Human-centric interpretability of deep learning for digital pathology

Mara Graziani

PhD student, Hes-so Valais and UniGe

UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE
Department of Informatics

Hes·so VALAIS WALLIS

PROCESS

European Commission | Horizon 2020 European Union funding for Research & Innovation

E-talk @Swiss Digital Pathology 17.09.2020

Who am I ?

PhD focus:

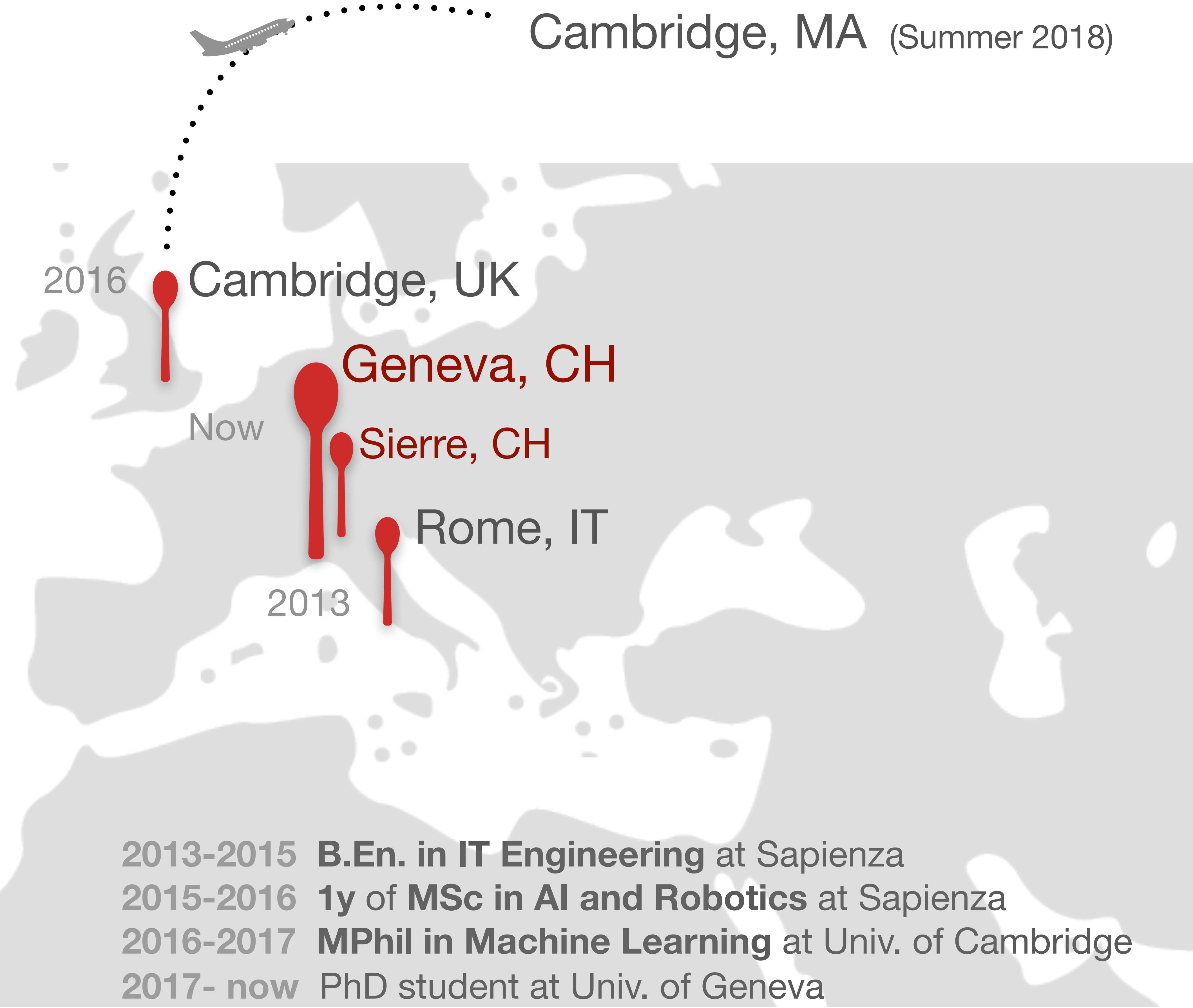
Interpretability of Deep Learning for Medical Imaging

Started in:

November 2017

Funded by:

EU H2020 PROCESS



mara.graziani@hevs.ch

My research question

Can we generate human-centric explanations of deep learning and can we use them to improve model performance?

My research question

Can we generate human-centric explanations of deep learning and can we use them to improve model performance?

Motivation:

Ease the interaction, improve models with little extra complexity, debug models, GDPR* right for explainability, improve trust and accountability, remove bias or data memorization, generate answers to “why” questions on model behaviour and decisions.

* General Data Protection Regulation

Outline

* Introduction and definition of **human-centric interpretability for deep learning**
* Presentation of research in this direction:
  * Evaluation of visualization tools
  * Concept-based interpretability with Regression Concept Vectors
  * Guiding CNNs with user-defined features
* Remarks
* Conclusions

Interpretability: What and why?

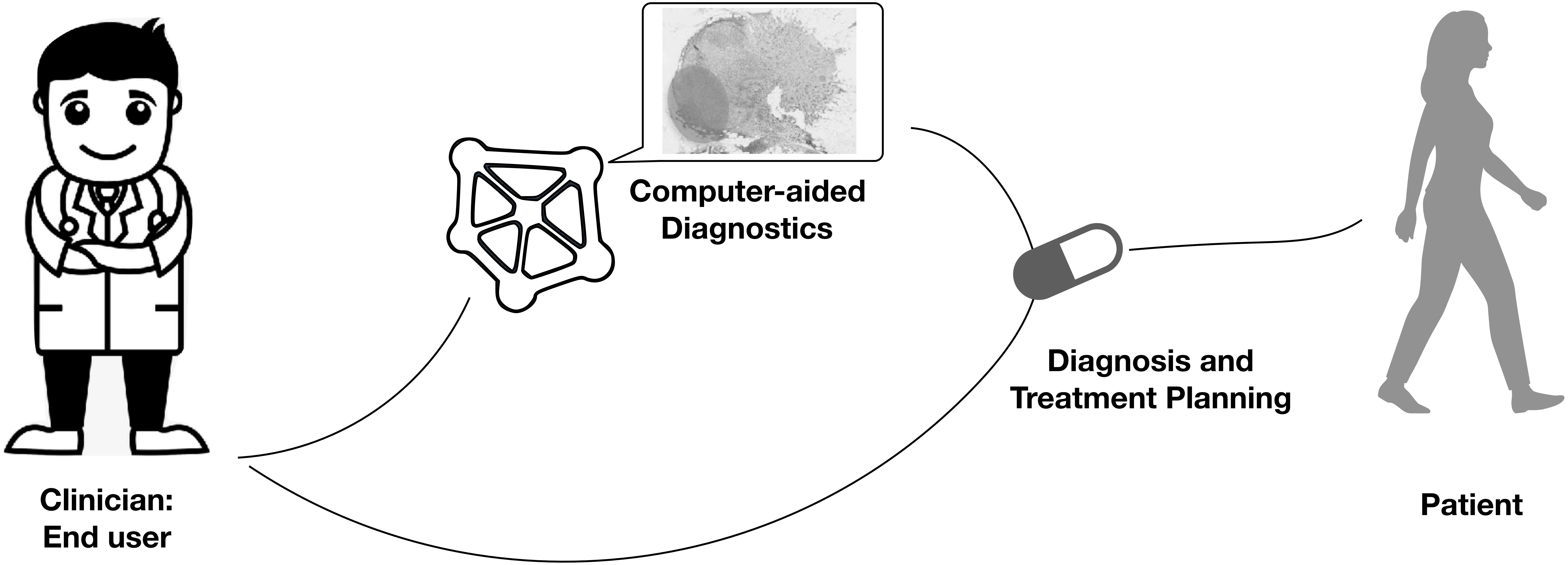
"Interpretability is the ability to explain or to present in understandable terms to a human*."

[Kim et al., 2018]

* not all humans are familiar with Machine Learning

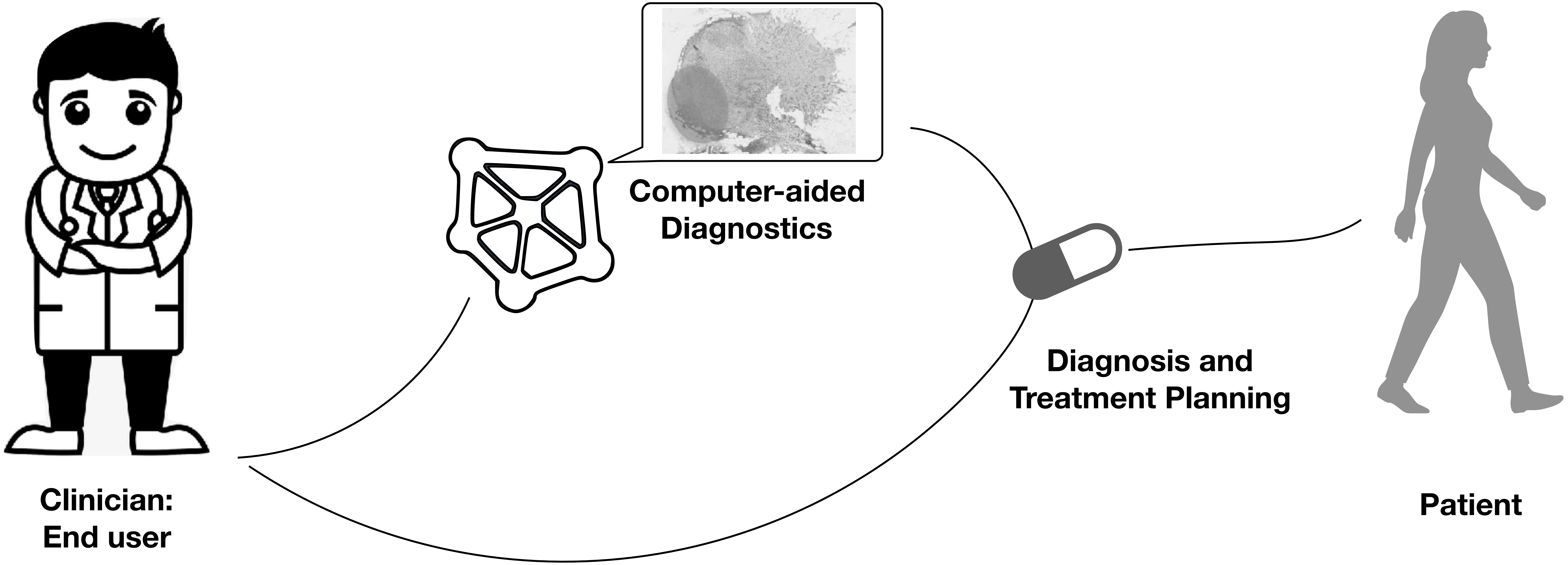
Example:

Human-Centric Interpretability for Cancer Diagnosis



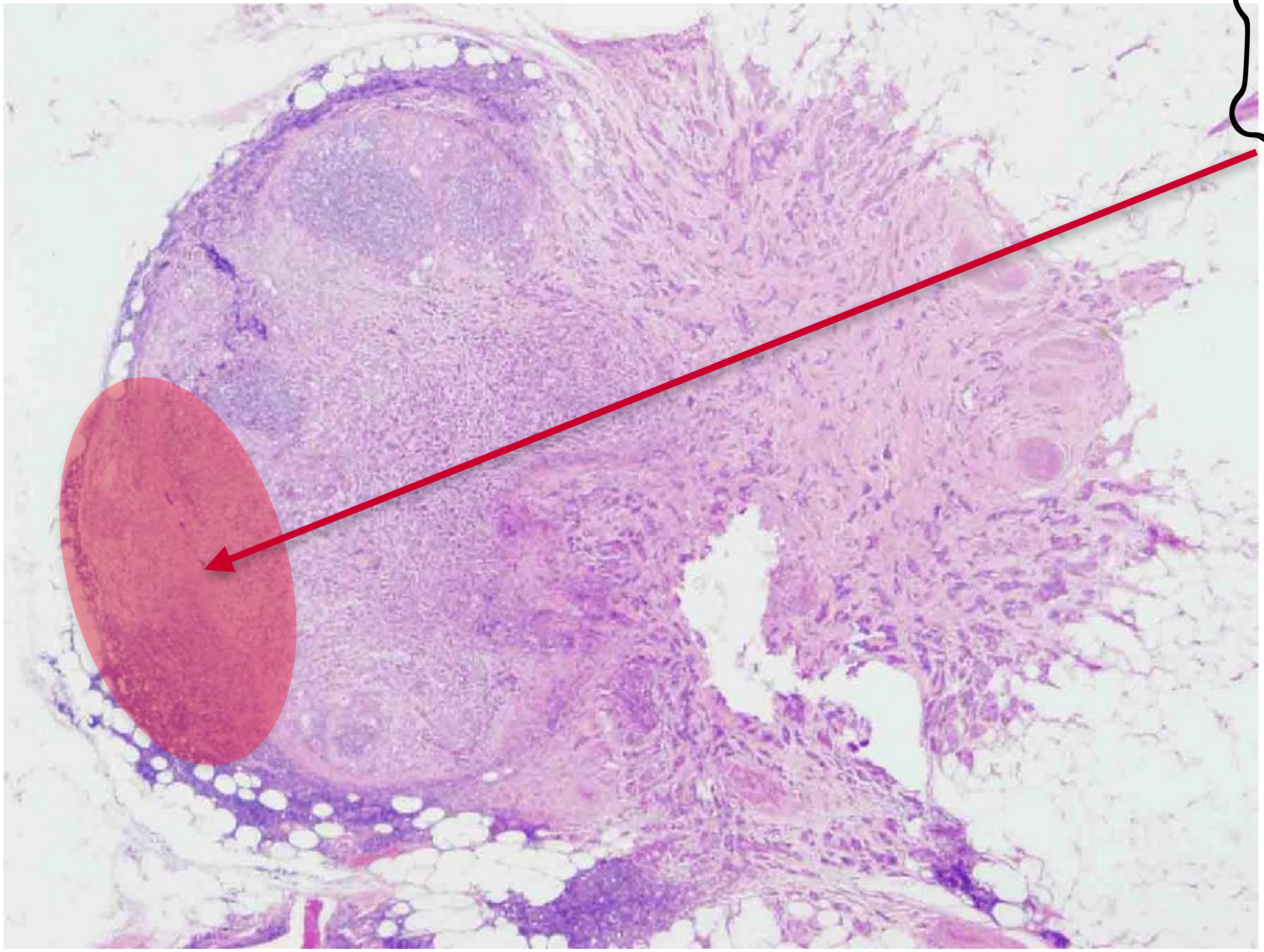
Example:

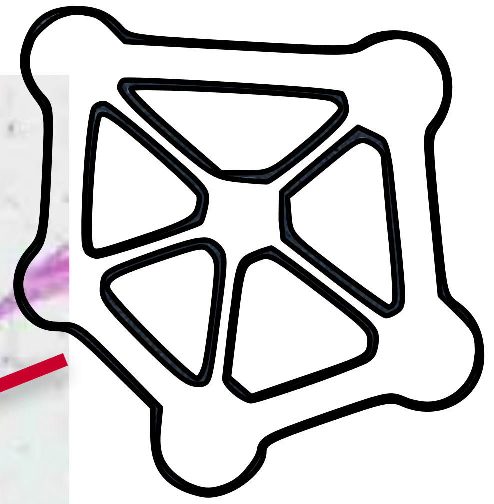
Human-Centric Interpretability for Cancer Diagnosis



1. *CNN for localization*

Clinician:
End user



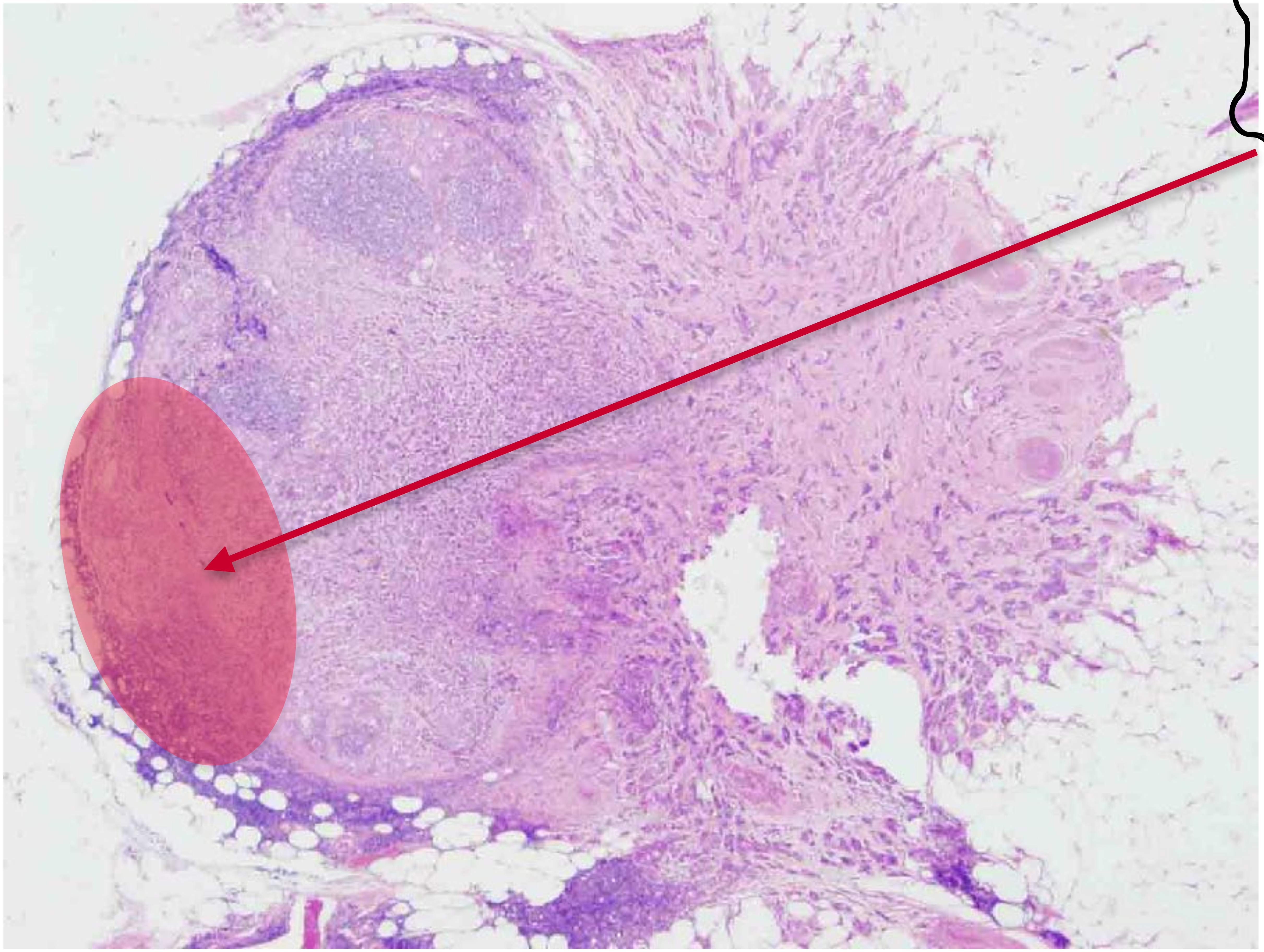


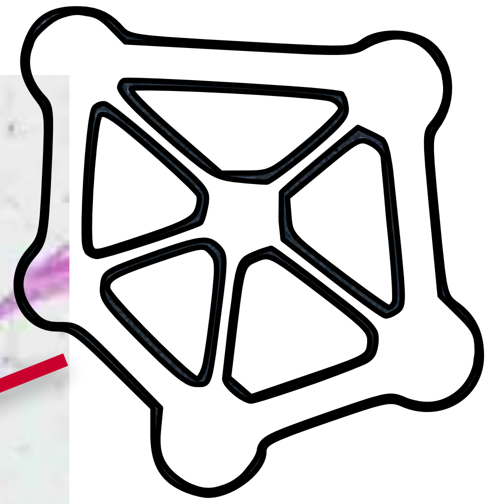
CNN

This is a high-grade tumor region!

2. CNN for localization with explanations of abnormalities

Clinician:
End user





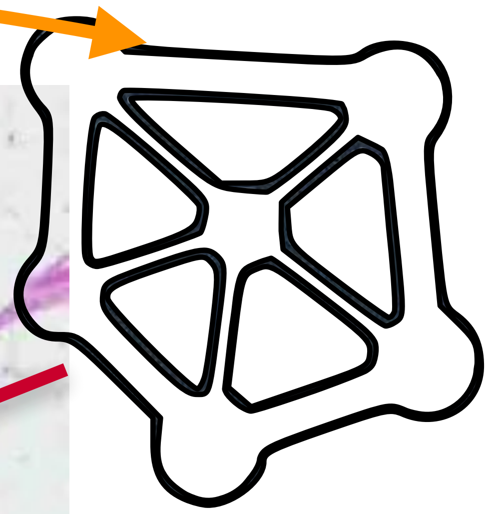
CNN
With
Interpretability

This is a high-grade tumor region:

1. The nuclei are 30% larger than non-tumor average
2. The nuclei texture appears vesicular (contrast is 40% larger than average)

Scenarios:

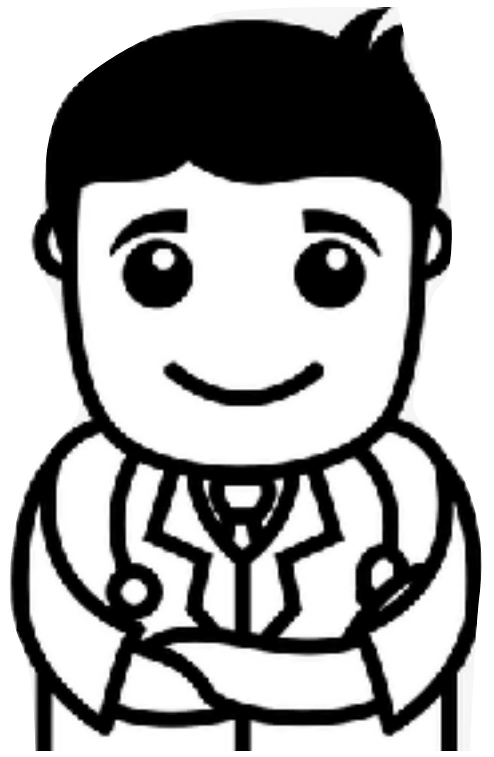
3. CNN for localization with explanations of abnormalities <u>and with guided feature learning by user-input</u>



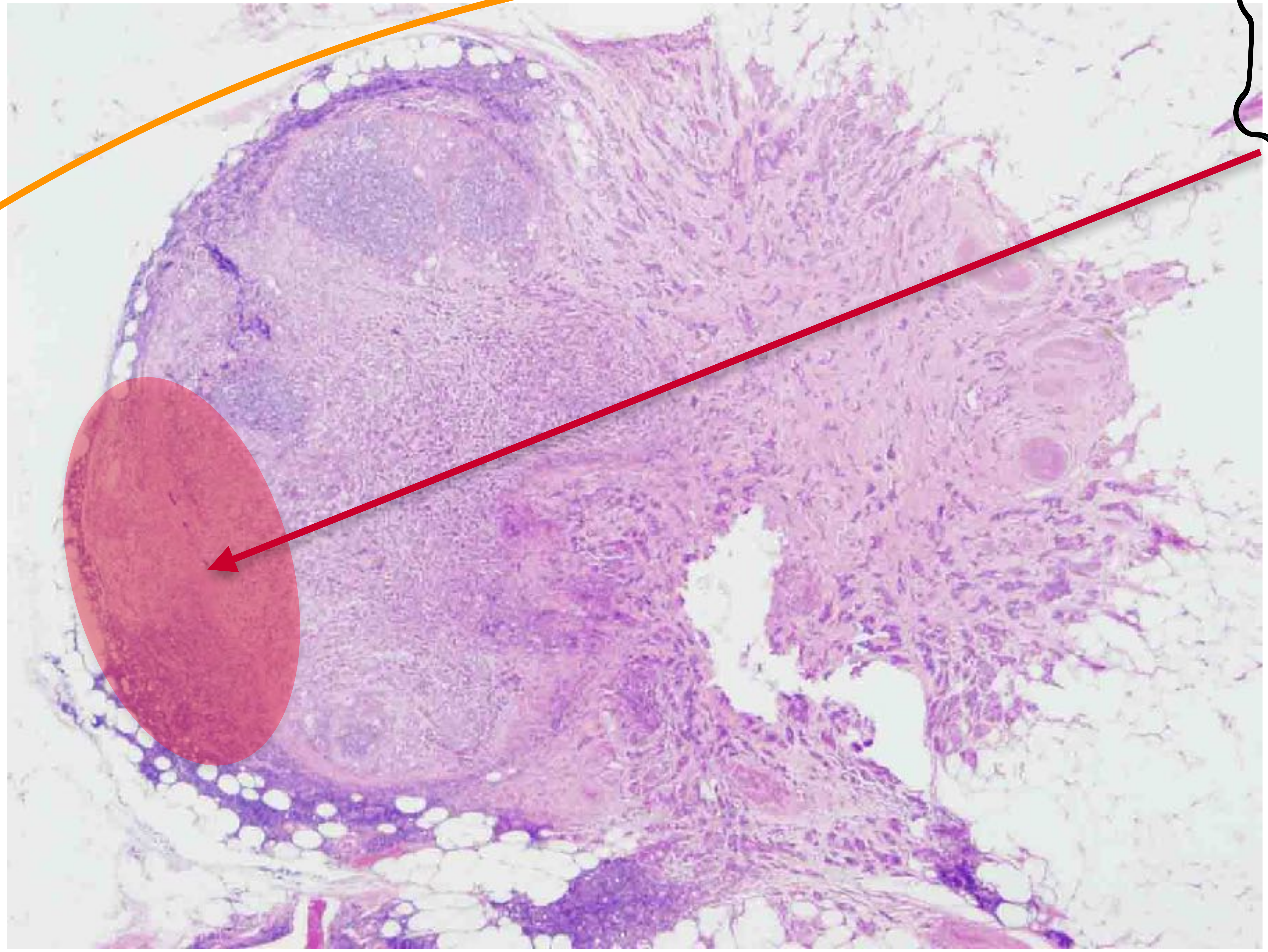
Guided CNN

With

Interpretability



Clinician:
End user

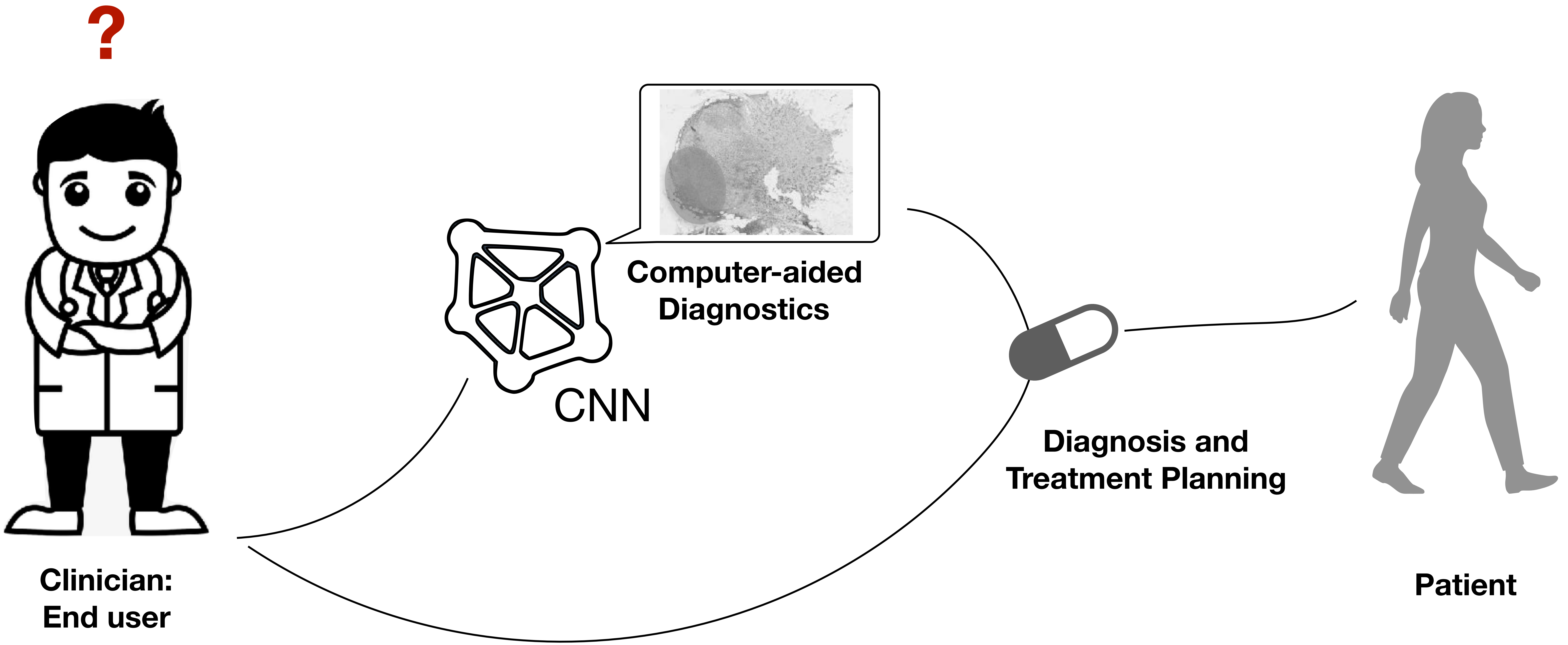


This is a high-grade tumor region:

1. The cells are 30% larger than non-tumor average
2. The nuclei texture appears vesicular (contrast is 40% larger than average)

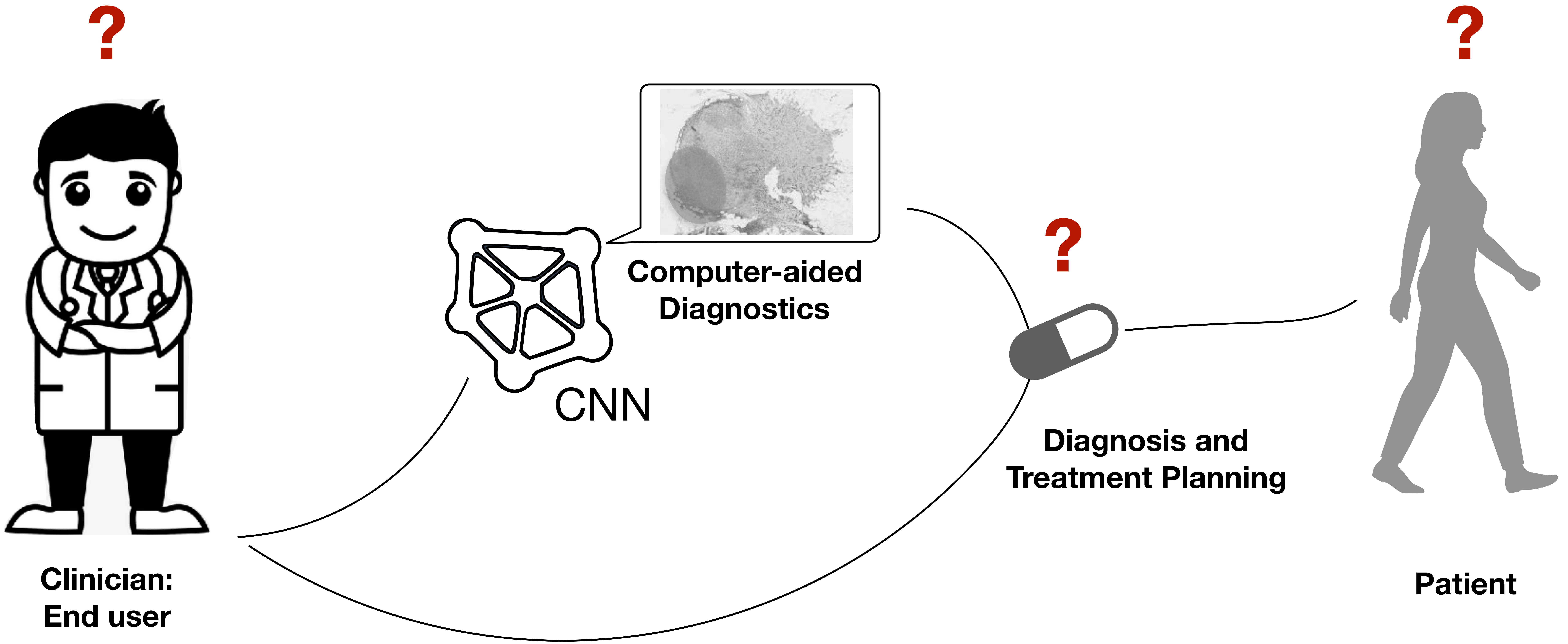
Let’s analyse the three scenarios

1. CNN for localization



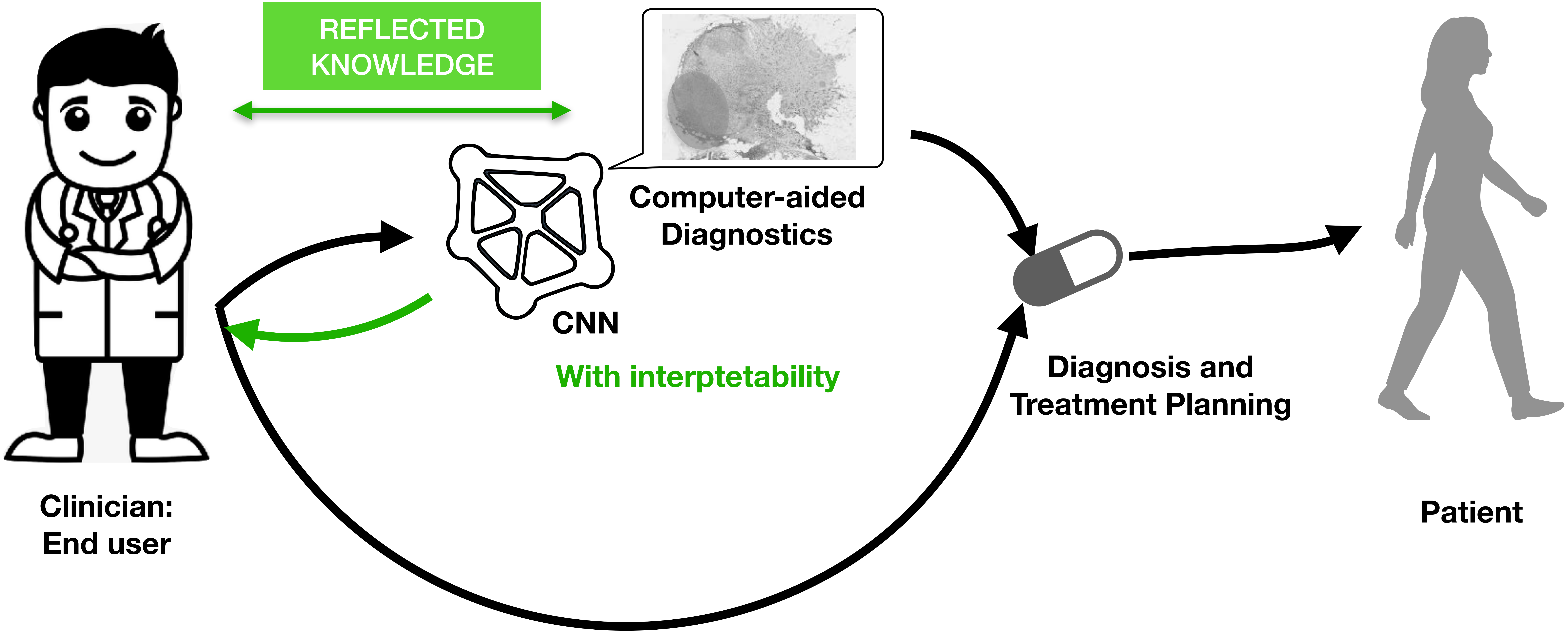
Scenarios:

1. CNN for localization



Scenarios:

2. CNN for localization with explanations of abnormalities

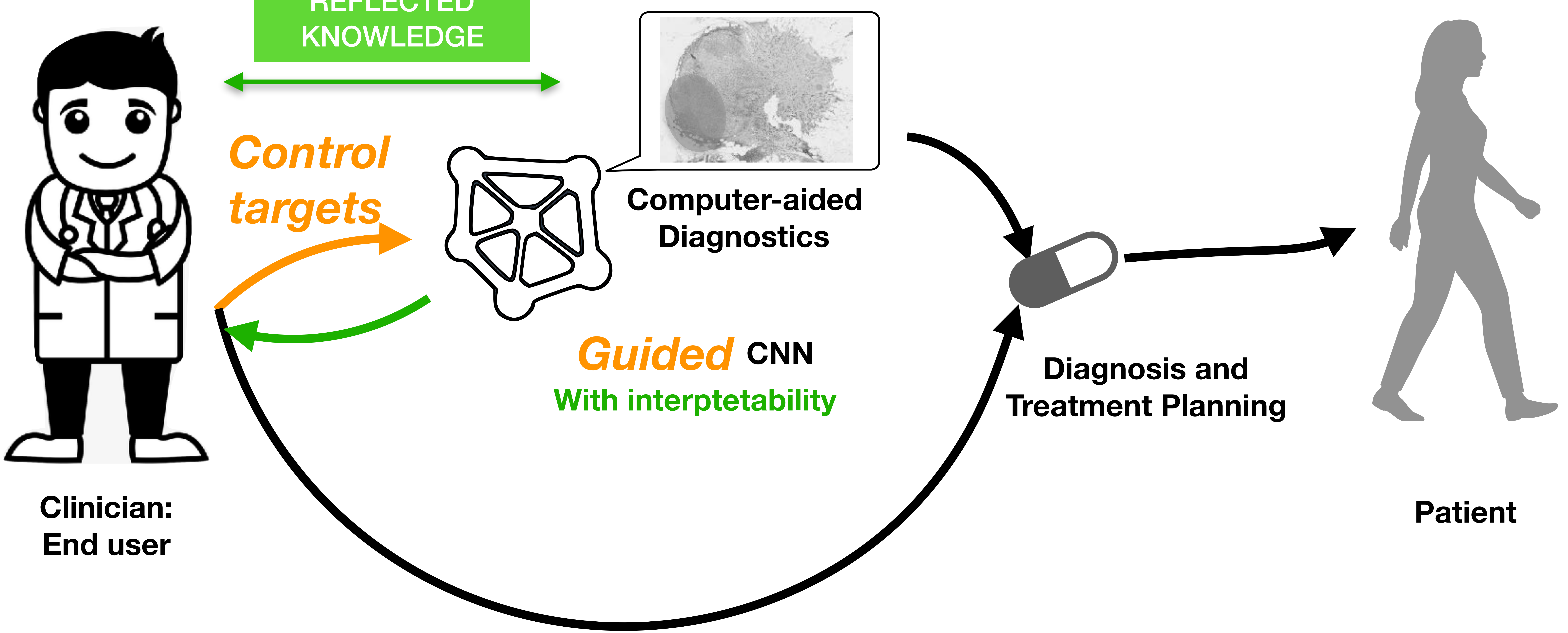


- Explaining the decisions of a complex model in understandable terms by doctors eases the interaction with AI and improves the quality of the diagnosis [Carrie J.C. et al., 2019].

Scenarios:

3. CNN for localization with explanations of abnormalities and with guided feature learning by user-input

Control Targets:
- Nuclei size is relevant
- Image domain is not relevant



To summarize

* CNNs for tumor localization can support pathologists in the diagnosis, but may leave them with **unanswered questions about the output** (scenario 1)
* **Interpretability should help** the clinician verify that the CNN decision making respects the guidelines and knowledge in the domain (scenario 2).
* The expertise of clinicians is a valuable input for the network training, that **could** be **guided** to ensure that certain visual features are taken into account and others are not (scenario 3).

Human-centric DL interpretability

=

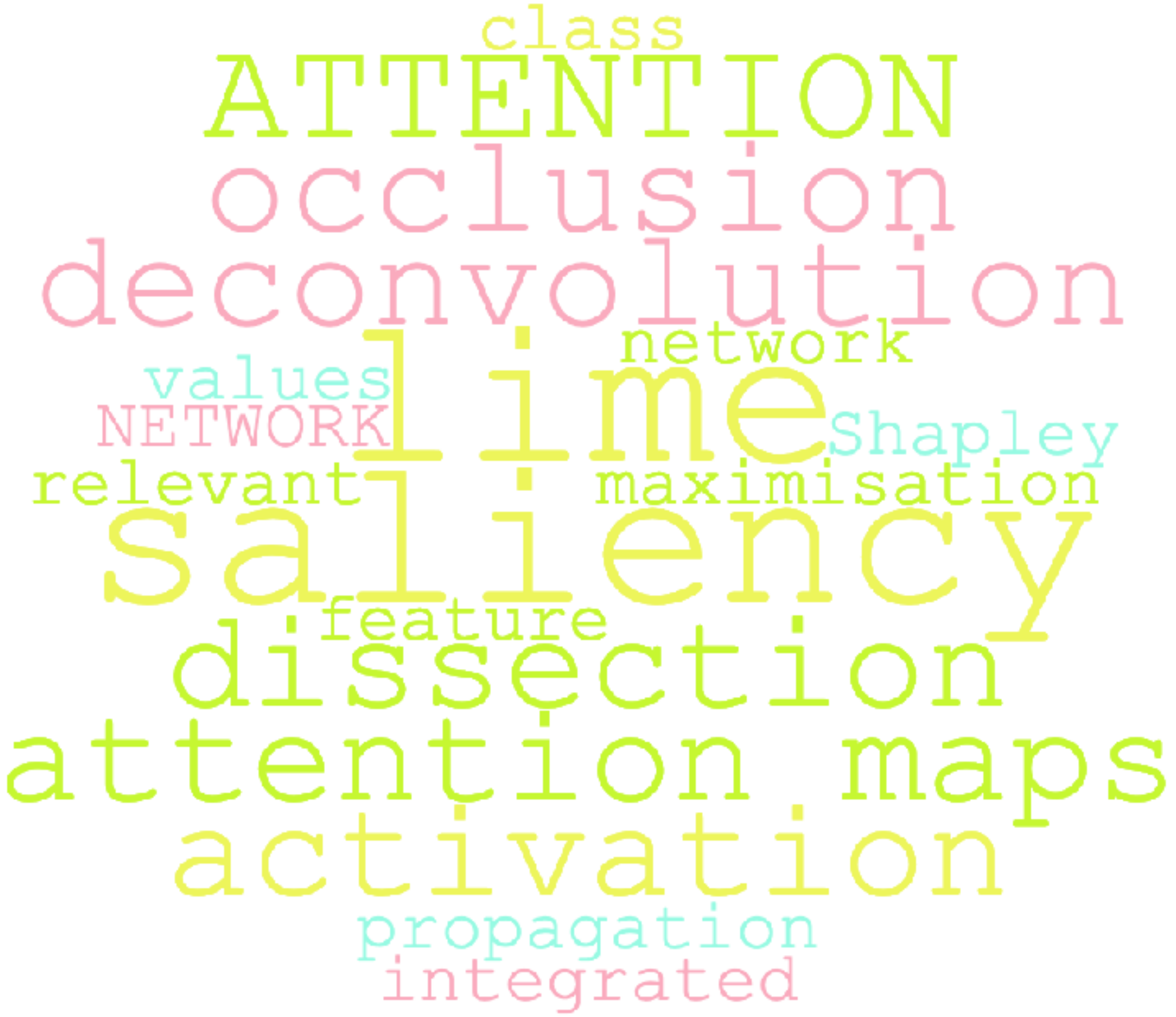
A tool that **supports** the pathologists in making decisions by providing **explanations** and allowing the introduction of **feedback** to refine training

Our work in this direction

* **Evaluation of visualization methods for histopathology**
* Concept-based interpretability of CNNs
* Guidable CNNs

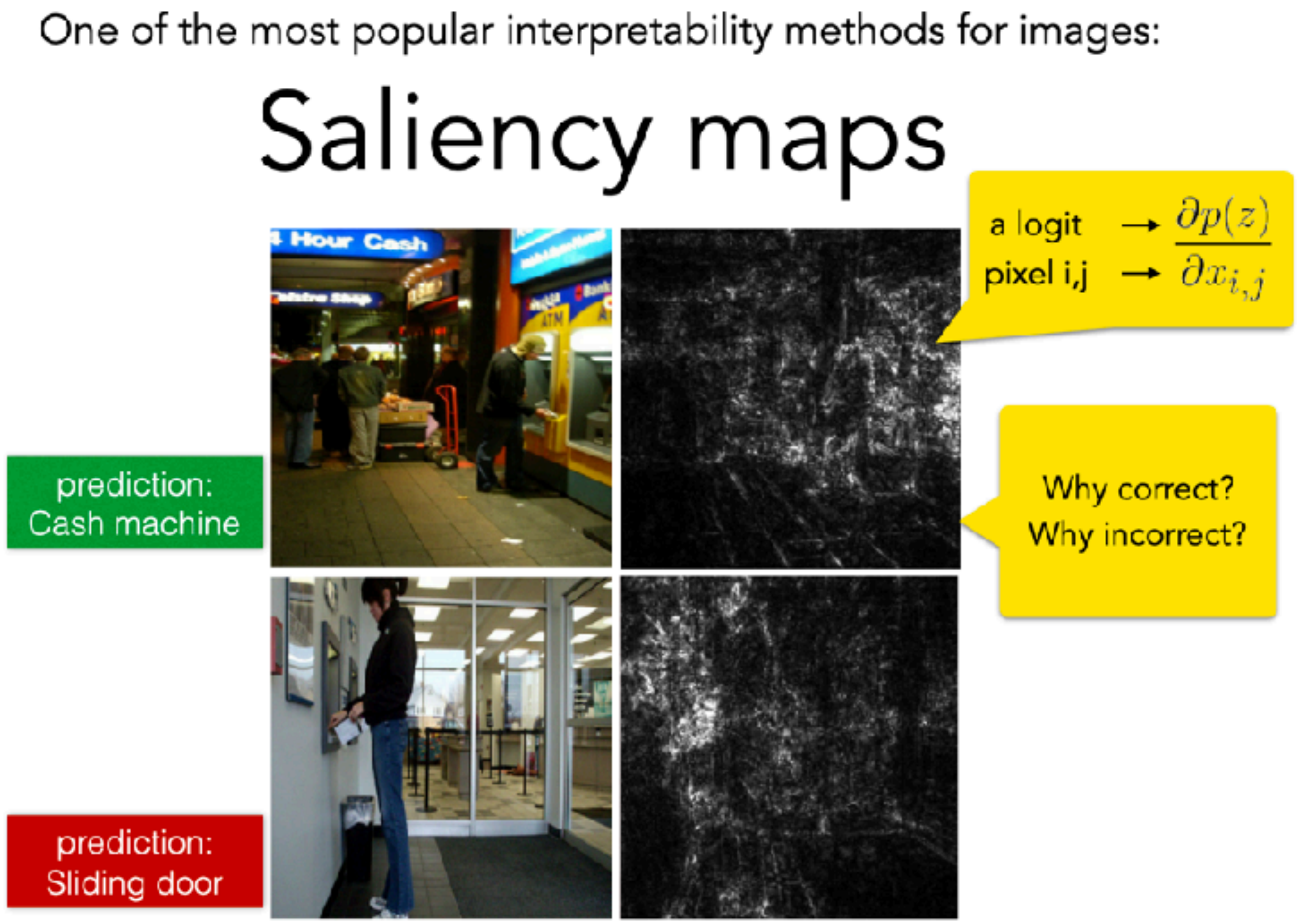
Feature-attribution:

Evaluation of visualization tools



e.g saliency

$$\frac{\partial output}{\partial input}$$



Slide credits: B. Kim

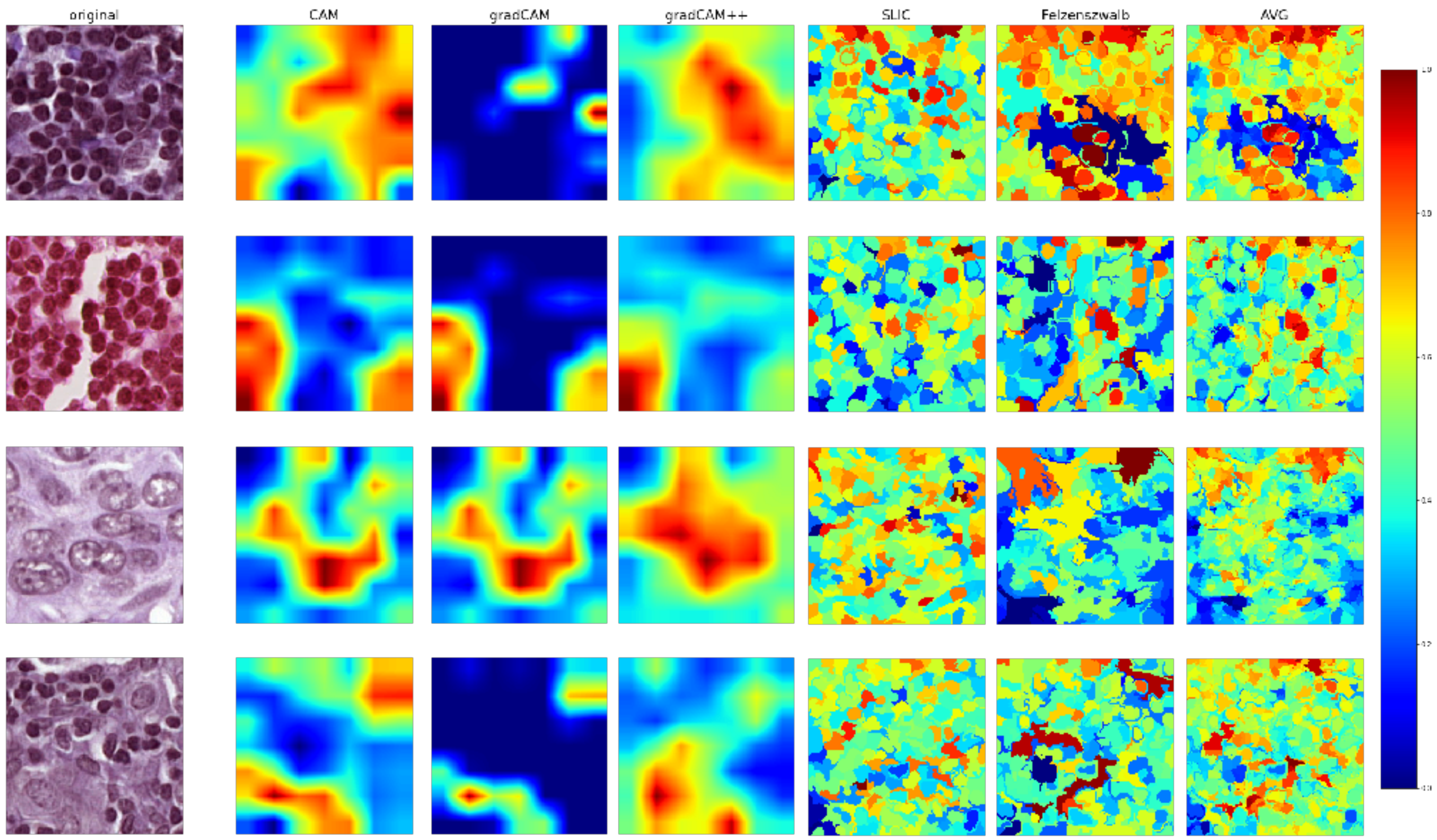
Issues:

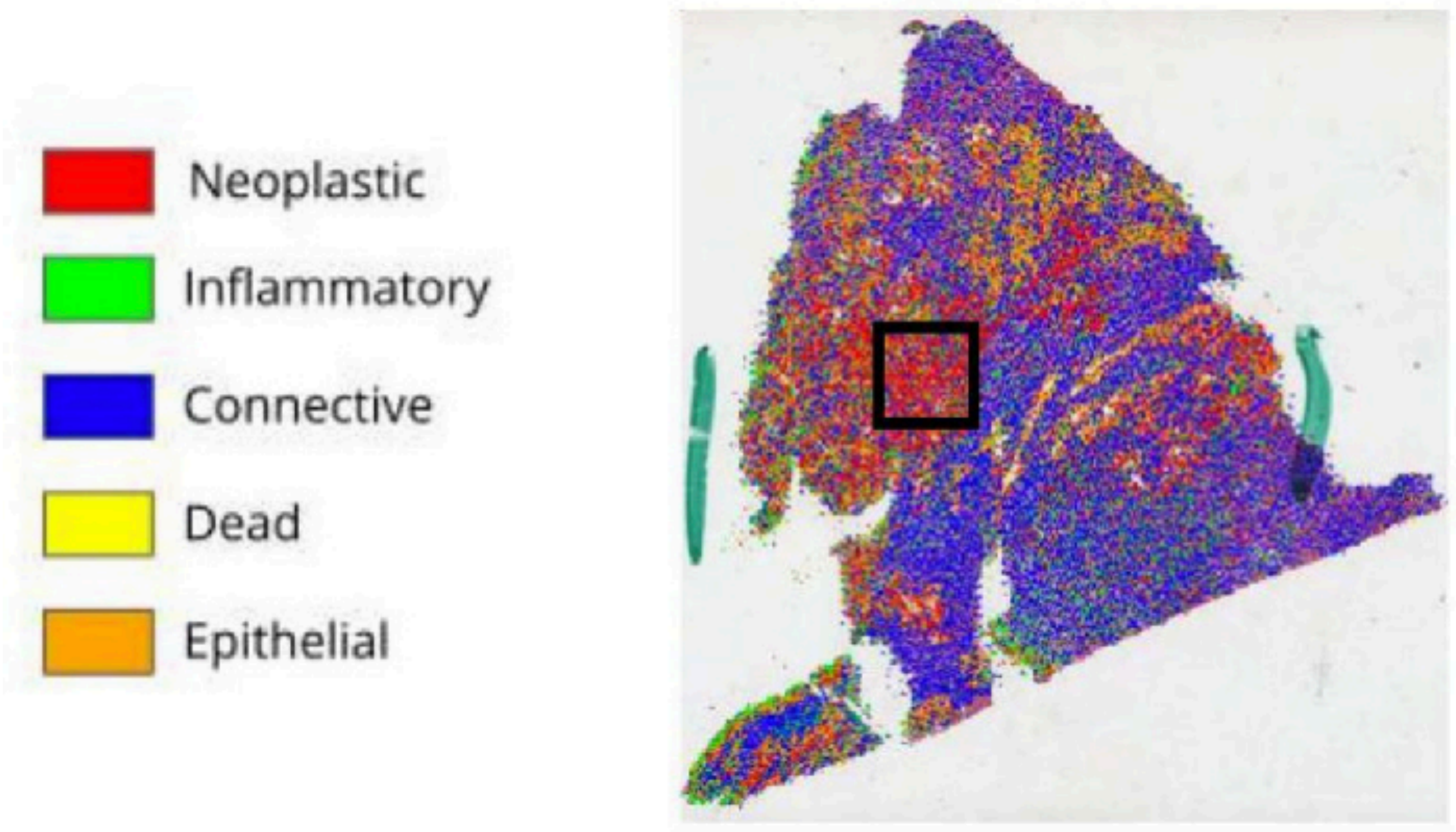
- Difficult Abstraction
- Sometimes Ambiguous [Kim et al., 2018]
- Consistency issues [Adebayo et al., 2018]

Hes·so// VALAIS WALLIS

Feature-attribution:

Evaluation of visualization tools





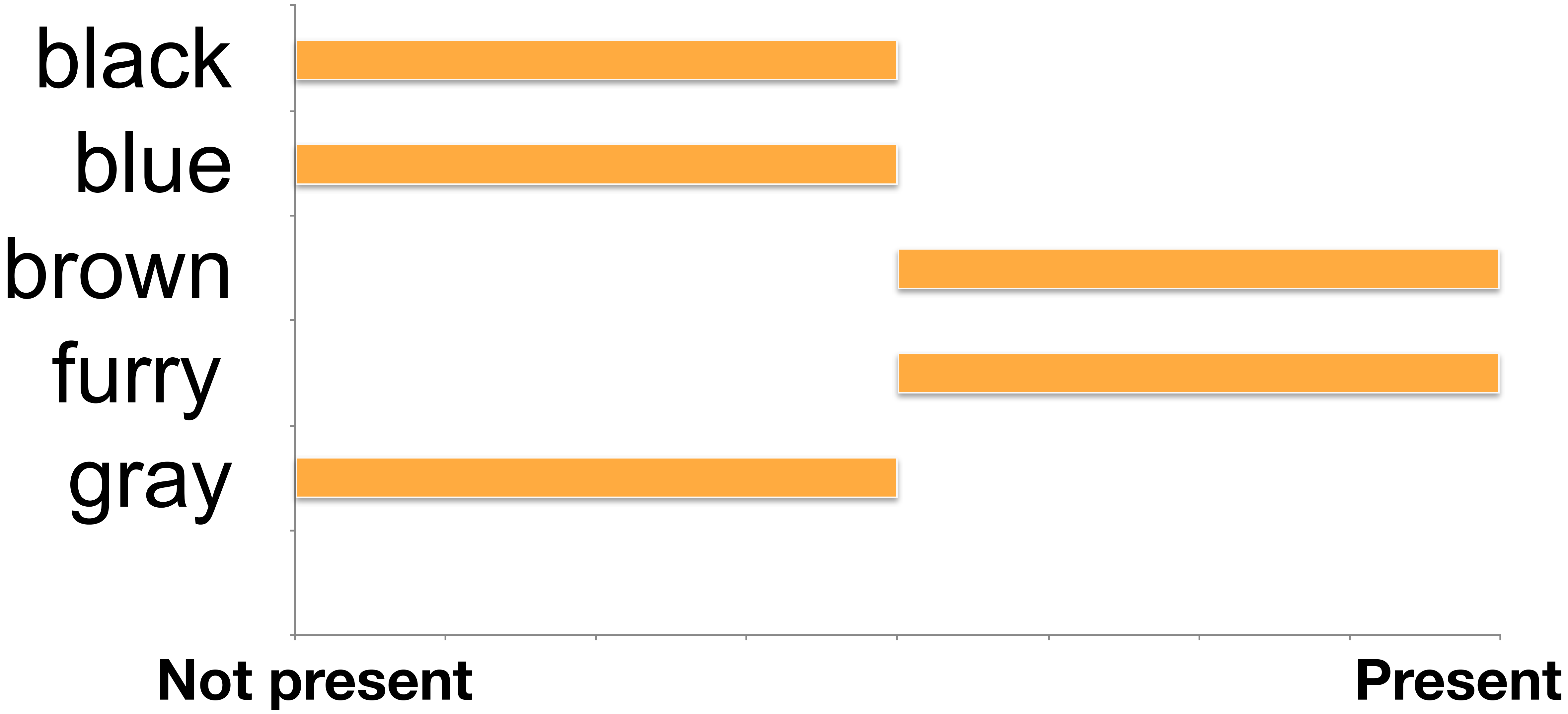
https://jgamper.github.io/PanNukeDataset/

Our work in this direction

* Evaluation of visualization methods for histopathology
* **Concept-based interpretability of CNNs**
* Guidable CNNs

Concept attribution with Regression Concept Vectors

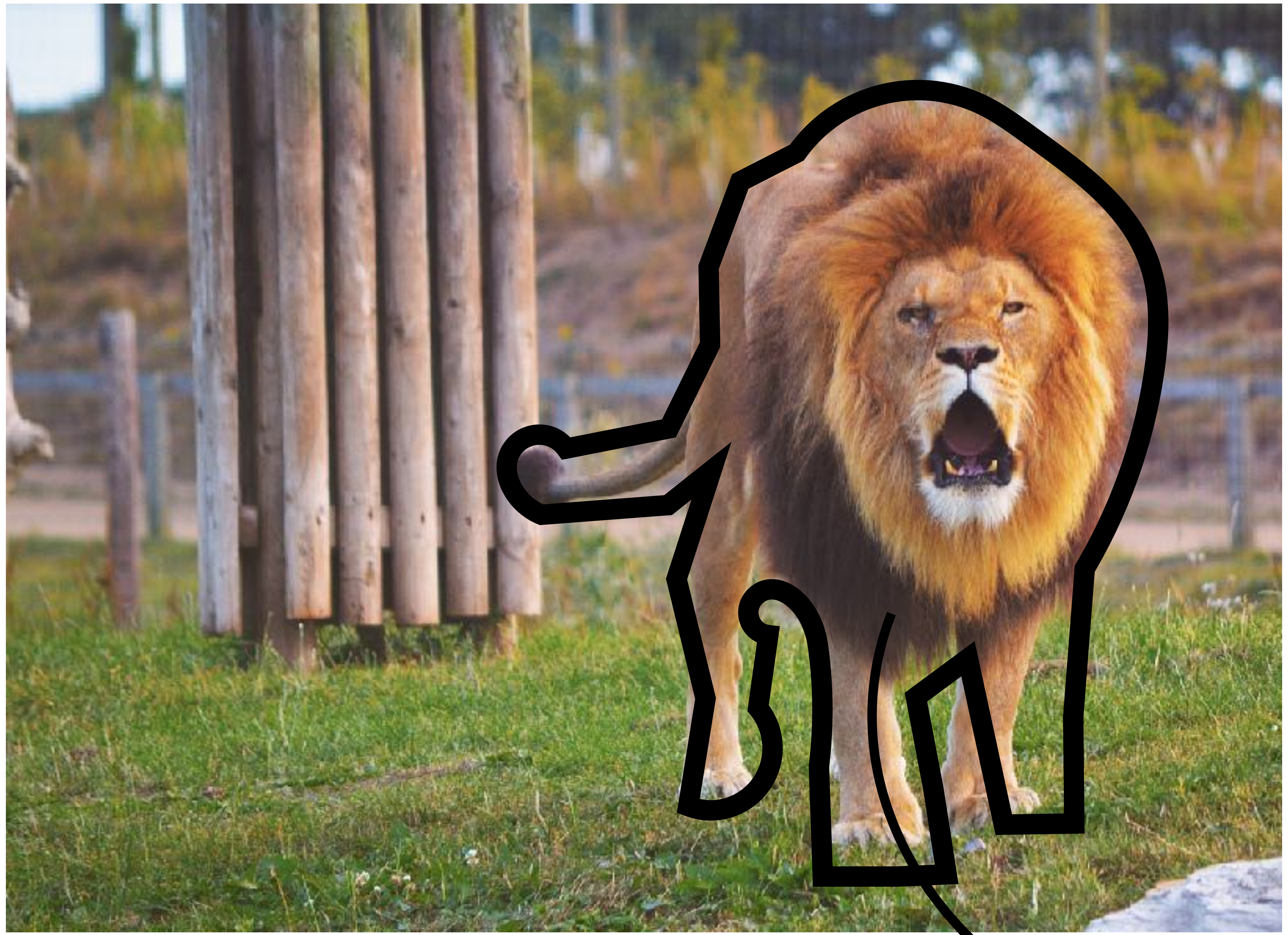
Taking inspiration from [Kim et al., 2018] on interpreting CNN activations with human-friendly binary concepts (presence vs absence).

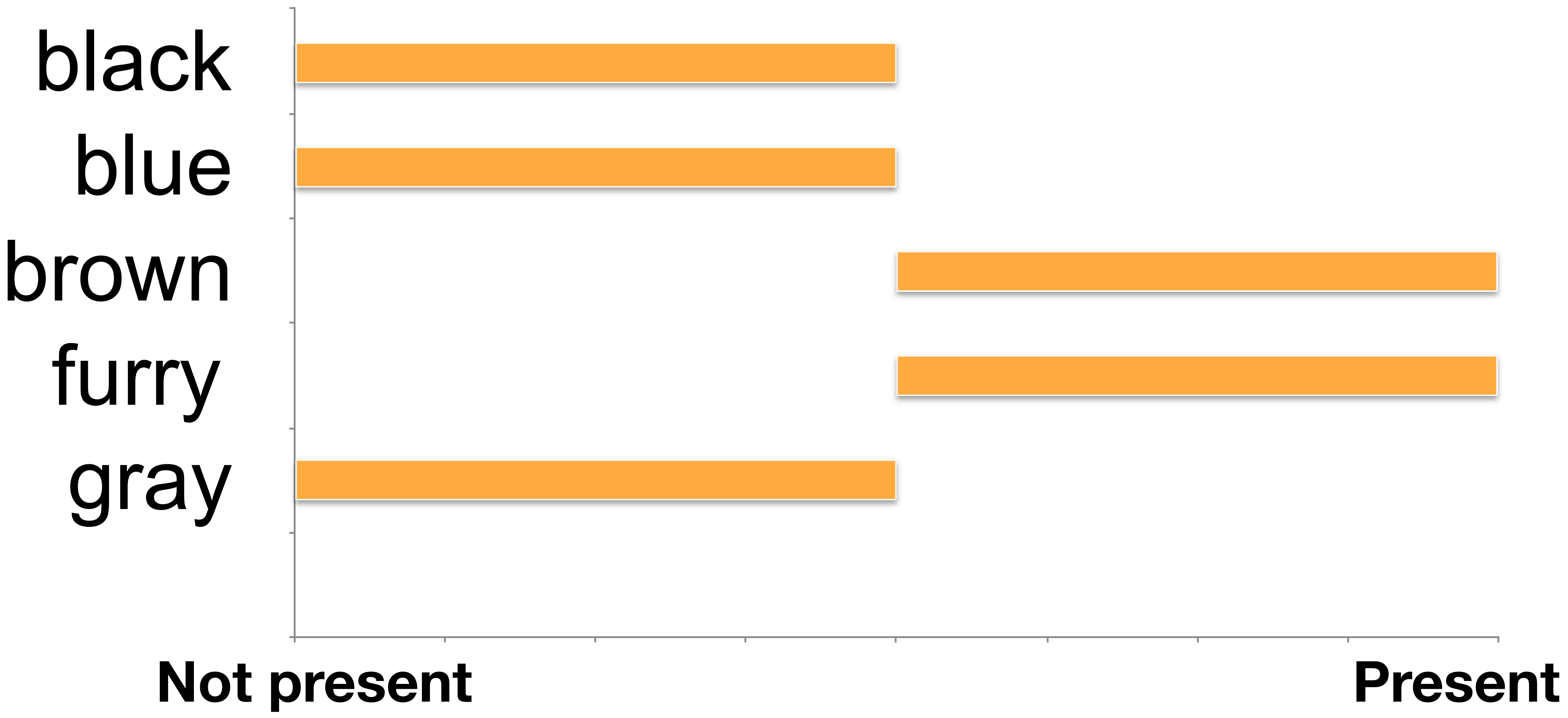


Concept-based interpretability:

Concept attribution with Regression Concept Vectors

Taking inspiration from [Kim et al., 2018] on interpreting CNN activations with human-friendly binary concepts (presence vs absence).

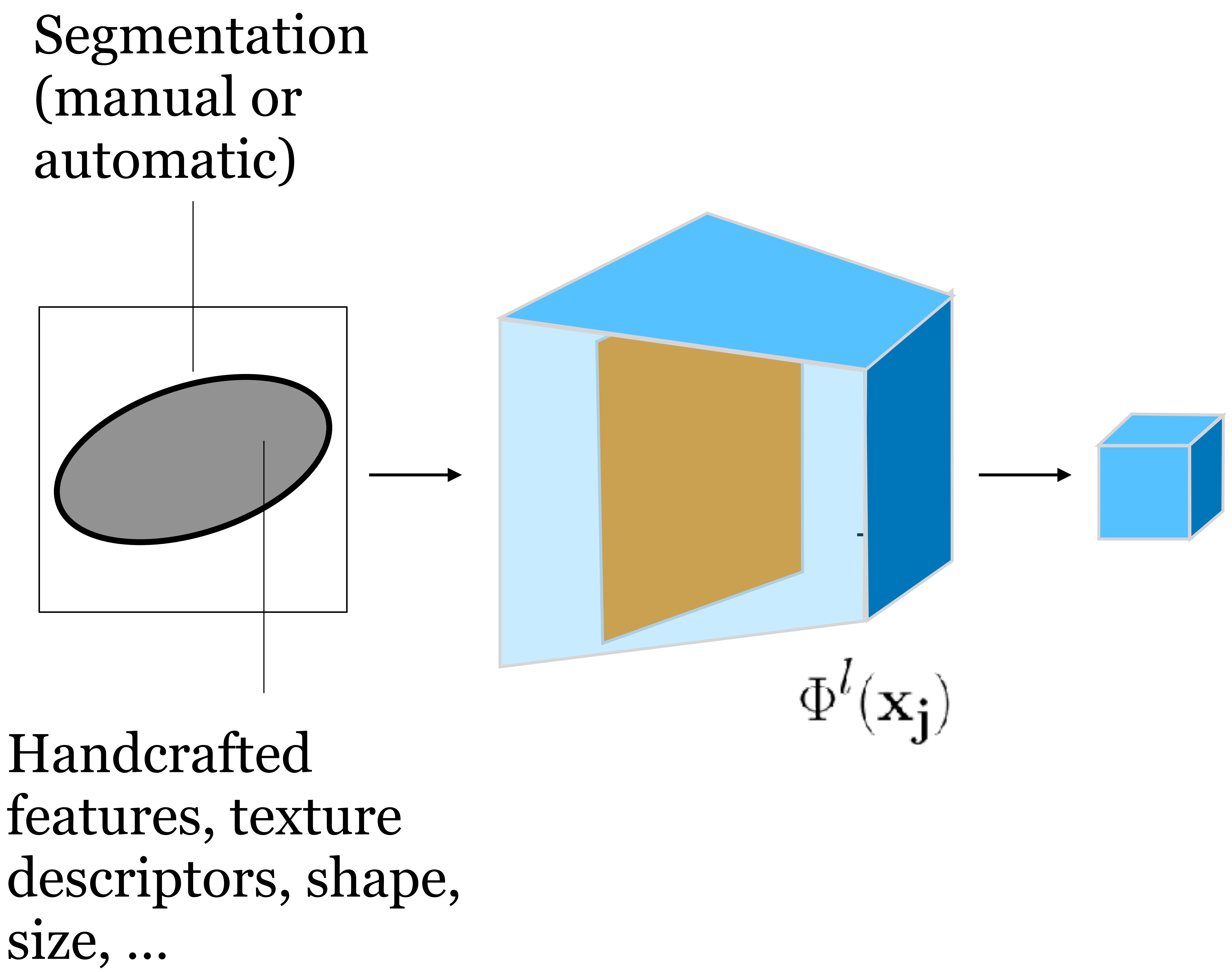




Measuring texture “**coarseness**”, “**brown-ness**”

Concept-based interpretability:

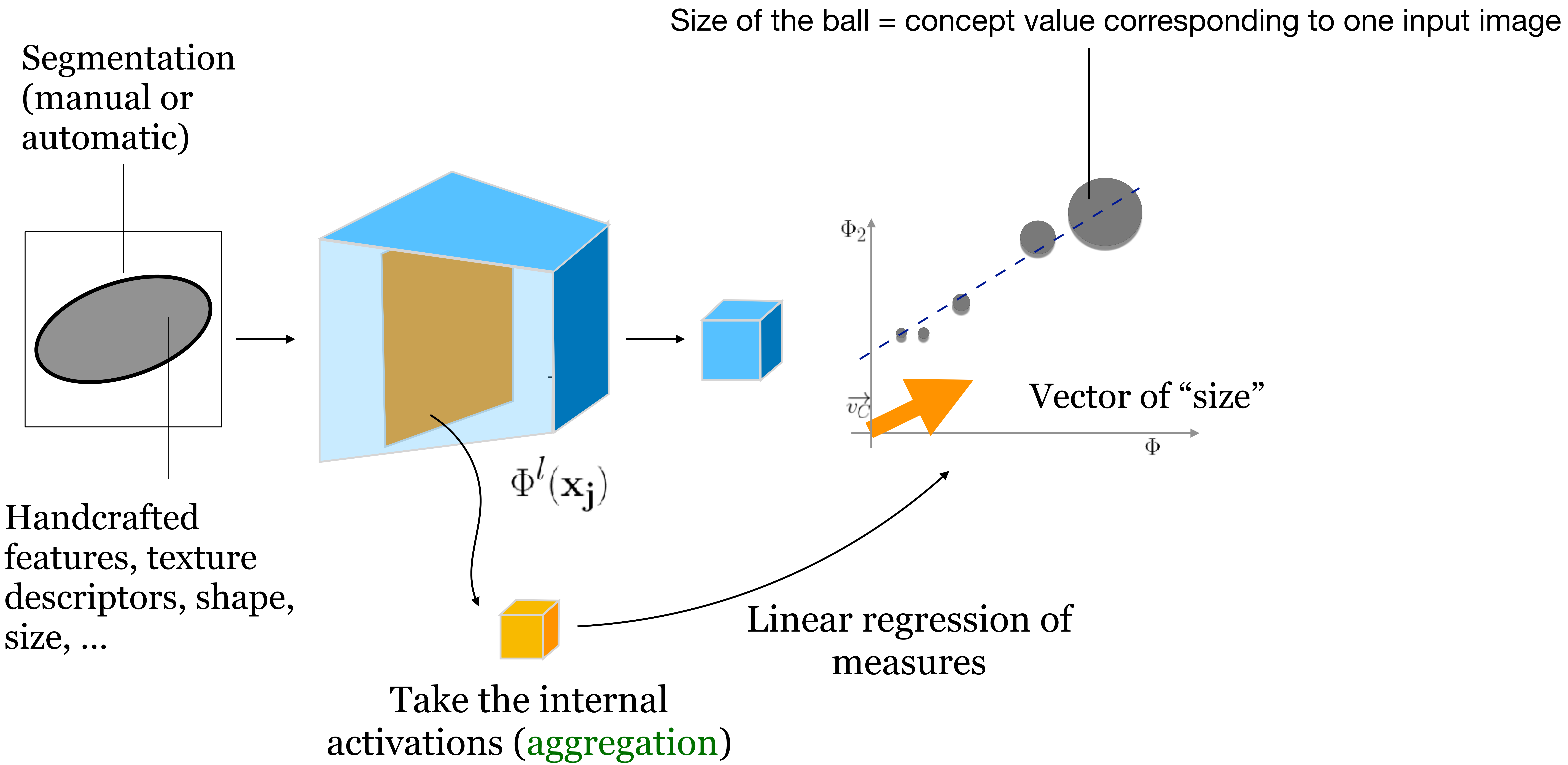
*Concept attribution with Regression Concept Vectors**



 Best paper award, iMIMIC, MICCAI 2018!

Concept-based interpretability:

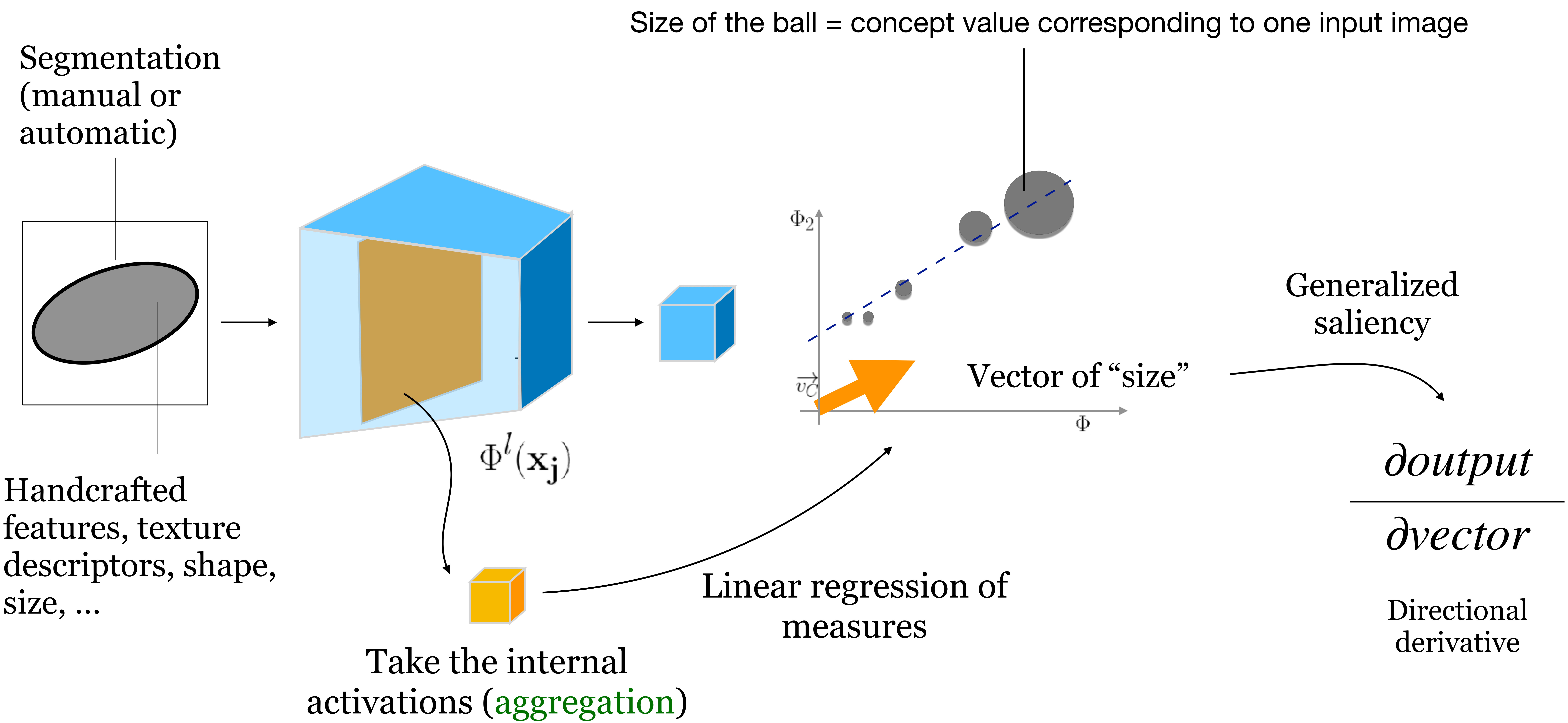
*Concept attribution with Regression Concept Vectors**



Best paper award, iMIMIC, MICCAI 2018!

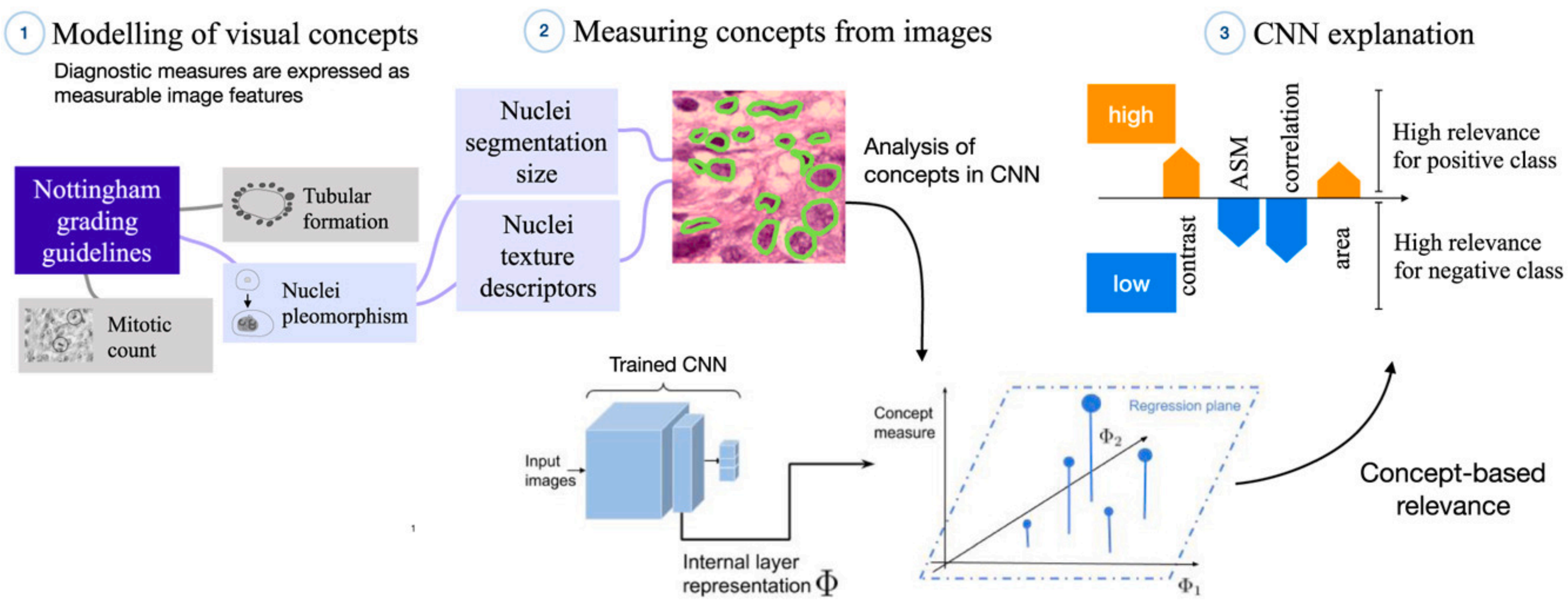
Concept-based interpretability:

*Concept attribution with Regression Concept Vectors**



Concept-based interpretability:

Regression Concept Vectors: application to histopathology



[Graziani et. al, 2020]

Remarks

* **Interpretability can be used to** verify that the CNN decision making respects clinical guidelines and knowledge in the domain
* **Visualizations of saliency heatmaps** give feedback on the relevant input pixels, while **concept-based** explanations use directly clinically relevant measures such as nuclei size and appearance.
* The expertise of clinicians can be used to **guide network training** by the combination of multitask and adversarial learning.

Remarks

* **Interpretability can be used to** verify that the CNN decision making respects clinical guidelines and knowledge in the domain
* **Visualizations of saliency heatmaps** give feedback on the relevant input pixels, while **concept-based** explanations use directly clinically relevant measures such as nuclei size and appearance.
* The expertise of clinicians can be used to **guide network training** by the combination of multitask and adversarial learning.

Q&A