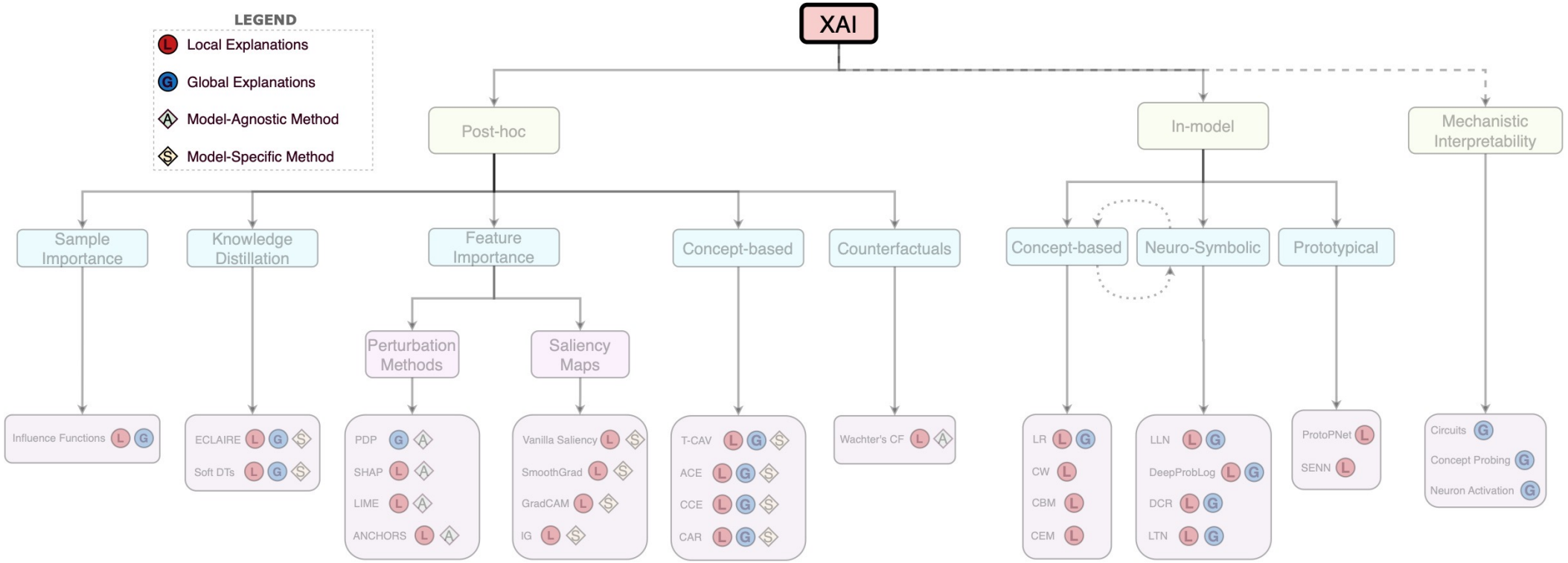
EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Conclusion - Lent 2025

UNIVERSITY OF CAMBRIDGE

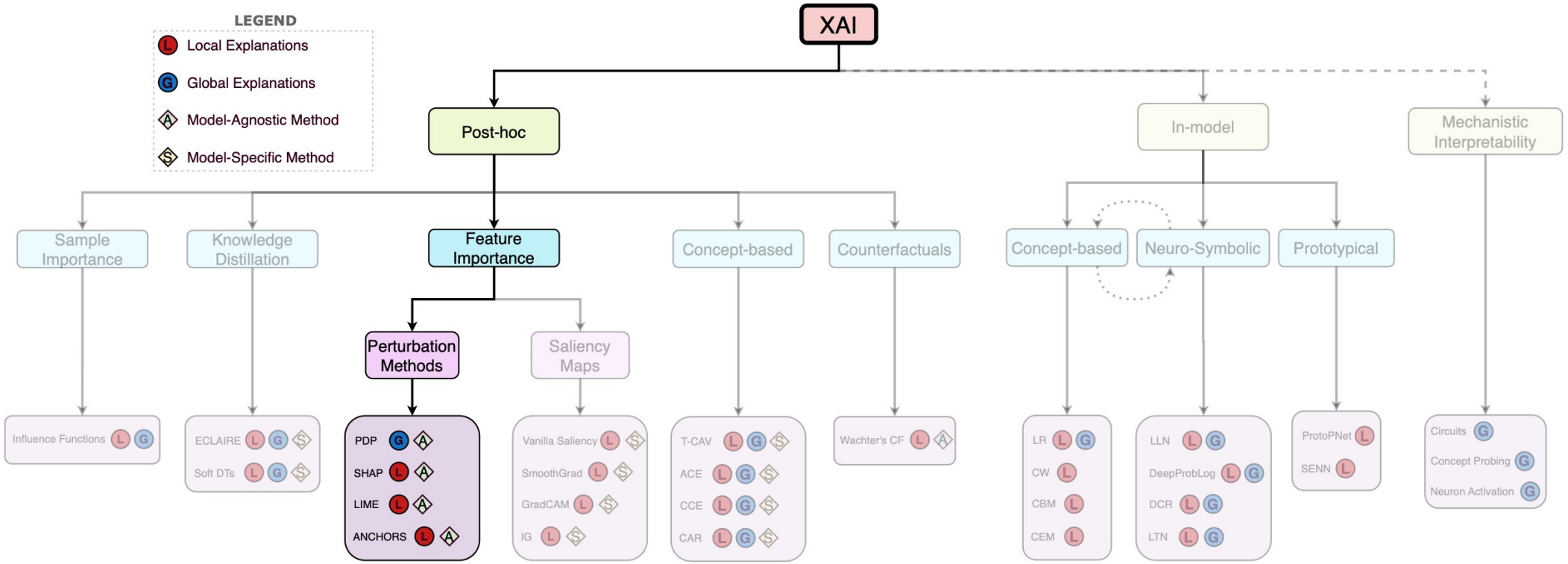
THE ROAD AHEAD IN XAI

BEFORE ANYTHING, A LITTLE SUMMARY



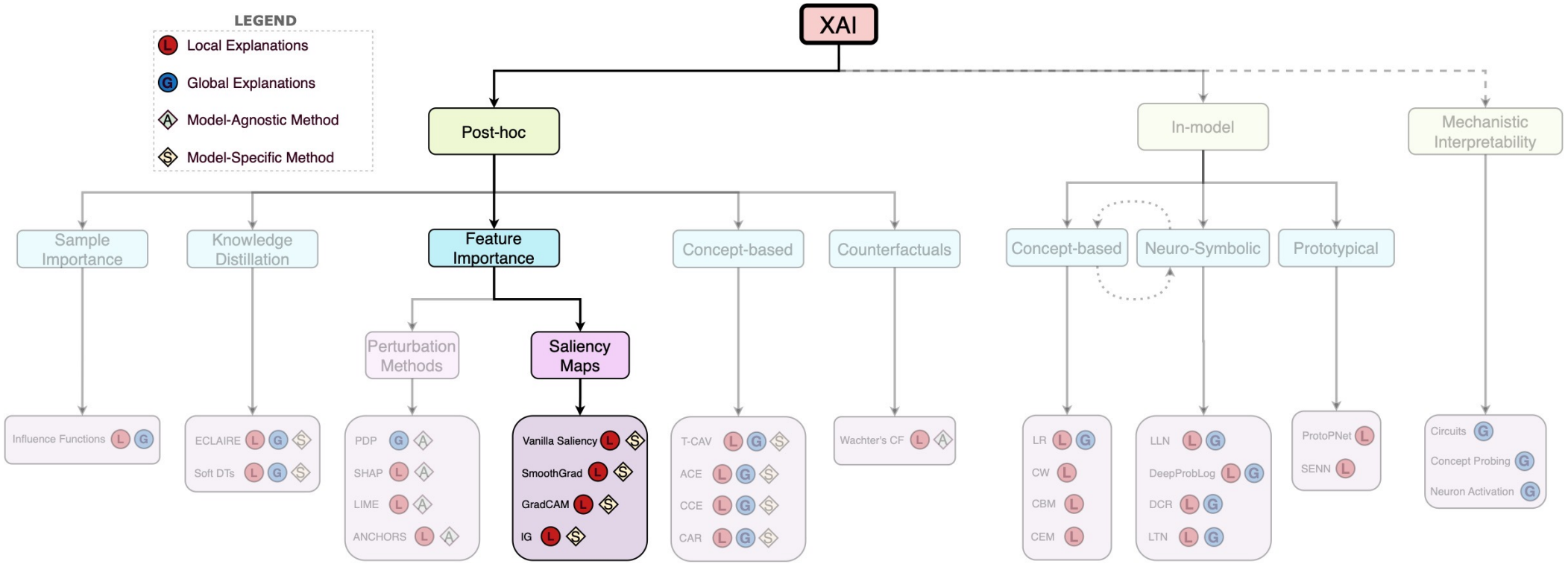
Let's take a step back and see the different areas of XAI that we discussed in the past eight weeks

BEFORE ANYTHING, A LITTLE SUMMARY



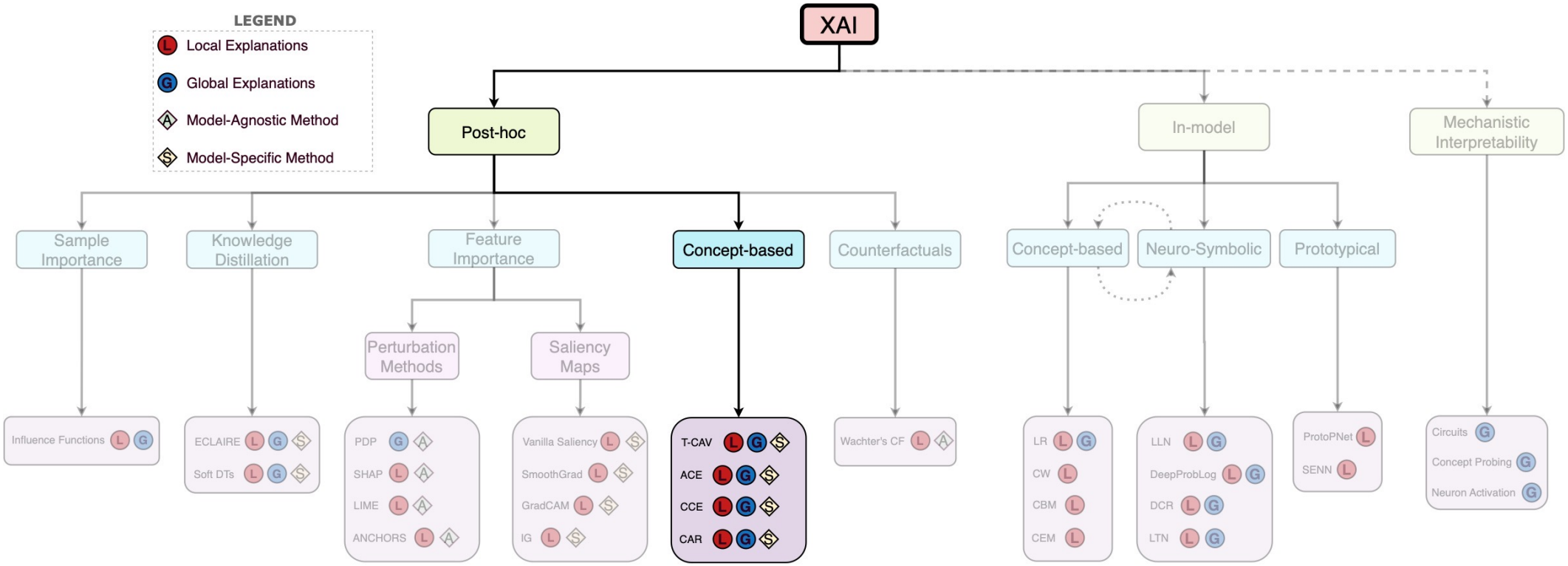
First, we discussed model-agnostic feature importance methods (so-called perturbation methods)

BEFORE ANYTHING, A LITTLE SUMMARY



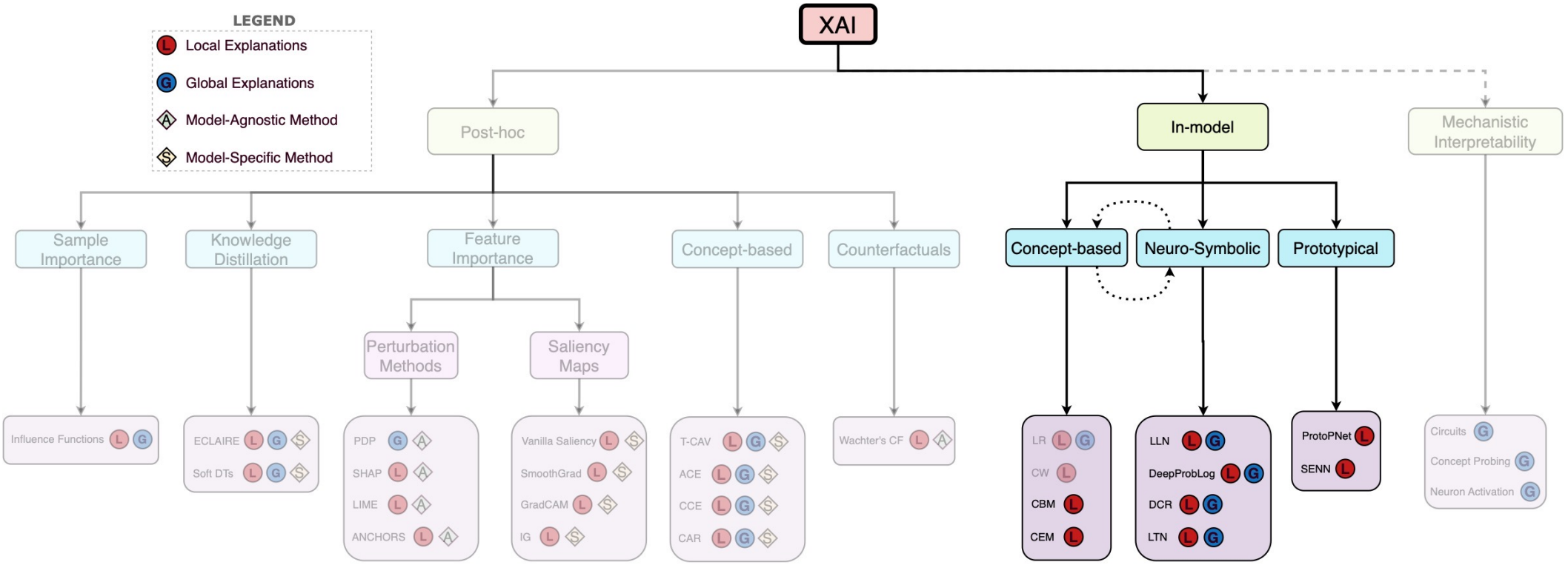
Then, we focused on model-specific feature importance methods like saliency maps

BEFORE ANYTHING, A LITTLE SUMMARY



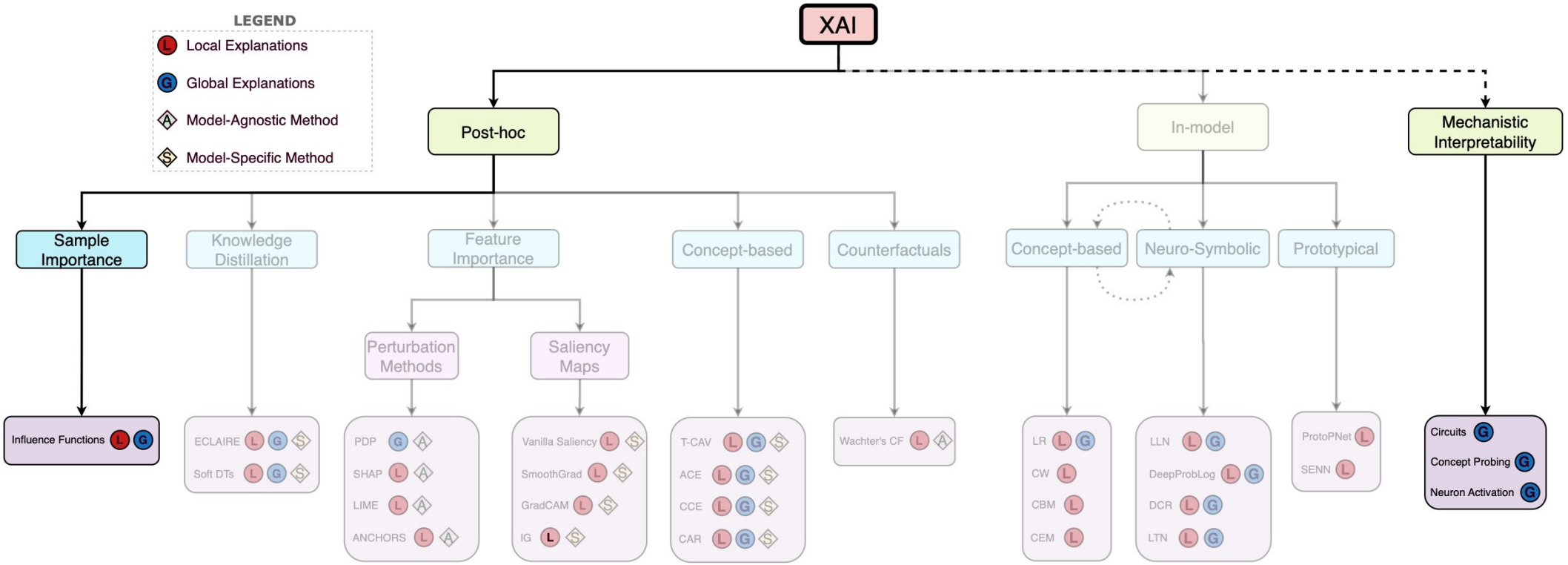
We followed up by exploring how to use human-understandable concepts to explain different models

BEFORE ANYTHING, A LITTLE SUMMARY



We discussed some limitations of post-hoc methods and introduced in-model explainable methods

BEFORE ANYTHING, A LITTLE SUMMARY



We concluded by briefly introducing recent trends such as influence functions and mechanistic interpretability

REMINISCING ABOUT THE FUTURE OF XAI

SOME REMINISCING ABOUT THE FUTURE OF XAI

With AI getting more and more intertwined with day-to-day activities, we conjecture that XAI will become a core component for:

1. **Regulation:** GDPR and the EU AI Act are just the beginning
2. **Standardisation:** most global efforts (telecommunications, internet, etc...) have been standardised under some international organisation, AI will soon follow
3. **Data science**: we will not analyse data, we will analyse models
4. **Scientific exploration:** Explaining and being explained is crucial for human understanding. Knowledge discovery can be exponentially improved if we are able to interrogate AIs



Thank you!

EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Explainable Artificial Intelligence

Lent 2025

UNIVERSITY OF CAMBRIDGE