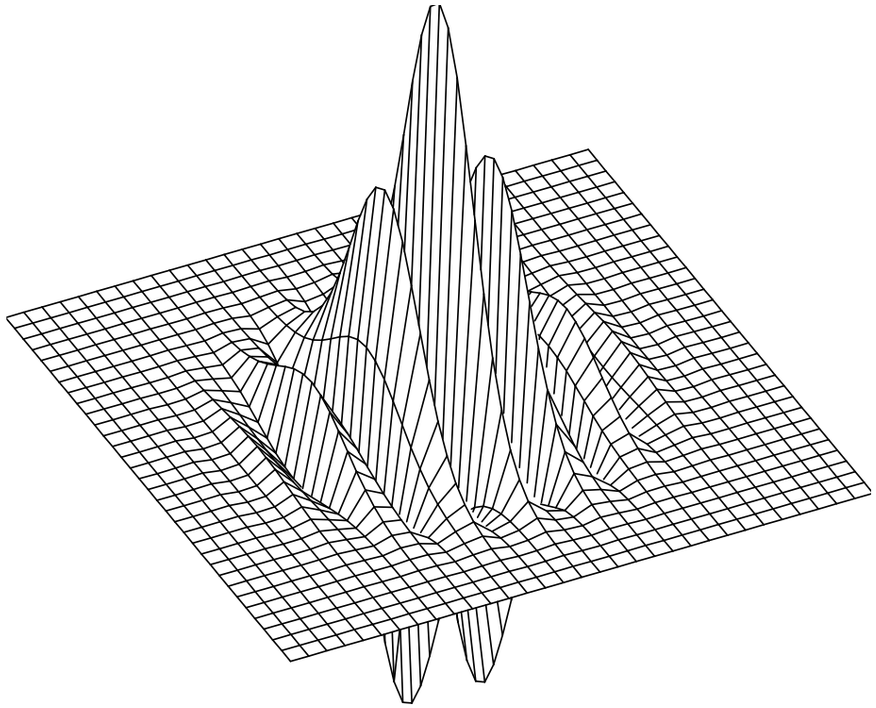


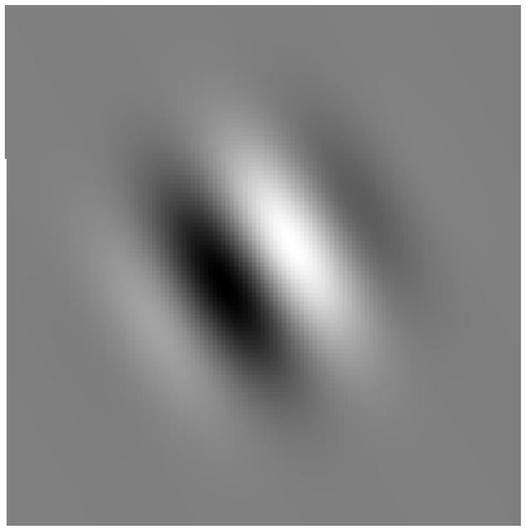
© A. Krizhevsky et al., 2012

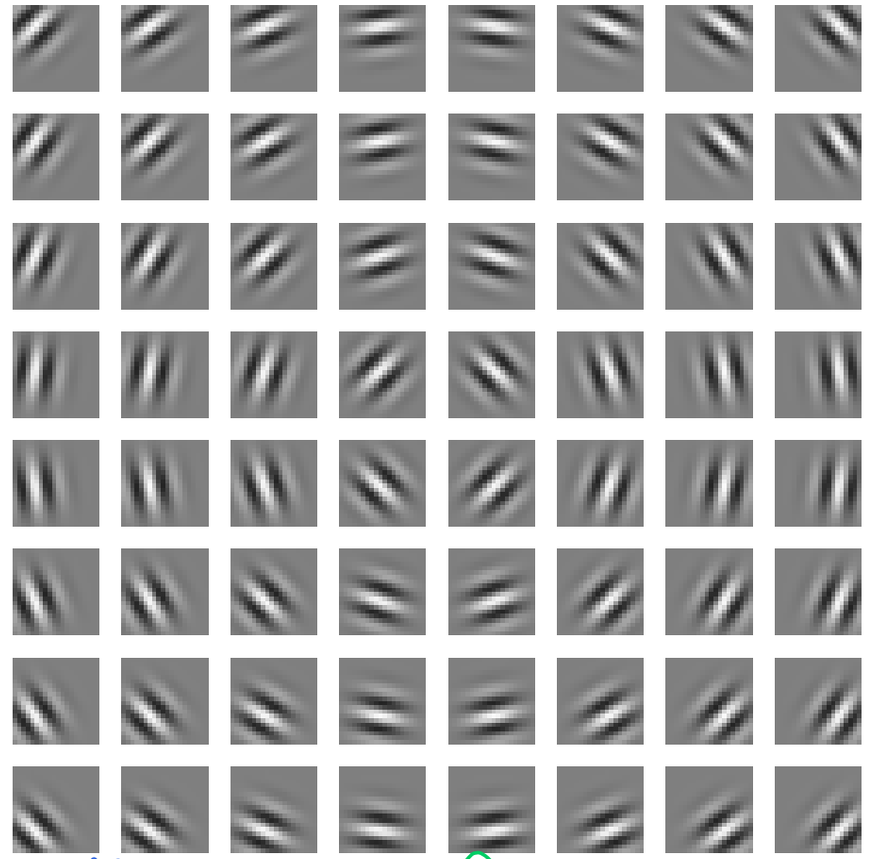
© S. Nayek

Visualisierung der Gewichte

Gabor-Filter







Filterbank

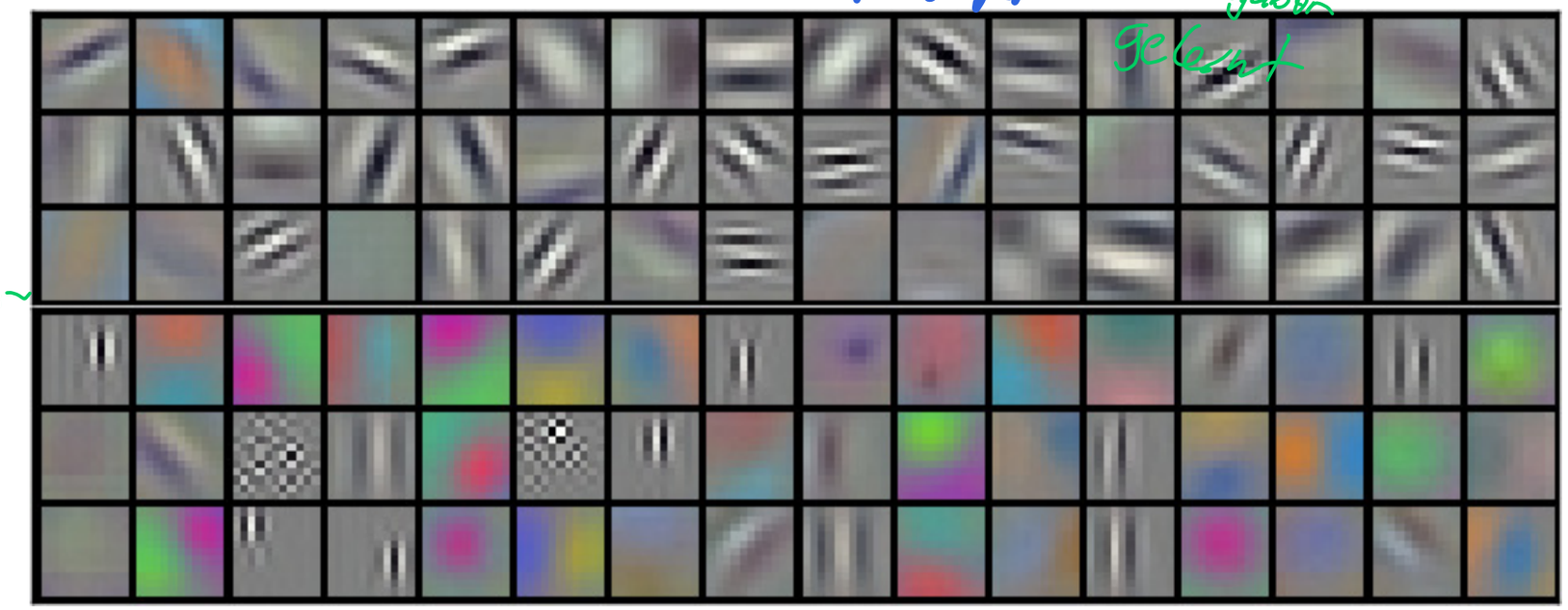
Linie wie Gabor gelernt

© Goodfellow

Gabor-Wavelet

Fouriertransformation

globale Operation, wo Bild in Frequenzraum überführt wird



© A. Krizhevsky et al., 2012

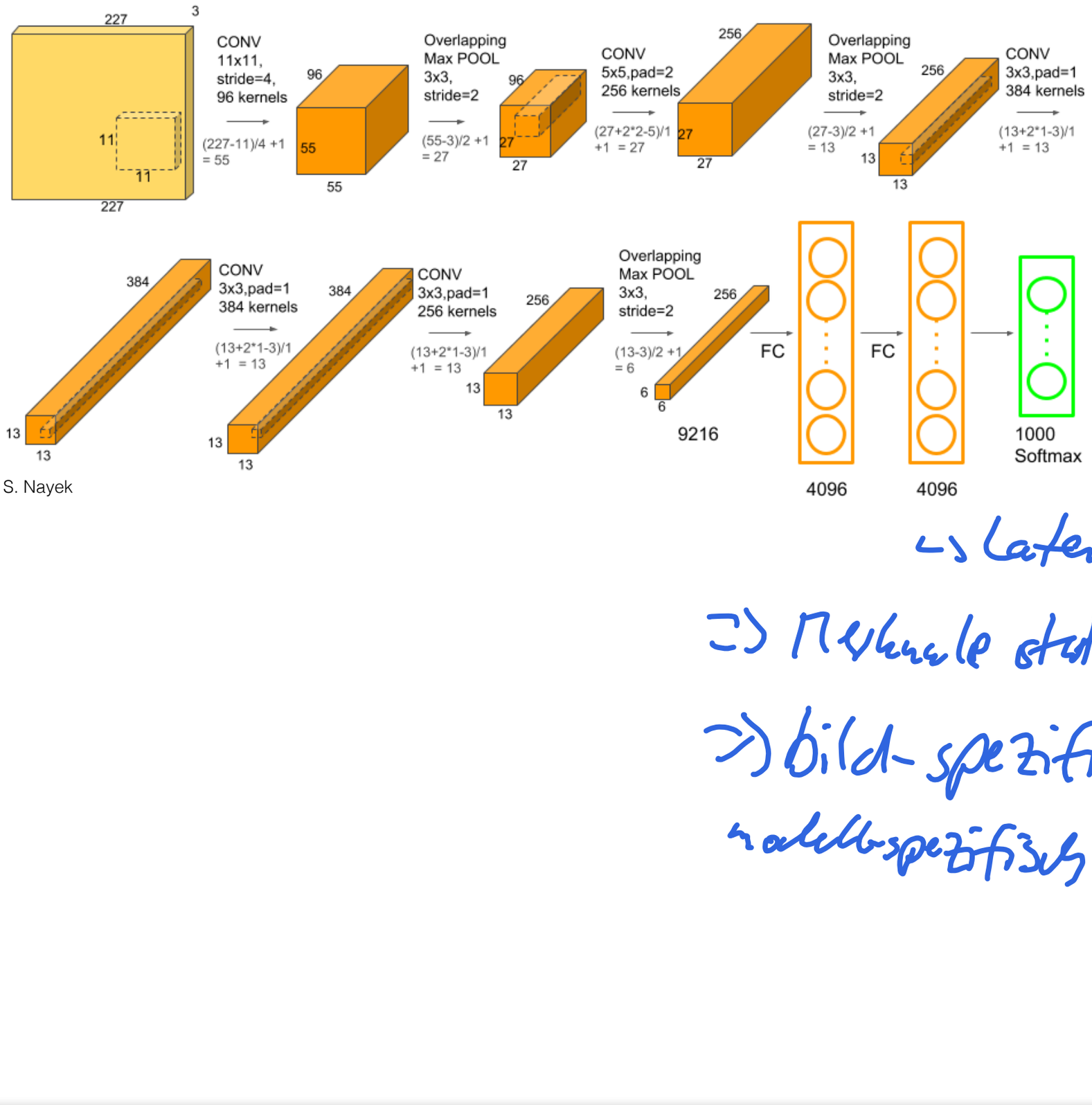
Visualisierung der Gewichte

Faltungskerne höherer Schichten

random noise-like

- ned so gute Idee die zu visualisieren

Visualisierung der Merkmale



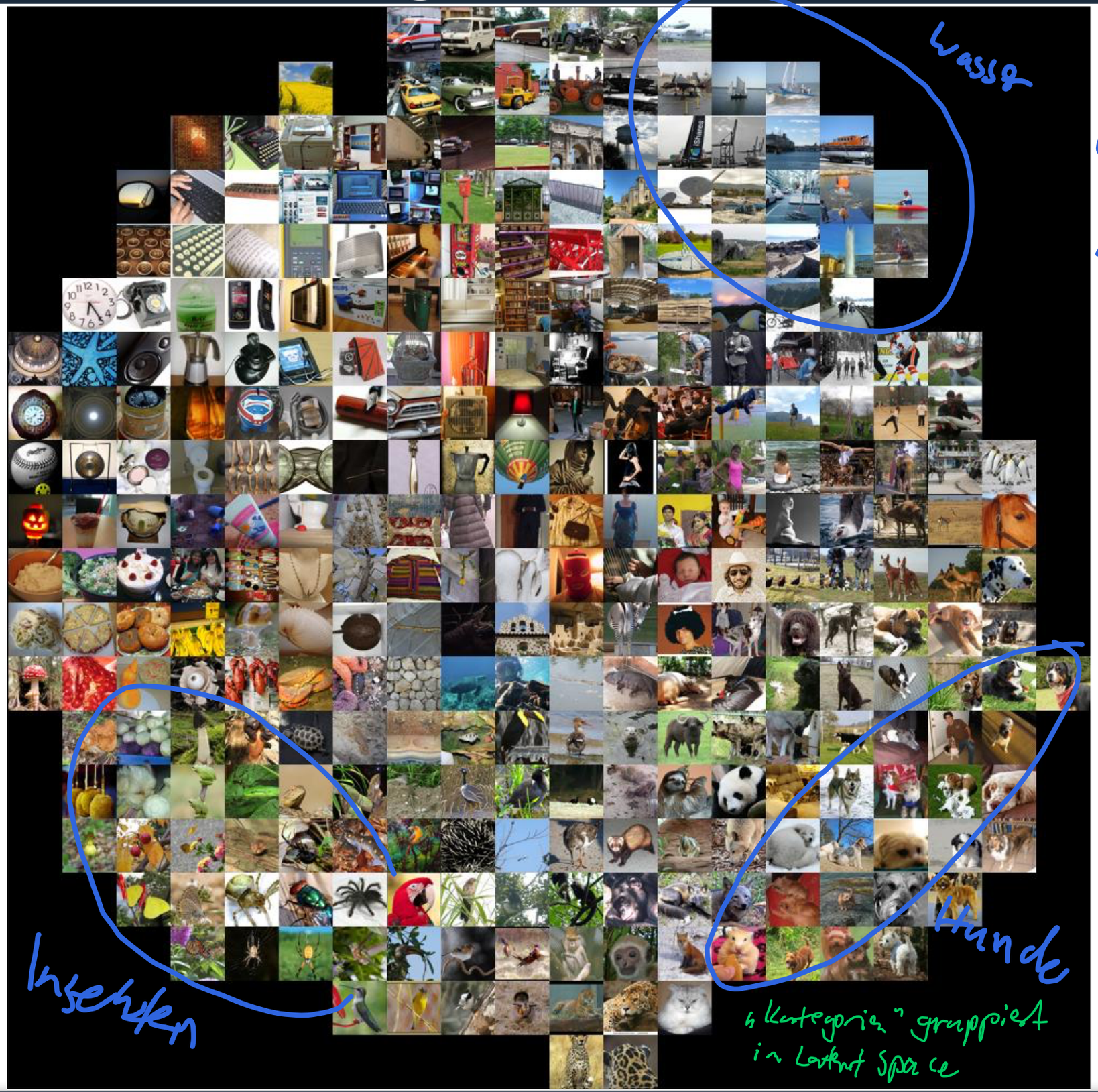
↳ Latent Space mal visualisieren?

⇒ Merkmale statt Gesichter ansehen

⇒ bild-spezifisch statt

modellspezifisch für jedes Bild anderer Datenpunkt im Latent Space

Visualisierung der Merkmale



t-SNE-Visualisierung des Latent Space des AlexNet

evo Datensatz (Label wurde des gesuchte Bild dargestellt davon)

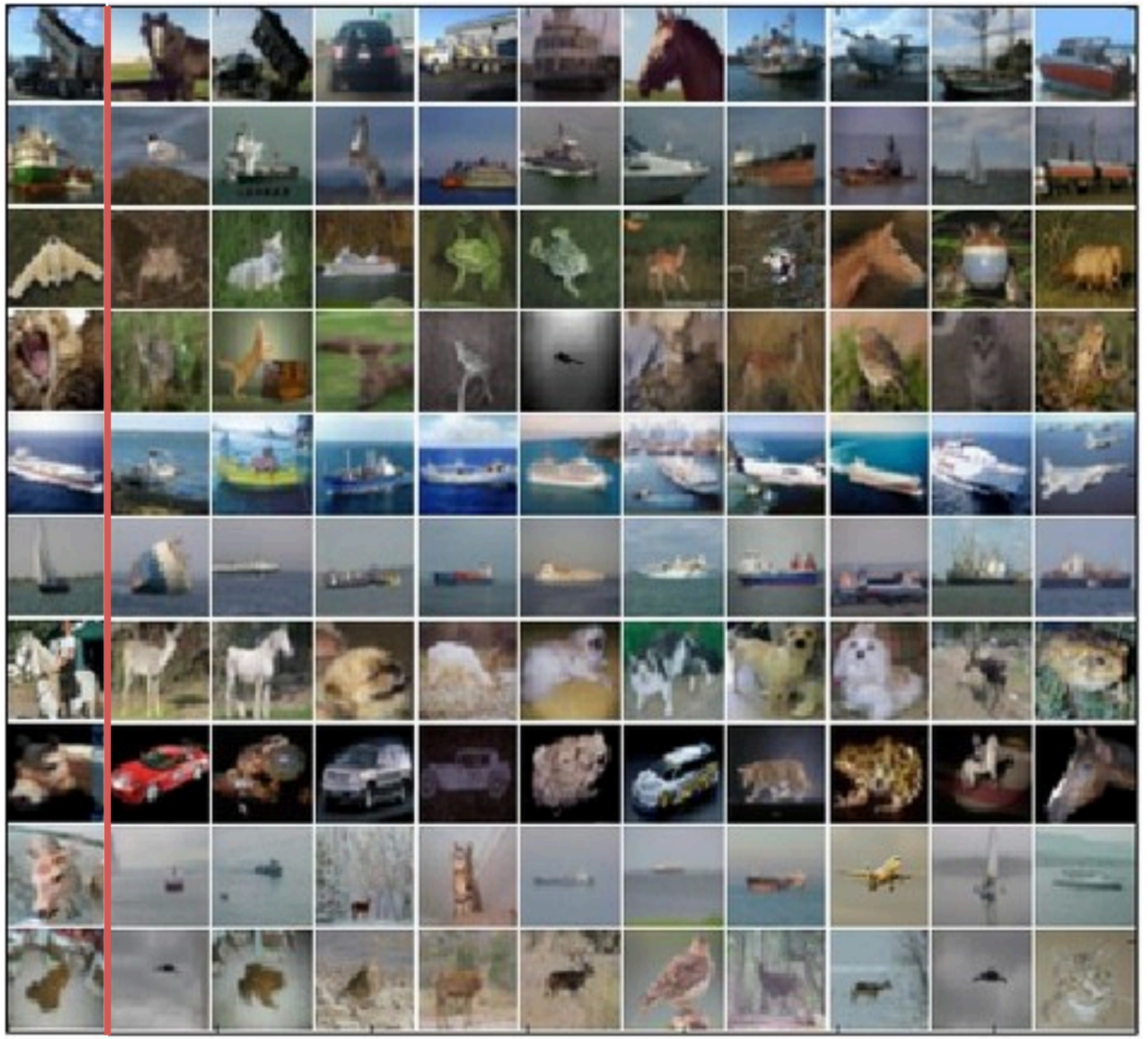
=) Bild als Datenpunkt

Visualisierung der Merkmale

andere Methode:

Testbilder

L_2-Norm auf Pixelbasis



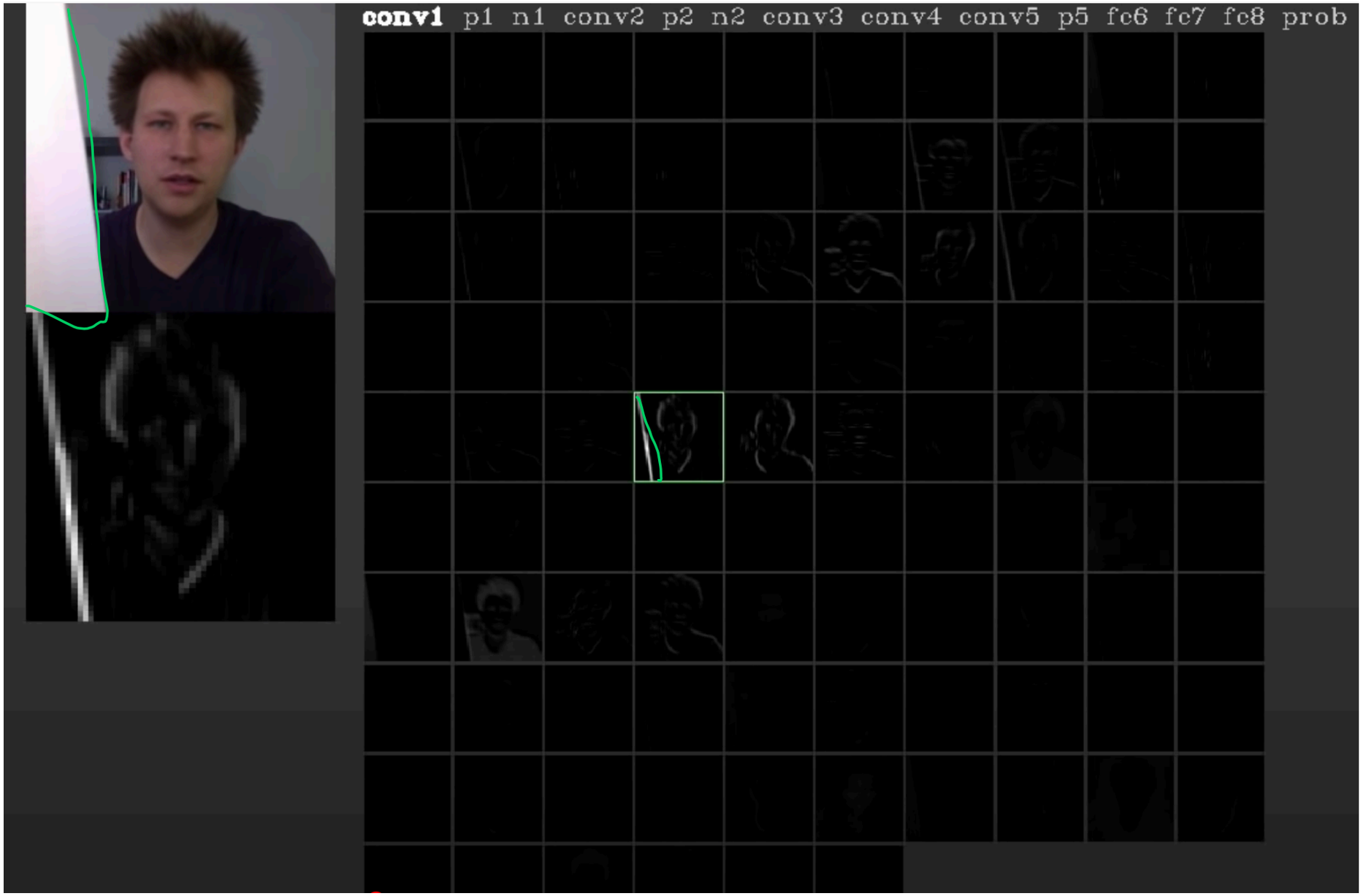
© Andrej Karpathy

Visualisierung der Merkmale

L_2-Norm auf Merkmalsebene des Latent Space

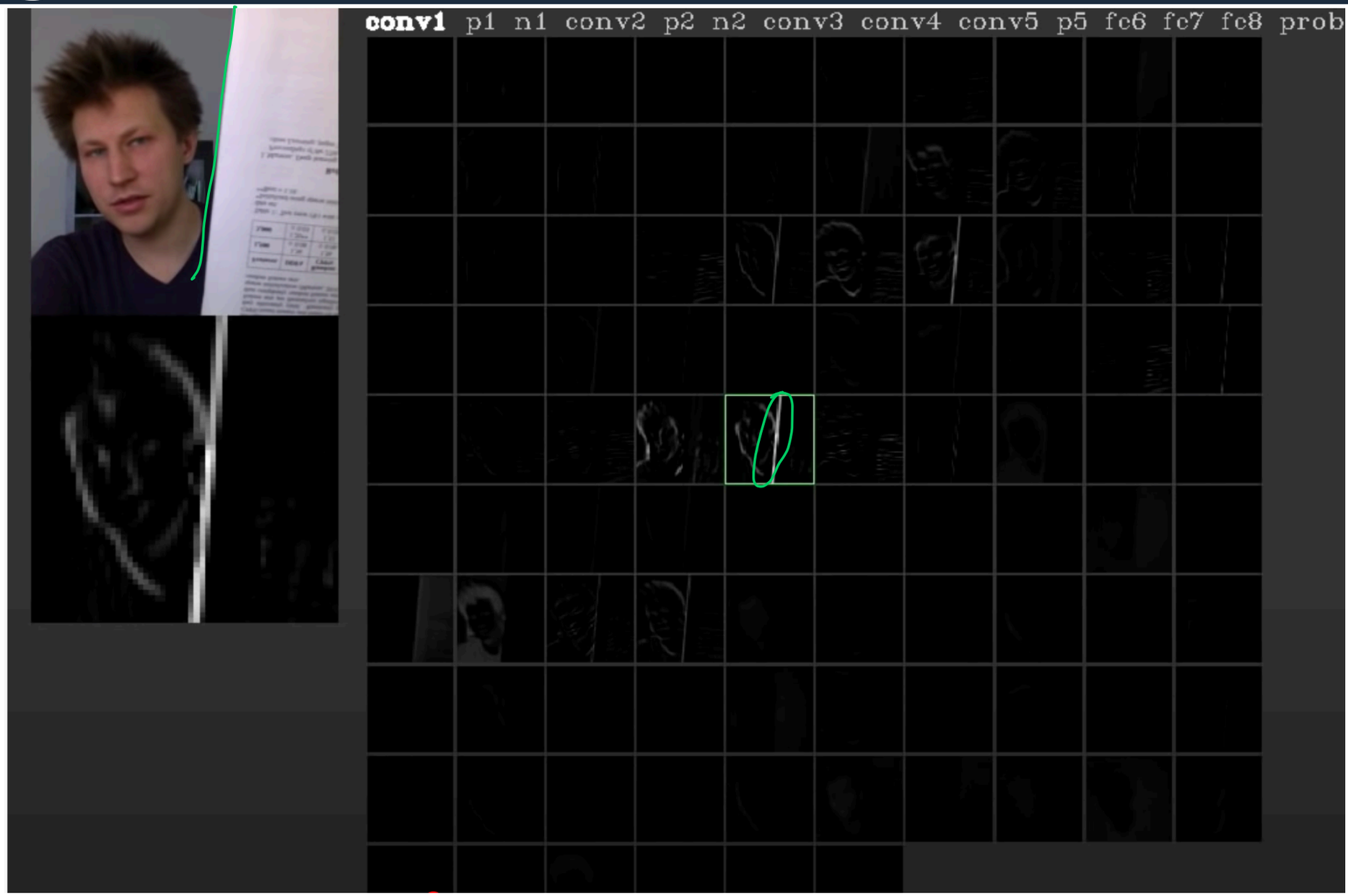
© A. Krizhevsky et al., 2012

Visualisierung der Merkmale



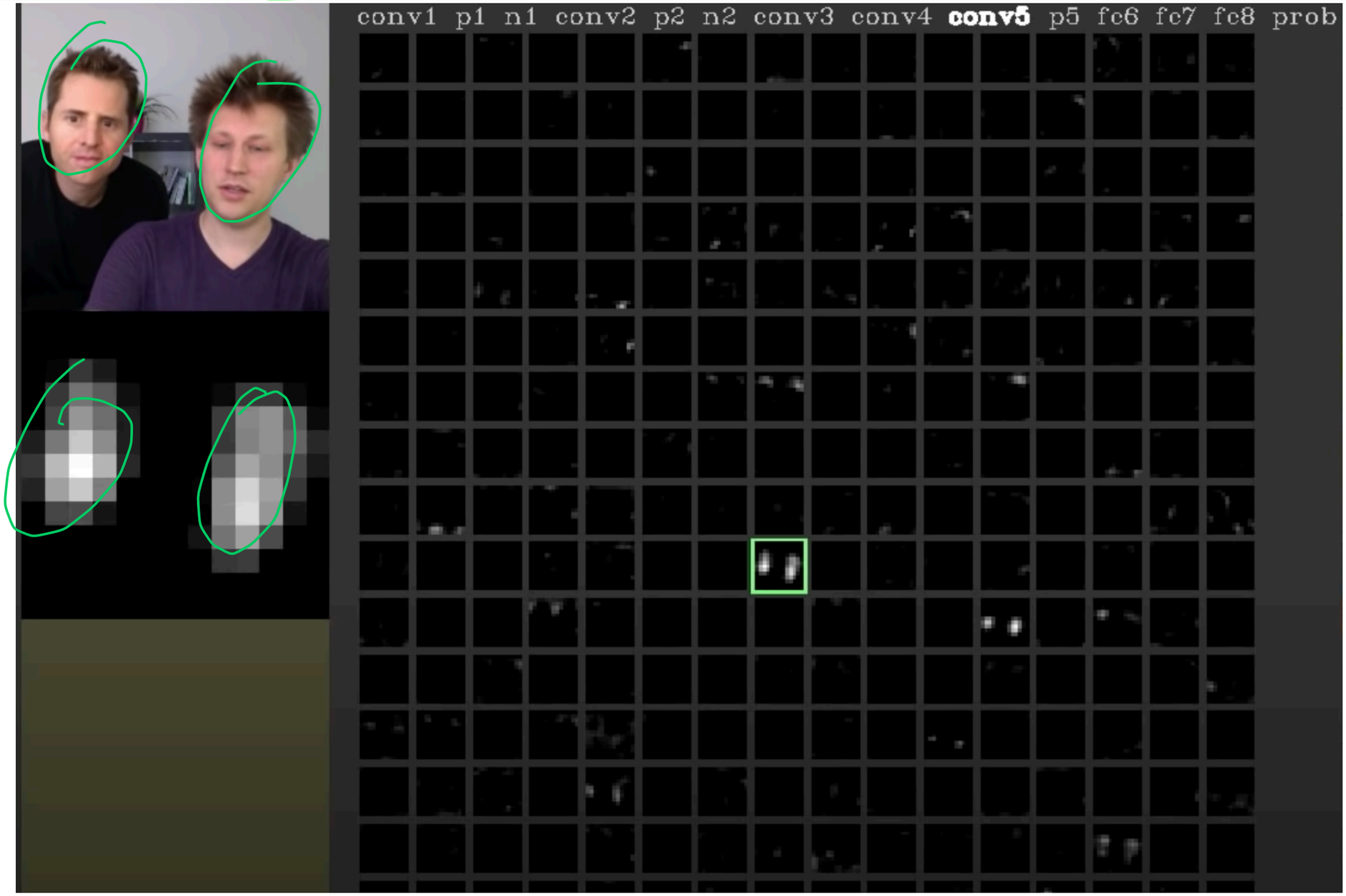
© Yosinski et al.

Visualisierung der Merkmale



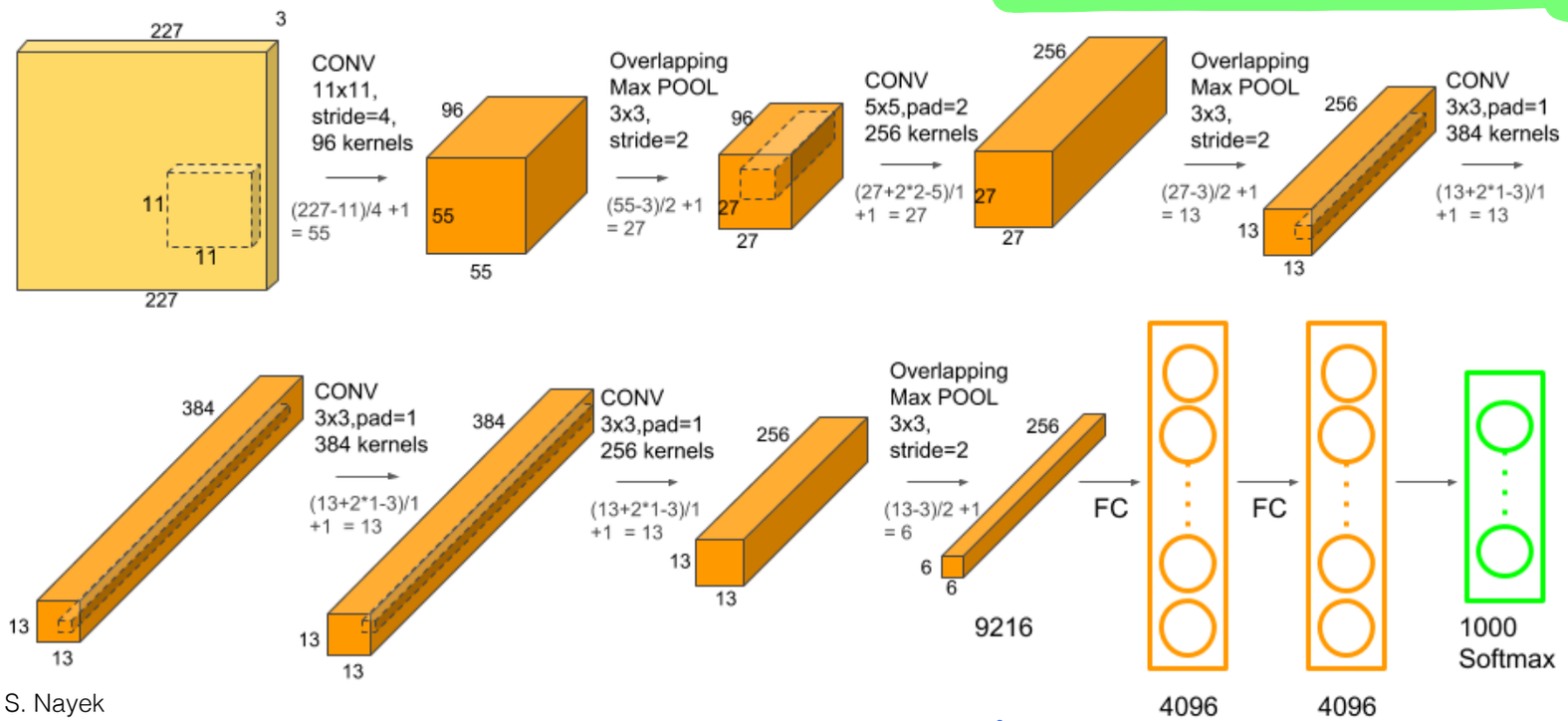
© Yosinski et al.

Visualisierung der Merkmale



↳ Interpretation im Nachhinein experimentell

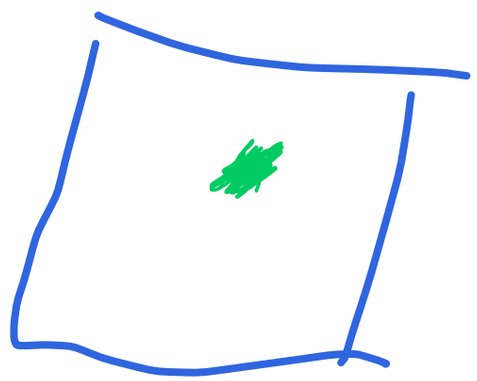
Visualisierung von relevanten Bildregionen



© S. Nayek

4 wenn ich den nicht hätte, würde ich damit zu der endgen Entscheidung kommen"

finden von maximal aktivierenden Bildausschnitte bzgl. einer Merkmalskarte



→ Maximal aktivierende Bildausschnitte zu bestimmter Merkmalskarte finden

(a) Merkmalskarte wählen

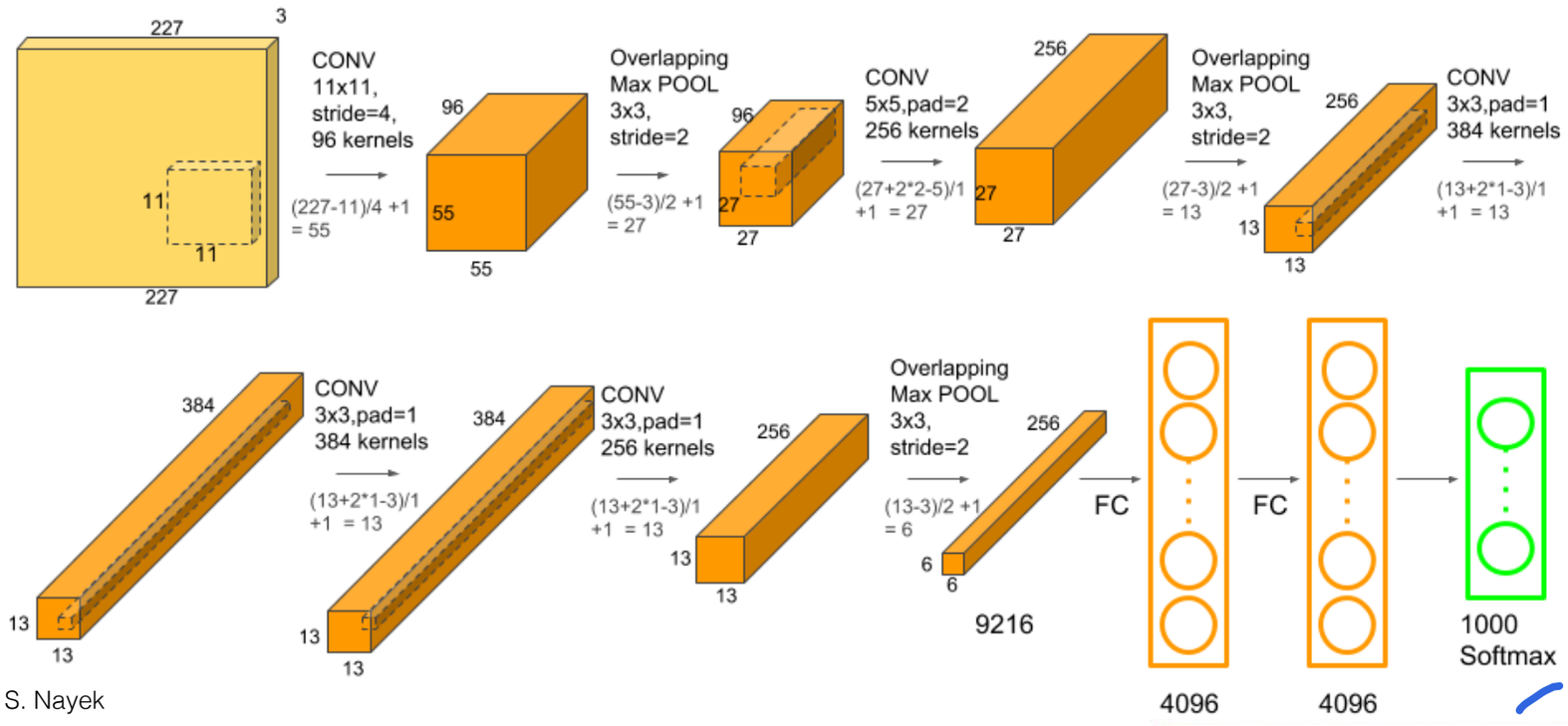
(b) Testbilder feed-forward verarbeiten

(c) Bilder und Position der maximalen Aktivität finden

(d) rückwärts aus Kenntnis der relevanten Positionen in Bild navigieren

© Yosinski et al.

Visualisierung von relevanten Bildregionen



© S. Nayek

Top 10

© Springenberg et al., 2015

Lineares Modell leicht erklärbar

$$y = g(\boldsymbol{x}) = \phi_0 + \sum_i \phi_i x_i$$

(Weight, Input)

$1{,}05 \cdot x_1 + 1{,}5 \cdot x_2 \Rightarrow 127.000$

Annahme: x_1, x_2 stochastisch unabhängig

$\Rightarrow$ 1,05 Bedeutung von x_1

$\Rightarrow$ 1,5 Bedeutung von x_2

Bestimmung: $R_i(x) = \frac{dy}{dx_i}$

$\Rightarrow$ Partielle Ableitung von y nach x_i

Saliency Maps

$$R_i^c(\boldsymbol{x}) = \left| \frac{\partial S_c(\boldsymbol{x})}{\partial x_i} \right|$$

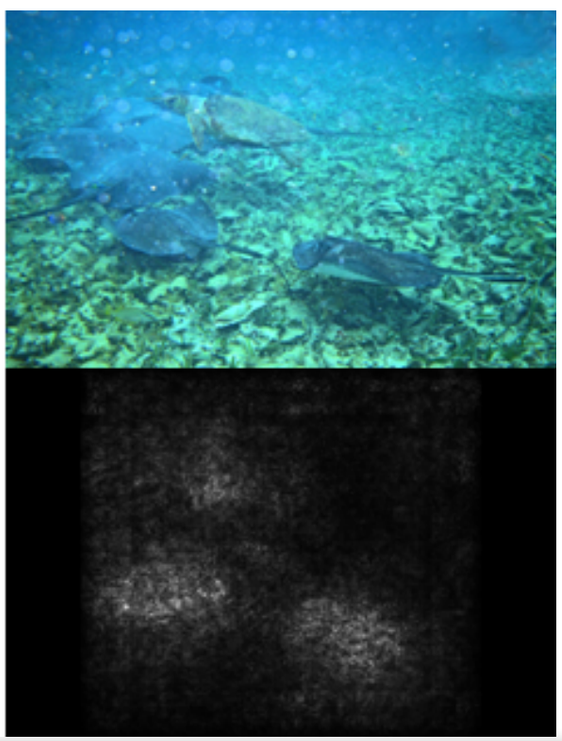
$S_c(x)$: Zieloutput des NN für Klasse c

x_i: Pixel

=) Welche Pixel müsste verändert werden, um den Score einer Klasse maximal zu beeinflussen

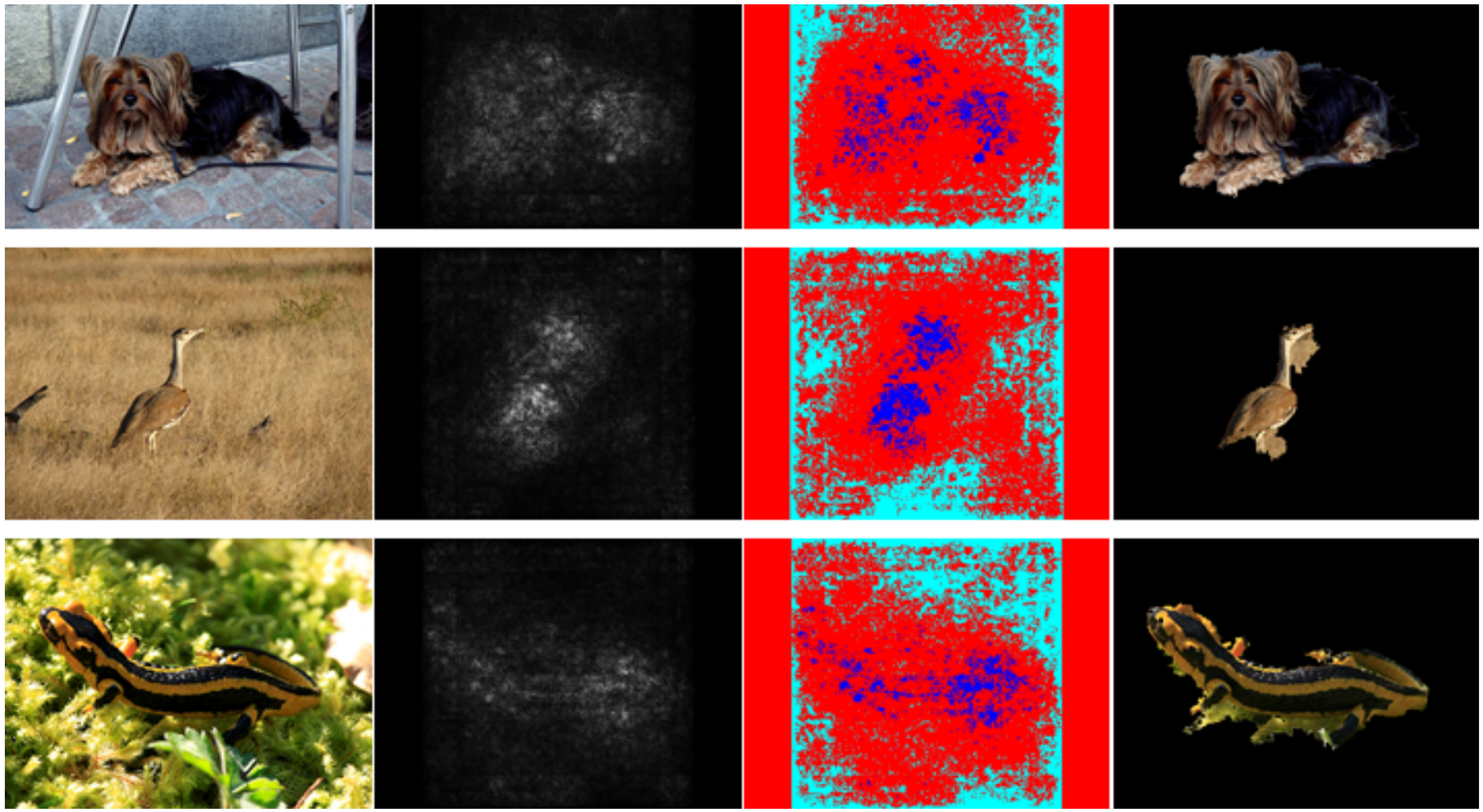
© Simonyan et al., 2014

Saliency Maps



© Simonyan et al., 2014

Saliency Maps



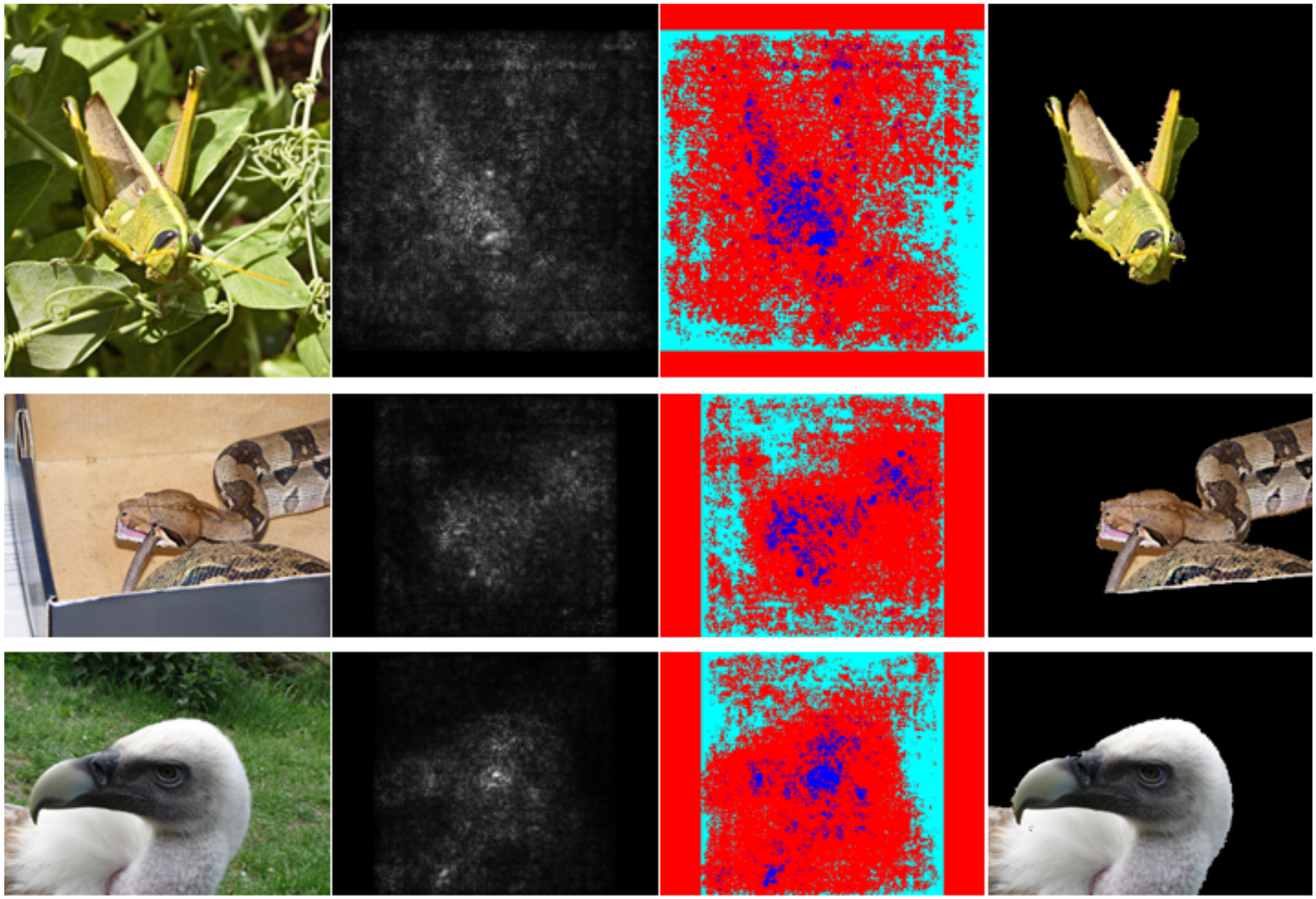
Original

Saliency map

Thresholding:
-> blau: Objekt
-> Cyan: Hintergrund
-> rot: unentschieden

Segmentierung
Graph-Cut-Algorithmus

Saliency Maps



$$y = g(\boldsymbol{x}) = \phi_0 + \sum_i \phi_i x_i$$

$y = 120.000$

$\phi_1 = 1{,}05$

$\phi_2 = 1{,}5$

$x_1 = 100.000$

$x_2 = 10.000$

weight ... ist fast nur von x_1 bestimmt

$\phi_1 \cdot x_1 = 105.000$

viel weniger

$\phi_2 \cdot x_2 = 15.000$

$R_i(x) = \underbrace{x_i}_{\text{weight}} \cdot \underbrace{\frac{\partial y_c}{\partial x_i}}_{\text{...}}$

Gradient x Input

$$R_i^c(\boldsymbol{x}) = \frac{\partial S_c(\boldsymbol{x})}{\partial x_i} \cdot x_i$$

Saliency · x_i angelängt







Gradient x Input

$$R_i^c(\boldsymbol{x}) = \frac{\partial S_c(\boldsymbol{x})}{\partial x_i} \cdot x_i$$




hier auch wieder

Lösung: Integrated Gradients ↓

Integrated Gradients

nicht um eine Intensität soll das Bild dominieren

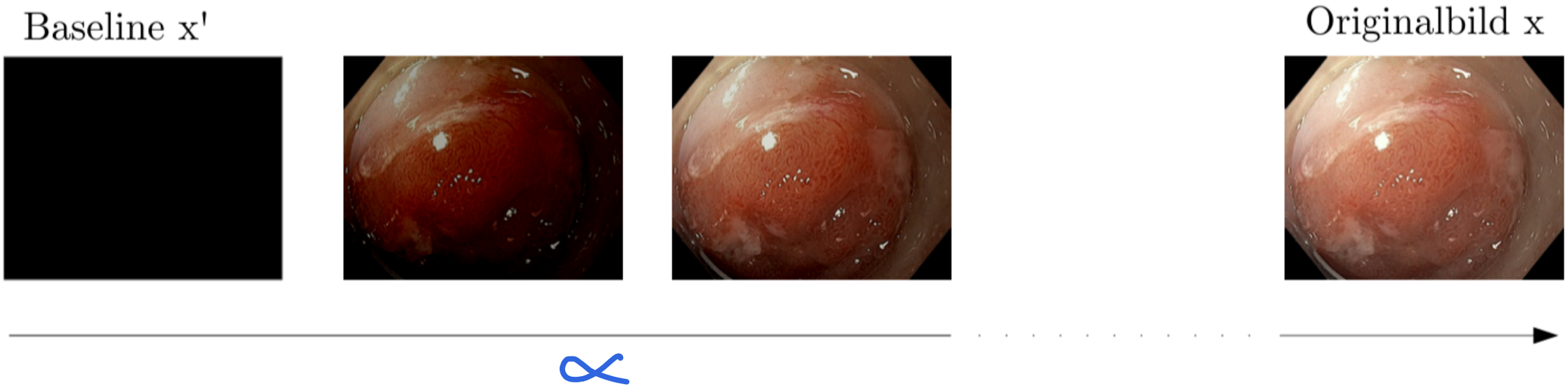
Input wird mit α skaliert

$$R_i^c(\boldsymbol{x}) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}'))}{\partial x_i} d\alpha$$

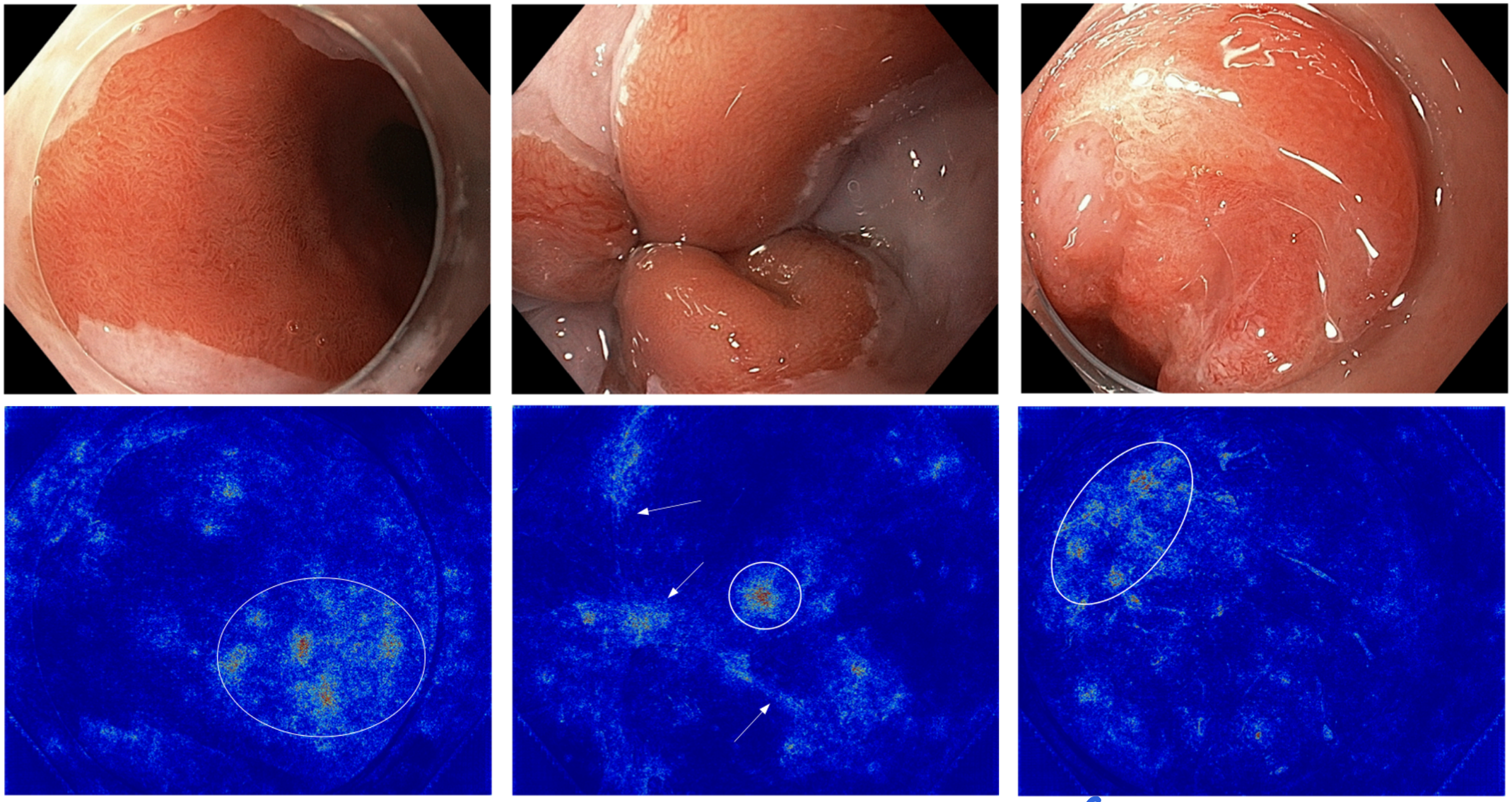
$$x_i \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\alpha \cdot x)}{\partial x_i} \cdot d\alpha$$

Baseline: Abwesenheit von Merkmalen

⇒ Wie stark erhöht sich die Zielgröße, wenn die Präsenz eines Merkmals verstärkt wird?



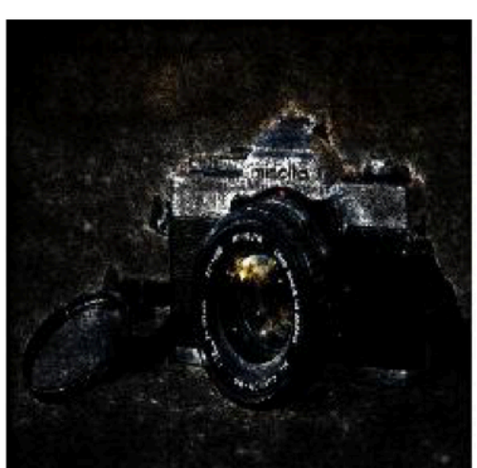
Integrated Gradients

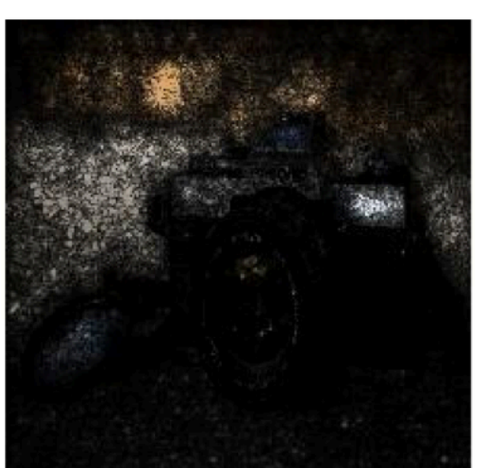


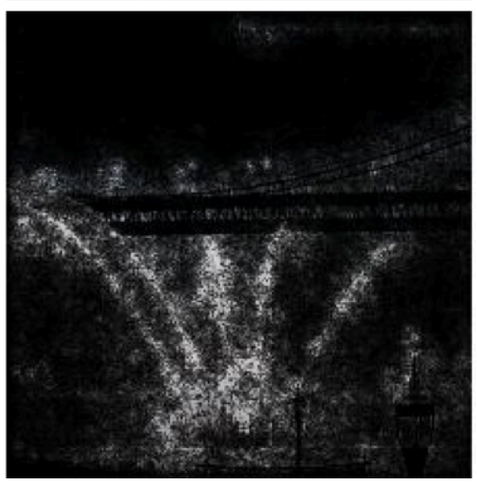
eher nicht so nützlich

Integrated Gradients

besser als vorher







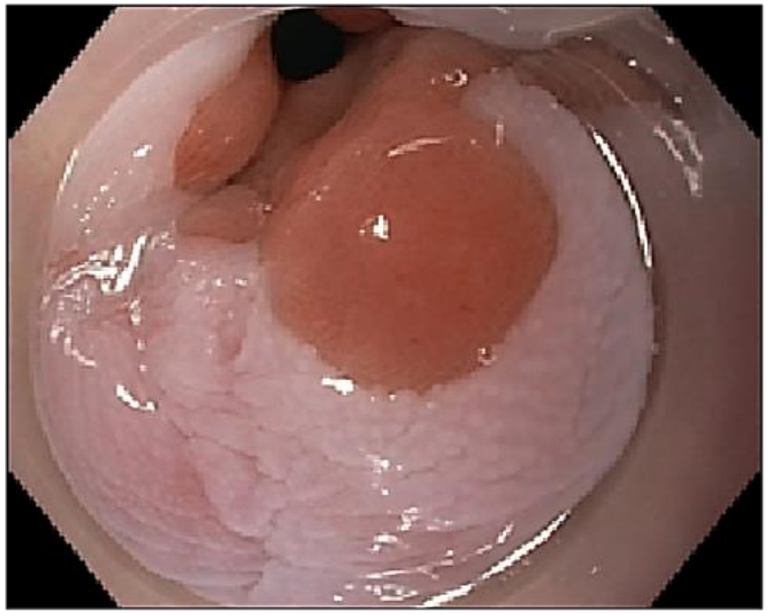
Integrated Gradients

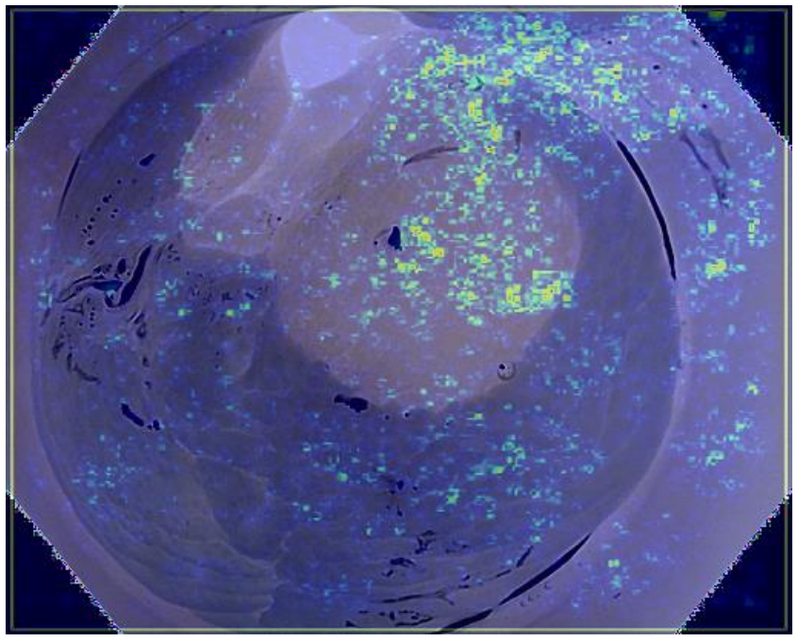
Gradient × Input

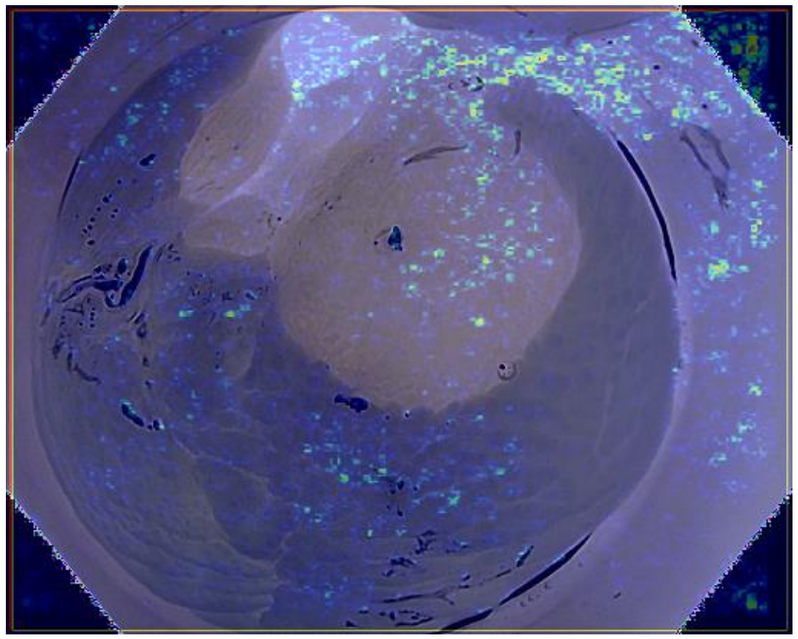
Integrated Gradients

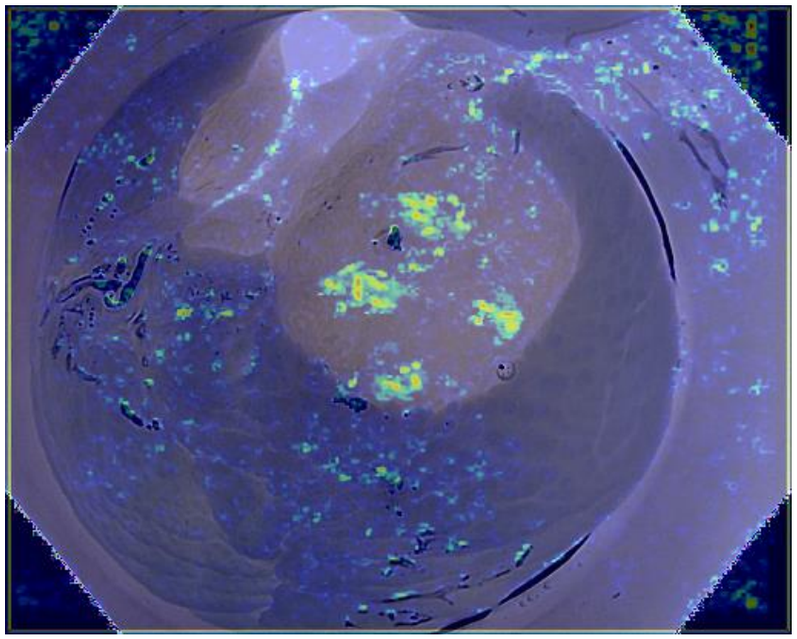
Integrated Gradients

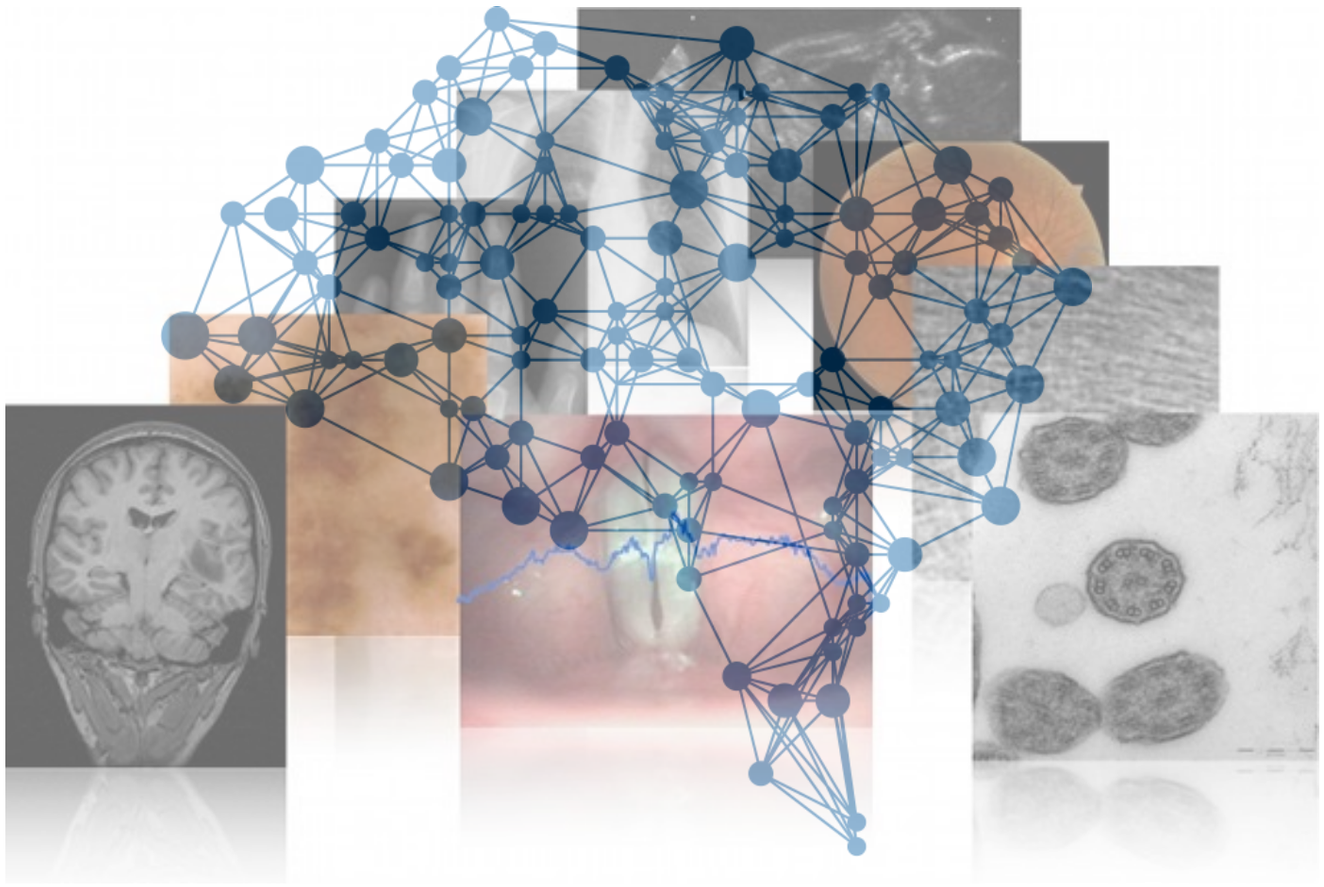
Überblick











Explainable AI

1. Einführung
2. Modell-spezifische Methoden
3. **Modell-unspezifische Methoden**

Fragen & Antworten

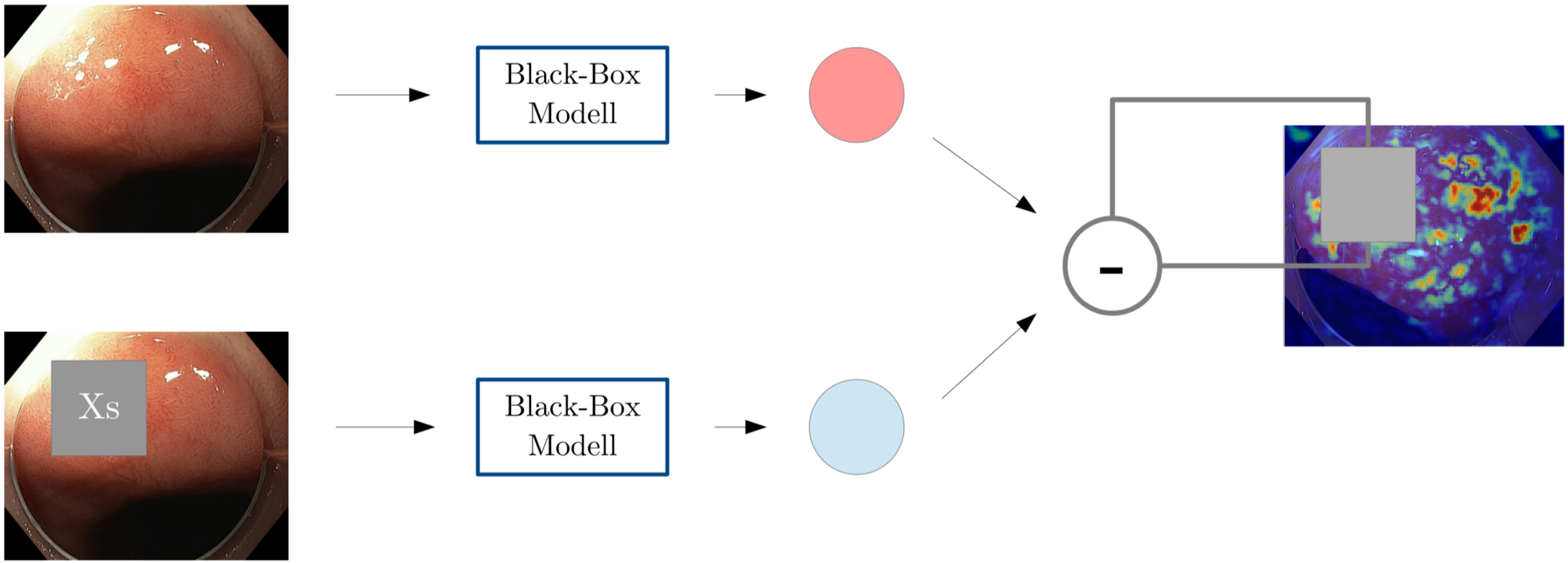
- Altklausuren nicht vorhanden
- Aktivierungsfunktion: macht aus linearem etwas nichtlineares
- fragt keine Formeln ab, stattdessen stellt er Formeln vor und man soll die dann erklären/beschreiben
- t-SNE Og (334) nicht erklären, wenn dann die angepasst mit $\frac{1}{4}$

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \rightarrow p_{ij} = \frac{+ p_{i|j}}{2N}$$

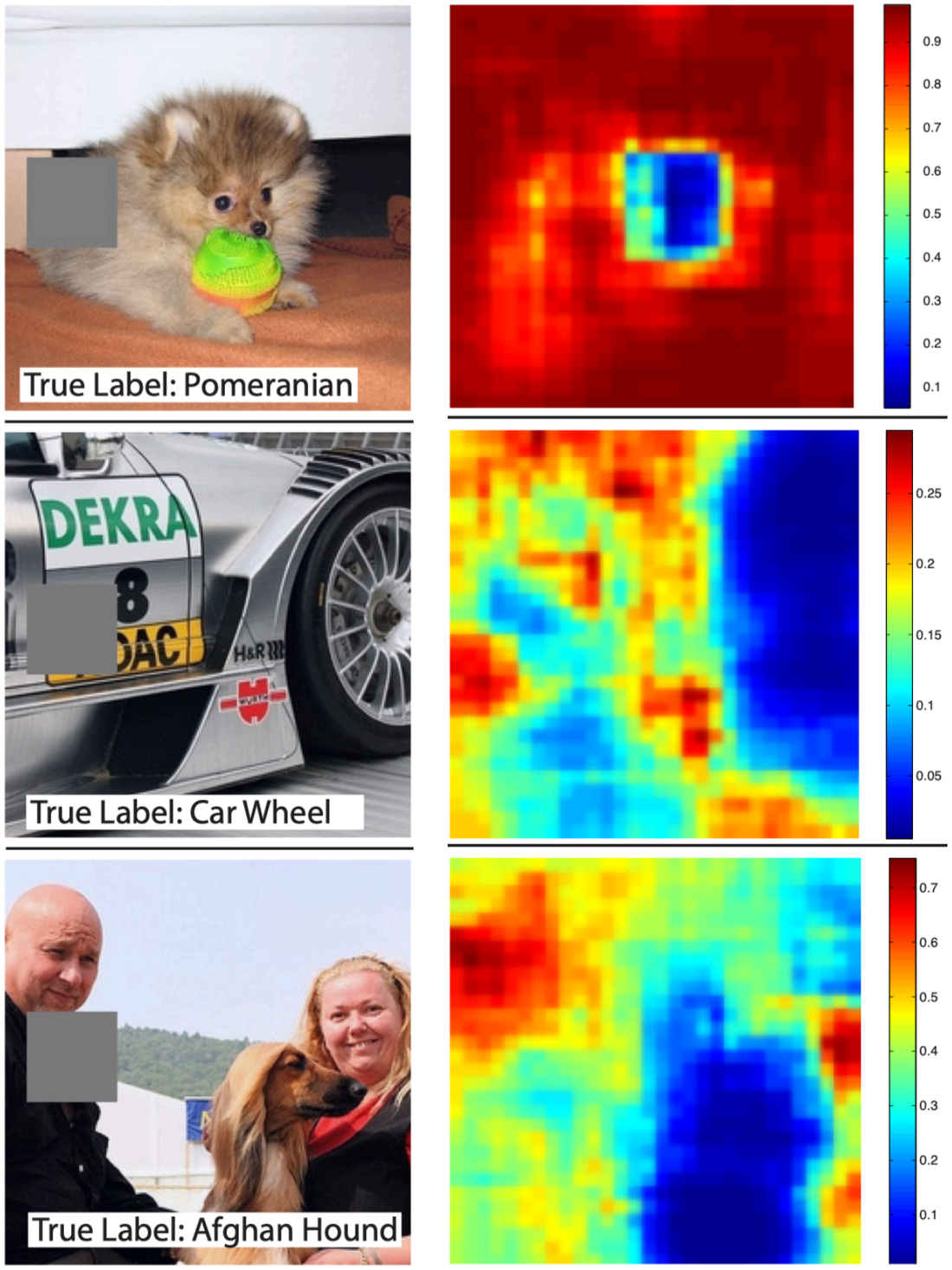
$\frac{1}{4}$ approximation

Okklusion

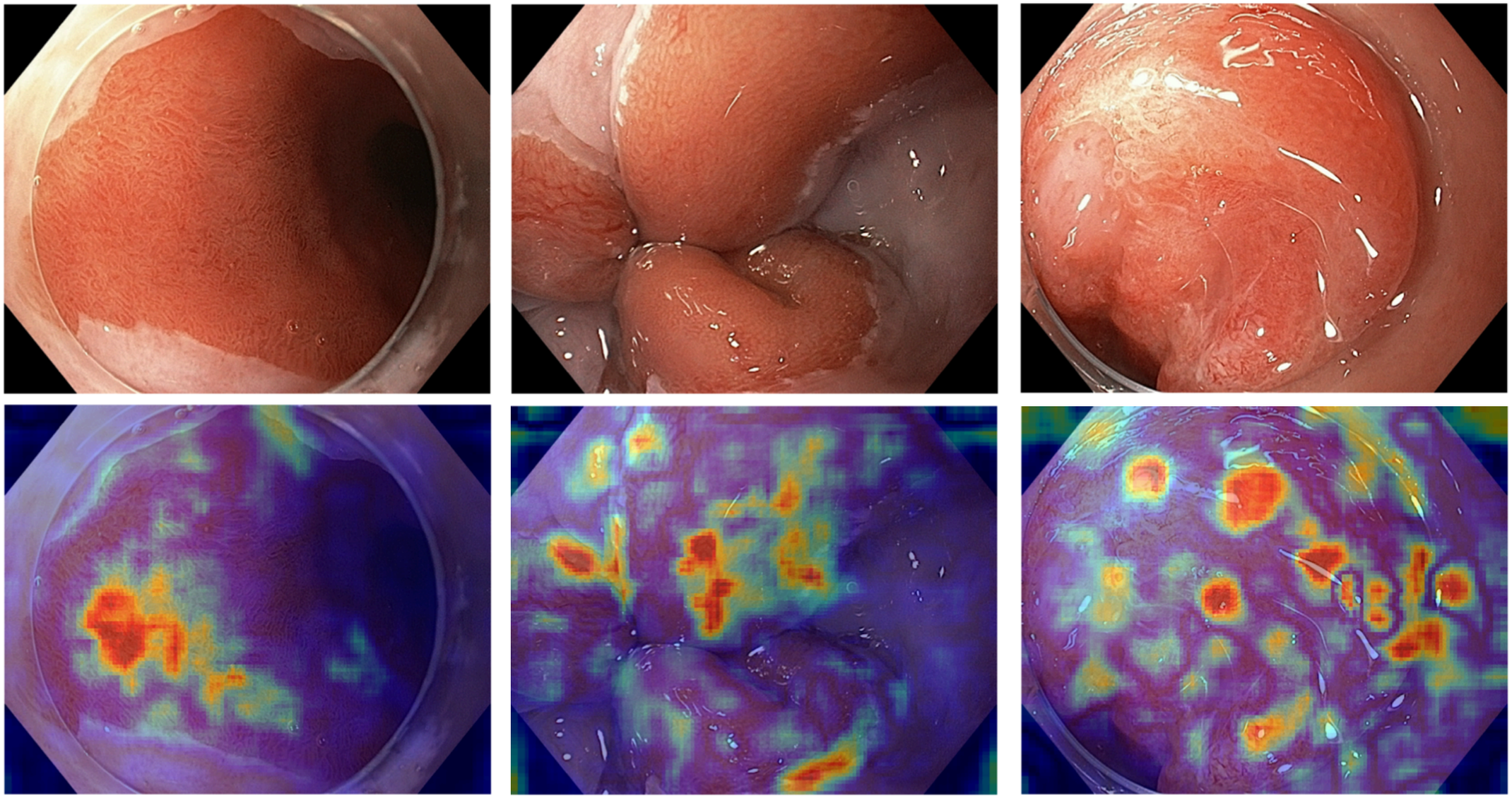
$$R_i^c(\boldsymbol{x}) = S_c(\boldsymbol{x}) - S_c(\boldsymbol{x} \setminus \boldsymbol{x}_s)$$



Okklusion



Integrated Gradients



LIME

local interpretable model-agnostic explanation

Lokal

Interpretiert das Black-Box Modell in lokaler Umgebung eines konkreten Beispiels

Interpretierbar

Approximiert das komplexe Black-Box Modell z.B. durch lineare Regression bzw. Logistische Regression

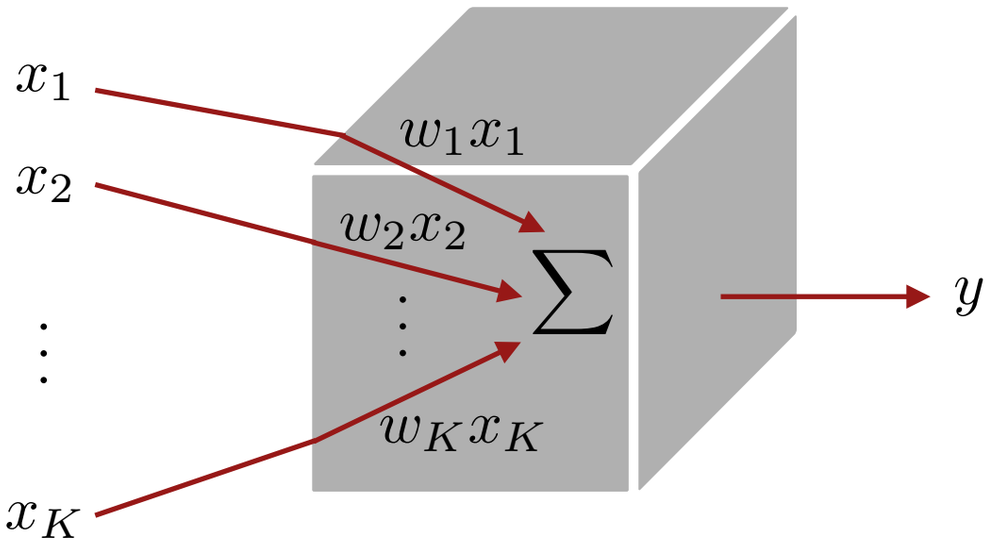
Model-Agnostic

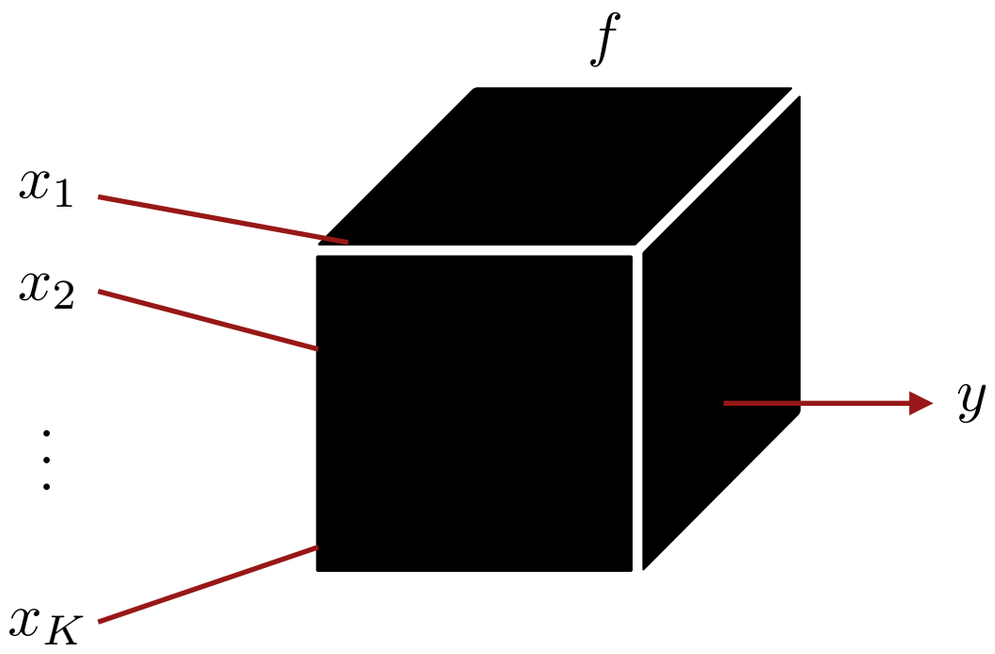
Das Black-Box Modell bleibt unberührt, keine Details müssen bekannt sein

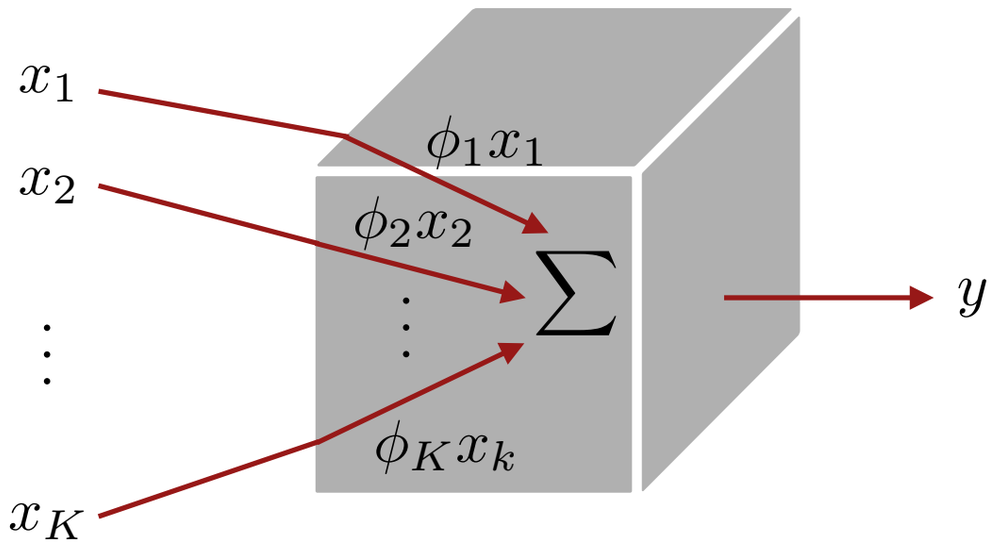
LIME

local interpretable model-agnostic explanation

Idee





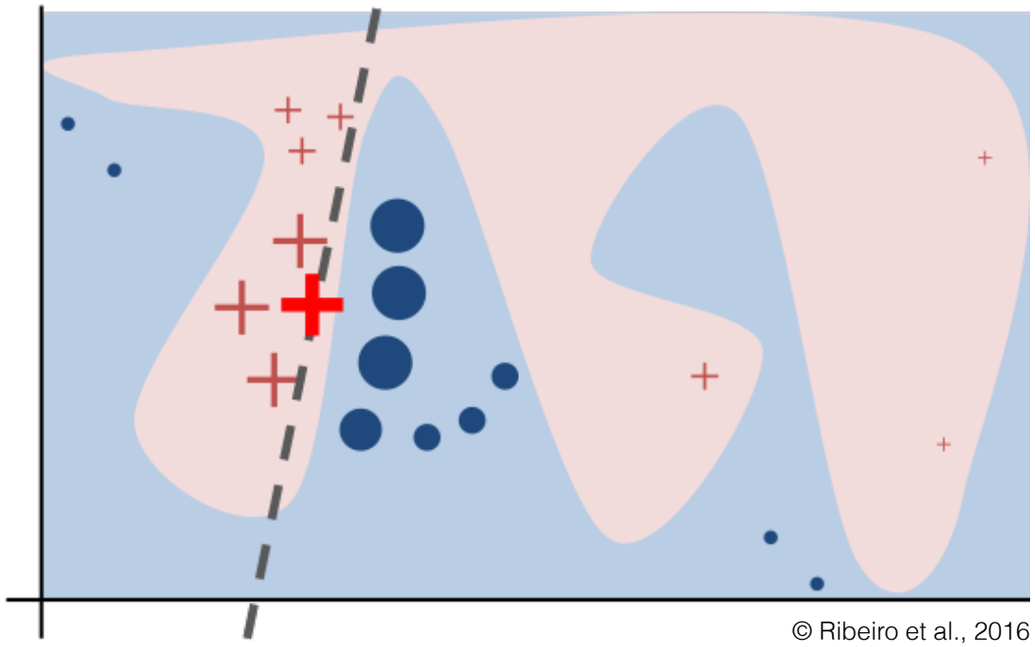


$$\phi_i = \psi_i\left(f, \boldsymbol{x}^{(n)}\right)$$

LIME

local interpretable model-agnostic explanation

Idee

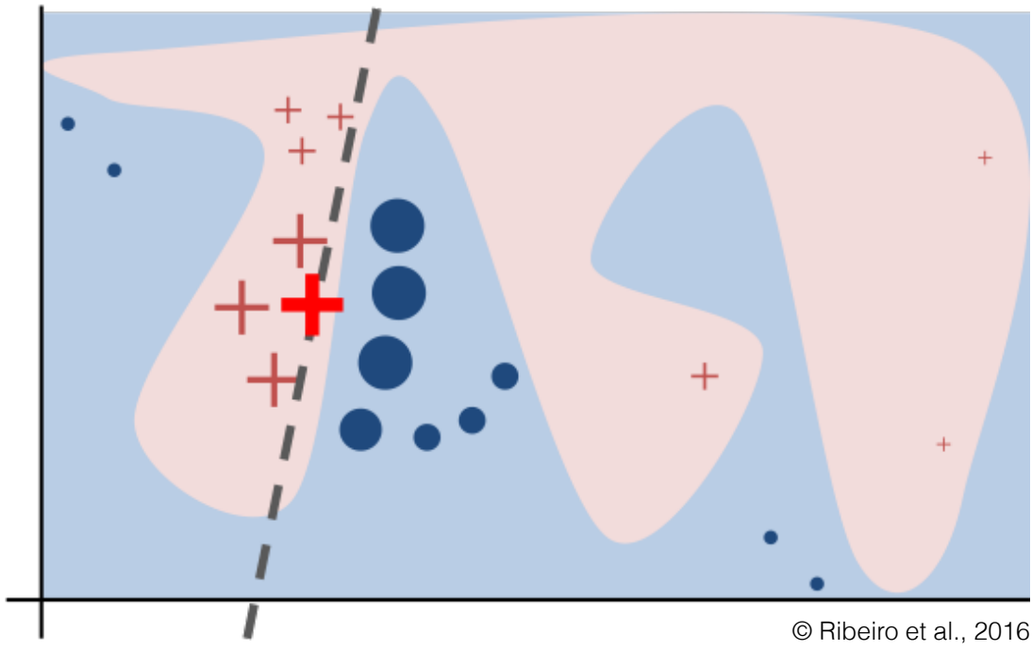


© Ribeiro et al., 2016

LIME

local interpretable model-agnostic explanation

Idee



Klassifikator f

interpretierbares Modell g

Originalbild x

interpretierbare Repräsentation x'

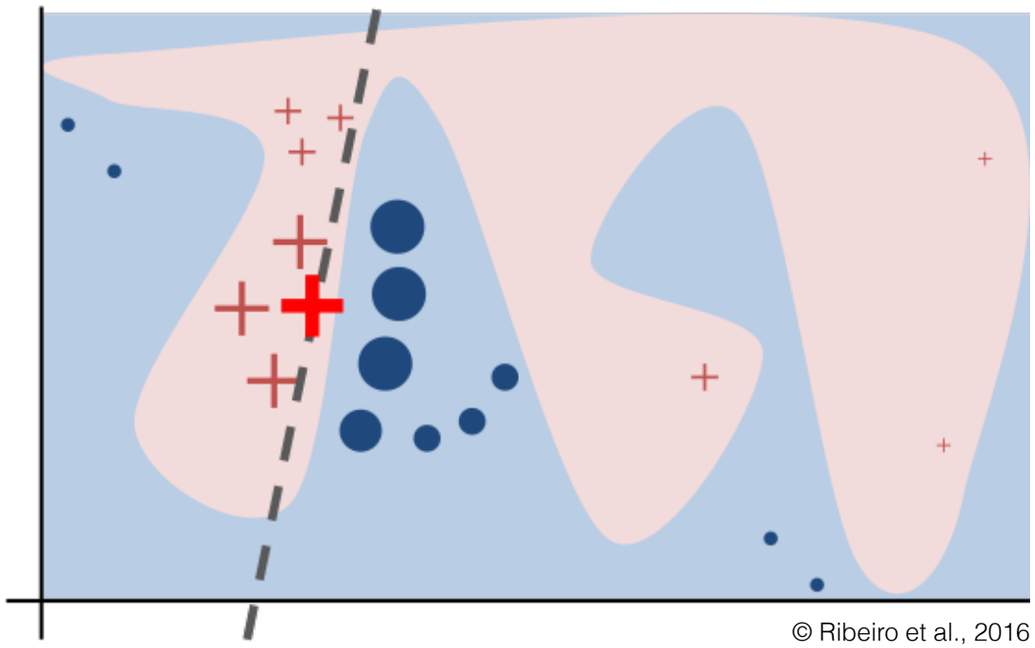
generierter Datenpunkt z

interpretierbarer generierter Datenpunkt z'

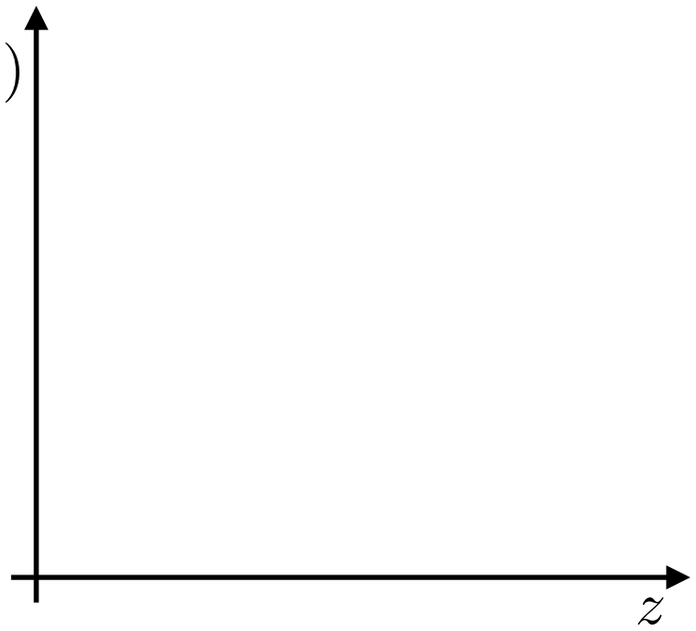
LIME

local interpretable model-agnostic explanation

Idee



© Ribeiro et al., 2016



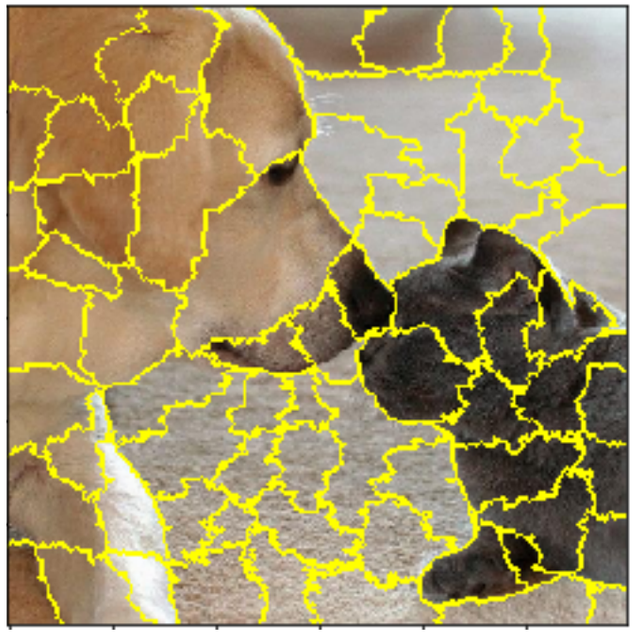
Maß für die Nähe von z zu x: π_x

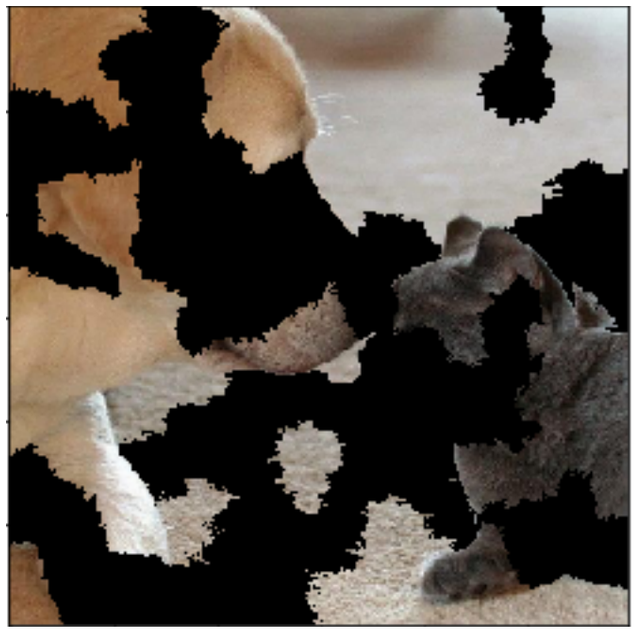
$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z'} \pi_x(z) \left(f(z) - g(z')\right)^2$$

$$g(z') = \phi_0 + \sum_i \phi_i z'_i$$

LIME

local interpretable model-agnostic explanation





LIME

local interpretable model-agnostic explanation

© Ribeiro et al., 2016

LIME

local interpretable model-agnostic explanation

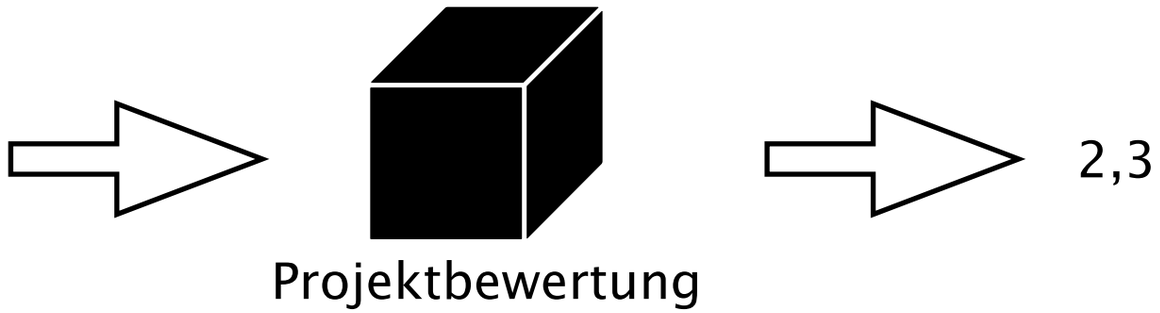
Offen

Wie kommen wir zu Superpixeln?

Kann man bestimmte Merkmale als insgesamt wichtiger als andere markieren?

SHAP

SHAP



SHAP

SHapley Additive exPlanations

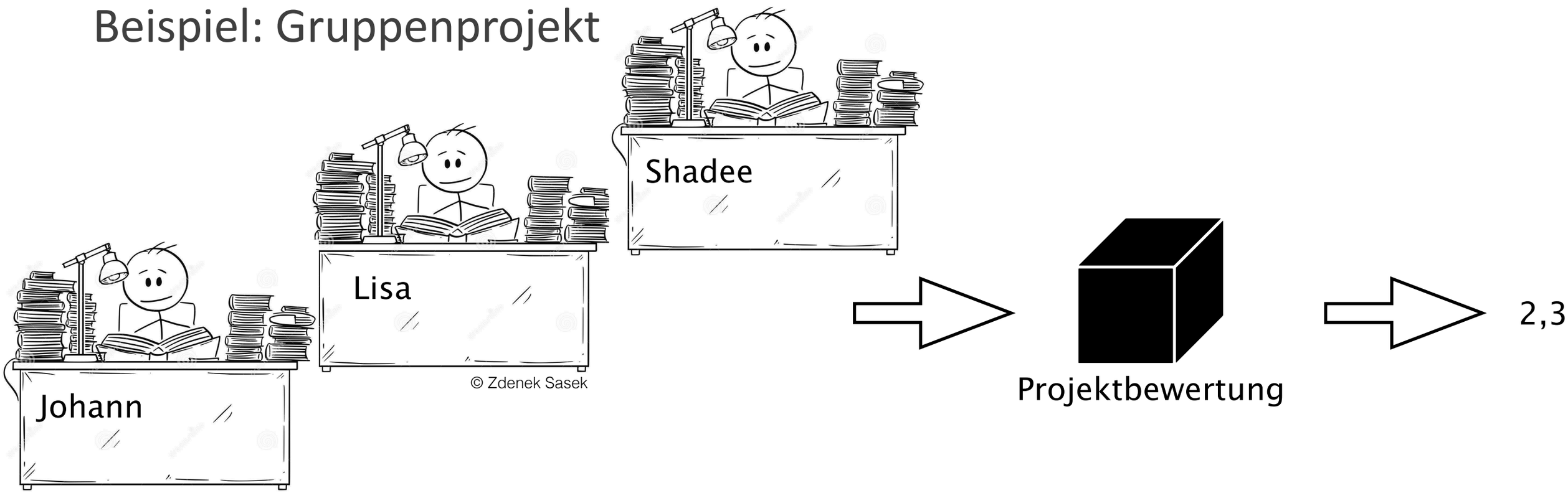
Shapley Werte

aus der Spieltheorie

liefern die Beiträge einzelner Aspekte zum Gesamtergebnis

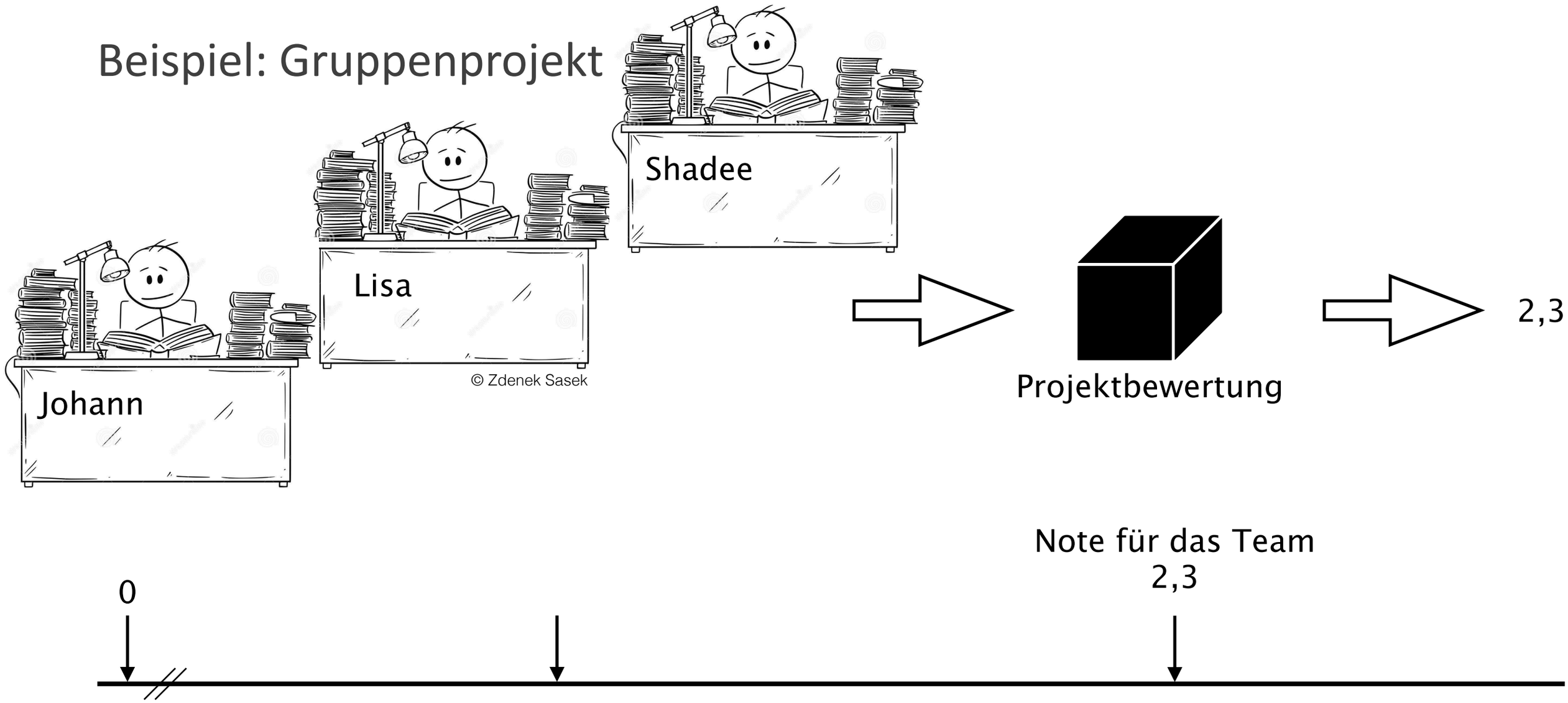
dazu: alle möglichen Zusammensetzungen betrachten

Beispiel: Gruppenprojekt



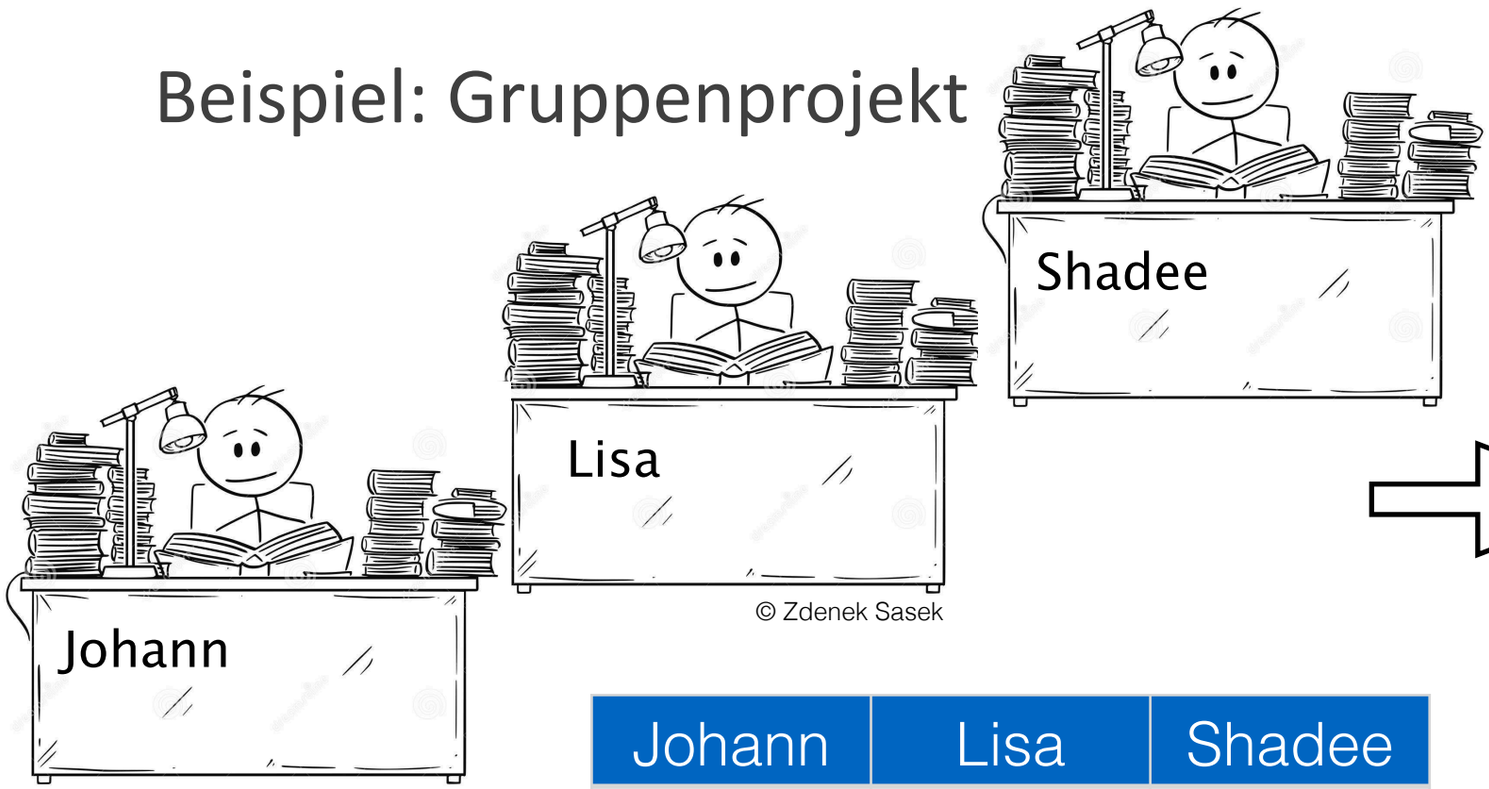
SHapley Additive exPlanations

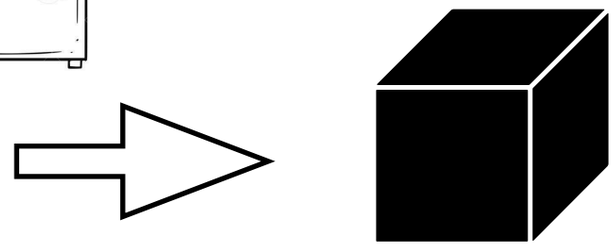
Beispiel: Gruppenprojekt



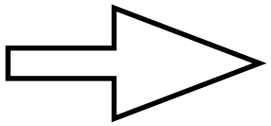
SHapley Additive exPlanations

Beispiel: Gruppenprojekt





Projektbewertung



2,3

Johann	Lisa	Shadee
X	-	-
-	X	-
-	-	X
X	X	-
X	-	X
-	X	X
X	X	X

SHAP

SHapley Additive exPlanations

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right)$$

Johann	Lisa	Shadee
X	-	-
-	X	-
-	-	X
X	X	-
X	-	X
-	X	X
X	X	X

SHAP

SHapley Additive exPlanations

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right)$$

$$\pi_x(z') = \frac{M - 1}{\binom{M}{|z'|} |z'|(M - |z'|)}$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z'} \pi_x(z') \left(f(z) - g(z') \right)^2$$

$$g(z') = \phi_0 + \sum_i \phi_i z'_i$$

SHAP

SHapley Additive exPlanations