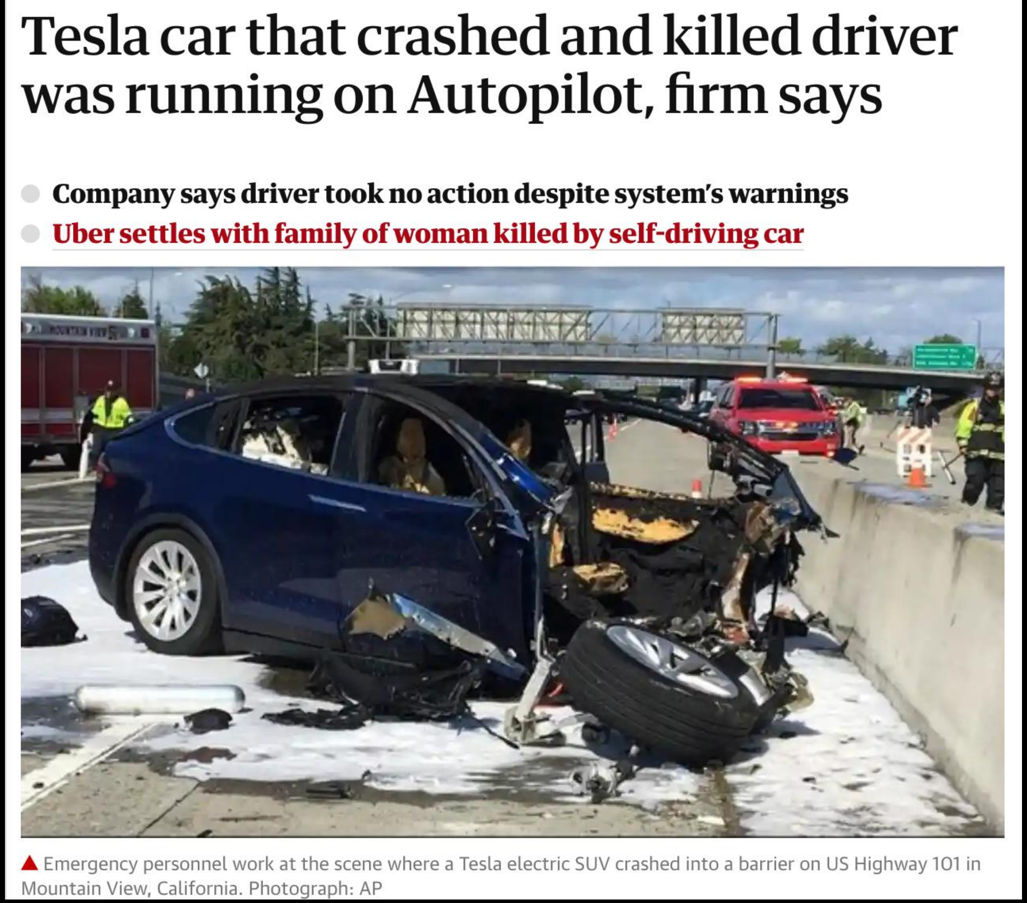
Some Shades of Grey

Interpretability and explanatory capacity of DNNs

Andreas Dengel @ TRILATERAL AI CONFERENCE 2024

dfki ai

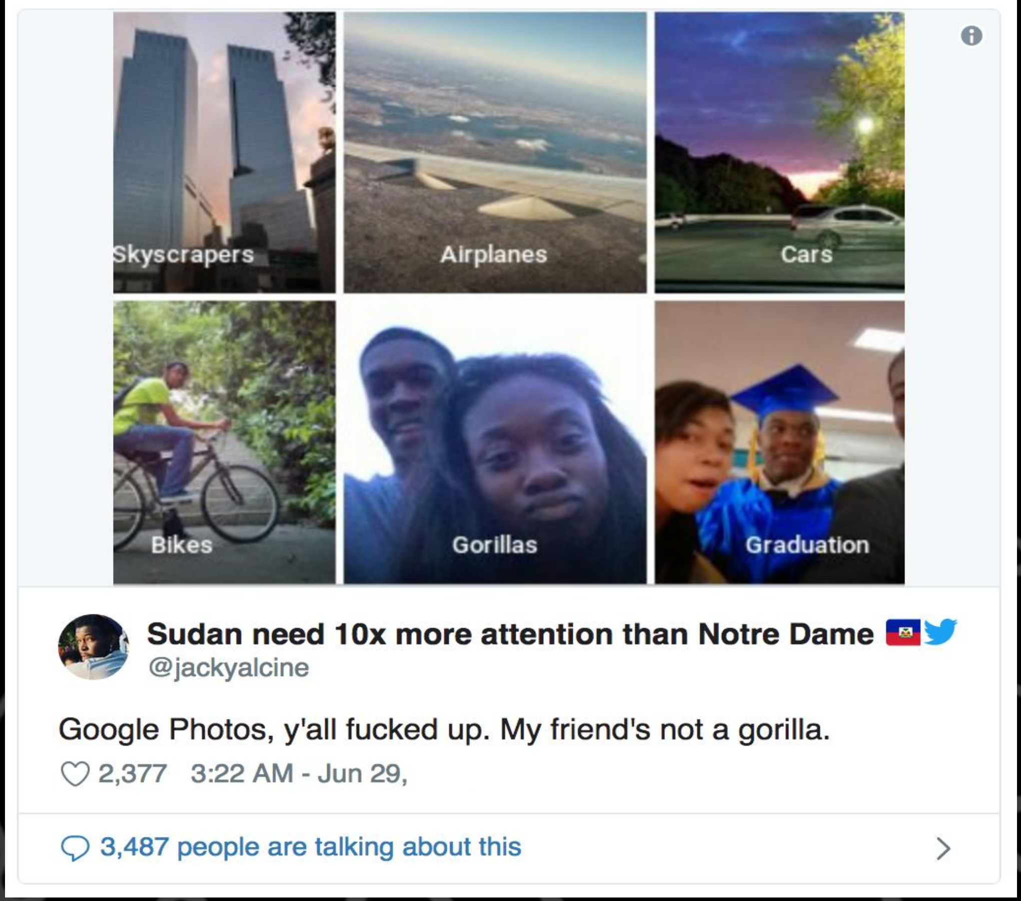
Negative examples lead to growing concerns regarding the usage of neural networks



Tesla car that crashed and killed driver was running on Autopilot, firm says

- Company says driver took no action despite system's warnings
- Uber settles with family of woman killed by self-driving car

▲ Emergency personnel work at the scene where a Tesla electric SUV crashed into a barrier on US Highway 101 in Mountain View, California. Photograph: AP



Sudan need 10x more attention than Notre Dame 🇭🇹
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

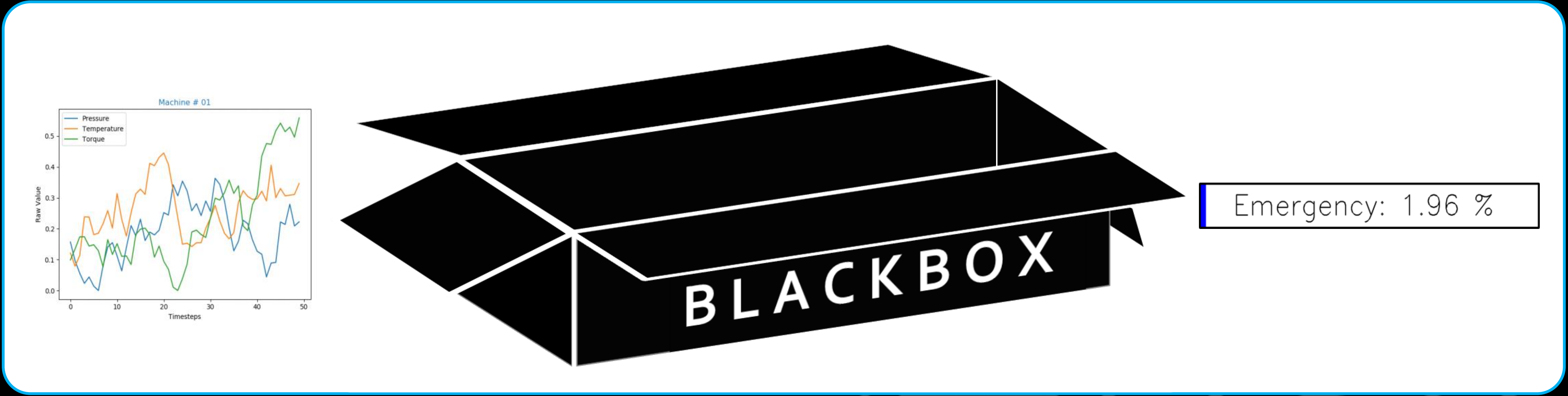
♡ 2,377 3:22 AM - Jun 29,

3,487 people are talking about this

This sensitiveness also hold for foundation models via adversarial or jailbreak attacks!

Deep Neural Networks (DNN) can be viewed primarily in terms of their input and output, without knowledge about internal processes

Black Box Problem



In many areas, insight into decision making is just as important as the decision itself, especially in risk-sensitive and safety-critical applications

The tradeoffs between system autonomy and transparency is growing as AI systems increase in their internal complexity!

The transparency and comprehensibility of decisions and forecasts are becoming more and more important*

Interpretability

Interpretability refers to the observation and representation of cause and effect within a system, without necessarily knowing why something happens

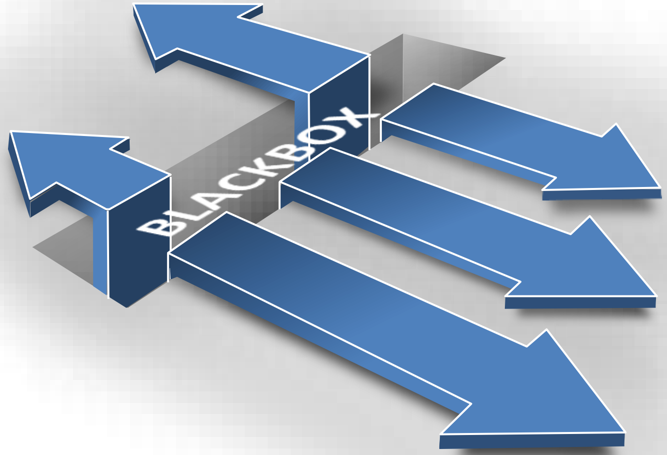
Explainability

Explainability, on the other hand, concerns the ability to explain the inner function of a system in human terms (e.g. by means of a given example)

responsible data

decision processes

Visualization



Explanation

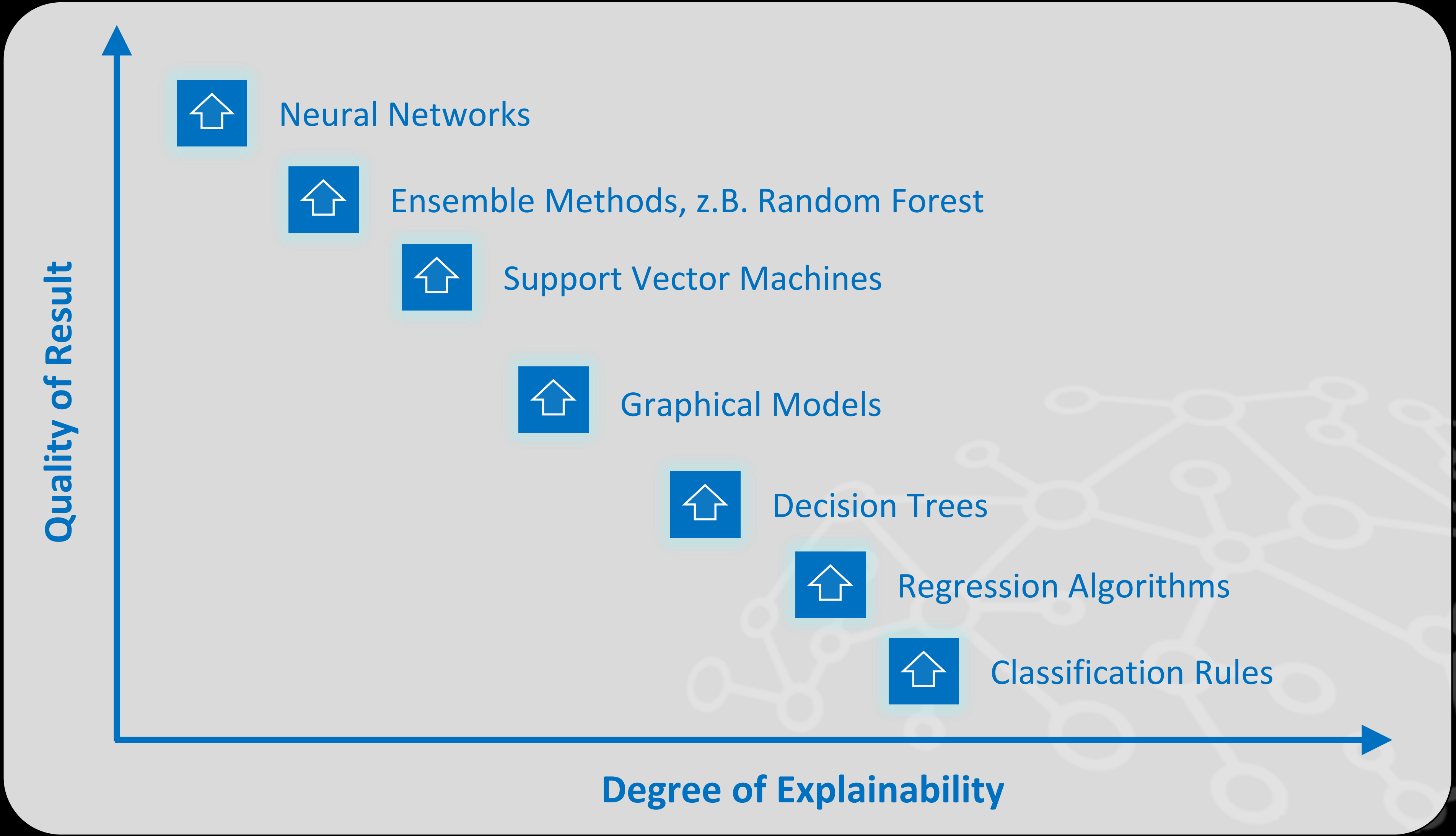
traceability

reliability

expressiveness

- *S. Palacio, A. Lucieri, M. Munir, S. Ahmed, J. Hees, and A. Dengel, XAI Handbook: Towards a Unified Framework for Explainable AI,* Proceedings ICCV, Responsible PR&MI 2021, *1st* International Workshop on Responsible Pattern Recognition and Machine Intelligence. *https://arxiv.org/abs/2105.06677.*

Although (deep) neural networks offer enormous advantages in terms of their accuracy, they lack the ability to explain their results



Neural Networks

Ensemble Methods, z.B. Random Forest

Support Vector Machines

Graphical Models

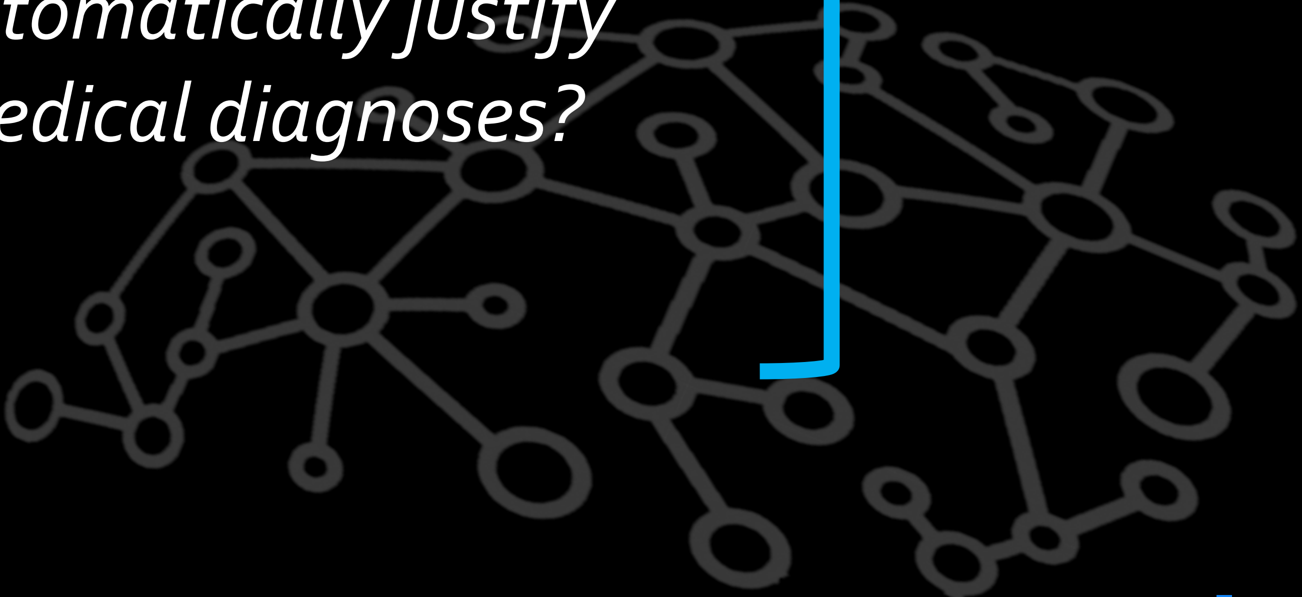
Decision Trees

Regression Algorithms

Classification Rules

Example

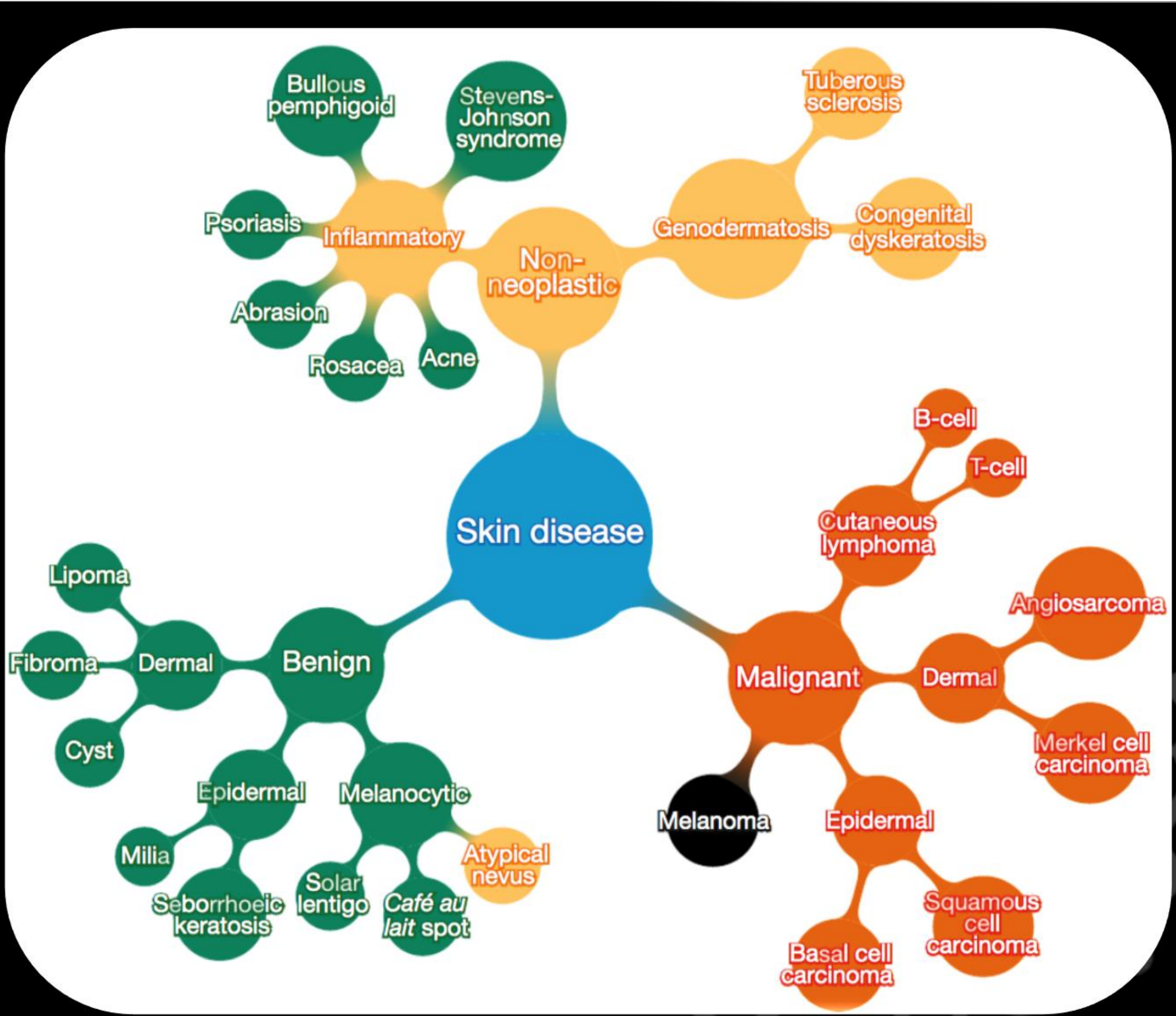
How can we automatically justify and explain medical diagnoses?



Due to their frequency, the analysis of lesions is one of the most important applications of AI in medicine

- The complete skin cancer taxonomy contains more than 2,000 diseases
- Taxonomy is organized based on the visual and clinical similarity of diseases

Red corresponds to malignant;
green corresponds to benign skin changes;
yellow can be both.



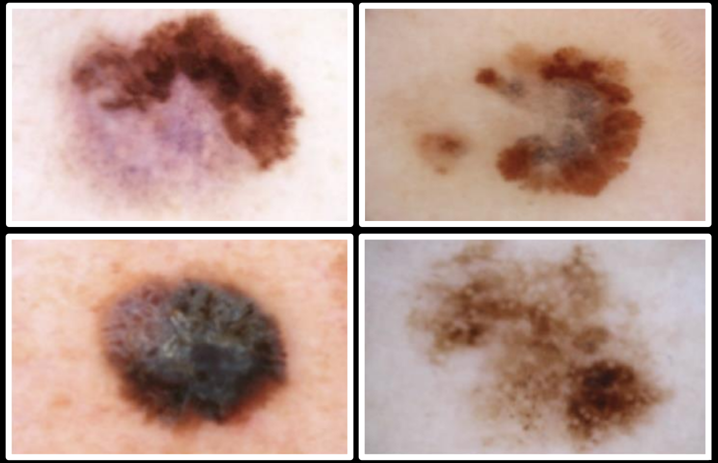
Source: https://www.nature.com/articles/nature21056.epdf?author_access_token=8oxIcYWf5UNrNpHsUHd2StRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPakXos4UhQAFZ750CsBNMMsISFHIKinKDMKjShCpHIlYPYUHhNzkn6pSnOCt0Ftf6

Skin lesions / moles can look very different and can be classified into different types (melanoma, nevus, seborrheic keratosis…)

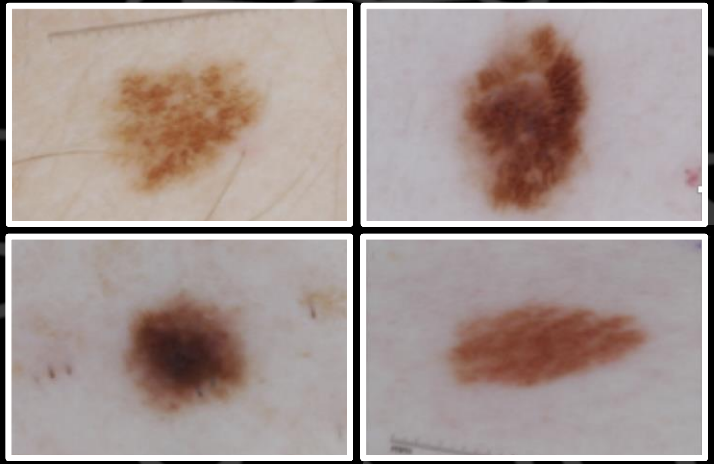
- Primary task is to distinguish between benign and malignant lesions

- Diagnosis is usually carried out by means of dermatoscopy, i.e., devices that enlarge and illuminate the moles in order to make the underlying skin layers visible

- Early detection is crucial for successful healing and requires expert knowledge and experience
  - 5-year survival rate for patients at 96 % but drops to 63% when they reach lymph nodes and 20% when they reach distant organs!*



Melanoma



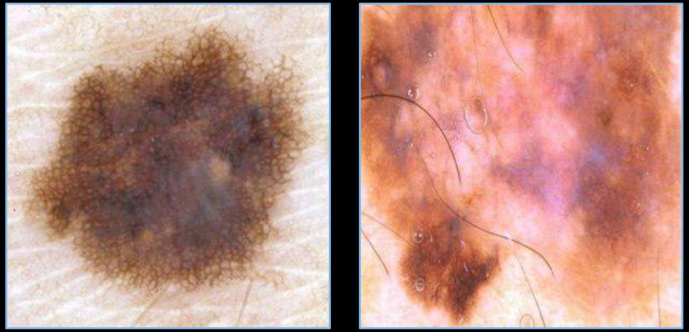
Nevus

* Source: https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/

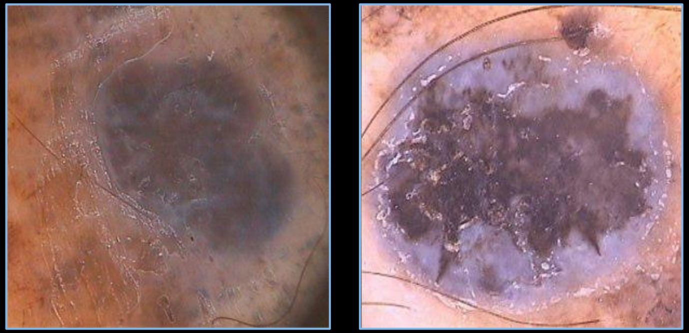
** Image Source: https://www.doccheckshop.de/diagnostik/fachspezifische-diagnostik/dermatoskopie/dermatoskope/10946/heine-mini-3000-dermatoskop?sPartner=google&number=157672&gclid=Cj0KCQjwl8XtBRDAARIsAKfwtxAROPNTVjTPvBoK600bNrUSvx5sys4_btrsBYqLOy17W6WafFb_GZkaAlBBEALw_wcB)

Diagnosis of skin lesions requires the recognition of highly complex structural features in the lesions

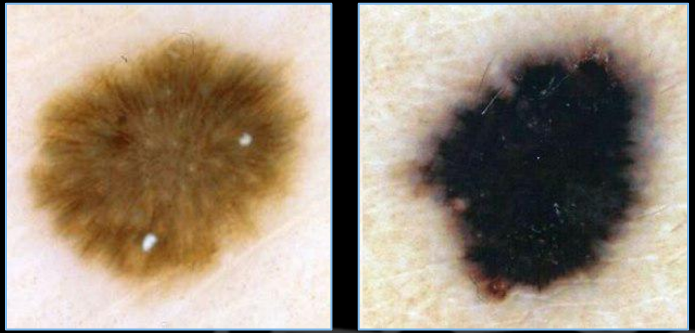
1 Regression structures with white or bluish coloring



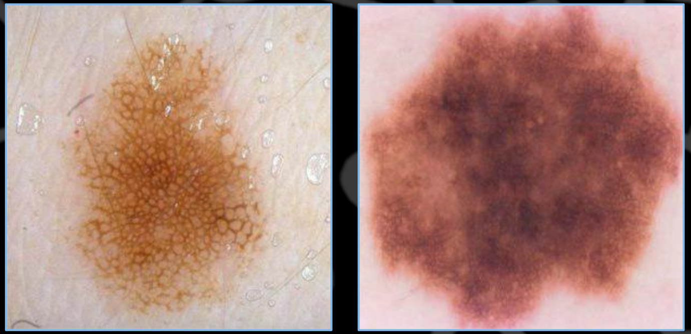
2 Complex lesions with three and more color variations

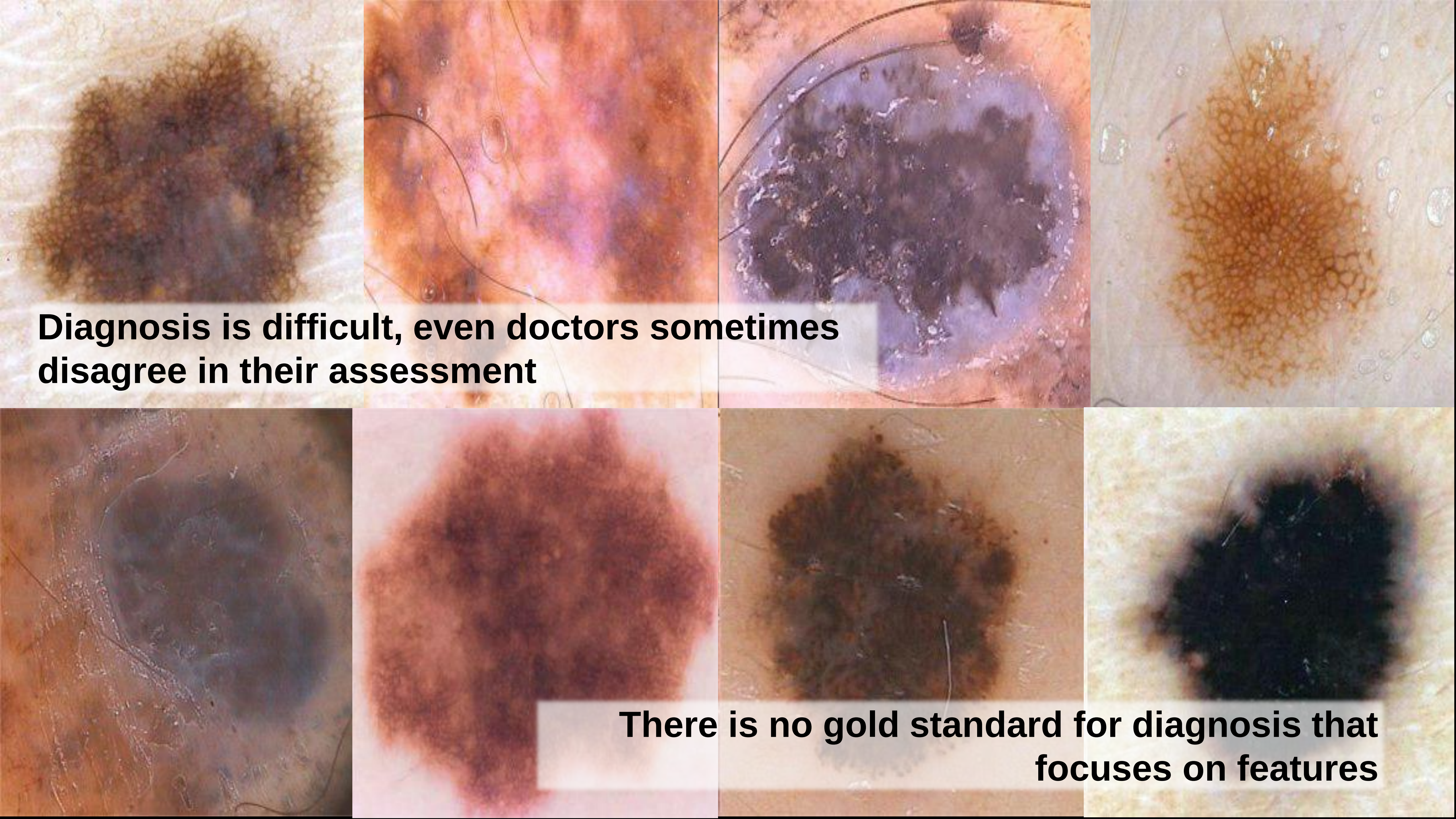


3 Radial extensions with regular or irregular distribution



4 Structural networks with typical and atypical pigmentation





Diagnosis is difficult, even doctors sometimes disagree in their assessment

There is no gold standard for diagnosis that focuses on features

Deep Neural Networks (DNN) assess images of skin lesions with certified dermatologists' accuracy

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

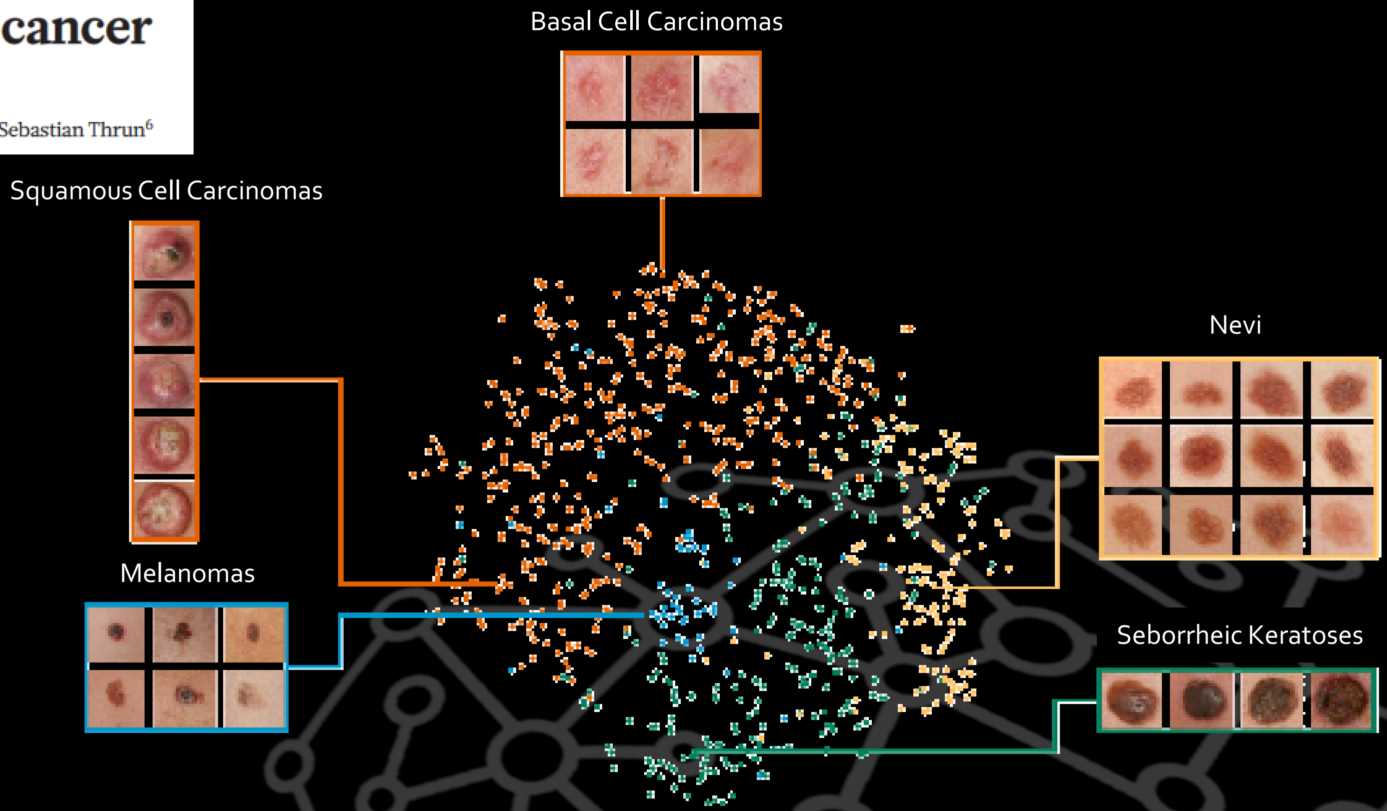
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

LESIONS LEARNT

Artificial intelligence powers detection of skin cancer from images PAGES 36 & 115

NATURE.COM/NATURE
2 February 2017 £10
Vol. 542, No. 7639

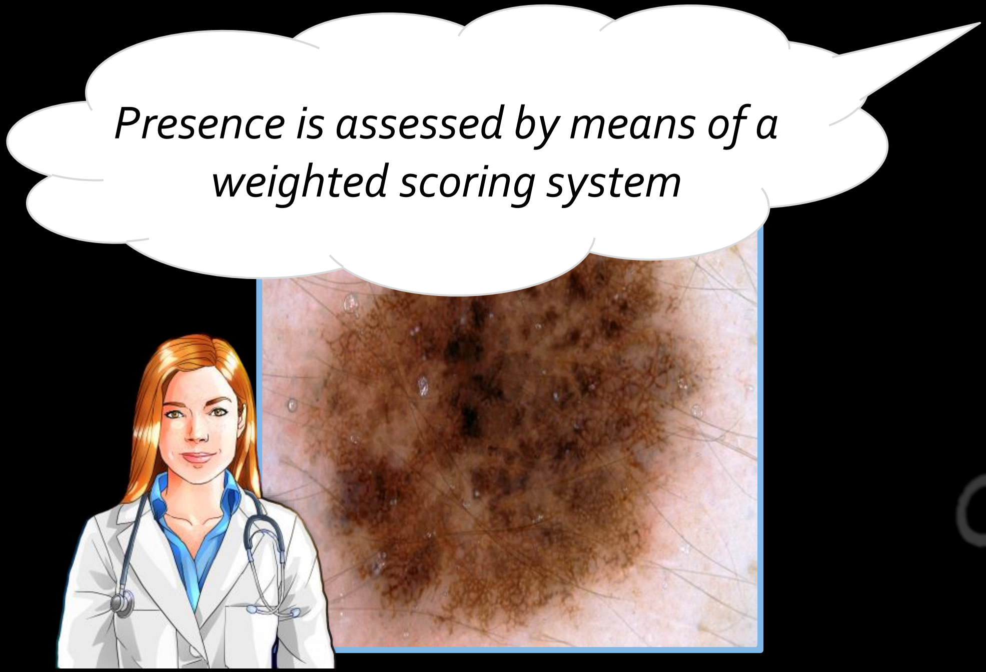


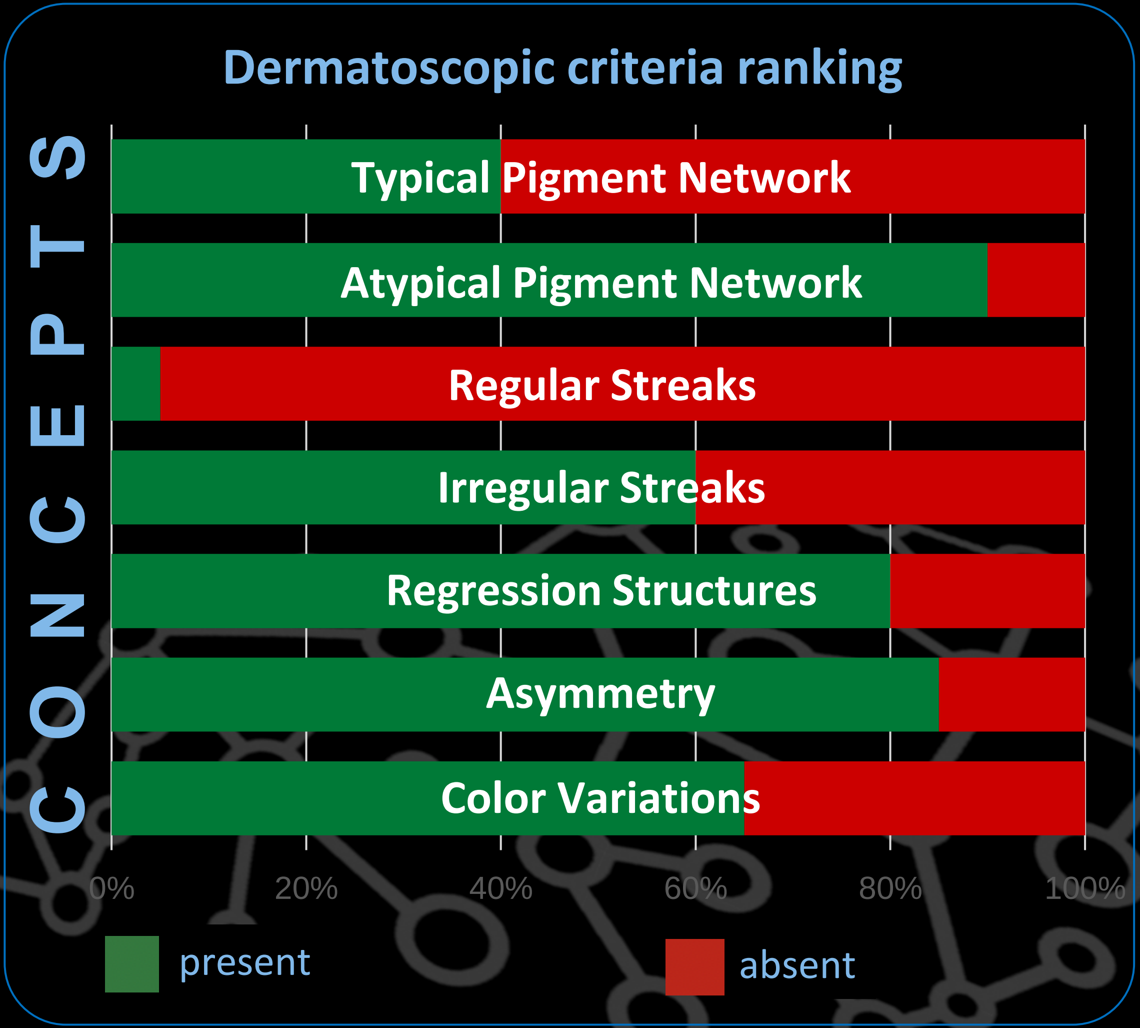
... it would be even better if DNN could justify their decision!

Source: https://cs.stanford.edu/people/esteva/nature/

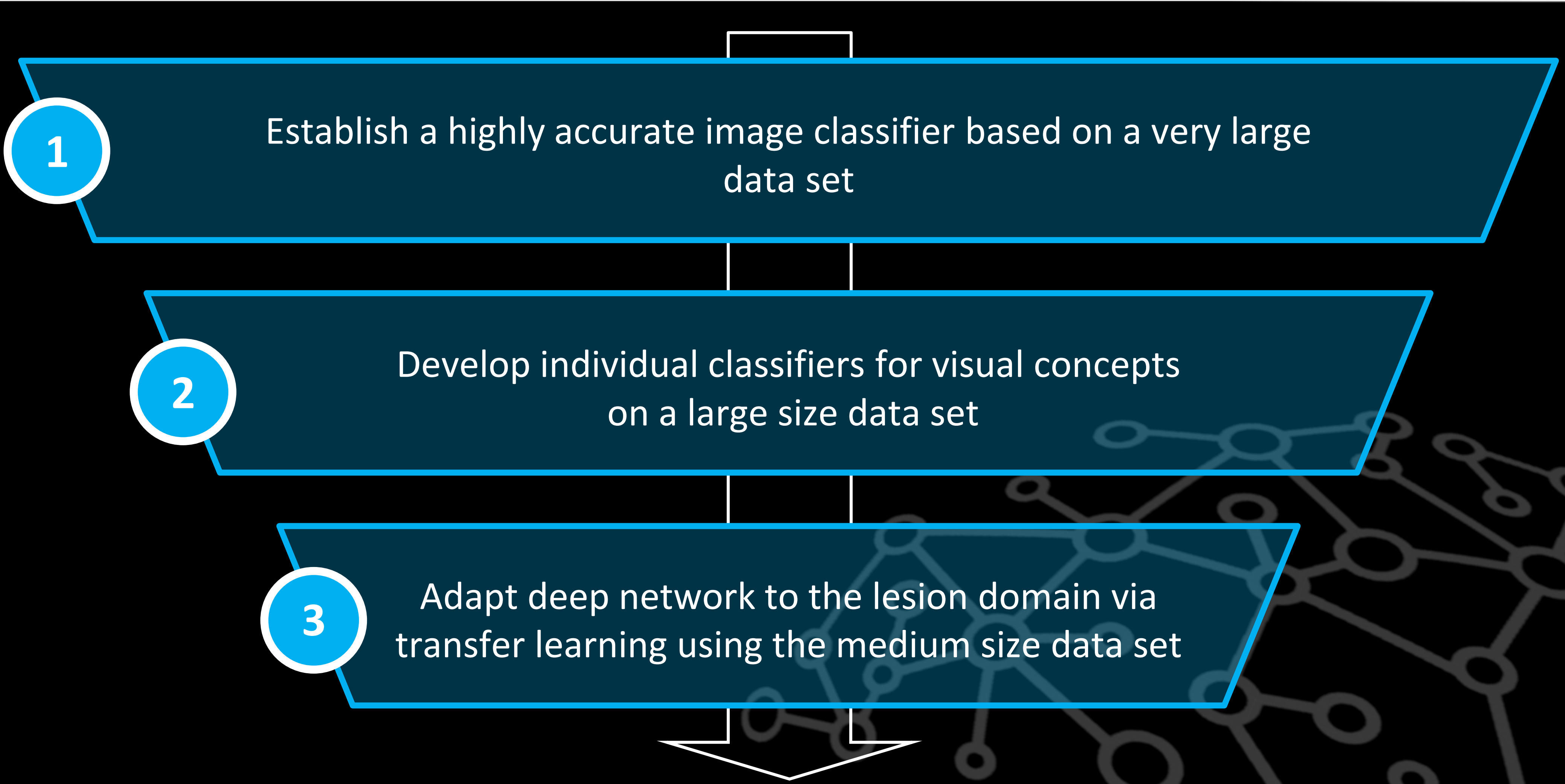
Such characteristics of skin lesions can also be queried by dermatologists using special checklists (dermoscopic criteria ranking).

- Example: 7-point checklist or ABCDE rule
  - Subjective verification of dermatoscopic criteria or concepts
  - If threshold is exceeded, then removal or further treatment.




Dermatoscopic criteria ranking

Although there is no gold standard on skin lesion data, we can build a workaround that provides a well working solution



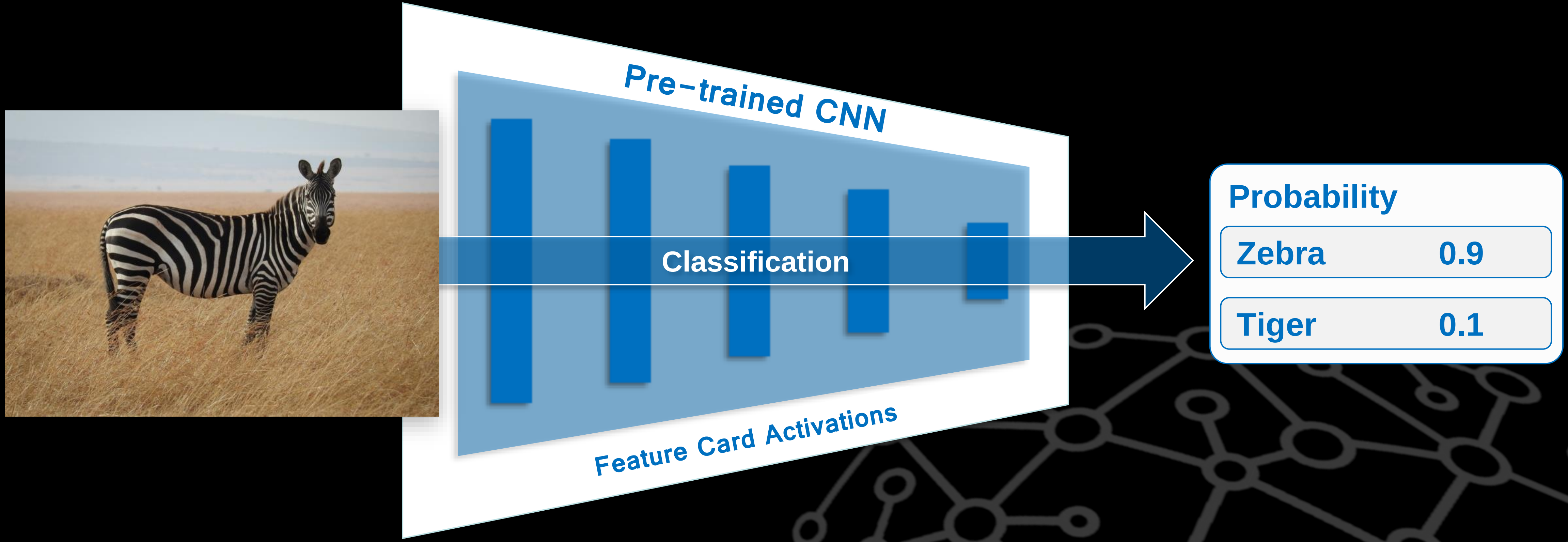
1 Establish a highly accurate image classifier based on a very large data set

2 Develop individual classifiers for visual concepts on a large size data set

3 Adapt deep network to the lesion domain via transfer learning using the medium size data set

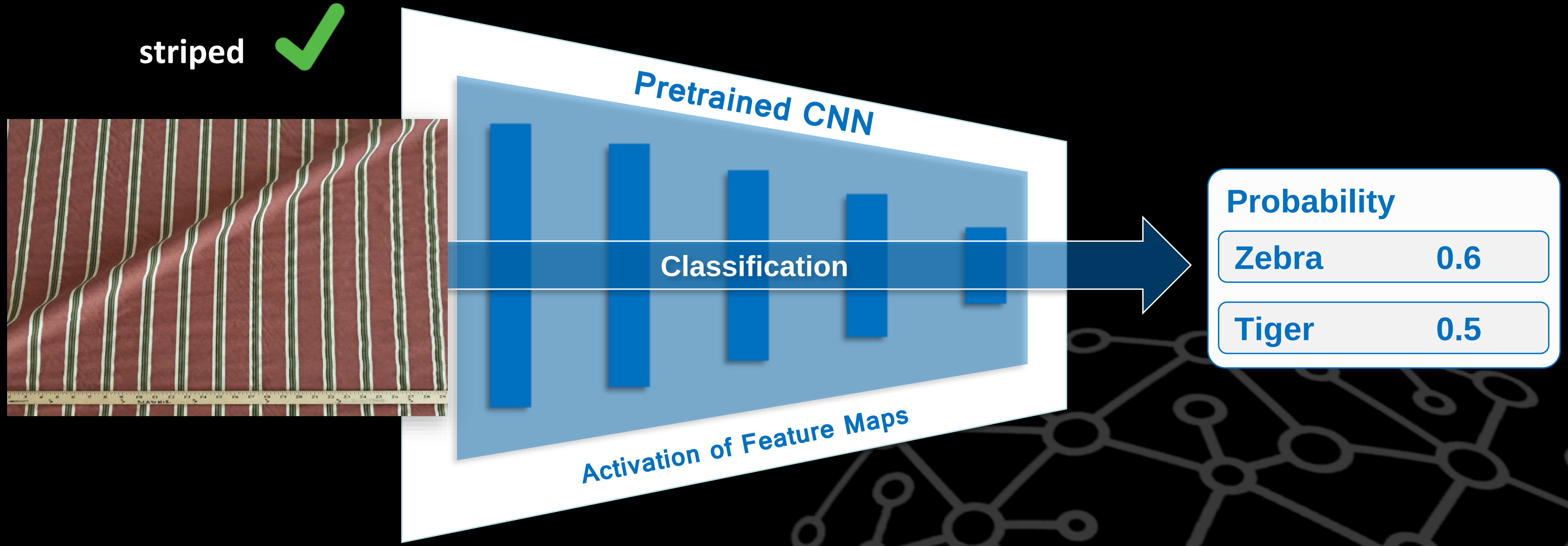
XAI-System for Skin Lesions

For skin lesion detection, we use a pre-trained, high-performance end-to-end CNN that did not learn any explicit domain concepts during training



Network was not explicitly trained on lesion concepts!

In order to apply concept mapping, we need a new dataset that contains concepts with corresponding labels



striped ✓

Pretrained CNN

Classification

Activation of Feature Maps

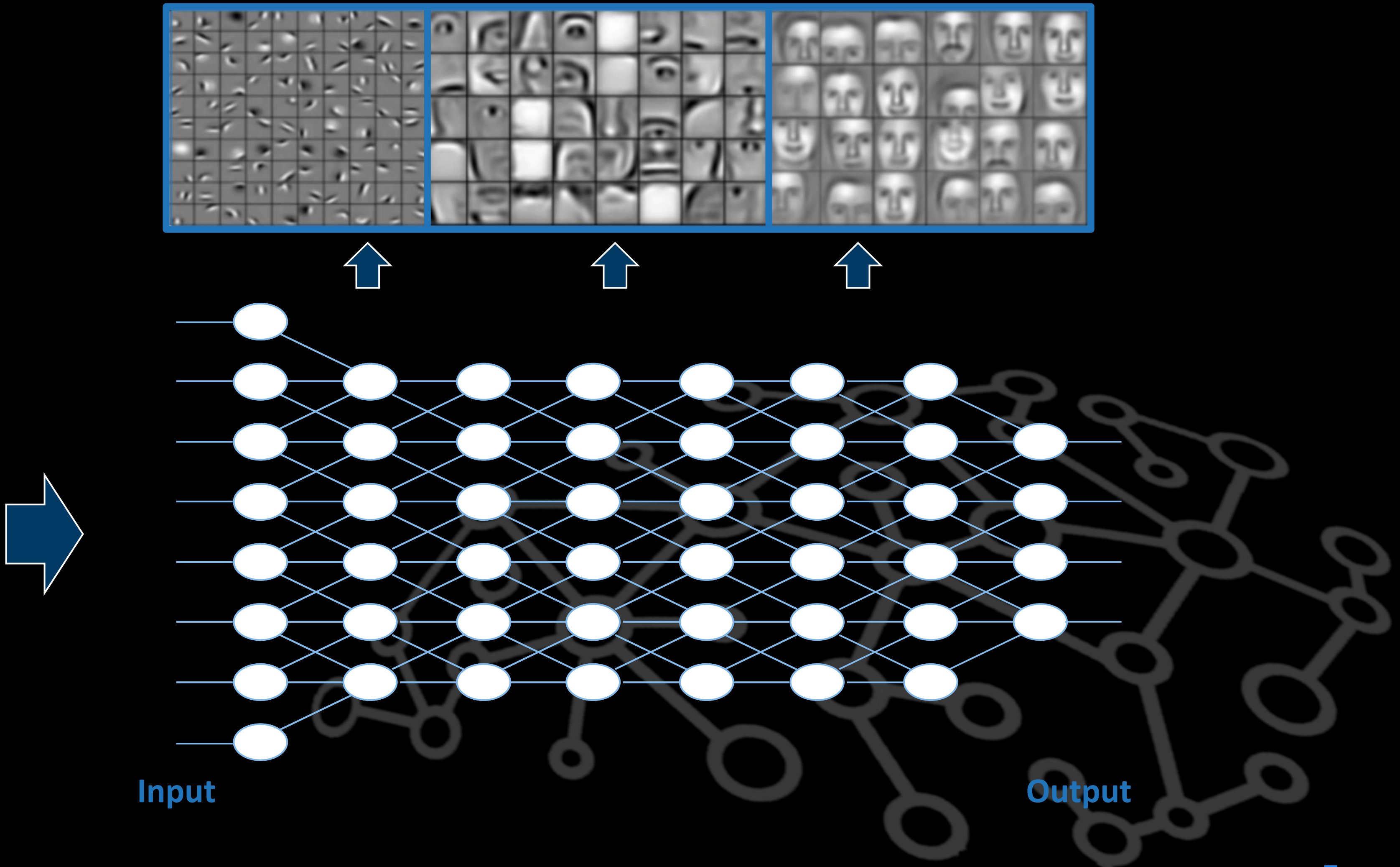
Probability	
Zebra	0.6
Tiger	0.5

Since we do not have enough data, we use a textile dataset 😉 !

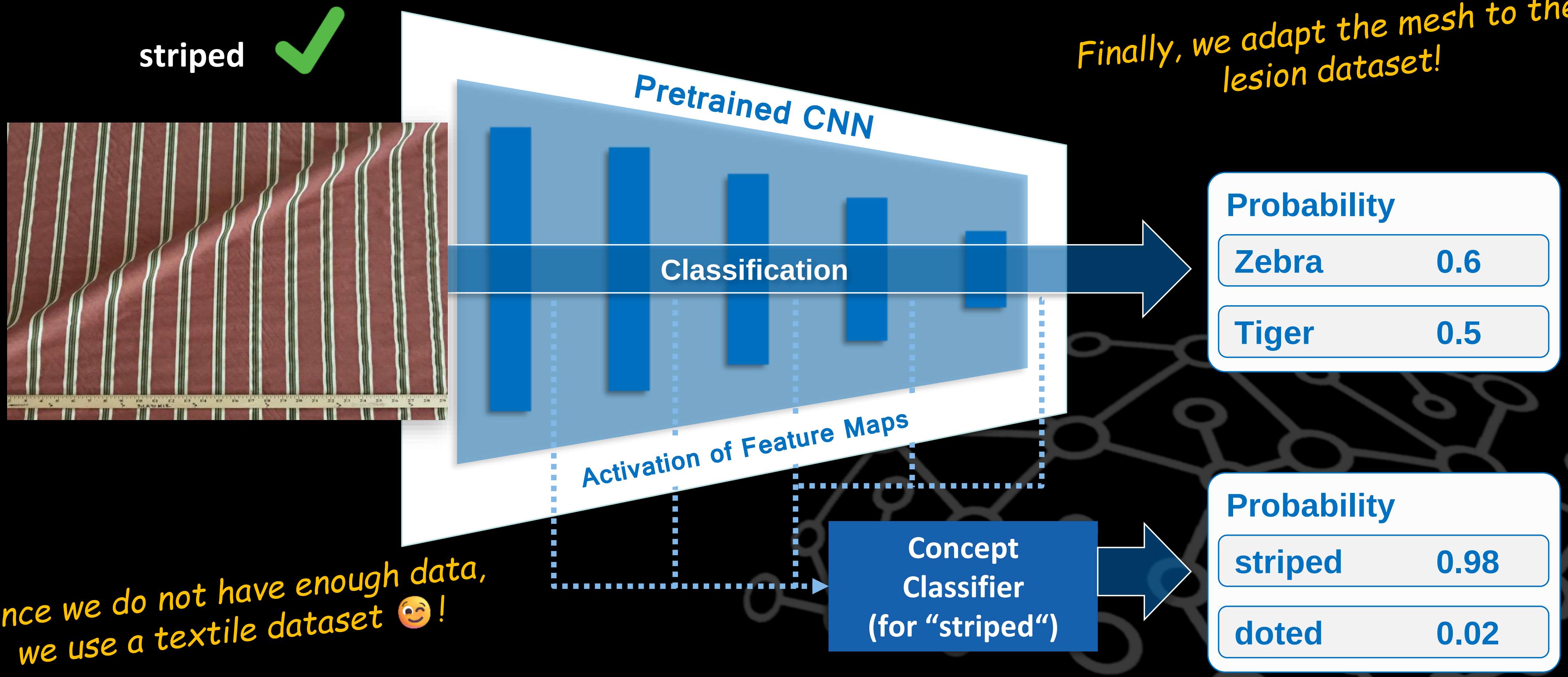
* Kim, Been, et al. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." International Conference on Machine Learning. 2018.

A Deep Neural Network learns a hierarchical structure from different feature abstractions (corners, edges, object properties up to whole objects).

We are employing this characteristic!



In order to apply concept mapping, we need a new dataset that contains concepts with corresponding labels

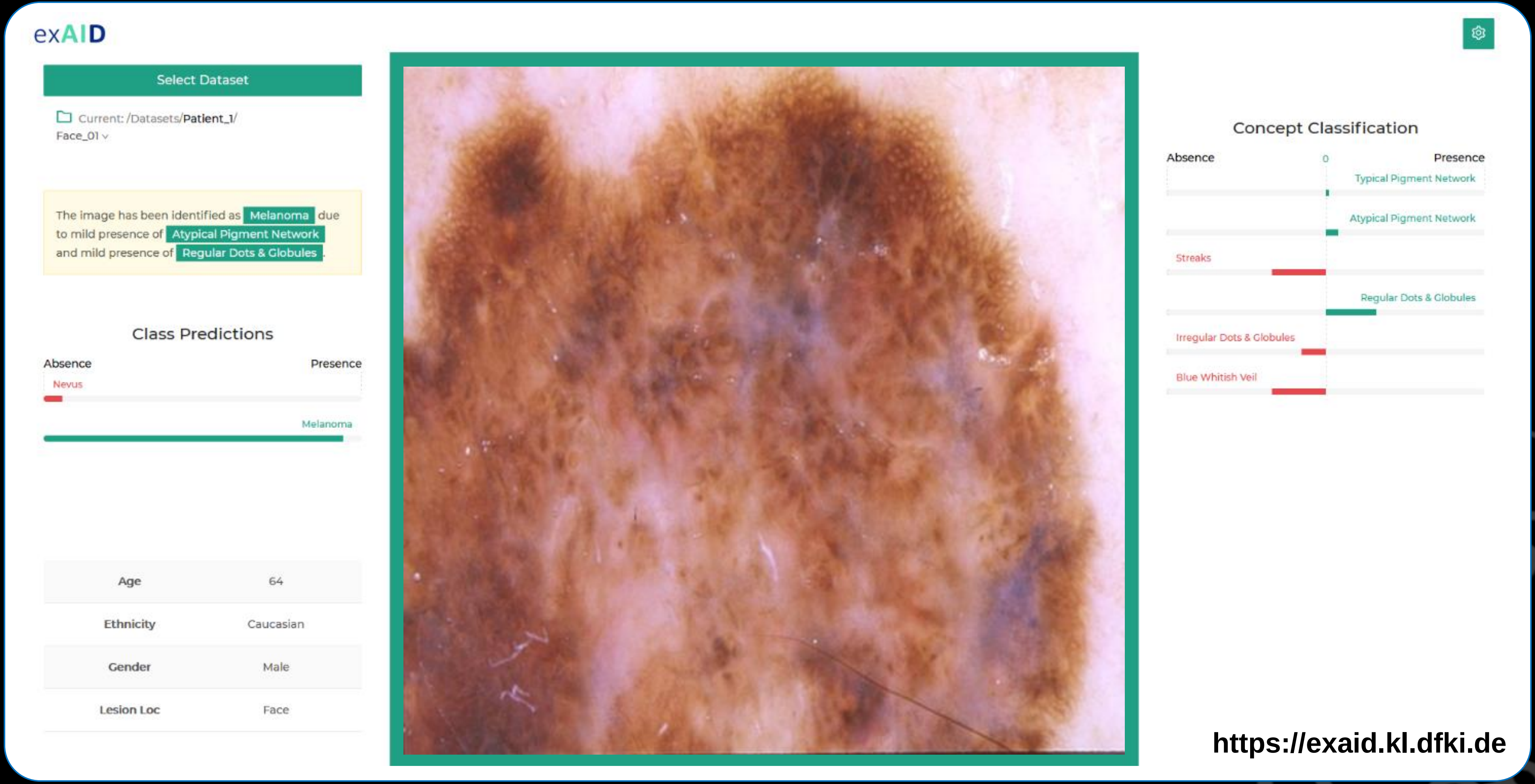


Finally, we adapt the mesh to the lesion dataset!

Since we do not have enough data, we use a textile dataset 😉 !

* Kim, Been, et al. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." International Conference on Machine Learning. 2018.

Via this approach, AI not only assesses images of skin lesions with certified dermatologists' accuracy, but generates additional visual explanation



Current: /Datasets/Patient_1/
Face_01

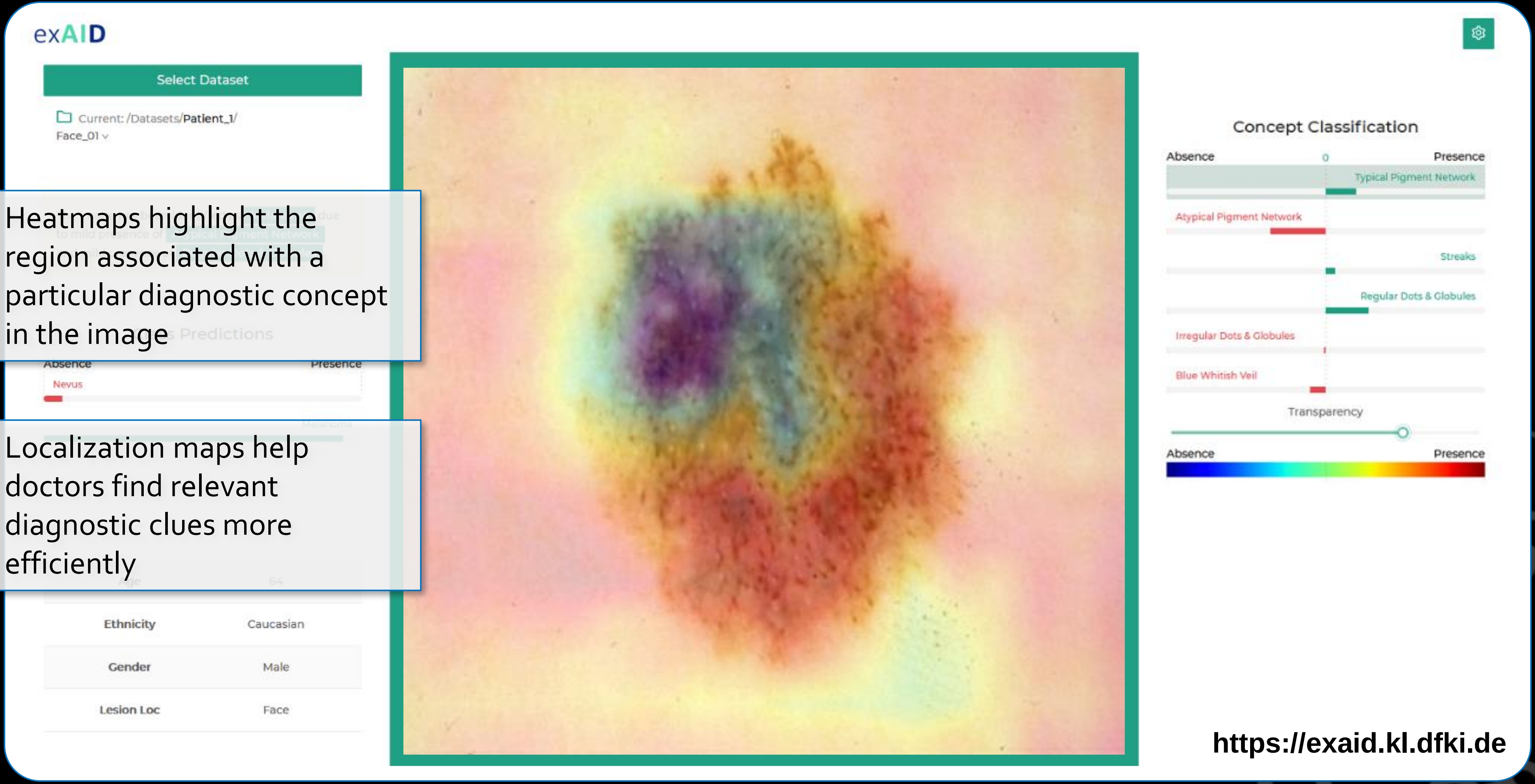
The image has been identified as Melanoma due to mild presence of Atypical Pigment Network and mild presence of Regular Dots & Globules.

Age	64
Ethnicity	Caucasian
Gender	Male
Lesion Loc	Face

https://exaid.kl.dfki.de

A. Lucieri, M. A. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, *On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors*, IJCNN 2020, Int'l Joint Conference on Neural Networks, Glasgow, Scotland (July 2020).

Concepts are displayed via localization map for specialist analysis



Current: /Datasets/Patient_1/
Face_01 ⌄

Ethnicity	Caucasian
Gender	Male
Lesion Loc	Face

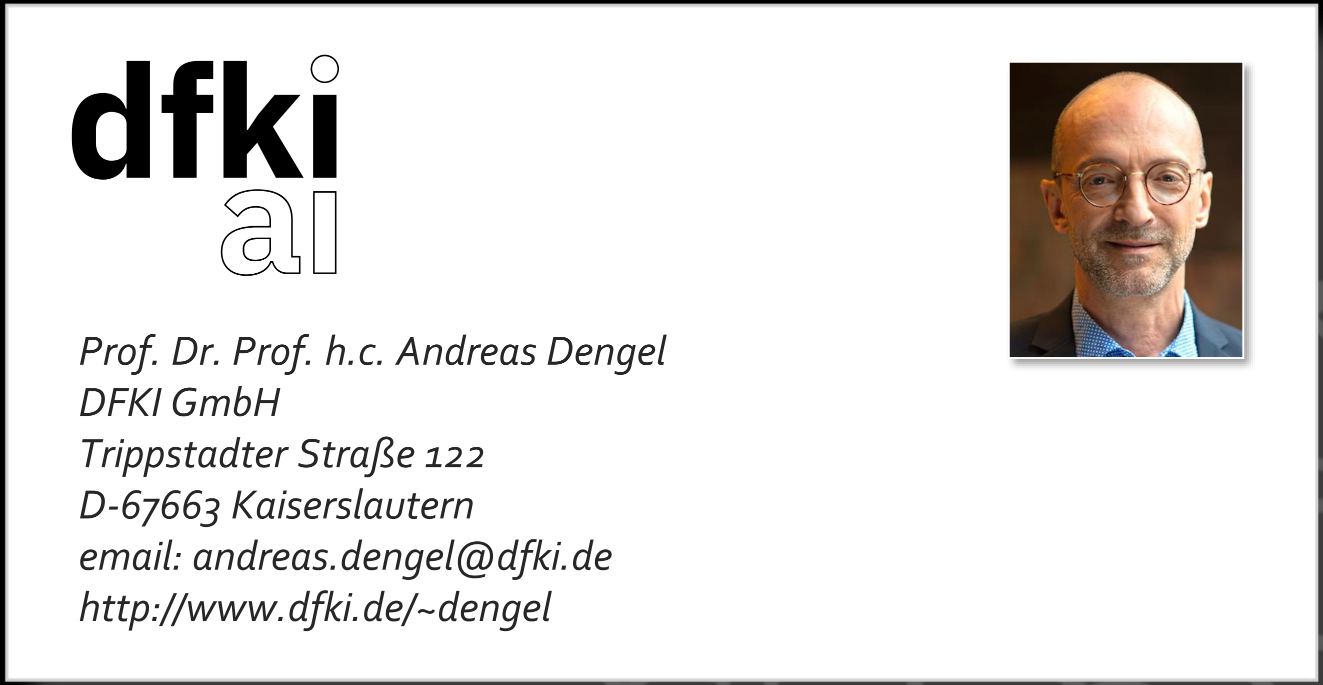
Heatmaps highlight the region associated with a particular diagnostic concept in the image

Localization maps help doctors find relevant diagnostic clues more efficiently

https://exaid.kl.dfki.de

A. Lucieri, M. A. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, *On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors*, IJCNN 2020, Int'l Joint Conference on Neural Networks, Glasgow, Scotland (July 2020).

Thank you for your time and attention!



Prof. Dr. Prof. h.c. Andreas Dengel
DFKI GmbH
Trippstadter Straße 122
D-67663 Kaiserslautern
email: andreas.dengel@dfki.de
http://www.dfki.de/~dengel