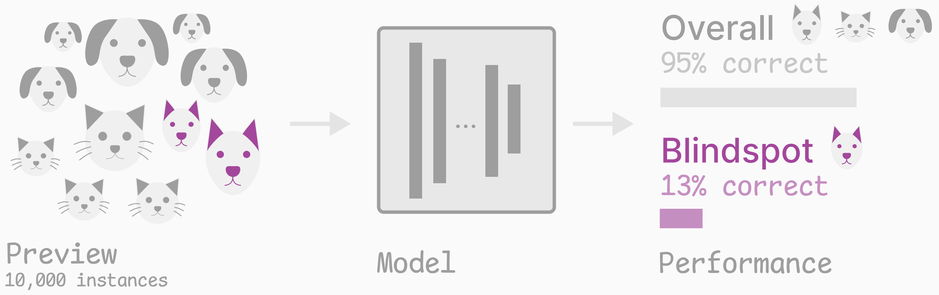
Mirror | Interactive Discovery of Blindspots in Machine Learning Models

Donald Bertucci, Ángel Alexander Cabrera, Nari Johnson, Gregory Plumb, Erica Fu, Adam Perer

Introduction

Your model may have great performance, but it may be failing on an important subset of the data: a Blindspot.



Who cares about Blindspots?

- Models have impact on your life
  In Healthcare, Self-Driving, Finance, Admissions, and more!
- Blindspots are unexpected and could have catastrophic consequences if not discovered

Discovering Blindspots

Existing Methods

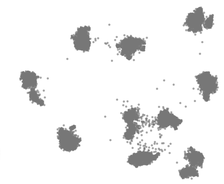
- Interactive (human)
  - Need meaningful metadata to start
  - Constrained by what you have (metadata)
- Automated (computer)
  - Many incoherent blindspots
  - Does not align with human in the majority of cases

Our Method

Interactive Blindspot Discovery Bringing humans into the loop with data projection.

Learned Metadata Create new metadata interpretable to humans, defined by humans.

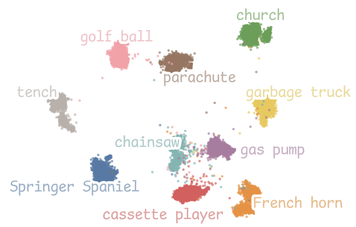
Interactive Blindspot Discovery



A **Project** the data and visualize in 2D

PROJECT

B **Create Awareness** from metadata and model outputs



Color By
Correct

Model Confused?

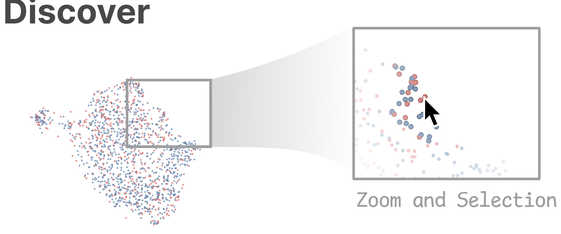
C **Filter** and **Project** again

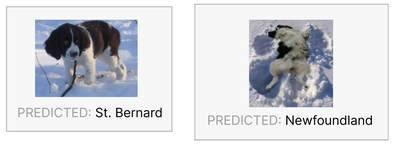
FILTER Springer Spaniel

PROJECT

New Insights

D **Discover**

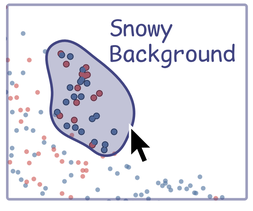




Pattern of Error

Learned Metadata

Create new labels meaningful to humans with a **Metadata Learner**.



FILTER Springer Spaniel → PROJECT → REGION LABELER

Metadata Learner

Rest
88% correct

Blindspot
55% correct

Zeno CMU DIG CMU HCII CMU ML