
ABACUS.AI

ABACUS.AI

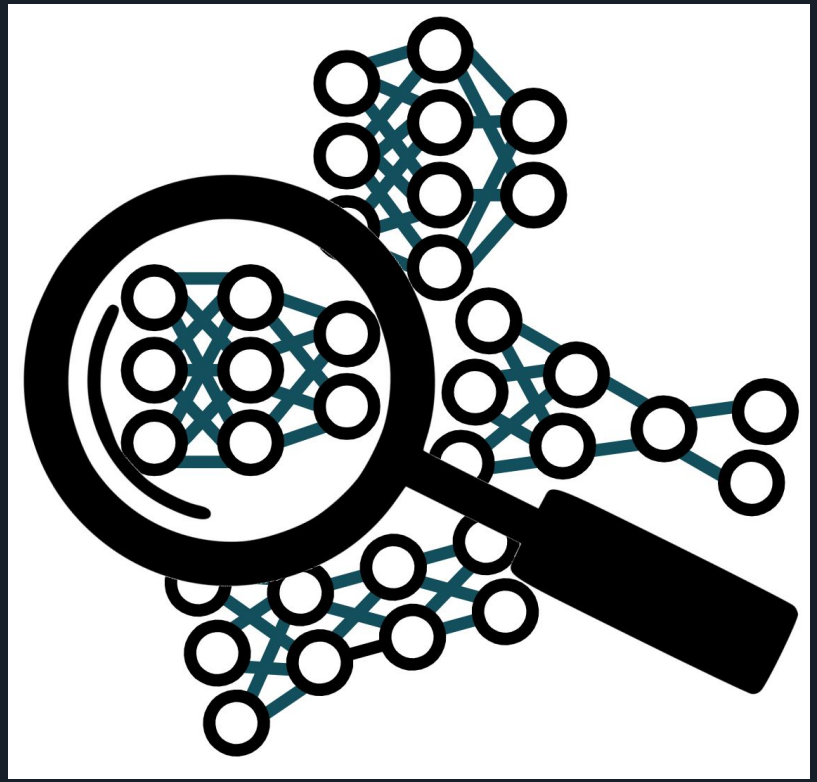
Research Methodology

Colin White
Abacus.AI

Outline

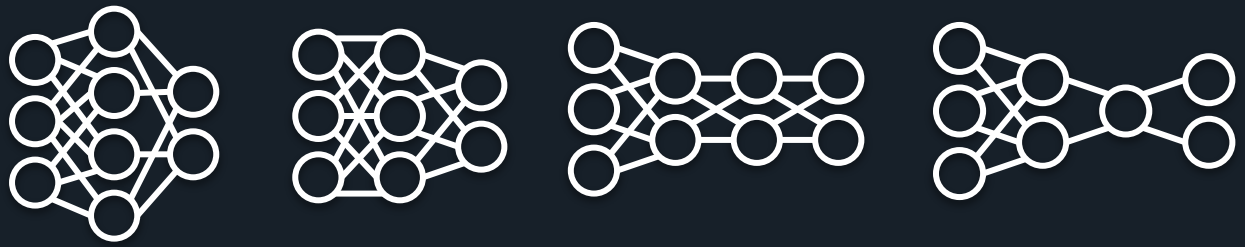
Part 1: My own research
- AutoML
- De-biasing ML
- Explainability in ML

Part 2: Research methodology
- Conducting research
- Writing papers
- Tips

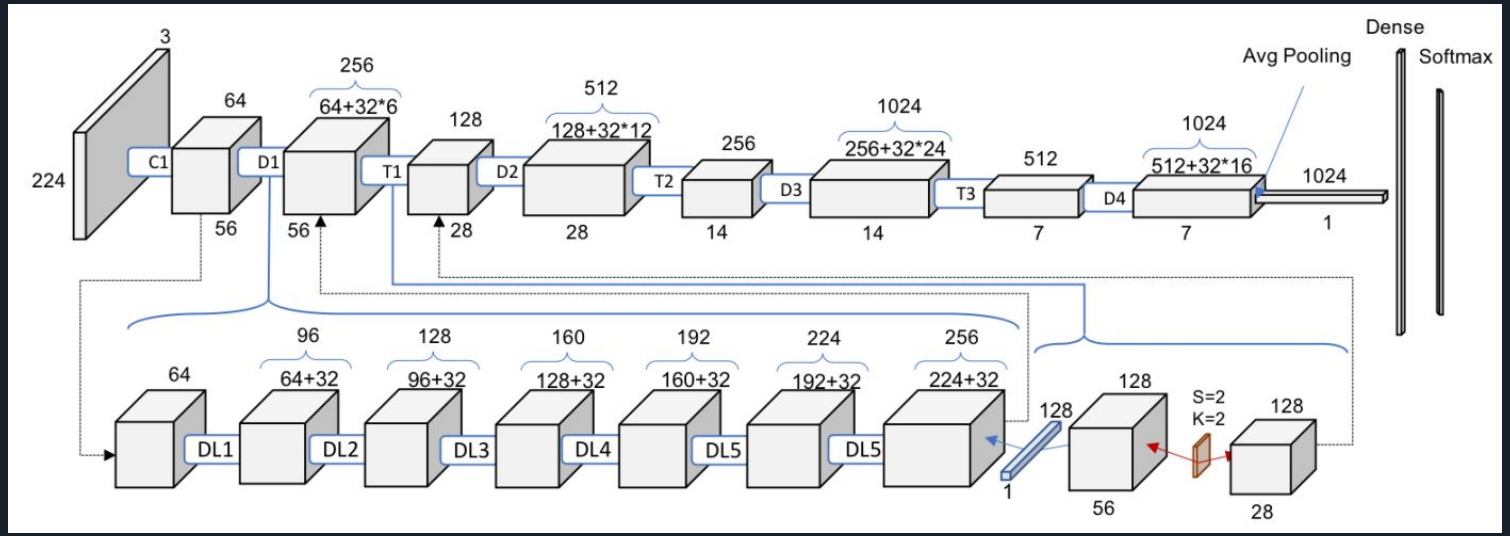


Deep learning

- Powerful machine learning technique
- Huge variety of neural networks for different tasks
- Becoming more specialized and complex
- Huge amount of hyperparameters



Deep learning

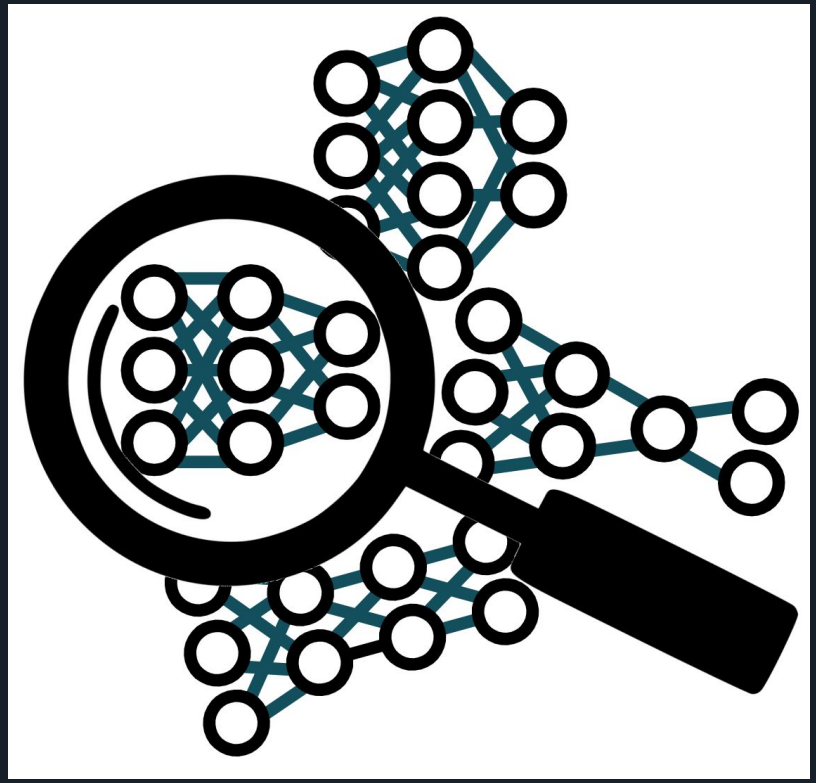


Large amounts of hyperparameters that must be tuned

Source: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

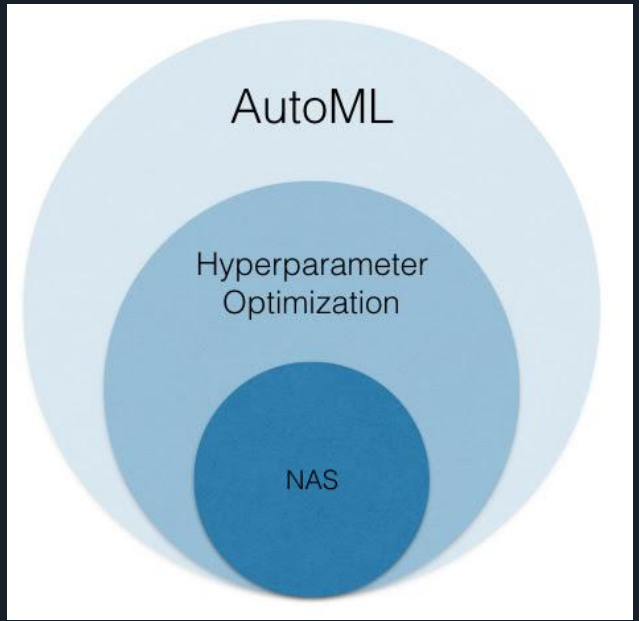
AutoML

- What if an algorithm could do this for us?
- AutoML (and neural architecture search) is a hot area of research
- Given a dataset, use an algorithm to find the best model for the dataset



Automated Machine Learning

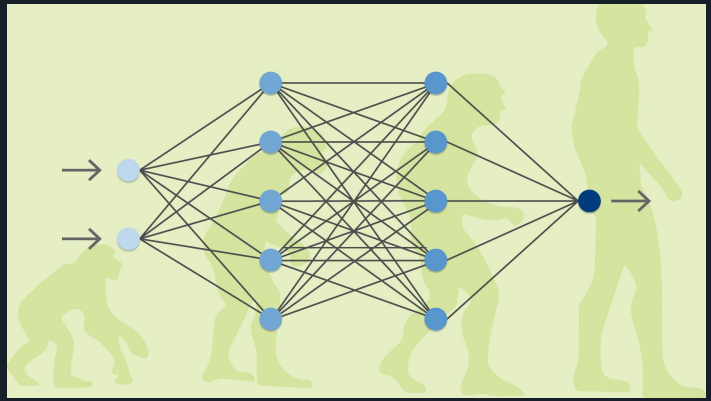
- Automated machine learning
  - Data cleaning, model selection, HPO, NAS, ...
- Hyperparameter optimization (HPO)
  - Learning rate, dropout rate, batch size, ...
- Neural architecture search (NAS)
  - Finding the best neural architecture

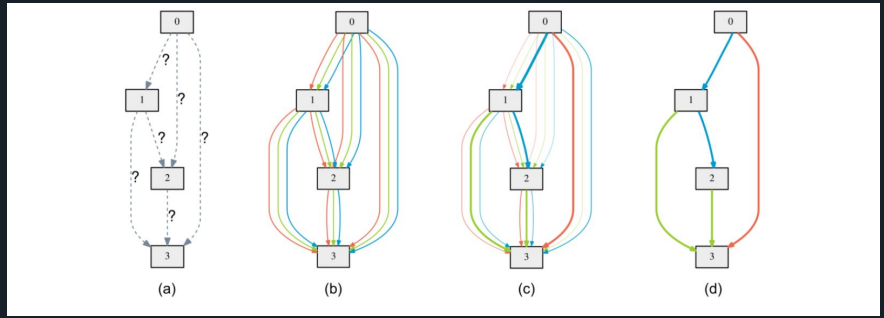


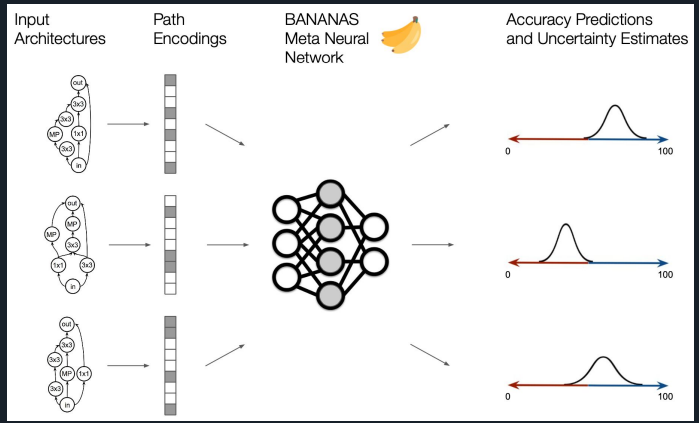
Source:https://determined.ai/blog/neural-architecture-search/

Search Algorithms for NAS

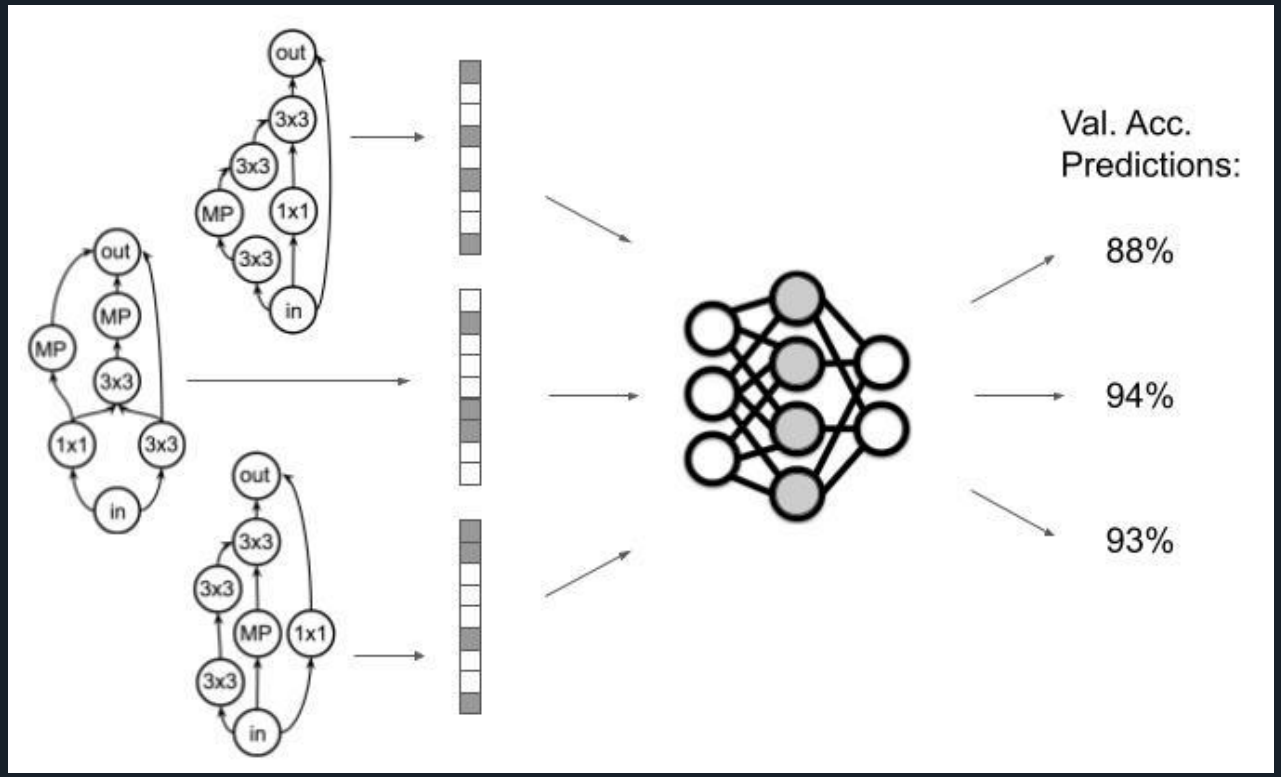
- Evolutionary search
- Reinforcement Learning
- Bayesian optimization with a GP
- BayesOpt with a neural predictor







Neural Network Predictor

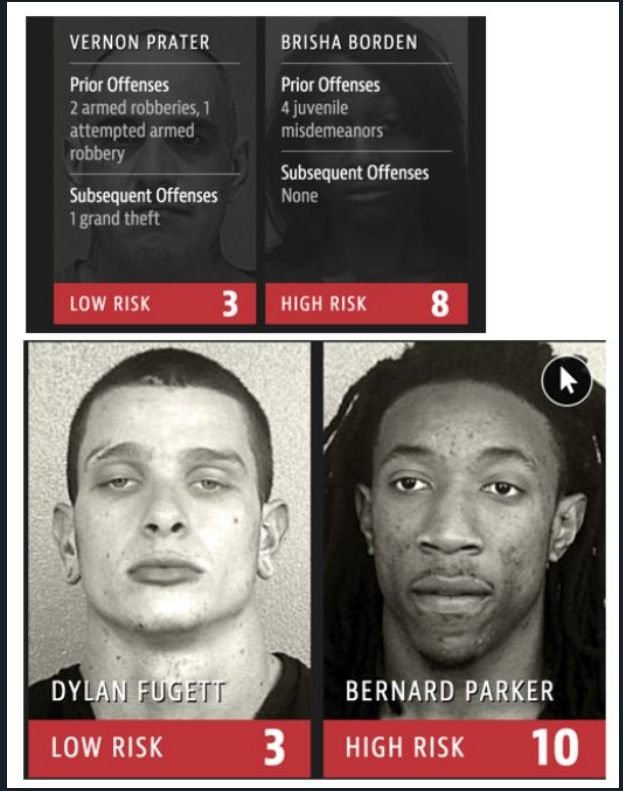


[W, Neiswanger, Savani 2019]

Bias in Machine Learning Models

COMPAS Assessment [Angwin et al. 2016]

- Used by US judges for bail, sentences
- Found to be racially biased



Bias in Computer Vision

1. **For one-to-one matching, the team saw higher rates of false positives for Asian and African American faces relative to images of Caucasians.** The differentials often ranged from a factor of 10 to 100 times, depending on the individual algorithm. False positives might present a security concern to the system owner, as they may allow access to impostors.
2. **Among U.S.-developed algorithms, there were similar high rates of false positives in one-to-one matching for Asians, African Americans and native groups** (which include Native American, American Indian, Alaskan Indian and Pacific Islanders). The American Indian demographic had the highest rates of false positives.

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

June 24, 2020 · 8:00 AM ET

BOBBY ALLYN

A US government study confirms most face recognition systems are racist

by Karen Hao December 20, 2019

A U.S. Customs and Border Protection officer helps a passenger navigate a facial recognition kiosk at the airport.
DAVID J. PHILLIP/AP

https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

Removing bias is not easy

- Cannot just exclude an attribute
- Bias comes from many sources
- **Sampling biases**
  - more policing in predominantly Black neighborhoods
- **Historical biases**
  - certain classes of people were biased against historically

https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4

Fairness definitions

Given a labeled dataset, and a classifier,

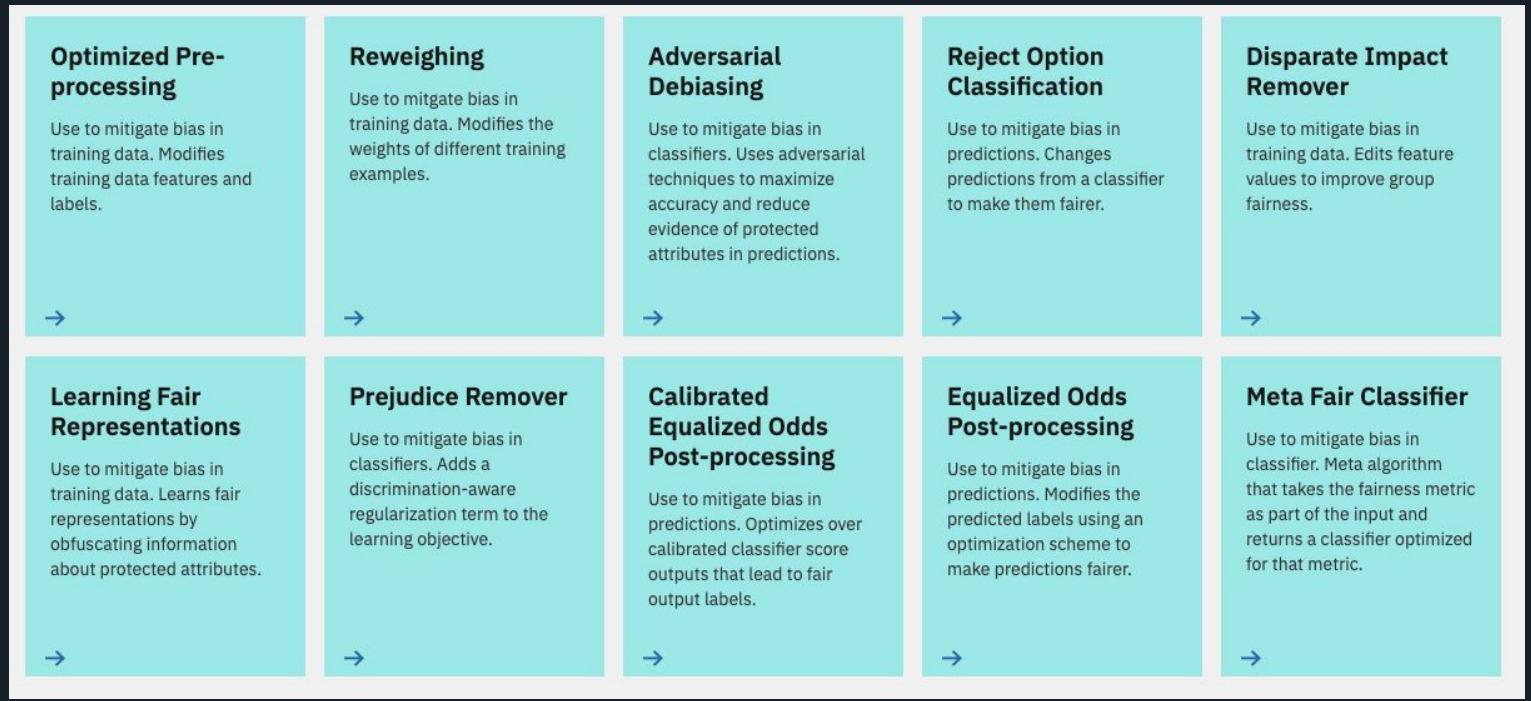
Statistical Parity: for all groups, percent who were predicted to commit a crime is equal

Equality of opportunity: for all groups, True positive rates are equal

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

[Dwork et al. 2012], [Hardt et al. 2016]

Bias mitigation algorithms



[Bellamy et al. 2018]]

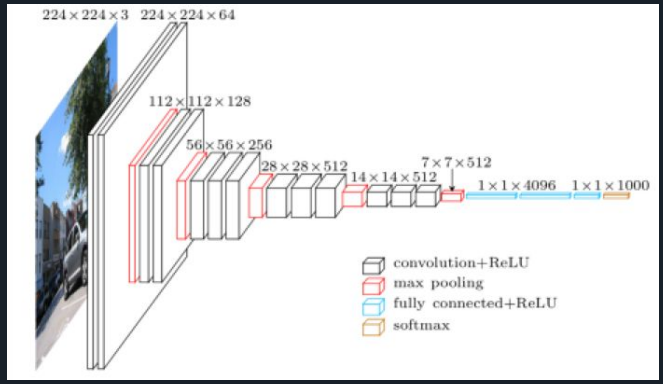
Intra-processing Algorithms

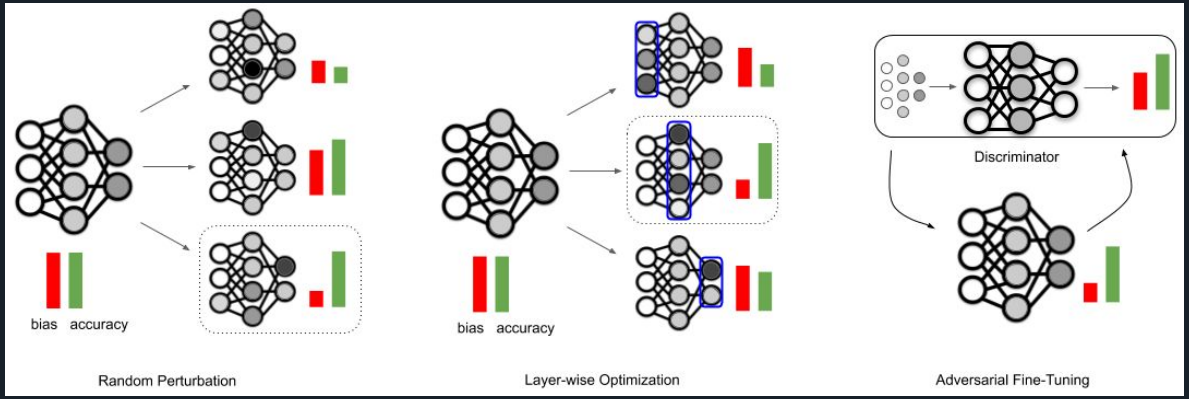
Given a pretrained model

- Typically trained on generic dataset, biased

Given a new, specific task, dataset

Goal: fine-tune the existing model and debias

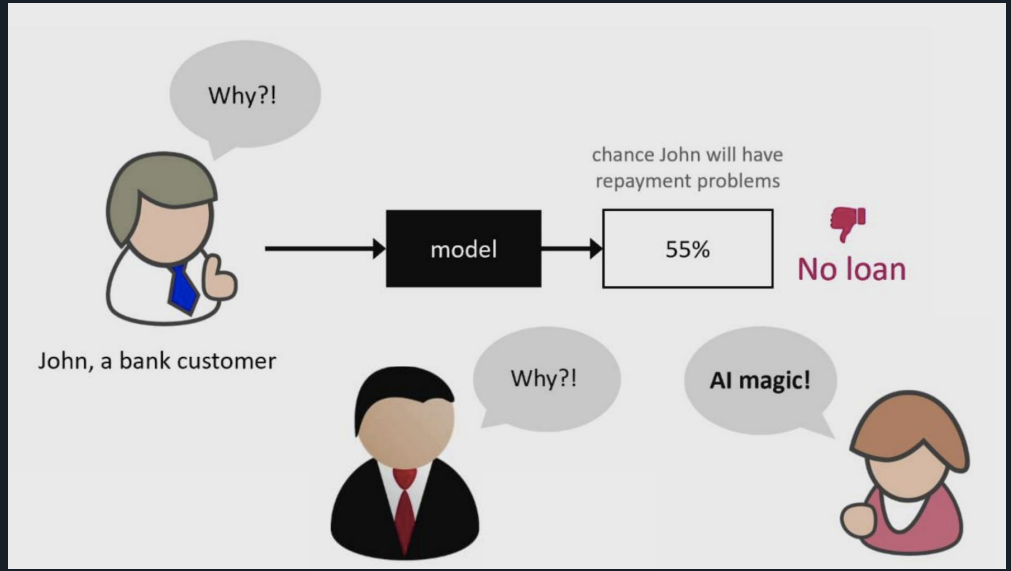




Explainability

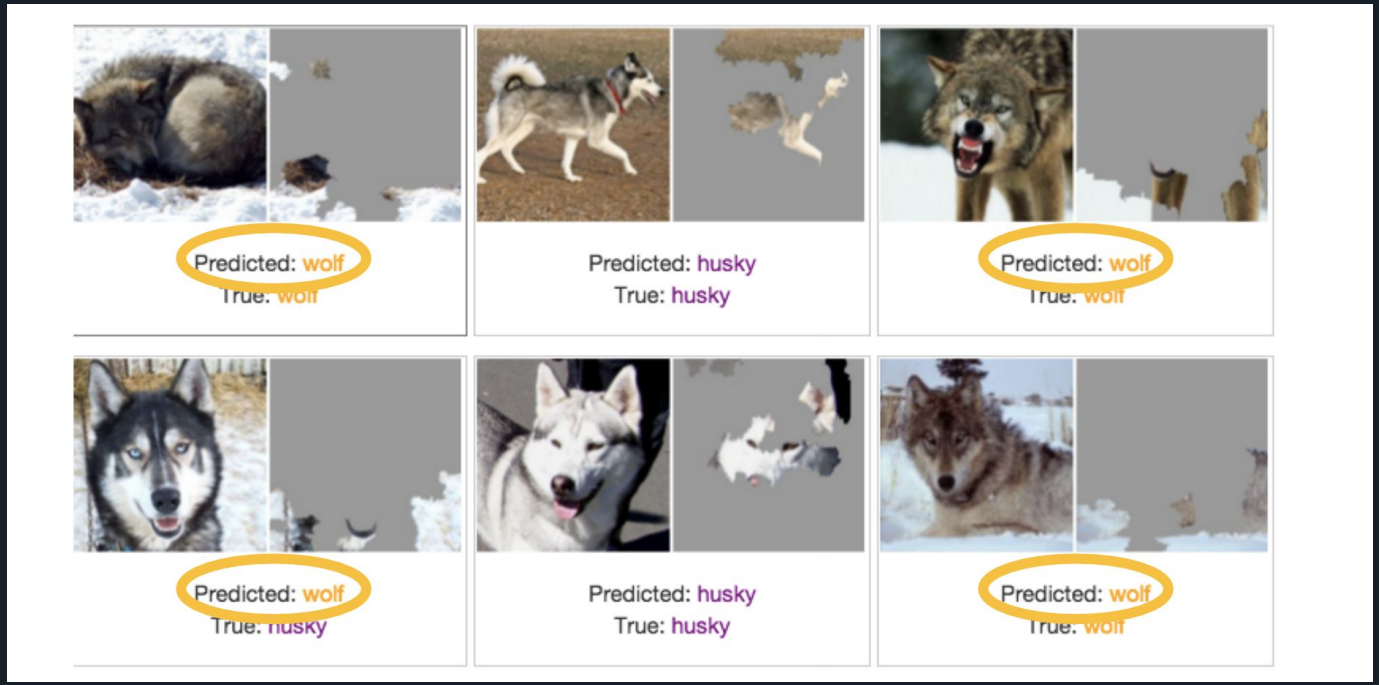
Why should our ML models be interpretable?

- Debugging
- Legal obligation
- Faster adoption
- Detecting bias



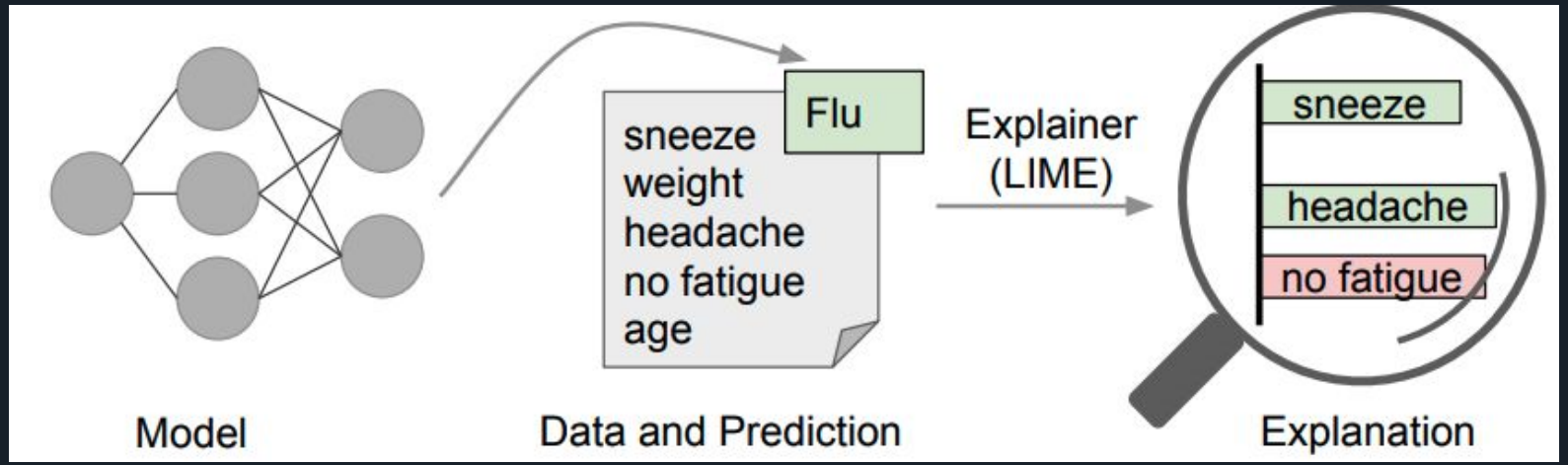
Source: https://www.microsoft.com/en-us/research/video/explainable-ai-for-science-and-medicine/

“We made a great snow detector”



Source: https://arxiv.org/pdf/1602.04938.pdf

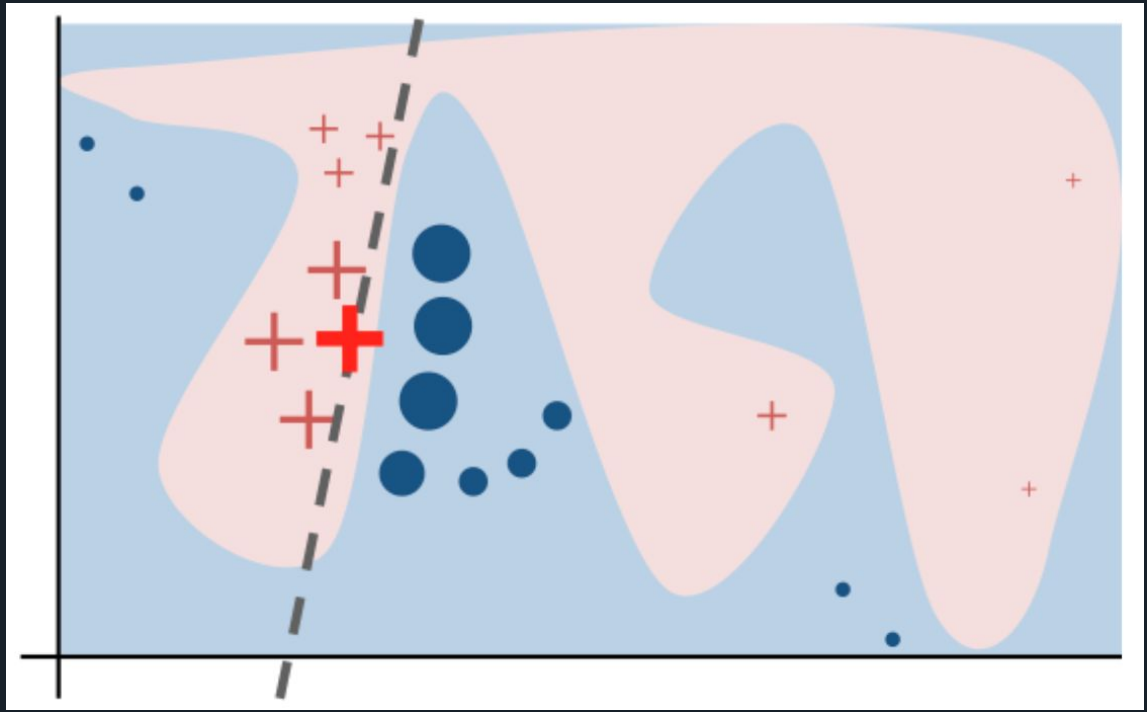
Feature Attribution



- LIME
- SHAP

Source: https://arxiv.org/abs/1602.04938

LIME [Ribiero, Singh, Guestrin ‘16]

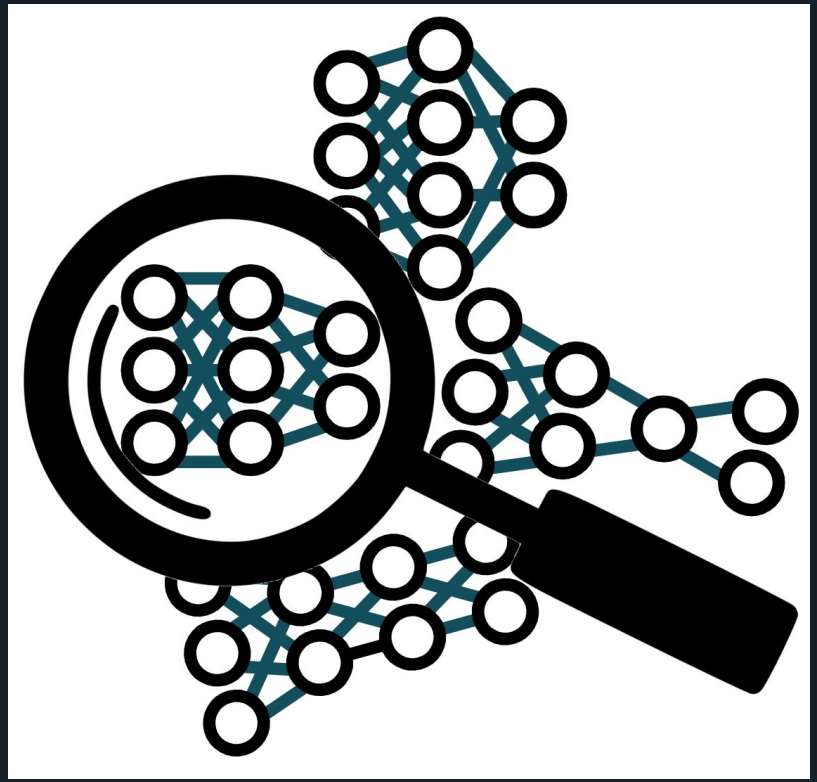


Source: https://www.microsoft.com/en-us/research/video/explainable-ai-for-science-and-medicine/

Outline

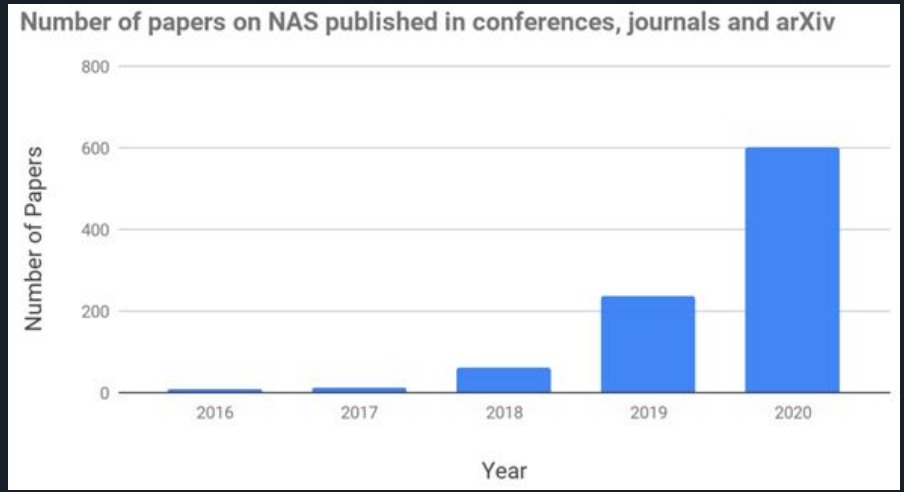
Part 1: My own research
- AutoML
- De-biasing ML
- Explainability in ML

Part 2: Research methodology
- Conducting research
- Writing papers
- Tips



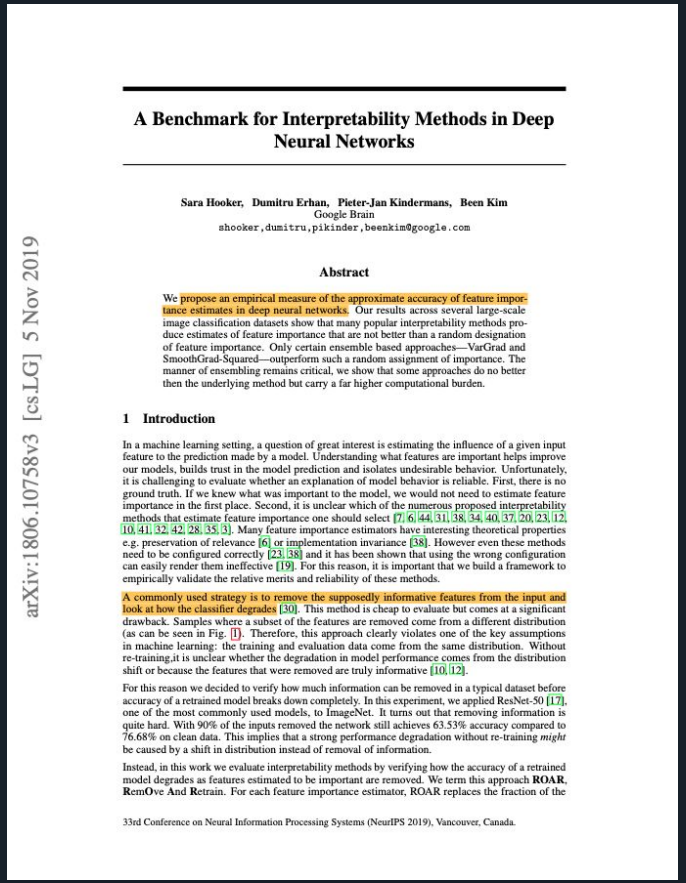
Starting a Research Direction

- Literature review
  - Start with blog posts or survey papers
- **There is a lot of research out there, more than you would expect!**
- Find a few papers that you like best, and read them closely
  - Look at papers that **cite / are cited by** these papers
- Write down questions or follow-up ideas



Reading research papers

- Read **actively**
- Highlight, write down notes and questions
- Open-source code?
  - Try to run the code



A Benchmark for Interpretability Methods in Deep Neural Networks

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, Been Kim
Google Brain
shooker,dumitru,pikinder,beenkim@google.com

Abstract

We propose an empirical measure of the approximate accuracy of feature importance estimates in deep neural networks. Our results across several large-scale image classification datasets show that many popular interpretability methods produce estimates of feature importance that are not better than a random designation of feature importance. Only certain ensemble based approaches—VarGrad and SmoothGrad-Squared—outperform such a random assignment of importance. The manner of ensembling remains critical, we show that some approaches do no better then the underlying method but carry a far higher computational burden.

1 Introduction

In a machine learning setting, a question of great interest is estimating the influence of a given input feature to the prediction made by a model. Understanding what features are important helps improve our models, builds trust in the model prediction and isolates undesirable behavior. Unfortunately, it is challenging to evaluate whether an explanation of model behavior is reliable. First, there is no ground truth. If we knew what was important to the model, we would not need to estimate feature importance in the first place. Second, it is unclear which of the numerous proposed interpretability methods that estimate feature importance one should select [7, 6, 44, 31, 38, 34, 40, 37, 20, 23, 12, 10, 41, 32, 42, 28, 35, 3]. Many feature importance estimators have interesting theoretical properties e.g. preservation of relevance [6] or implementation invariance [38]. However even these methods need to be configured correctly [23, 38] and it has been shown that using the wrong configuration can easily render them ineffective [19]. For this reason, it is important that we build a framework to empirically validate the relative merits and reliability of these methods.

A commonly used strategy is to remove the supposedly informative features from the input and look at how the classifier degrades [30]. This method is cheap to evaluate but comes at a significant drawback. Samples where a subset of the features are removed come from a different distribution (as can be seen in Fig. 1). Therefore, this approach clearly violates one of the key assumptions in machine learning: the training and evaluation data come from the same distribution. Without re-training,it is unclear whether the degradation in model performance comes from the distribution shift or because the features that were removed are truly informative [10, 12].

For this reason we decided to verify how much information can be removed in a typical dataset before accuracy of a retrained model breaks down completely. In this experiment, we applied ResNet-50 [17], one of the most commonly used models, to ImageNet. It turns out that removing information is quite hard. With 90% of the inputs removed the network still achieves 63.53% accuracy compared to 76.68% on clean data. This implies that a strong performance degradation without re-training *might* be caused by a shift in distribution instead of removal of information.

Instead, in this work we evaluate interpretability methods by verifying how the accuracy of a retrained model degrades as features estimated to be important are removed. We term this approach **ROAR**, **R**em**O**ve **A**nd **R**etrain. For each feature importance estimator, ROAR replaces the fraction of the

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Fleshing out a research question

- Questions/follow-ups from the papers you read most closely
- Goal: *interesting*, *useful* question that has not been studied before
- Is there open-source code available?
  - Run this code
- What dataset(s) can you use?
- What do you expect to happen?
- Try out initial experiments

Research conferences

- **Networking is important** for collaborations and opportunities
- Many academic conferences are virtual right now!
- Neural Information Processing Systems (NeurIPS)
  - $25 for students
  - Find a paper that looks interesting, and chat with the authors

Dec 6 - 14
neurips.cc

Networking

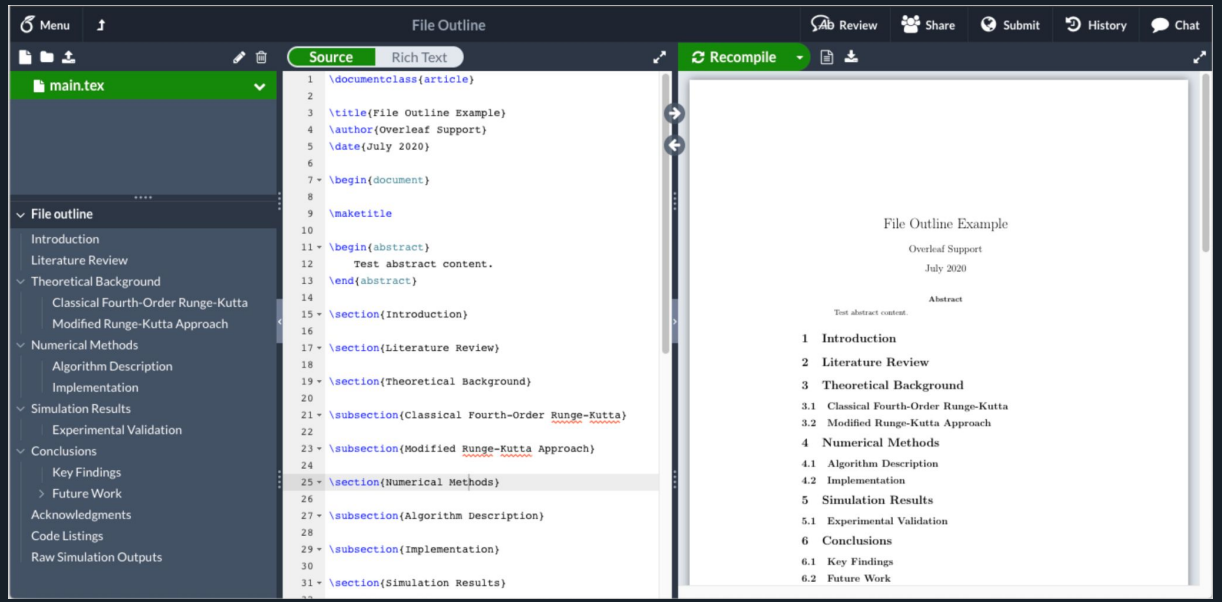
- Emailing a professor or researcher
  - Read their papers
  - Come up with ideas or follow-up questions
  - Open a GitHub “issue”

- Other initiatives
  - ML collective
  - ICLR 2022 initiative supporting first-time ICLR submitters

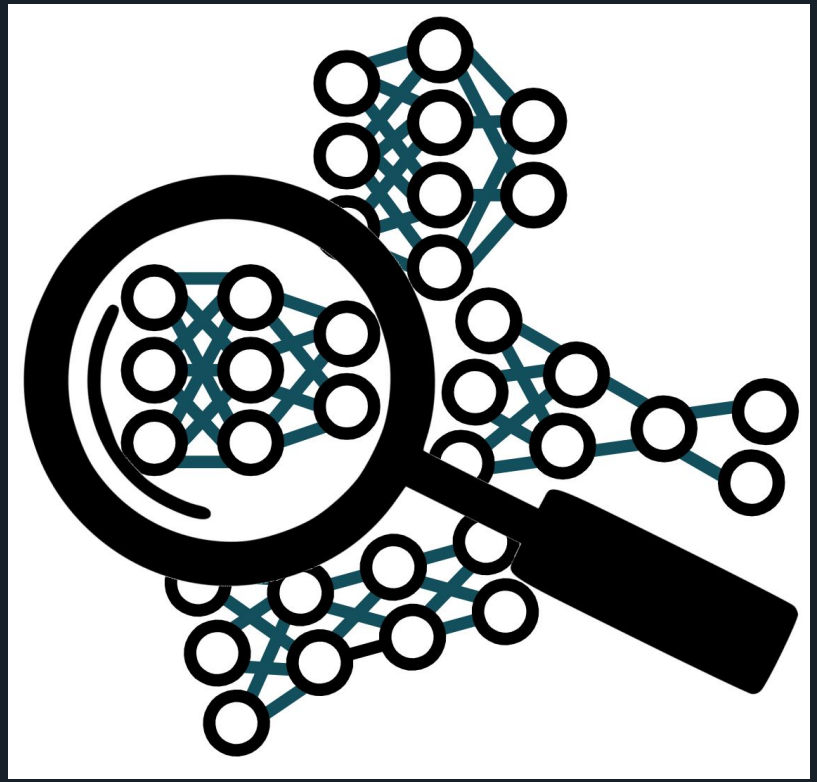
ABACUS.AI

Writing a research paper

- Closely look at the papers most similar to your project
- Come up with a “story”
- Write the abstract and introduction first
- Overleaf is a great tool for LaTeX



Thanks! Questions?




ABACUS.AI