Interpretability through **Explanations**

- Explain **one** prediction at a time

- Explanations are usually **influence scores** (~saliency maps)

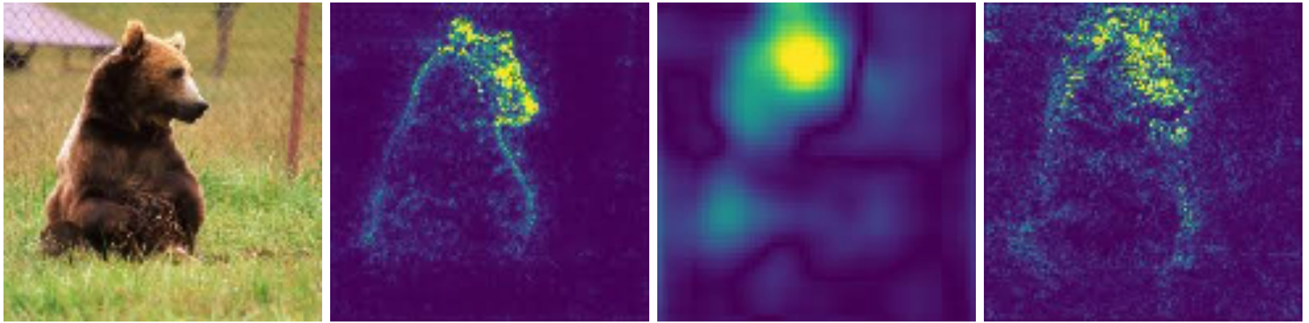
- Two main approaches to estimate influence:

- Perturbations

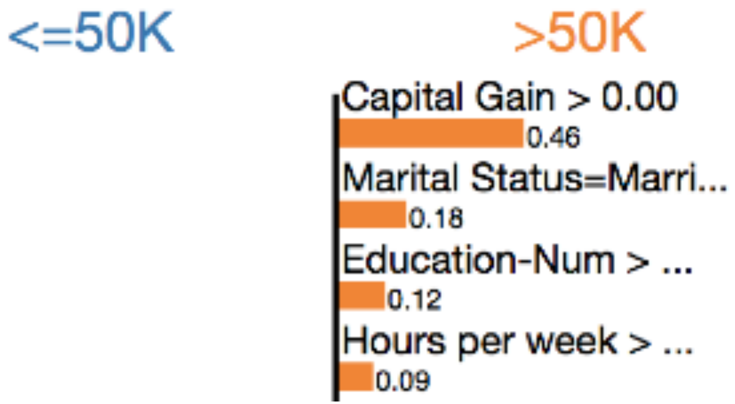
- Gradients/Relevance Propagation

86

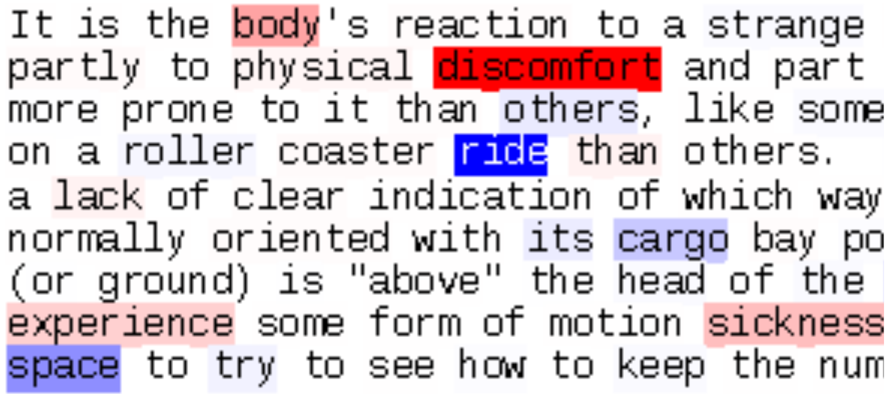
[Ribeiro et al., 2016]



[Hara et al., 2018]



[Ribeiro et al., 2016]



```
It is the body's reaction to a strange
partly to physical discomfort and part
more prone to it than others, like some
on a roller coaster ride than others.
a lack of clear indication of which way
normally oriented with its cargo bay po
(or ground) is "above" the head of the
experience some form of motion sickness
space to try to see how to keep the num
```

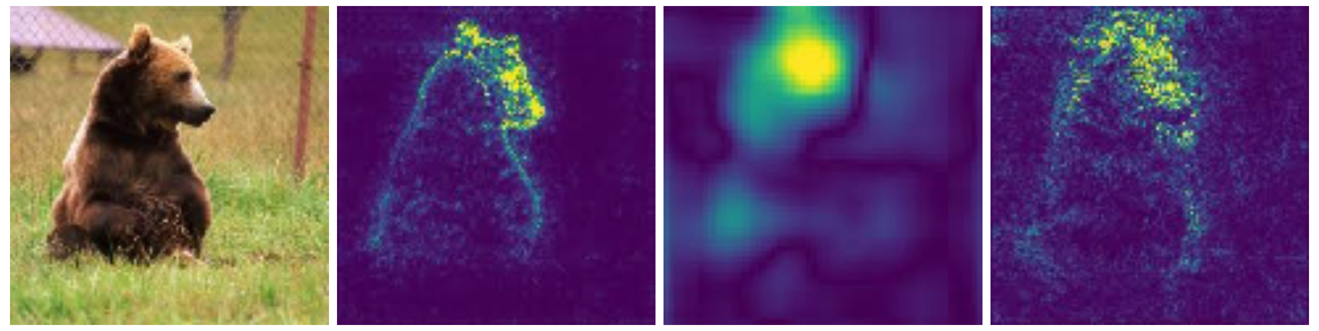
[Arras et al., 2017]

Both require computation "after-the-fact":

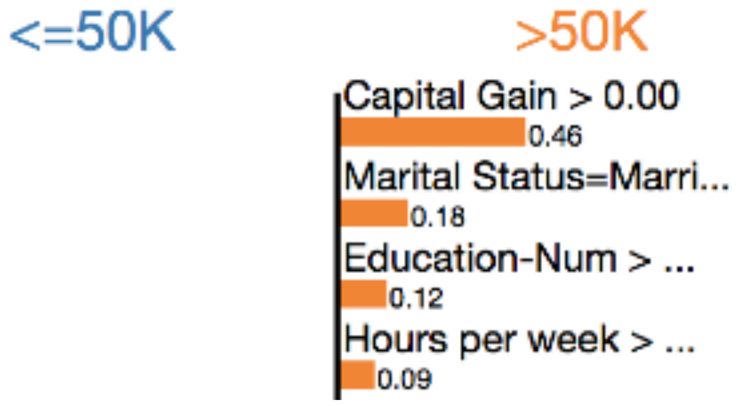
post-hoc interpretability

Interpretability through **Explanations**

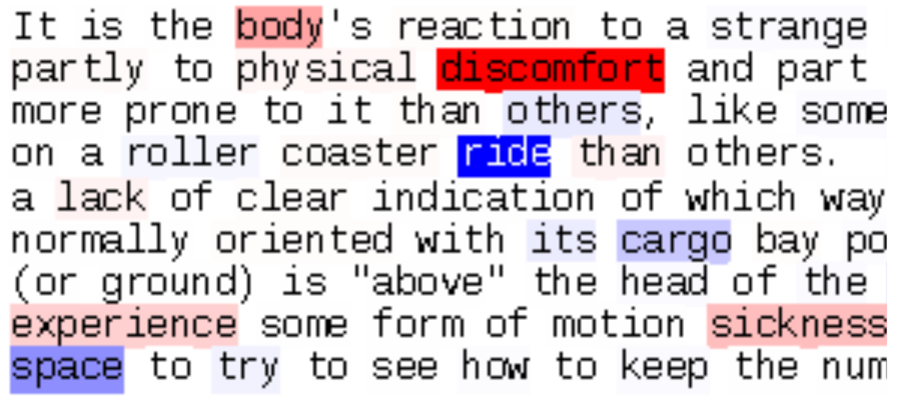
- Explain **one** prediction at a time
- Explanations are usually **influence scores** (~saliency maps)



[Hara et al., 2018]



[Ribeiro et al., 2016]



```
It is the body's reaction to a strange
partly to physical discomfort and part
more prone to it than others, like some
on a roller coaster ride than others.
a lack of clear indication of which way
normally oriented with its cargo bay po
(or ground) is "above" the head of the
experience some form of motion sickness
space to try to see how to keep the num
```

[Arras et al., 2017]

- Two main approaches to estimate influence:
  - Perturbations
  - Gradients/Relevance Propagation

Both require computation "after-the-fact":
post-hoc interpretability

Interpretability through **Explanations**