Reverse Engineering Machine Learning

DONALD MINER
Miner & Kasch

Intro: Supervised Machine Learning

In general, ML learns a function that maps observations to an outcome. Once trained, you can pass the "model" a new observation and see what the model thinks will happen.

outcome = f(observations)

Most supervised machine learning process follows three steps:

1. Gather training data
2. Train a model
3. Use the model

Example: Email Scam Detection

Dear Friend.

My name is Peter Lawson,a merchant in Dubai, in the U.A.E.I have been diagnosed with Esophageal Cancer which was discovered very late,due to my laxity in carrying for my health. It has defiled all forms of medicine, and right now I have only about a few months to live, according to medical experts.

I have not particularly lived my life so well, as I never really cared for anyone not even myself but my business. Though I am very rich, I was never generous, I was always hostile to people and only focus on my business as that was the only thing I cared for. But now I regret all this as I now know that there is more to life than just wanting to have or make all the money in the world. I believe when God gives me a second chance to come to this world I would live my life a different way from how I have lived it.

Now that God ! has called me, I have willed and given most of my properties and assets to my immediate and extended family members and as well as a few close friends. I want God to be merciful to me and accept my soul and so, I have decided to give arms to charity organizations and give succour and confort to the less priviledged in our societies, as I want this to be one of the last good deeds I do on earth.

The last of my money which no one knows of is the huge cash deposit of twenty four million dollars that I have with a Security Company in Europe for safe keeping. I will want you to help me collect this deposit and disburse it to some charity organizations and to the less priviledged.

Please send me a mail to indicate if you will assist me in this disbursement.

I have set aside 10% for you for your time and patience.

You can e-mail me at:plawson@hknetmail.com

Remain blessed.

Mr.Peter Lawson

SCAM OR
NOT SCAM?

Example: Facial Recognition

DONALD MINER OR
NOT DONALD MINER?

DONALD MINER OR
NOT DONALD MINER?

Intro: Gather Training Data

OBSERVATIONS

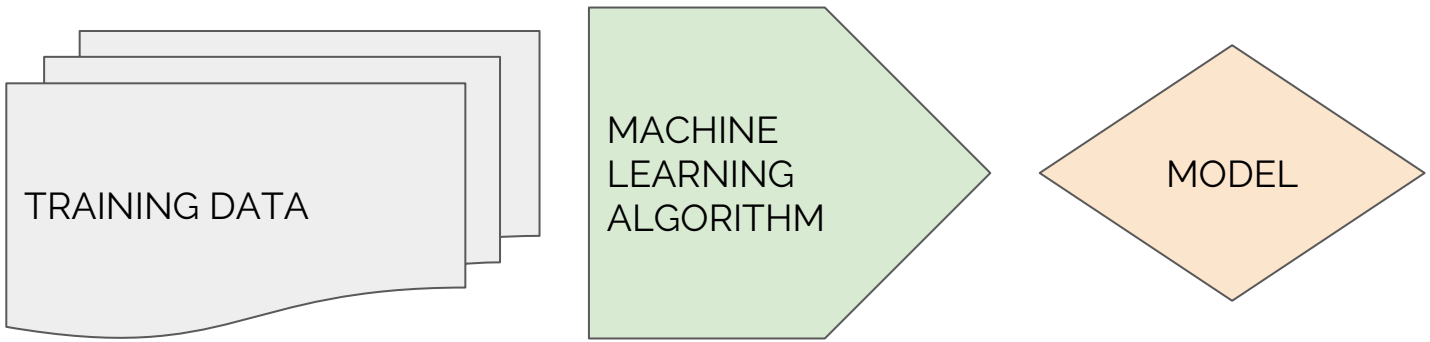
"THIS IS RICKY!"

OUTCOMES

By looking at historical observations and what happens from those observations, we can start building a data set that maps observations to outcomes.

These are "base truths" that represent observations of some hidden true function between these two things.

Intro: Train a Model

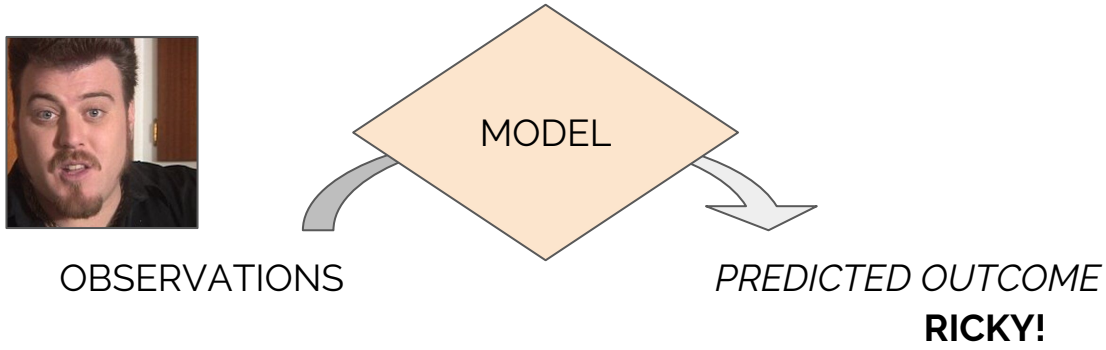


We take the training data and use a machine learning algorithm to try to learn a mapping from the original observations to the original outcomes.

There are numerous types of machine learning algorithms that are better at learning different kinds of mappings.

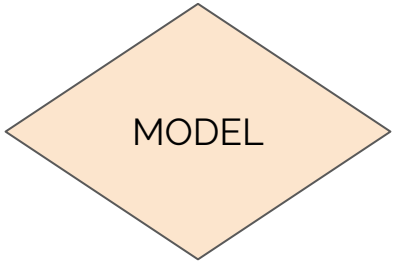
The output as a “model” that can be used to make predictions about new observations we’ve never seen before.

Intro: Use the Model



Now, instead we pass the new observations into our model and see what it says it thinks the outcome is, instead of having to observe it ourselves.

Attack: Reverse Engineering



Once trained, a model can be a black box.

We don't know how it works, but we can interact with it.

Attack: Reverse Engineering

MADE UP OBSERVATIONS

MODEL

42% Niels, 33% Don, 25% Ricky

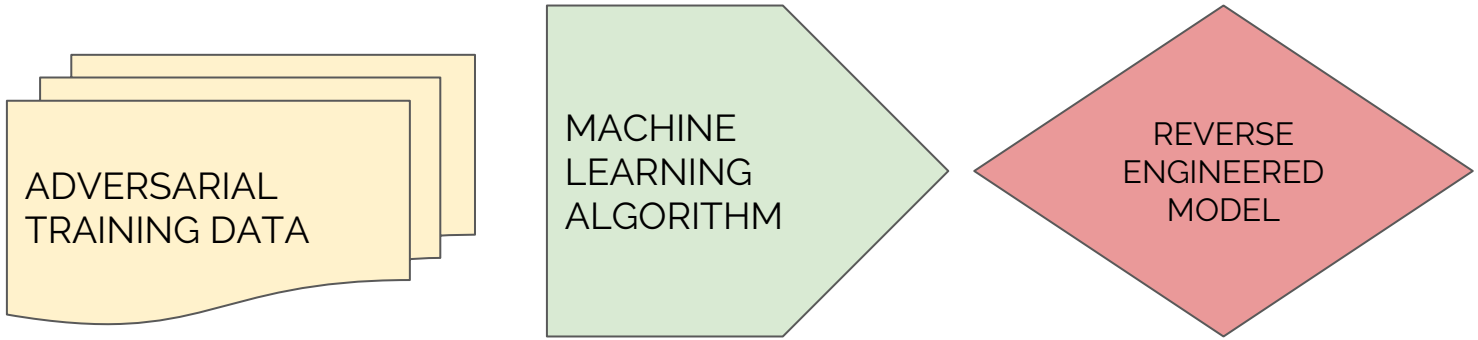
OUTCOMES

What we can do is pass the model generated examples and see what it does.

We'll do this a bunch of times to build a training data set on how the model works.

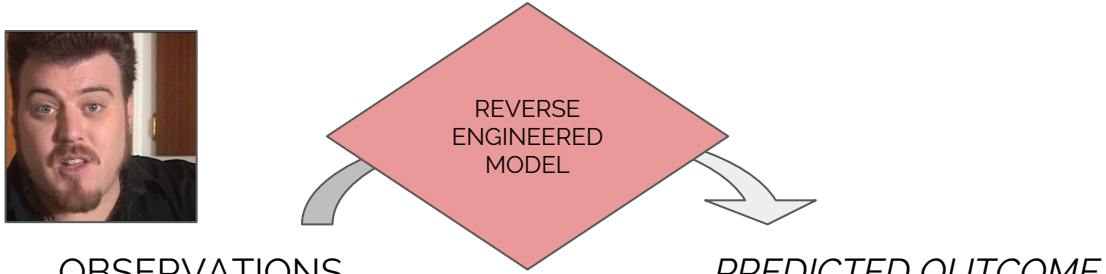
ADVERSARIAL TRAINING DATA

Attack: Reverse Engineering



We use machine learning just like how we did before, but instead we are now *training a model to behave like the black box model.*

We now have a "model of the model" and can interact with it however we want.



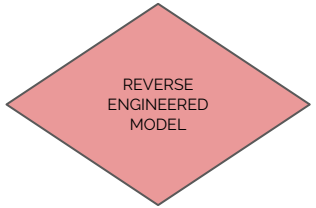
“Distilling the Knowledge in a Neural Network”

- It has been shown that a neural network can be trained to mimic the performance of other networks
  - Use the output of the network as the true labels to create a new training set.
  - The new neural network is trained to match this new training set.
  - By training on the output the new network learns to mimic the old.
  - Results show that the mimic network was very close in behavior to the model it was mimicing.

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Hinton et al. https://arxiv.org/pdf/1503.02531.pdf

Attack: Reverse Engineering

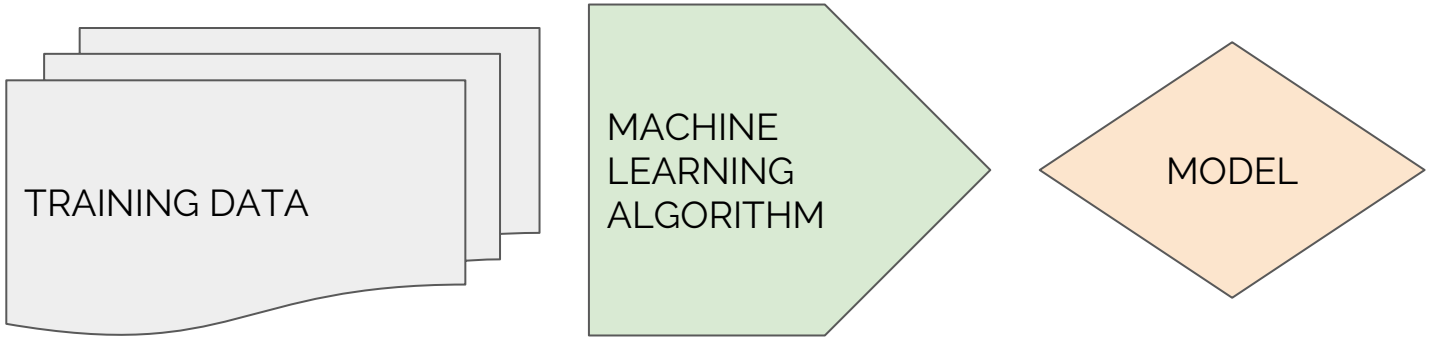


Use cases

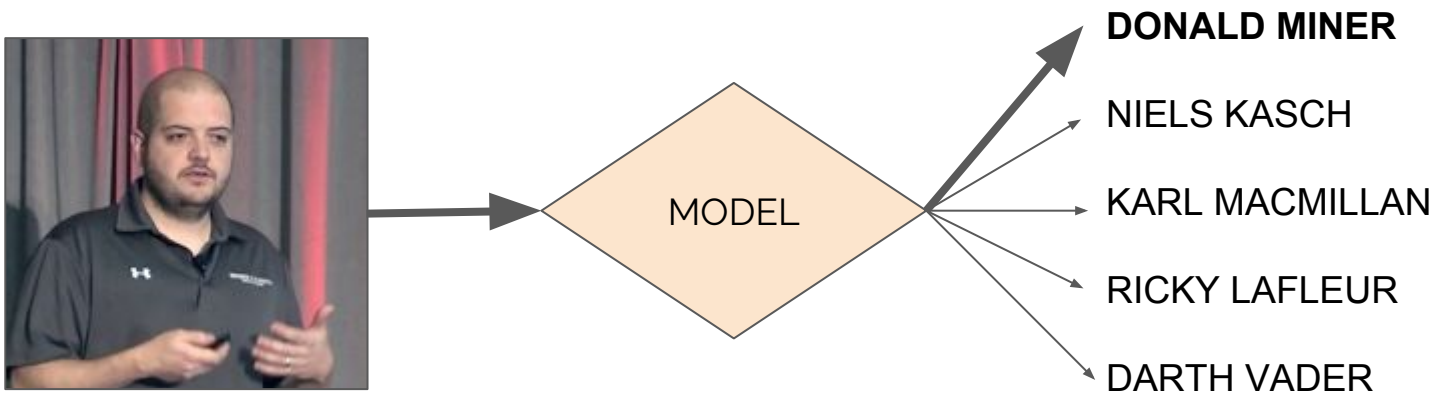
- Run experiments on how to trick the model without letting the original model know we are trying to trick it.
  - Email scams: generate tons of emails until we are able to get the model to mark it as a non-scam, instead of just spamming the server and seeing what goes through.
  - Facial recognition: keep trying different faces until we can pass through the facial recognition without having to keep standing in front of the security camera with paper face printouts to see what works.

- Predict what a ML-driven autonomous agent system will do before it does it
  - Take the observations of the real world and by knowing what the agent may do you may be able to antagonize it easily.

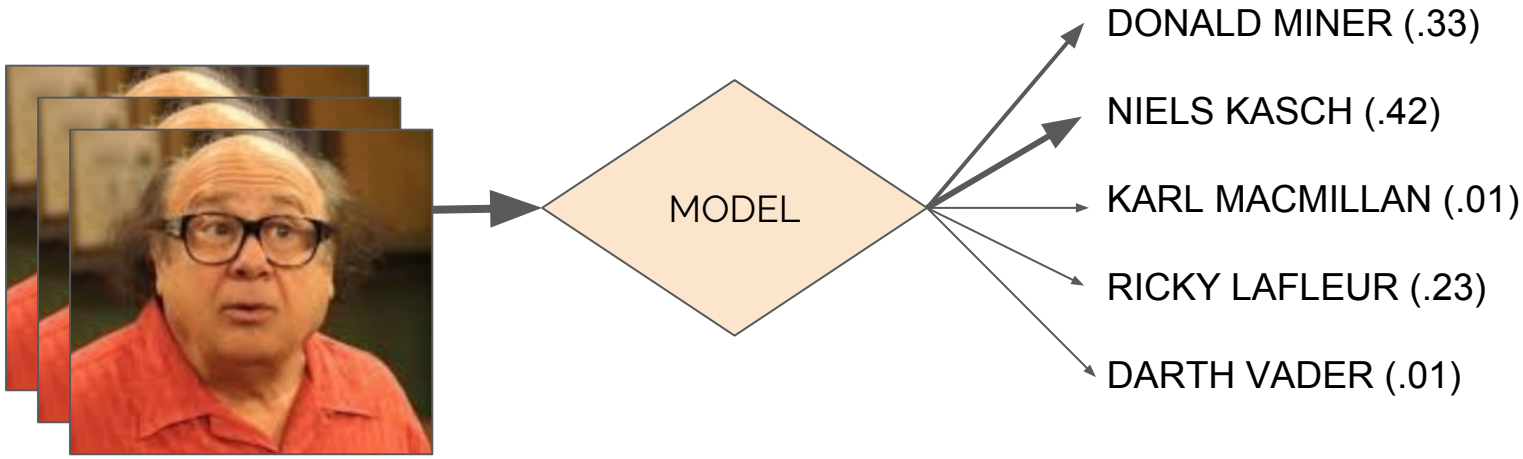
Attack: Inverting Models



Machine learning is trained from an training data. For example, in image classification that is the original images of the faces tied to their names.



Attack: Inverting Models



If we pass in a bunch of random photos of never before seen people, we can get the confidence scores of each prediction.

By taking these, we can layer them on top of each other to figure out what your face might look like.

“Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”

- Authors built an “inversion attack” that attempts to extract original training images from a model.
- For example, they used a facial recognition model to extract some of the faces the model has been trained on.

Frederickson et al., https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

Attack: Reverse Engineering

Use cases

- Extract original training data from the training data, which may be private or confidential.
  - Extract original segments of text
  - Extract certain features like survey results from individuals
  - Extract reconstructions of images

- Extract specific private/protected details of the input features that are not part of the output
  - Which keywords are keying to detect a scam?
  - Does the person we are looking for have glasses?

Recap

Machine learning can be reverse engineered

Sometimes information about the original training data can be extracted

What can we do?

Reverse Engineering Machine Learning

DONALD MINER
Miner & Kasch