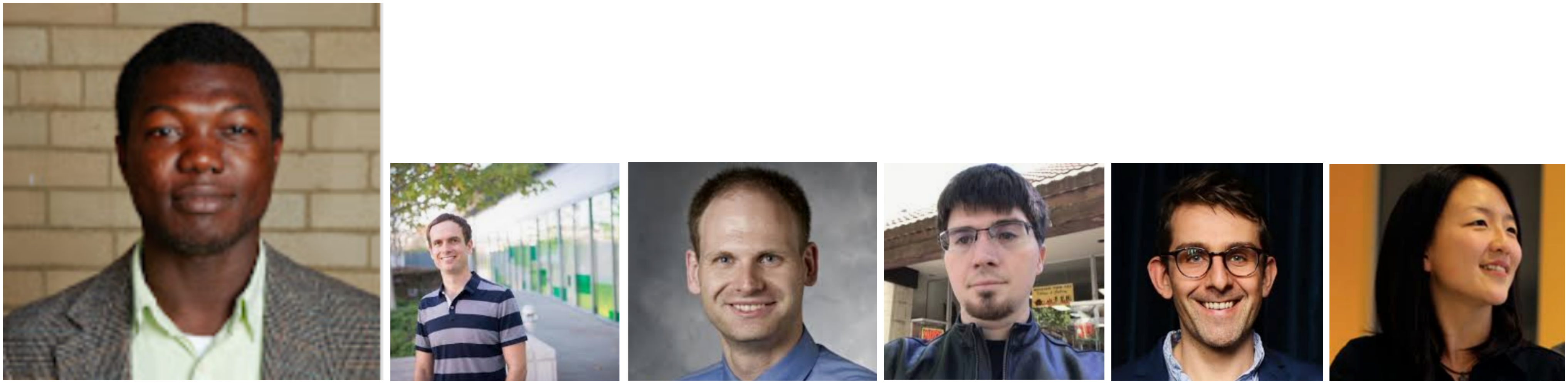
# Sanity Checks for Saliency Maps



Julius Adebayo[*+], Justin Gilmer[#], Michael Muelly[#], Ian Goodfellow[#], Moritz Hardt[^#], Been Kim[#]

[*]Work was done during the Google AI residency program, [+]MIT, [^]UC Berkeley, [#]Google Brain.

# Interpretability

To use machine learning more **responsibly**.

# Investigating post-training interpretability methods.

Given a fixed model, find the **evidence** of **prediction**.

$$\text{Explanation}|\text{Model}$$

# Investigating post-training interpretability methods.

A trained machine learning model (e.g., neural network)

$p(z)$

**Junco Bird-ness**

Given a fixed model, find the **evidence** of **prediction**.

Why was this a Junco bird?

One of the most popular techniques:

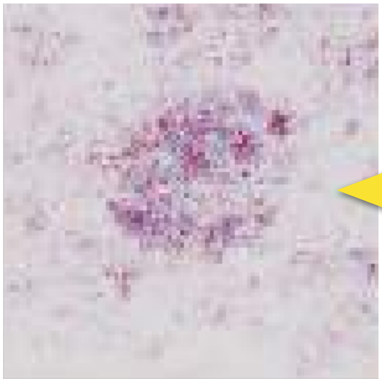
# Saliency maps

A trained
machine learning model
(e.g., neural network)

$p(z)$

**Junco Bird-ness**

Caaaaan do!



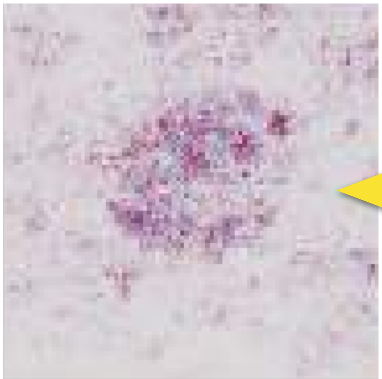
The promise:
these pixels are the **evidence** of **prediction.**

# Sanity check question.

A trained
machine learning model
(e.g., neural network)

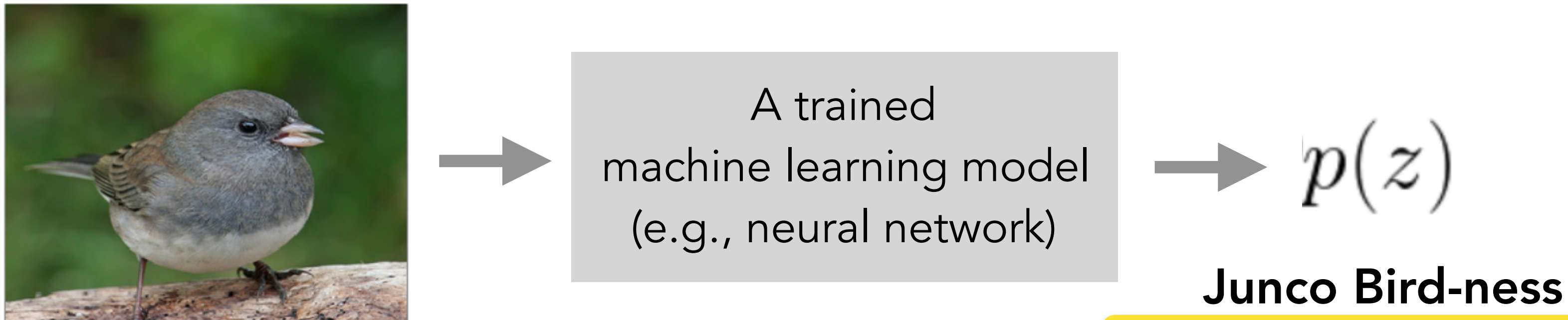
$p(z)$

**Junco Bird-ness**



The promise:
these pixels are the **evidence** of **prediction.**

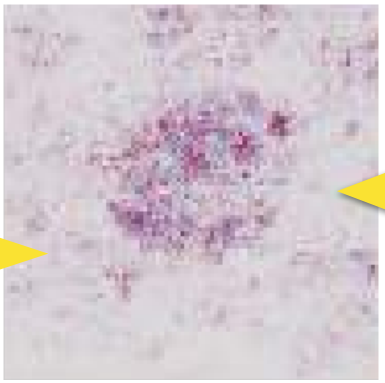
# Sanity check question.



A trained machine learning model (e.g., neural network)

$p(z)$

**Junco Bird-ness**

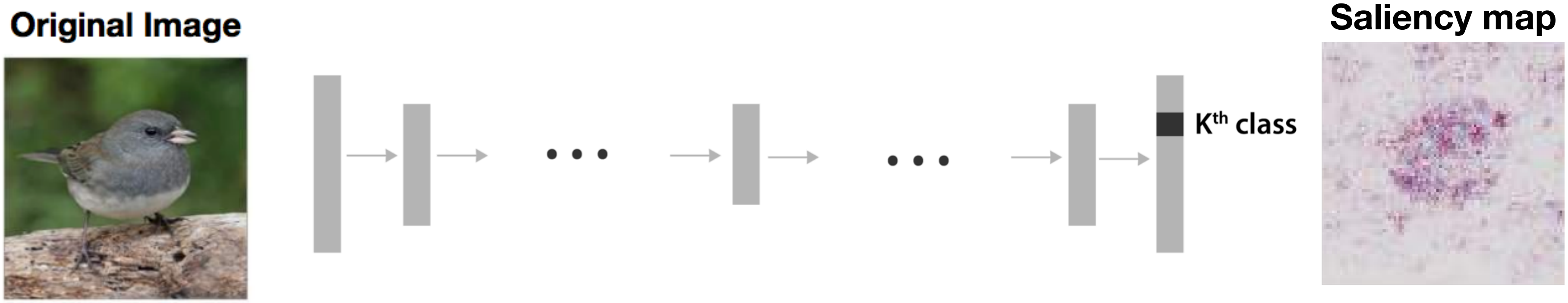


The promise: these pixels are the **evidence** of **prediction.**

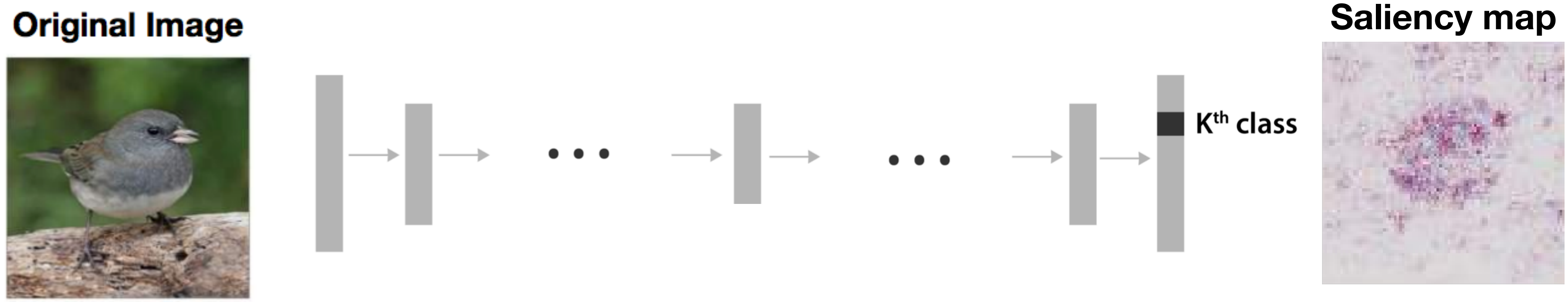
If so, when **prediction** changes, the explanation should change.

Extreme case:
If **prediction** is random,
the **explanation** should
**REALLY** change.

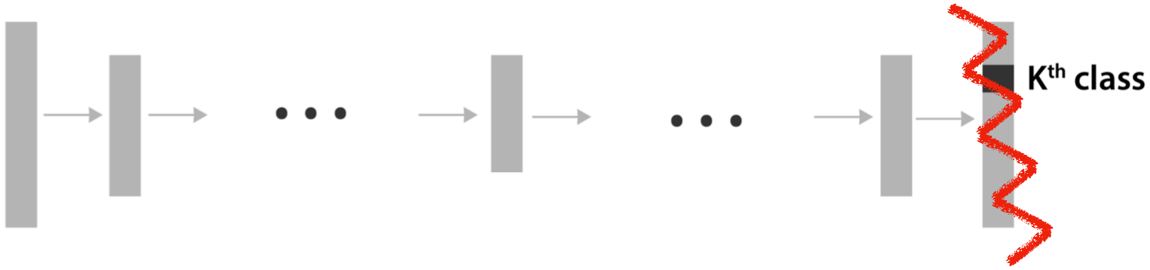
# Sanity check:
# When prediction changes, do explanations change?



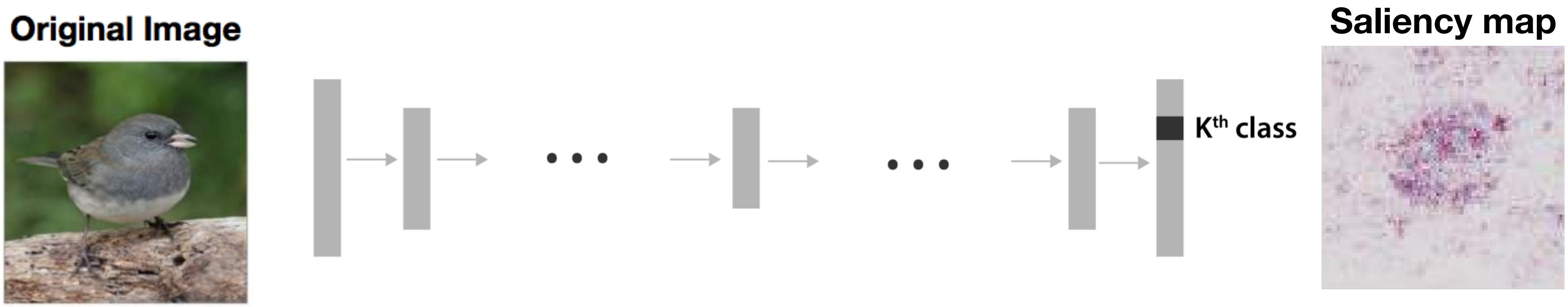
# Sanity check:
# When prediction changes, do explanations change?



Randomized weights!
Network now makes garbage predictions.

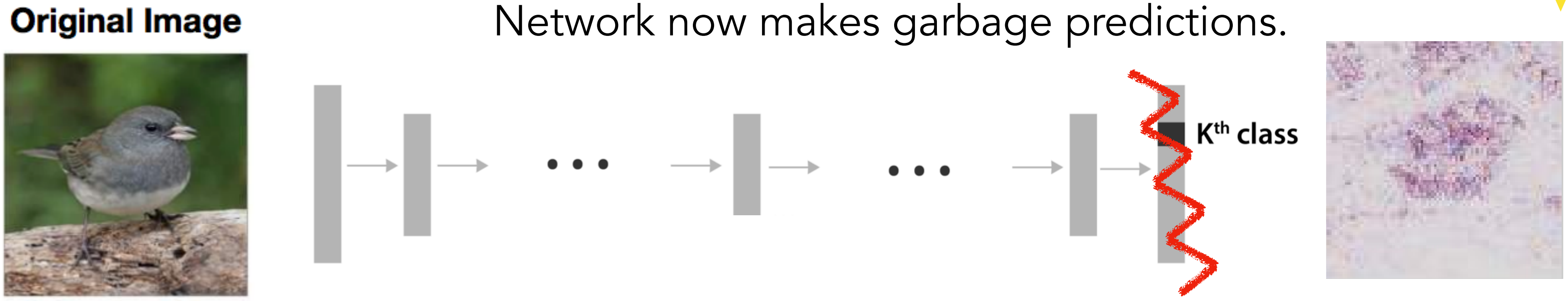


# Sanity check:
# When prediction changes, do explanations change?

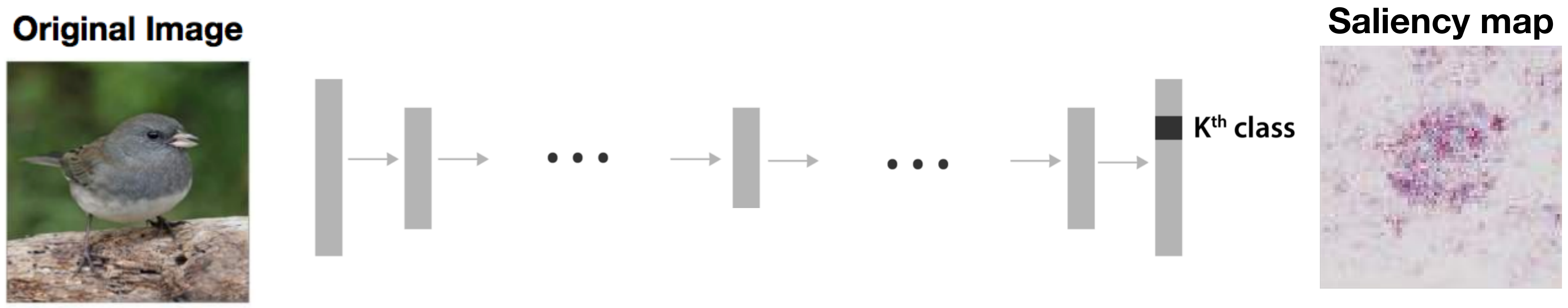


Randomized weights!
Network now makes garbage predictions.

!!!!!???!?

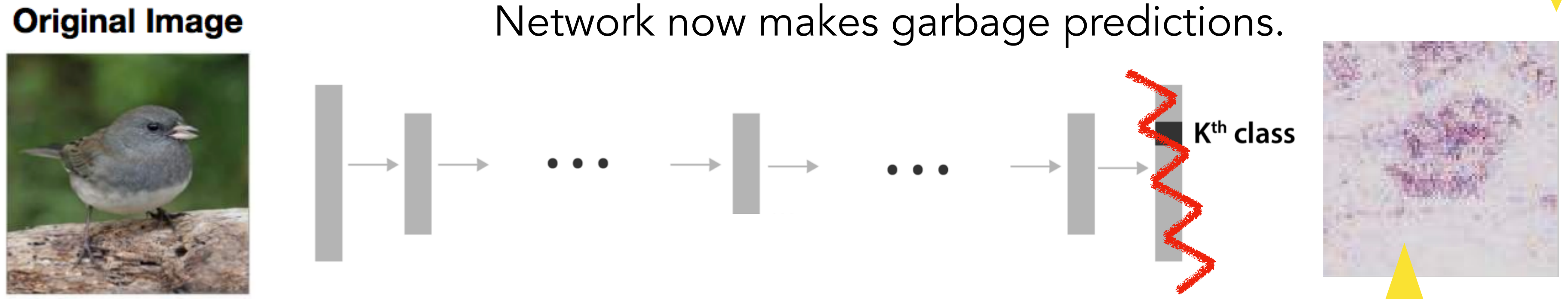


# Sanity check:
# When prediction changes, do explanations change?



!!!!!???!?

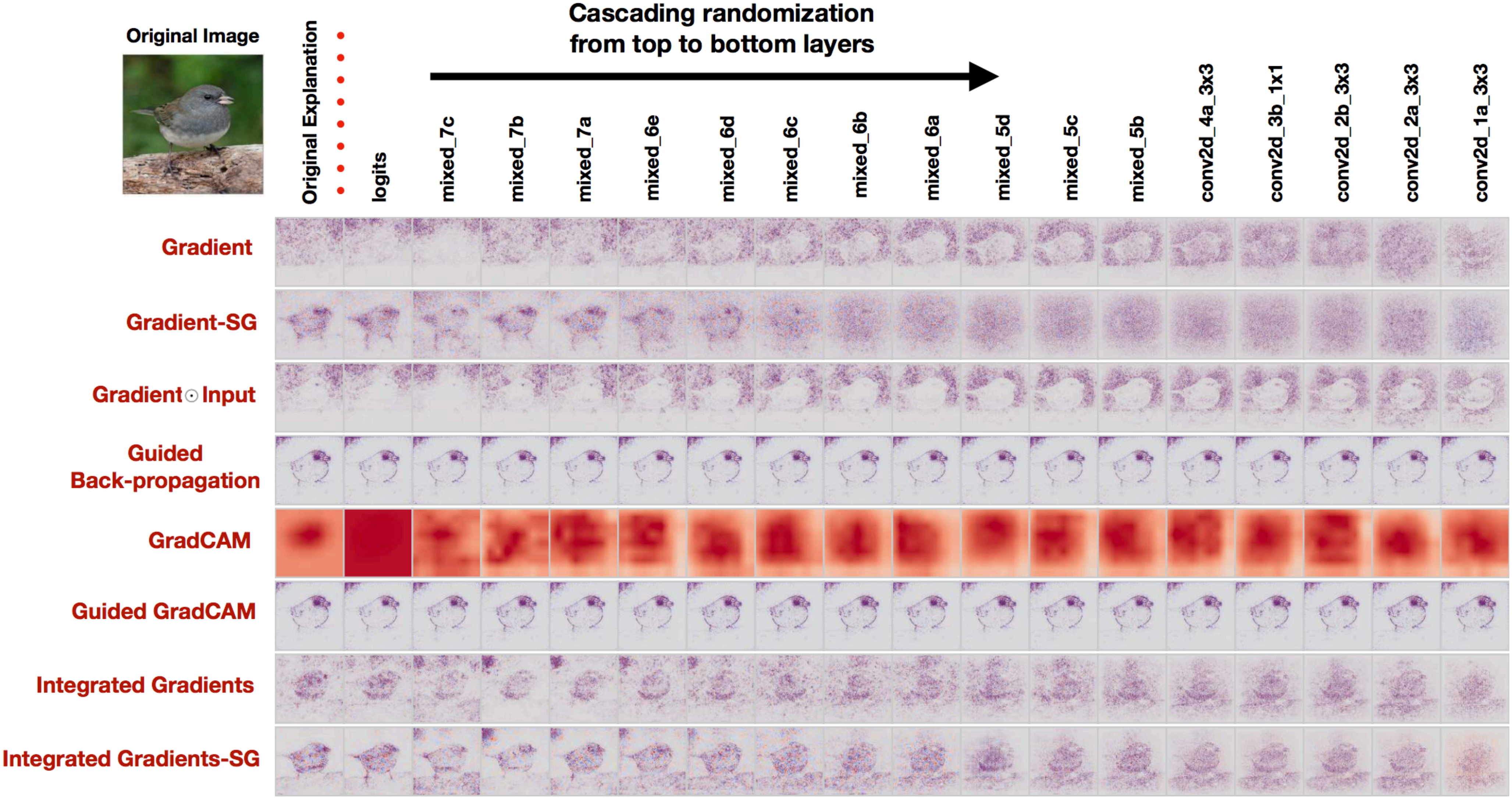
Randomized weights!
Network now makes garbage predictions.



the **evidence** of **prediction?????**

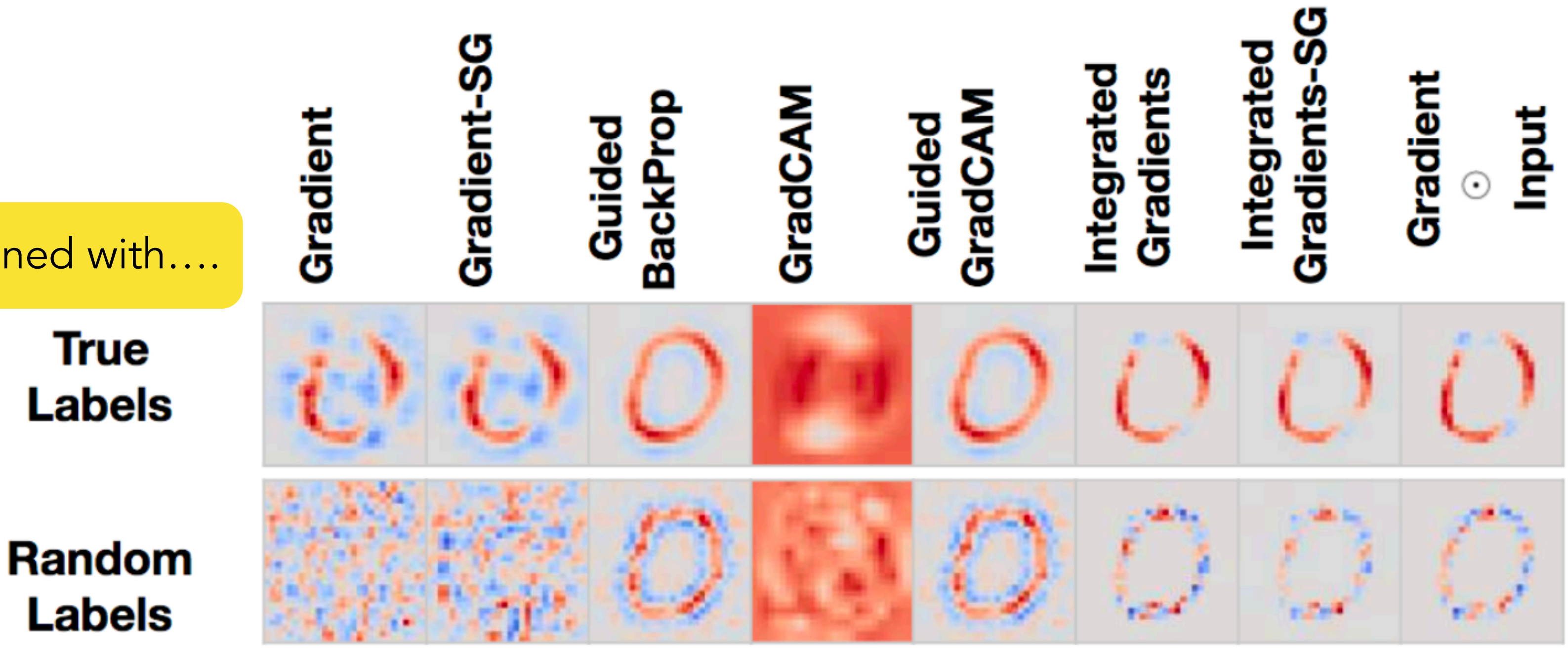
# Sanity check1:
# When prediction changes, do explanations change?

**No!**



# Sanity check2:
# Networks trained with true and random labels,
# Do explanations deliver different messages?

**No!**



Networks trained with….

# Conclusion

- **Confirmation bias**: Just because it "makes sense" to humans, doesn't mean it reflects the evidence for prediction.

- Do sanity checks for your interpretability methods!
  (e.g., TCAV [K. et al '18])

- Others who independently reached the same conclusions:
  [Nie, Zhang, Patel '18] [Ulyanov, Vedaldi, Lempitsky '18]

- Some of these methods have been shown to be useful for humans. Why? More studies needed.

Poster #30 10:45am - 12:45pm

@Room 210