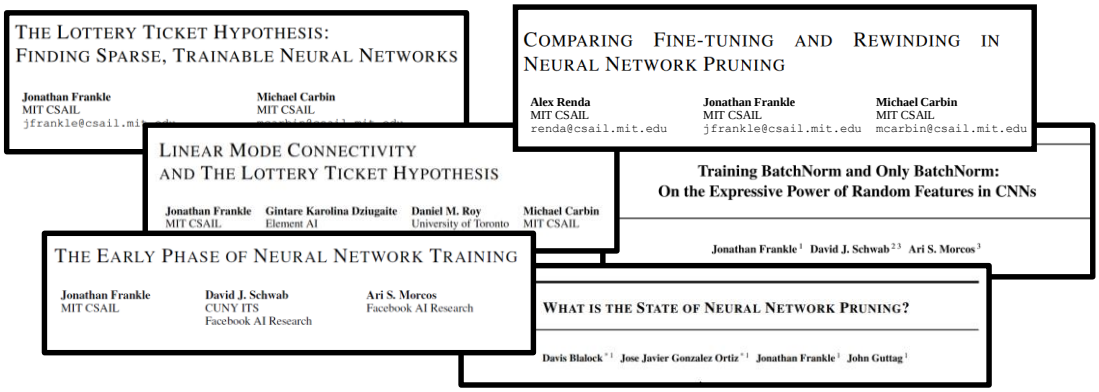
Artificial Intelligence for Policymakers

Jonathan Frankle
jfrankle@mit.edu

Who am I?

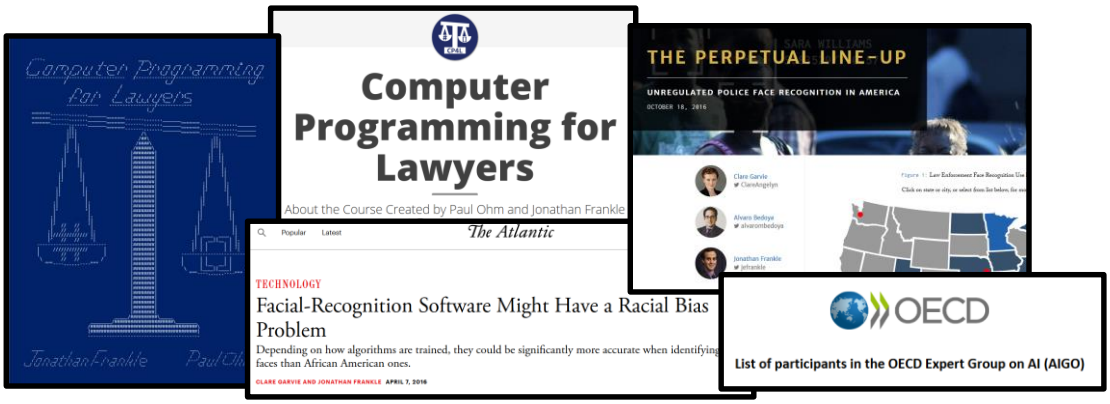
Who am I?

Fourth-year PhD student at MIT studying AI



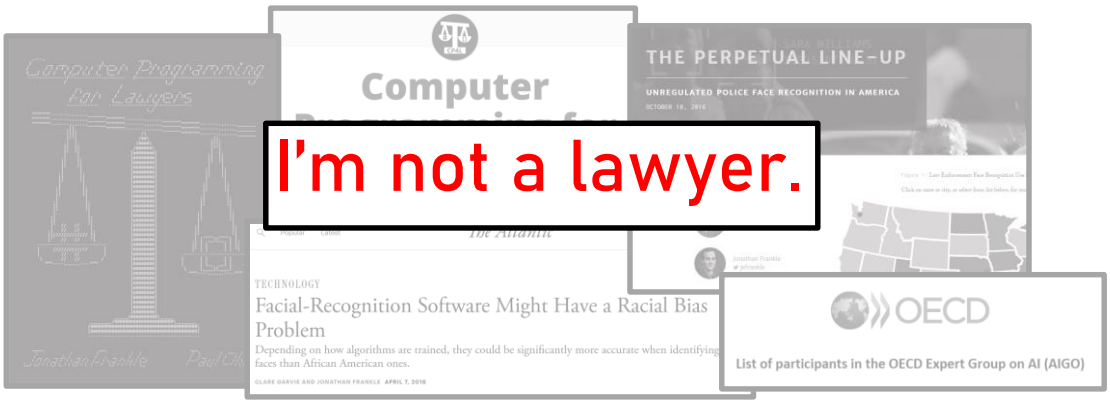
Who am I?

Adjunct Professor of Law at Georgetown Univ.



Who am I?

Adjunct Professor of Law at Georgetown Univ.



Let me just say up front that I'm not a lawyer. I'm not a legal expert, and I'm not qualified to answer legal questions. I leave that to the experts.

Goals

What is artificial intelligence?

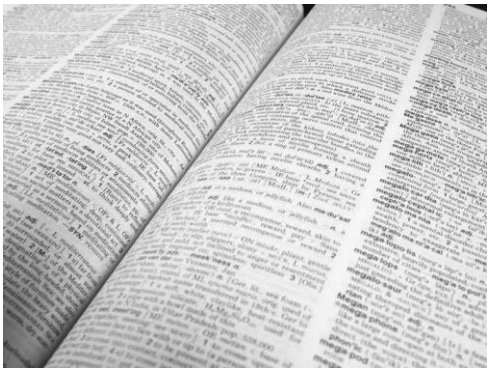
How are AI systems built?

What should we worry about?

Agenda

1. Definitions
2. How to build a model
3. Deep learning
4. How AI systems fail
5. What we should do about it (policy)

Definitions



Before we dig into definitions, let me ask. Why do we bother with definitions? In my mind, there is one big reason.

We want to make policy. To make policy, we have to be able to describe the systems that we're using to make policy.

Defining Artificial Intelligence

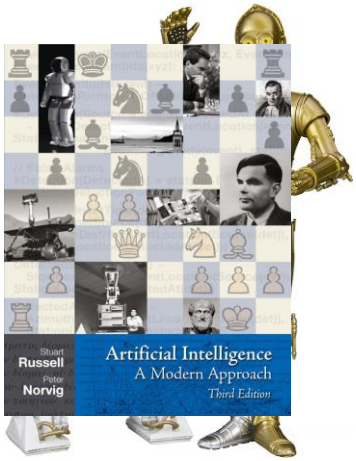
So when I say Artificial Intelligence, what do you think of?

Defining Artificial Intelligence

You might imagine robots from a movie. That's what I imagine.

And actually, this is how some very smart people define AI.

Defining Artificial Intelligence



If I open up my textbook on AI, I get the following definition.

Defining Artificial Intelligence

The exciting new effort to make computers think…[as] machines with minds, in the full and literal sense. (Haugeland, 1985)

So this definition is certainly one way of describing artificial intelligence.

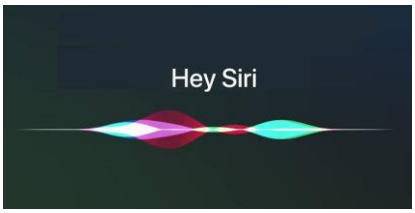
But nothing we have today comes close to this. None of our AI systems can think like humans. The biggest problem is that they make stupid mistakes, not that they're smart.

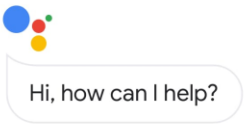
So this definition is too narrow. But it forces us to think about what it means to be artificially intelligent. What is the threshold when something becomes intelligent?

Defining Artificial Intelligence

These are definitely intelligent.

Defining Artificial Intelligence





Do we think these are intelligent? They're certainly not intelligent in a human way, but they can respond to us in a way that might be intelligent enough to seem like AI.

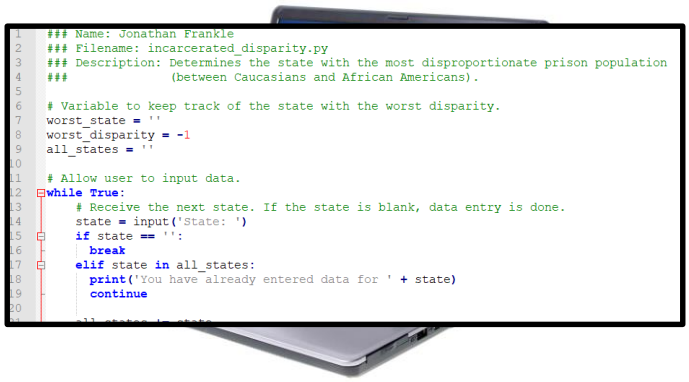
Defining Artificial Intelligence

What about a self-driving car? It's not very intelligent in the way that we are as humans, but I think this is still artificial intelligence.

Defining Artificial Intelligence

Is everything that runs on a computer AI?

Defining Artificial Intelligence



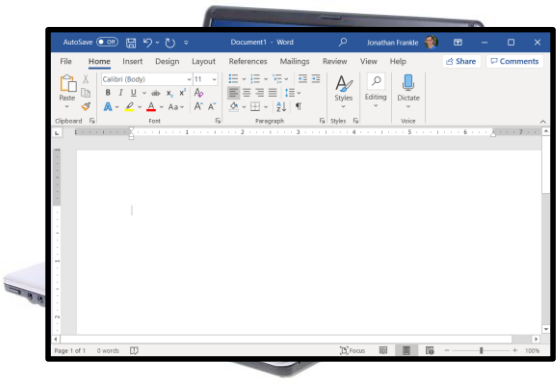
```
### Name: Jonathan Frankle
### Filename: incarcerated_disparity.py
### Description: Determines the state with the most disproportionate prison population
###              (between Caucasians and African Americans).

# Variable to keep track of the state with the worst disparity.
worst_state = ''
worst_disparity = -1
all_states = ''

# Allow user to input data.
while True:
    # Receive the next state. If the state is blank, data entry is done.
    state = input('State: ')
    if state == '':
      break
    elif state in all_states:
      print('You have already entered data for ' + state)
      continue
```

Are all computer programs AI?

Defining Artificial Intelligence

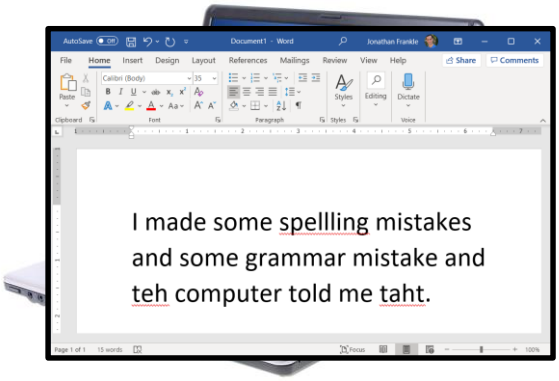
That would mean every website is AI

Defining Artificial Intelligence



Even this one?

Defining Artificial Intelligence



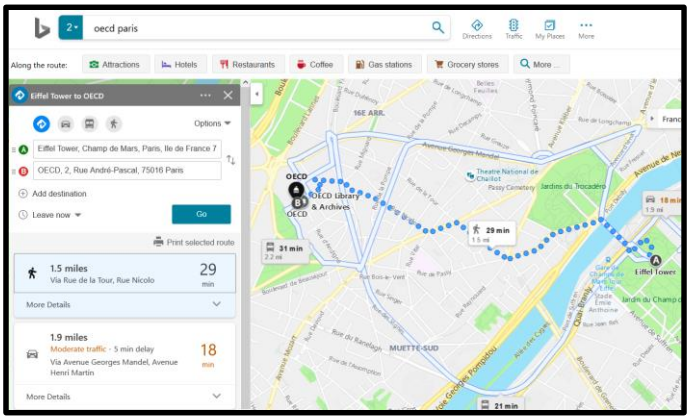
What about simple programs like spell checking? Is that artificial intelligence?

Defining Artificial Intelligence

Does AI even need to involve a digital computer?

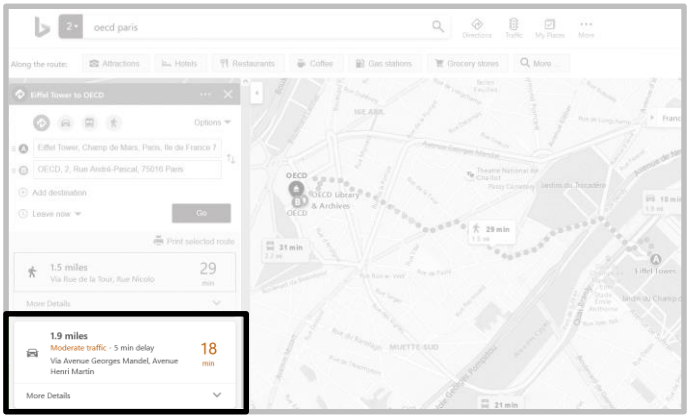
Do you see my point here – it's very hard to draw a line as to when something is intelligent. Let me give you a few more examples.

Defining Artificial Intelligence



Is mapping software artificial intelligence? If it were just an algorithm I wrote by hand, would it be?

Defining Artificial Intelligence



If it learned traffic patterns, would it be artificial intelligence?

Defining Artificial Intelligence

I hope you understand my point here. AI is far too imprecise a concept. It's very hard to come up with a good definition.

Defining Artificial Intelligence

Computer-aided reasoning (Jonathan Frankle, 2020)

Personally, this is my definition of AI. The first definition I showed you wouldn't describe anything we consider to be AI today. This definition includes everything that we do on a computer. And this is the big problem with defining AI. It's really hard to come up with a precise definition that draws the line in the right place, because we don't really know where that line is.

How do we proceed?

Okay, I've just told you to give up on defining AI. So how do we proceed?

I'm going to give you two answers.

How do we proceed?

1. Use another word.

The first answer is that may we should use a different word. Something more specific. I'm going to advocate for "machine learning." I'm going to define that for you now and tell you about why I think it's a better place to start.

How do we proceed?

1. Use another word.
2. Talk about what it does, not how it works.

And the second option, which I'll talk about later in the talk, is to avoid trying to describe how the system at all. I don't care if it's a computer program or a neural network or some very fancy machine. What I care about is what it does. For example, instead of

How do we proceed?

1. Use another word.
2. Talk about what it does, not how it works.

AI Policy

For example, instead of talking about AI policy

How do we proceed?

1. Use another word.
2. Talk about what it does, not how it works.

~~AI Policy~~

For example, instead of talking about AI policy

How do we proceed?

1. Use another word.
2. Talk about what it does, not how it works.

   Automated Decisionmaking

For example, instead of talking about AI policy

How do we proceed?

1. Use another word.

But we're going to begin by talking about a different word: machine learning.

Machine Learning

Machine learning. So to introduce machine learning, I want to give you an example.

Machine Learning

We have a picture of a dog. And I want you to build an algorithm that tells me whether it's dangerous or not. Now this dog doesn't look very dangerous.

Machine Learning

But this one looks a little scarier

Machine Learning

And this dog has three heads. It's very scary.

Machine Learning

How do we build an algorithm to predict whether a dog will bite?

So how do we do this? How do we build this algorithm?

Machine Learning

How do we build a model to predict whether a dog will bite?

And one thing I'll note here – I'm going to replace the word "Algorithm" with "model." When we're trying learn how to predict something, we usually refer to that as a model. Algorithm is a much broader term that could refer to any computer program.

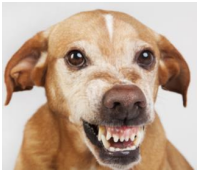
Machine Learning

So if we're going to do this the machine learning way, here's what we do. First, we need a lot of data.

1. Data Collection

So this is step 1: data collection.

1. Data Collection


This isn’t enough data.

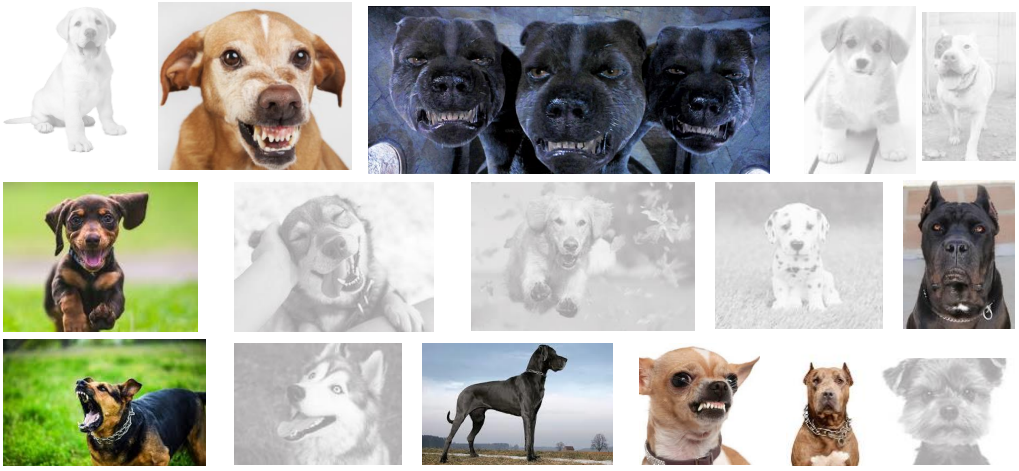
1. Data Collection

So we have more data. And this still isn't enough data, but it's a good start. If we were really going to train a model, I'd want 10,000 images or more.

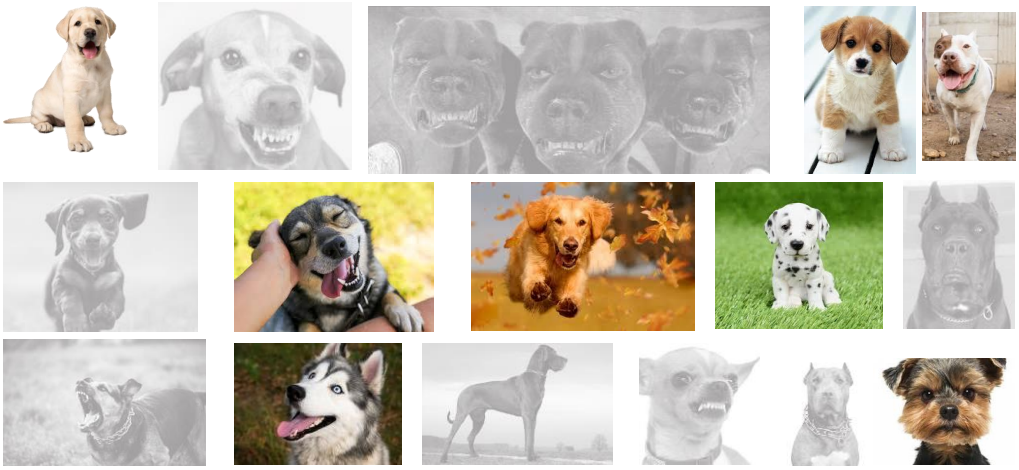
Now just having this data isn't enough. Right now, all I have are pictures of dogs. But I'm trying to learn to predict which ones are going to bite me. But in order to learn, I need you to teach me. I need you to tell me which dogs have actually bitten people. So we don't just need data – we need LABELS for that data.

I need you to tell me that these are the dogs that have bitten in the past.

1. Data Collection



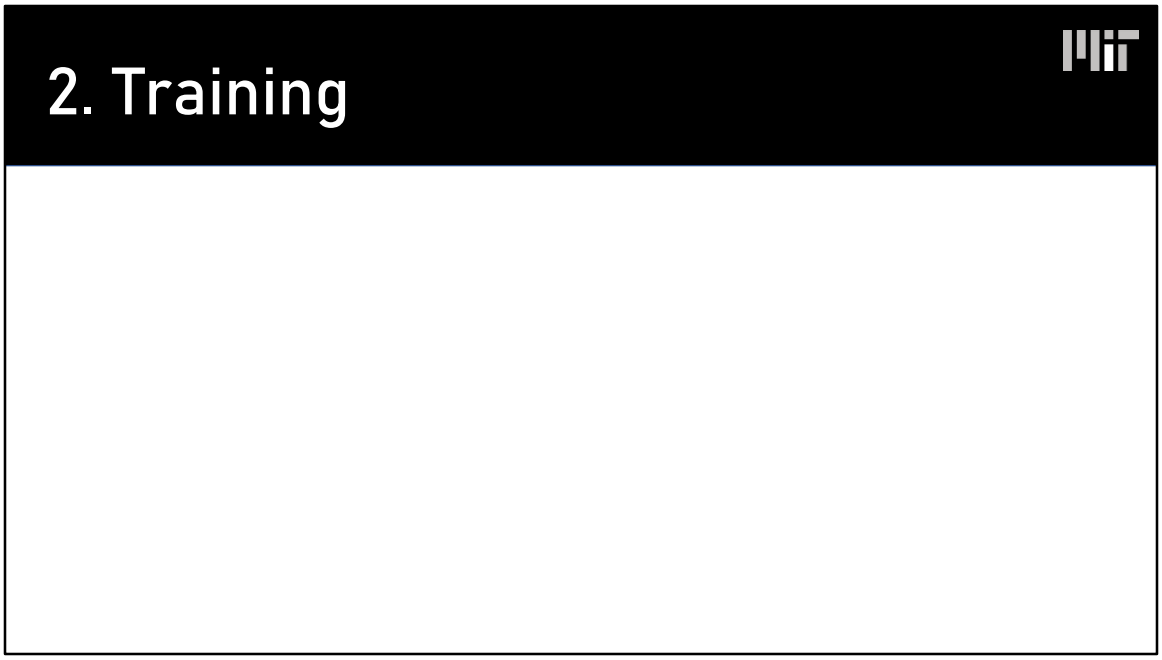
1. Data Collection



Now, this process, which we call “data collection,” can be expensive. It’s often hard to get good data, and it’s even harder to label it properly. And there are all sorts of problems we’ll talk about later, like privacy issues and concerns about bias.

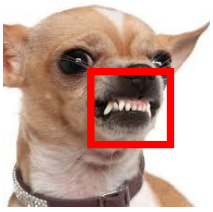
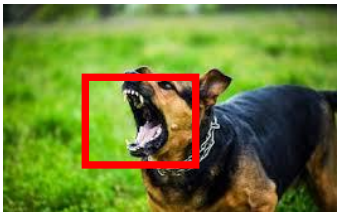
Now, we have our data. The next step is to learn from it, which we call training.

2. Training



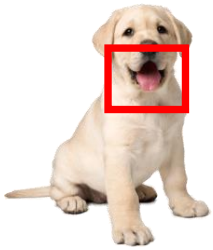
So during training, our goal is to understand what distinguishes dogs that bite from dogs that don't bite. We're looking for patterns in the data that we can use to make predictions in the future.

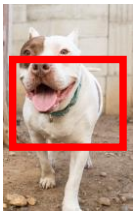
2. Training

For example, dogs that bite tend to show their teeth.

2. Training





Dogs that don’t bite seem to show their tongues.

2. Training



Dogs that bite tend to be big.

2. Training

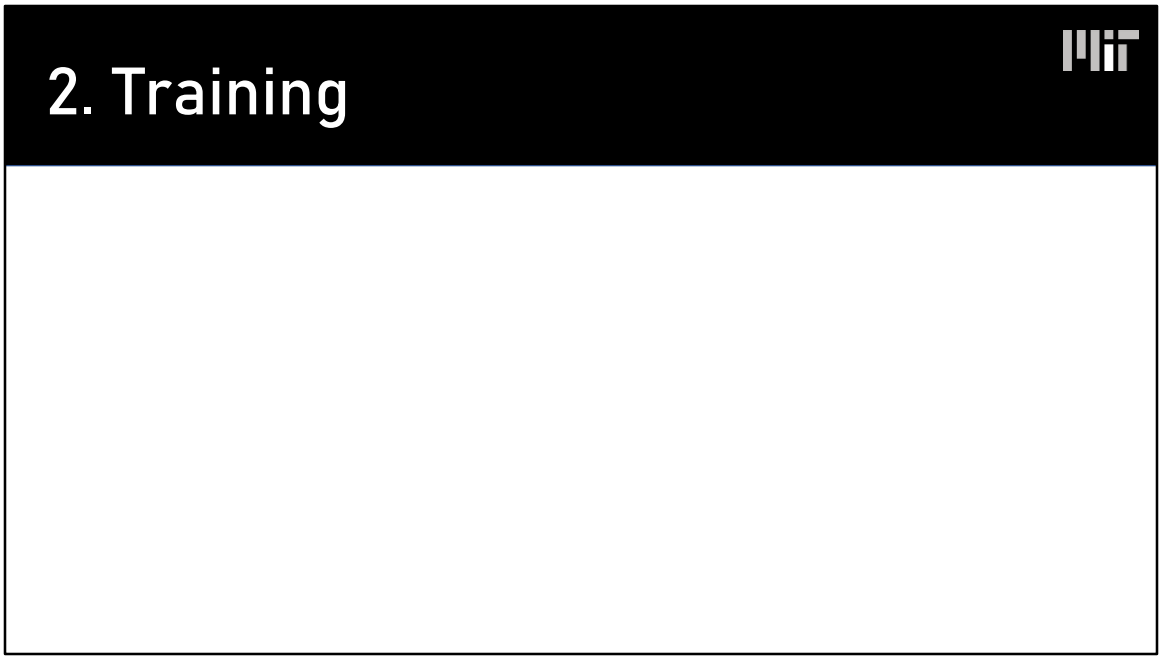


Dogs that don’t bite tend to be small

2. Training

But not always!

2. Training



And slowly but surely, we learn which qualities are associated with dogs that bite and dogs that don't bite.

2. Training

Features

Each of these qualities is known as a “Feature”

2. Training

Features:

- Is the dog showing teeth?
- Is the dog showing its tongue?
- How big is the dog?
- How many heads does it have?

Here are a few of the features we just discussed. Every time we get a new picture of a dog, we ask ourselves these questions, and based on the answers, we decide whether the dog is going to bite. That's our model.

So now we have a model, right? We're done, right? Nope, there are a few other important steps.

3. Evaluation

Before we can release this model into the world, we need to evaluate it. What does it mean to evaluate? We

3. Evaluation

Test the model on data and see how it performs

We test the model on data and see how it performs

3. Evaluation

Test the model:

- Is it accurate enough?

Are there certain kinds of dogs that it's bad at? Often, after we evaluate, we'll go back and change the model in some way. Maybe one of our features isn't very predictive. Maybe we're bad at recognizing corgis.

3. Evaluation

Test the model:

- Is it accurate enough?
- Is it bad at certain kinds of dogs?

Are there certain kinds of dogs that it's bad at?

3. Evaluation

Test the model:

- Is it accurate enough?
- Is it bad at certain kinds of dogs?

Maybe it's bad at corgis and we need more data.

3. Evaluation

Test the model:

- Is it accurate enough?
- Is it bad at certain kinds of dogs?
- Do we have any bad features?

Are our features useful? Are any of them causing us to make mistakes? During this process, we'll often go back and re-train the model a few times, changing various things about the way we train to get something that behaves better.

3. Evaluation

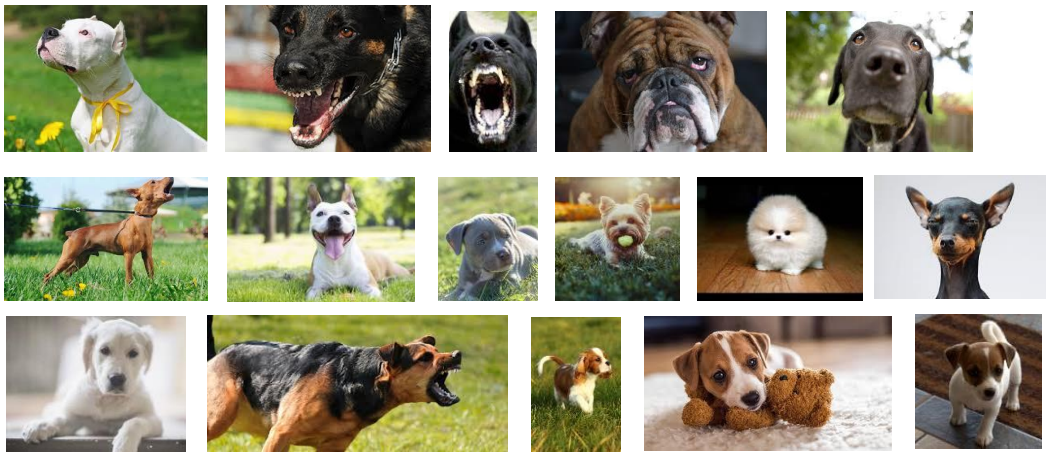
Test the model:

- Is it accurate enough?
- Is it bad at certain kinds of dogs?
- Do we have any bad features?
- What about on new, unseen data?

Now an important thing we need to do is try the model on new data. Data it has never seen before. Data that it wasn't trained on. The idea is that the model has seen the training data a lot already. What if the features we learned are specific to the training data? What if they don't generalize to new data? What if we simply memorized the training data?

To deal with this issue, we'll need even more new data.

3. Evaluation



This data is very important. It tells us how we expect our model to do on new data it hasn't seen during training.

Okay, so we now have a model. And we think it's pretty good. There's still one more step.

4. Deployment

Step 4 is deployment. This is when we actually send it out into the world.

4. Deployment

Efficiency

One big concern here is efficiency. Can the model run on my cell phone?

4. Deployment

Efficiency

One big concern here is efficiency. Can the model run on my cell phone?

4. Deployment

But it's also important to focus on how the model performs in practice. Often, problems may arise that you didn't expect.

4. Deployment

For example, Google's image classification algorithm labeled black people as gorillas.

Machine Learning

Okay, so to review, machine learning takes place in four steps.

Machine Learning

(Decide on the problem)

Well, really five. The first step is to decide on the problem you want to solve. This isn't always easy. And often, you know the problem, but you don't know how to put it into a concrete machine learning application. Maybe the problem is even impossible to solve with the data available. But assuming it is, our first step

Machine Learning

(Decide on the problem)

1. Data collection

Is to collect data.

Machine Learning

(Decide on the problem)

1. Data collection
2. Training

Then we train a model, trying to find patterns in the data that allow us to make good predictions. Trying to find useful features.

Machine Learning

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation

Then we evaluate the model to make sure it's working to our liking.

Machine Learning

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation
4. Deployment

And finally, we release it to the world. But even after we do that, we need to continue to monitor it to make sure it's behaving properly.

Okay, so now that you know what machine learning is, let me give it a formal definition. I define it as

Machine Learning

Developing a model of a process by learning from data to make predictions in new situations.

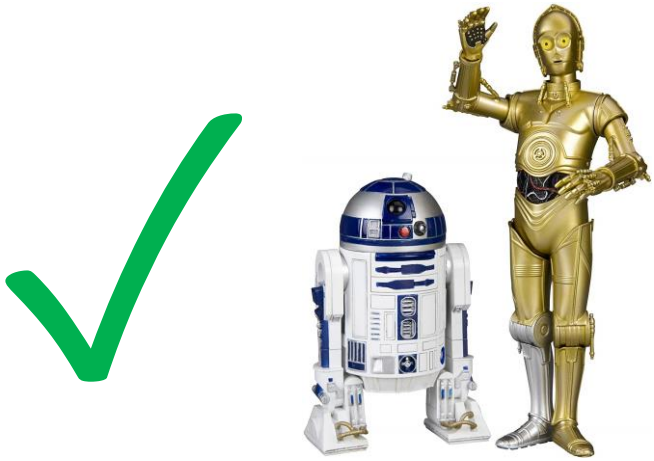
(read the definition). You can see that it's all about collecting data and training on that data to get a model that we can use in practice. It's all four of those steps in a single definition.

Now, what I like about that definition is that it's really clear. It's easy to say which things are machine learning and which things aren't.

Machine Learning

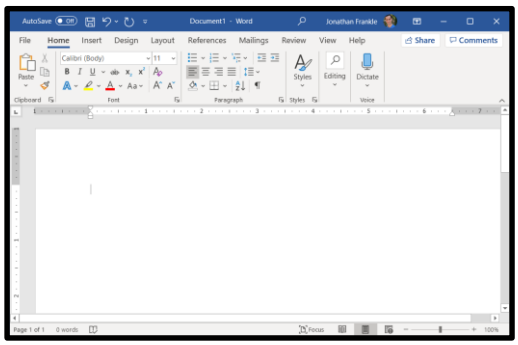
Our robots from star wars?

Machine Learning



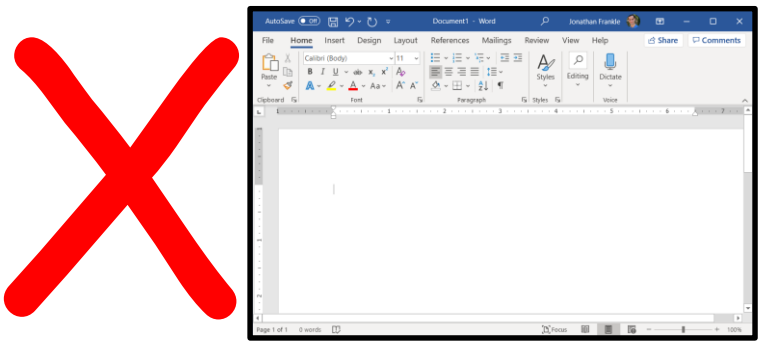
Machine learning.

Machine Learning



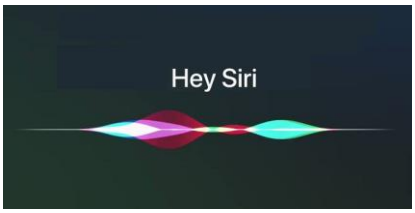
Microsoft word?

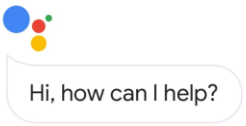
Machine Learning



Not machine learning

Machine Learning



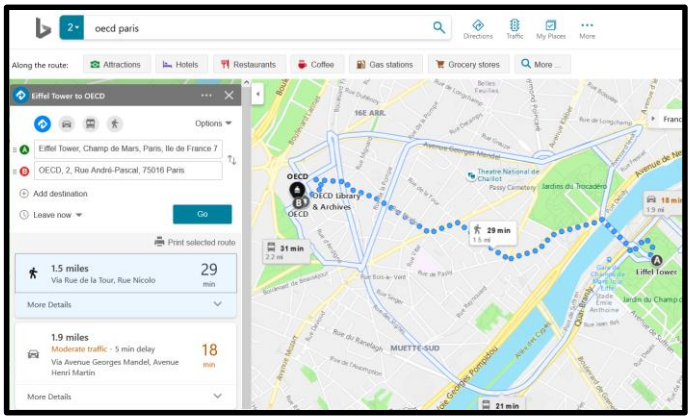


Voice assistants

Machine Learning

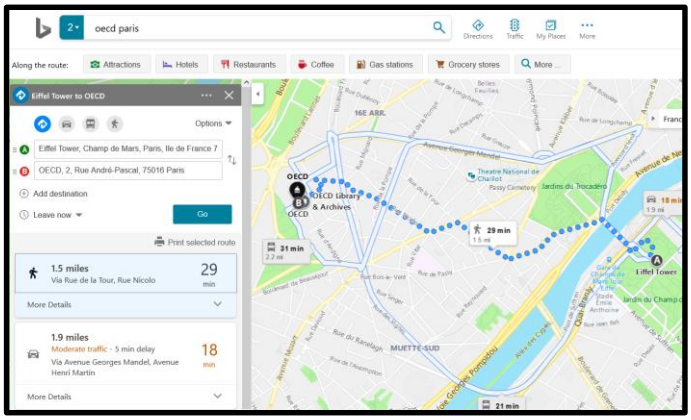
Machine learning

Machine Learning



Mapping software?

Machine Learning

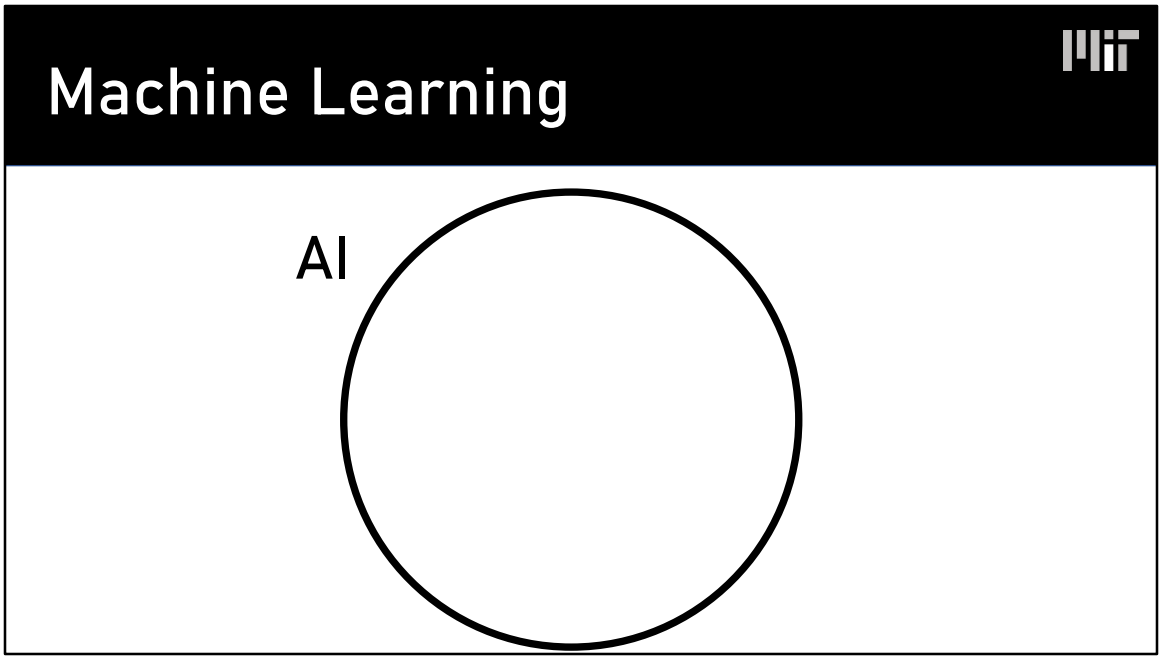


Well, it's complicated. If it's just using a hand-written algorithm to mapping, probably not. If it's learning based on traffic patterns and which routes work best in its past experience, then yes. But it's easy to analyze which aspects of it are and aren't machine learning.

Machine Learning

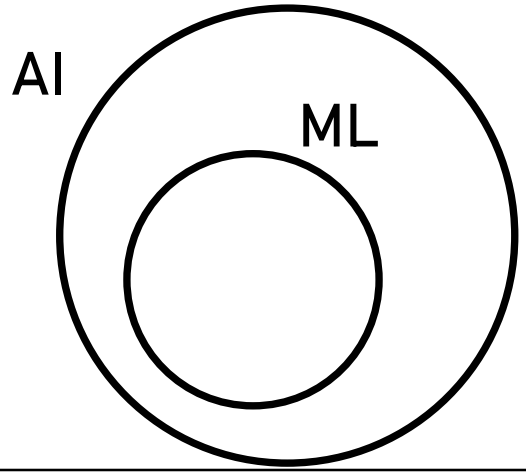
Now, not all AI uses machine learning. I think of AI as a big circle

Machine Learning



And

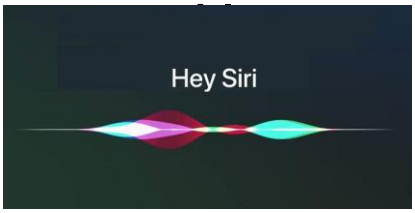
Machine Learning

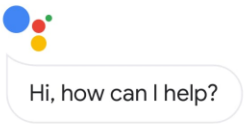


And machine learning is a smaller circle inside it. So the definition of ML doesn't cover everything.

But the nice thing is that all of the modern AI systems that are giving us headaches use machine learning.

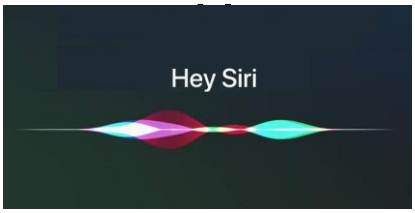
Machine Learning

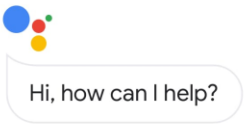




That’s voice assitants

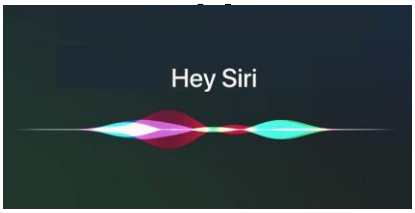
Machine Learning





Self-driving cars

Machine Learning



Facial recognition

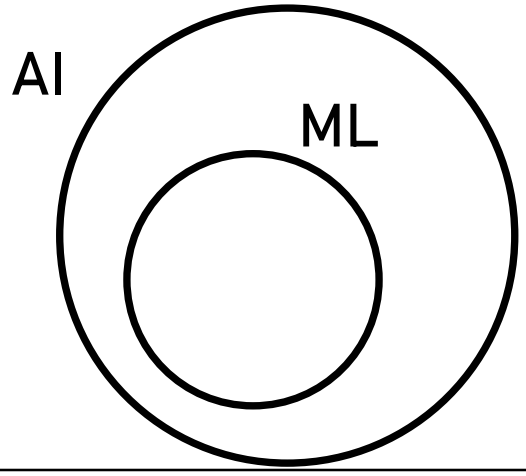
Machine Learning

Social media news feeds.

Machine Learning

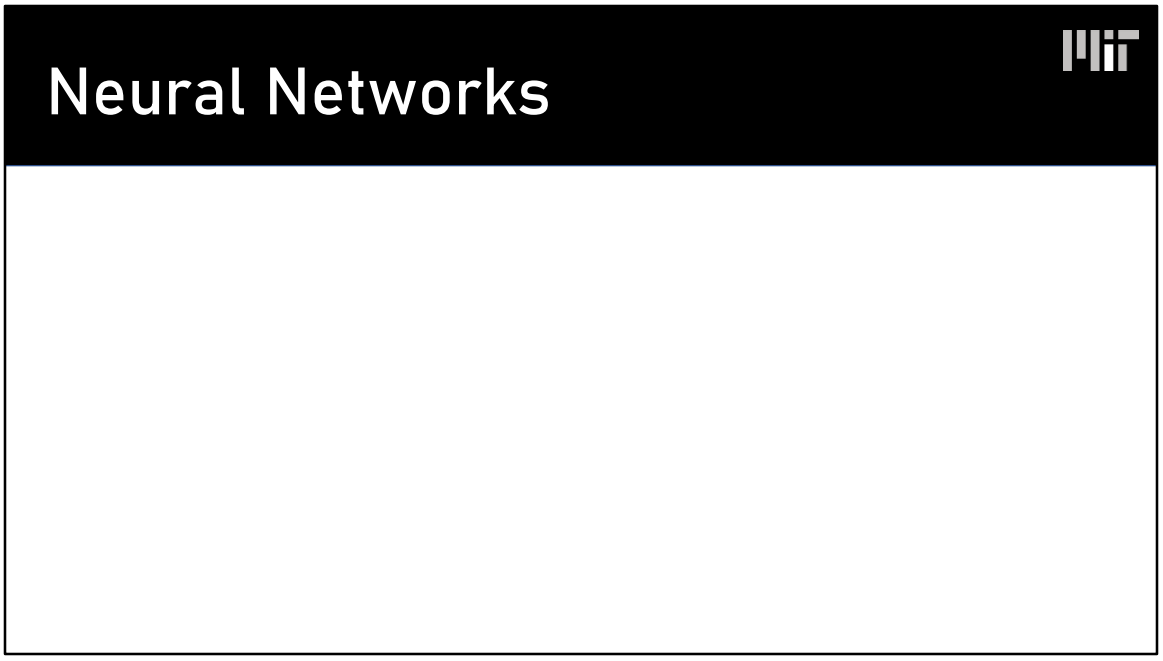
Credit scoring. So

Machine Learning



Although machine learning doesn't cover absolutely everything that might broadly be considered AI, it covers pretty much all of the important applications that are exciting and creating policy challenges today.

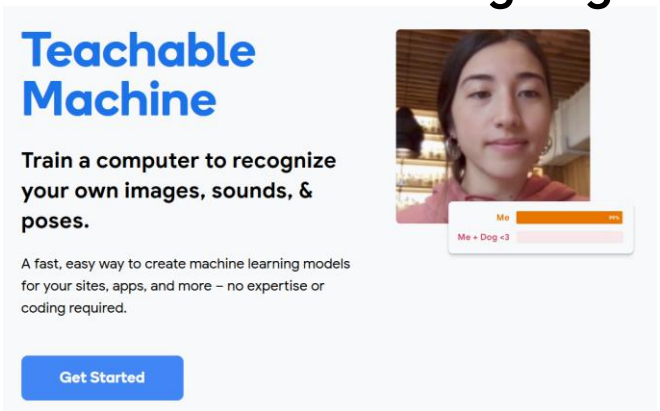
Neural Networks



And the most exciting of those technologies is something called “neural networks,” or “deep learning.” But before we discuss what those are, I want to do a quick demo of machine learning.

Demo

teachablemachine.withgoogle.com



We're going to use a cool tool that Google created call the "teachable machine." It allows you to create a small machine learning model in your browser. You can search for "teachable machine" if you want to follow along with me as I do the demo.

Demo:
1. Describe the task: let's do some basic facial recognition to tell apart two people
2. Call up two people and collect data. Data collection.
3. Train the model.
4. Evaluate on them to see how well it worked.
5. Try adding a third person.
6. Leave a couple of minutes for them to play with it on their computers.

Demo

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation
4. Deployment

As a reminder of the steps we went through with teachable machine

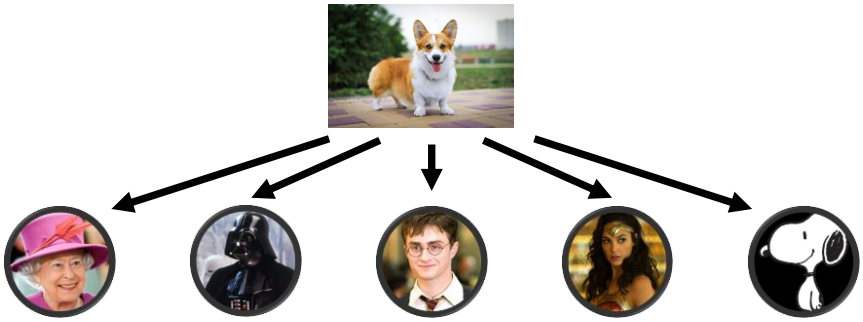
Neural Networks

So let's get back to neural networks. Now that we've gotten a good understanding of machine learning, I want to talk about one particular kind of machine learning model in a bit of detail. That's a neural network. You'll often hear the word "deep learning" used to discuss neural networks, and those two terms are largely interchangeable at this point. Neural networks are responsible for a lot of the recent advances that have brought you here - computer vision, self driving cars, natural language processing, deep fakes, everything like that. This is also my research specialty

Neural Networks

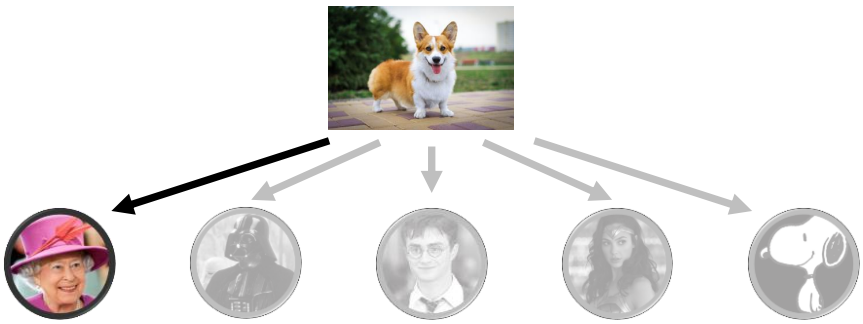
And we're trying to figure out whether it's going to bite us or not. Now, one way to do this is to consult with many experts on the subject.

Neural Networks



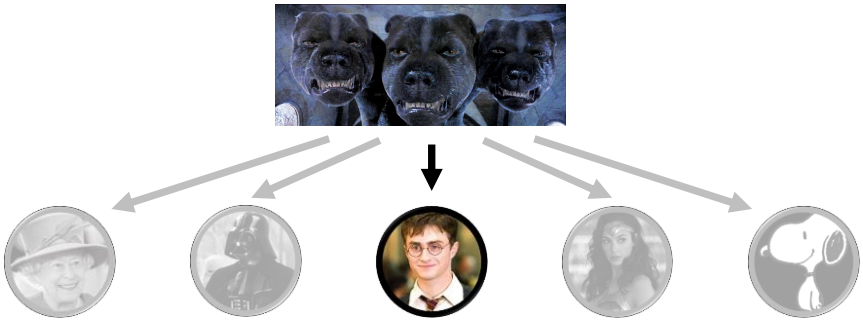
So we ask these experts what they think. Now why do we have lots of experts? Some of them will be better at some things, and some will be better at others. Queen Elizabeth will be better at Corgis.

Neural Networks



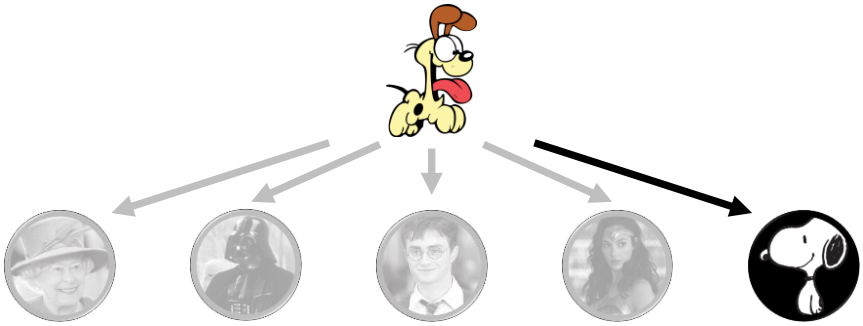
And Harry Potter is going to be better at three headed dogs

Neural Networks



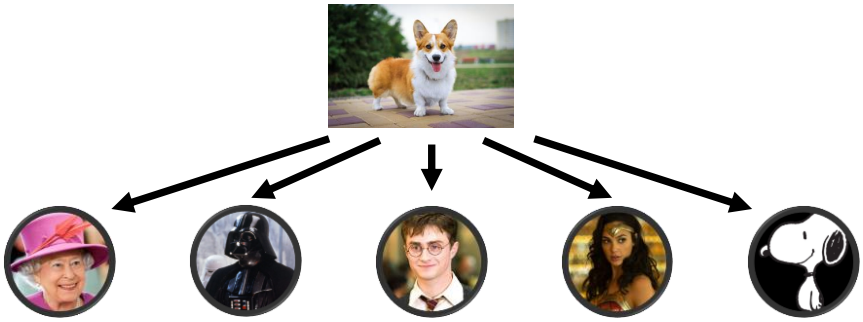
And Harry Potter is going to be better at three headed dogs

Neural Networks



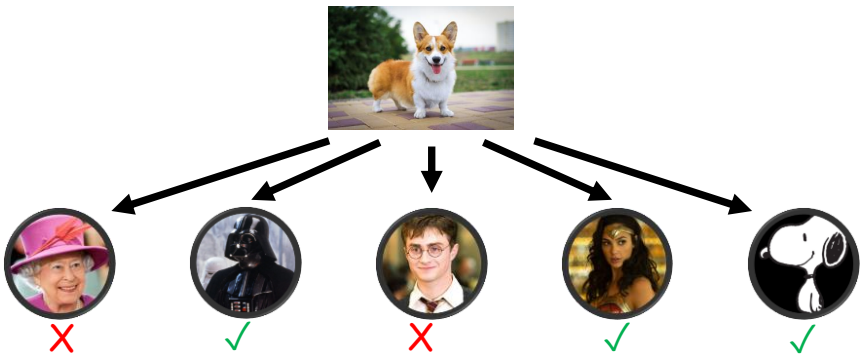
And maybe snoopy is better at cartoons.

Neural Networks



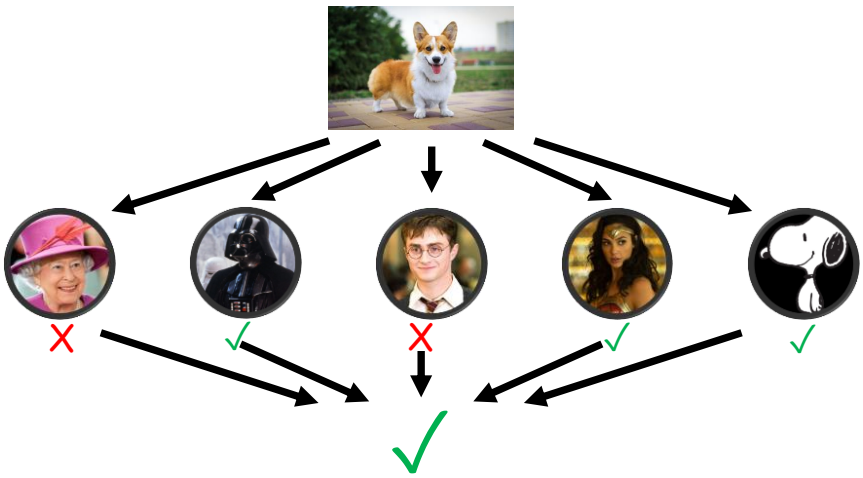
And slowly, we're going to try to learn how much we should trust each of these experts and how we should combine their opinions to make a final decision.

Neural Networks



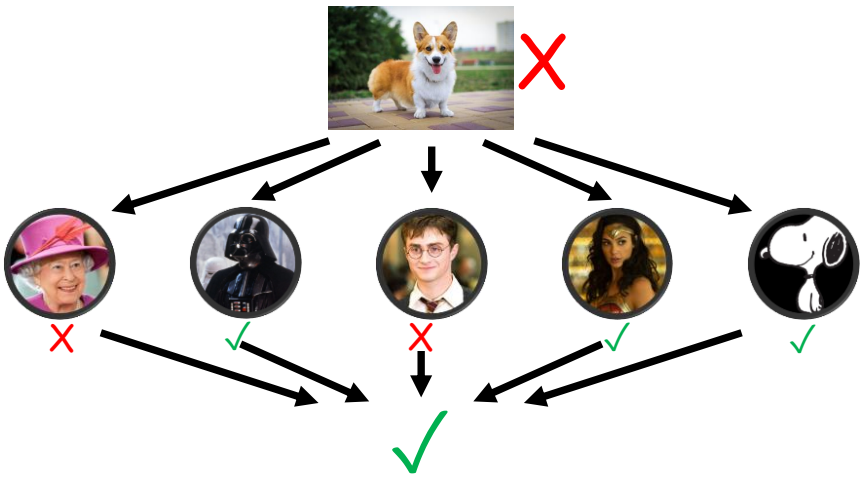
We get their input.

Neural Networks



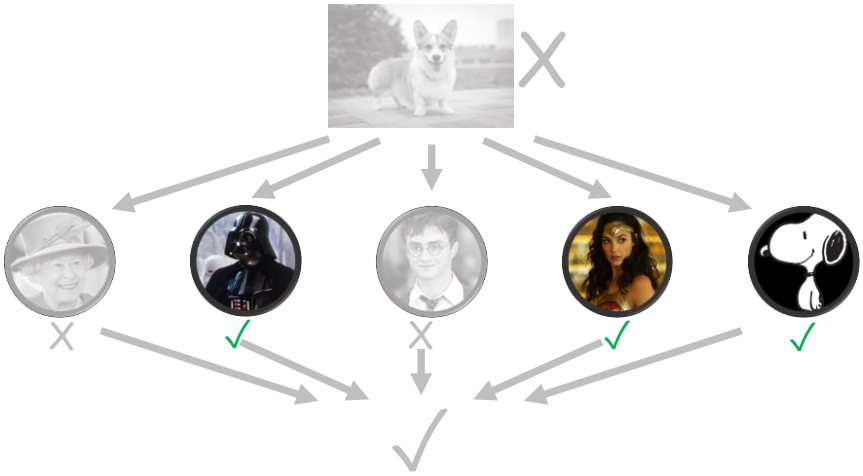
And then make our decision based on that.

Neural Networks



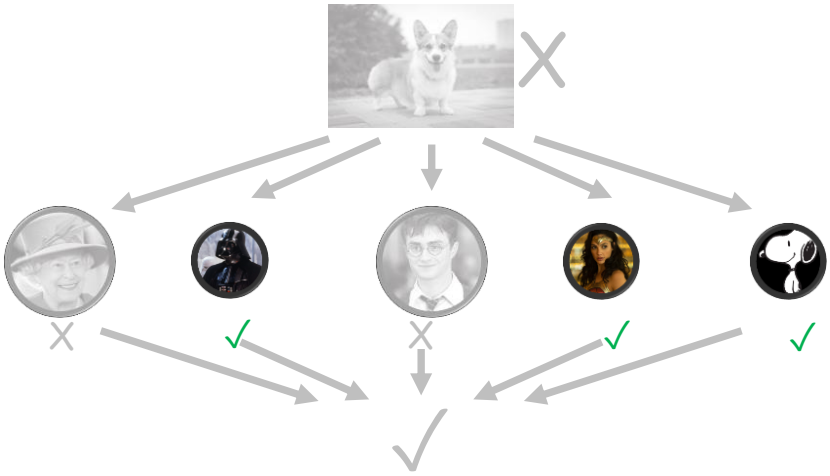
Then, we look at the correct answer. So we got it wrong.

Neural Networks



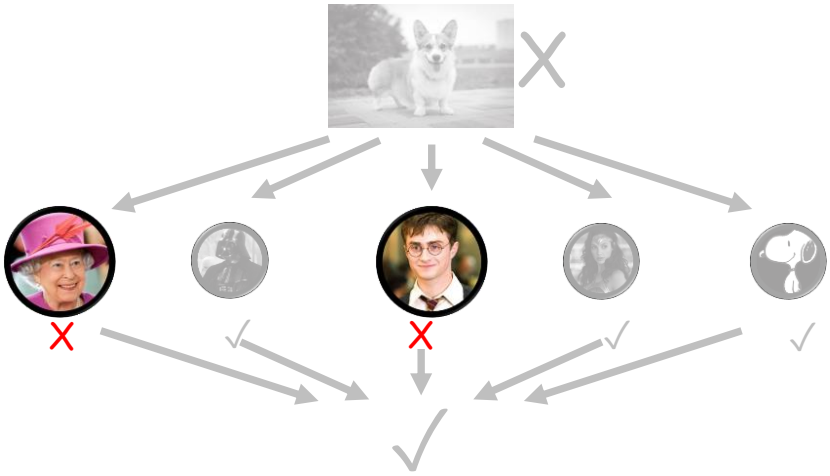
We take all the experts who got it wrong.

Neural Networks



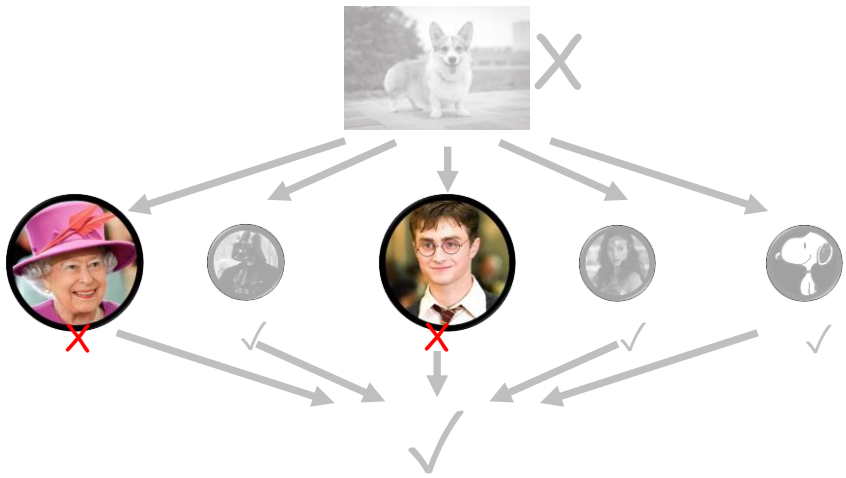
And we reduce their influence.

Neural Networks



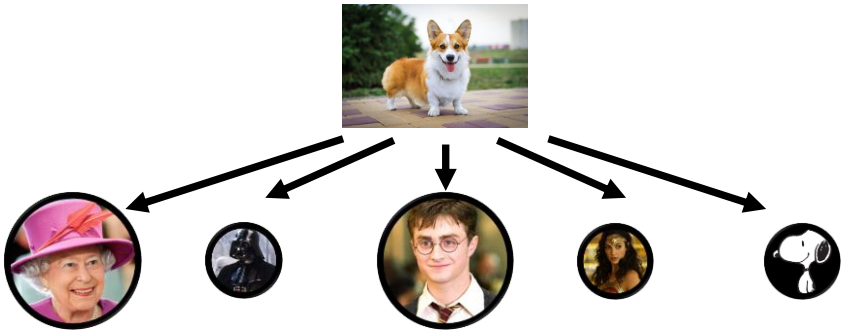
And we take the ones who got it right

Neural Networks



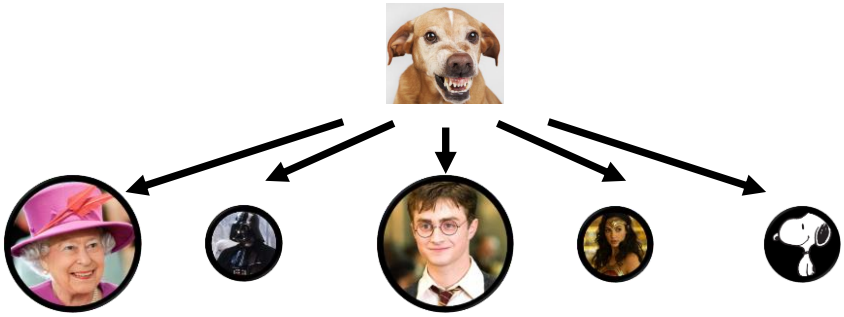
And we increase their influence.

Neural Networks



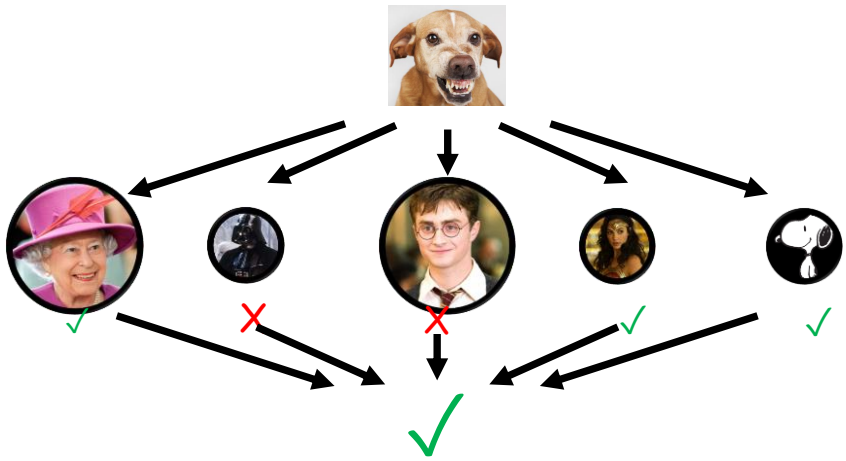
So now we have our new network. And we do this again.

Neural Networks



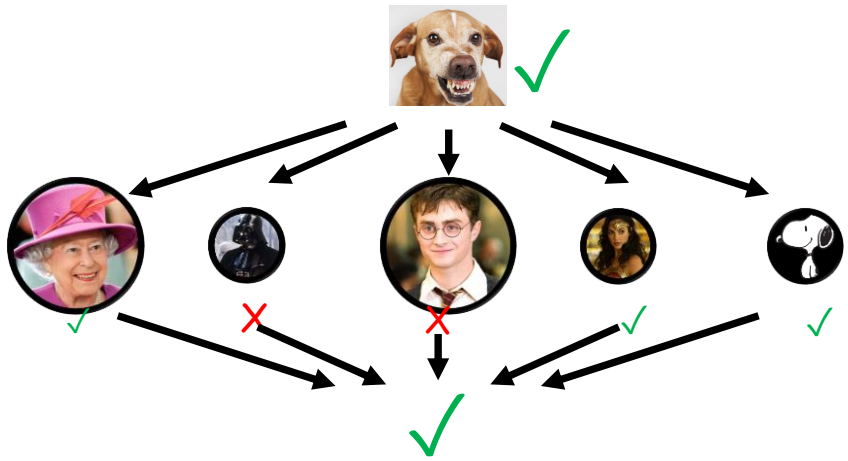
We take another input.

Neural Networks



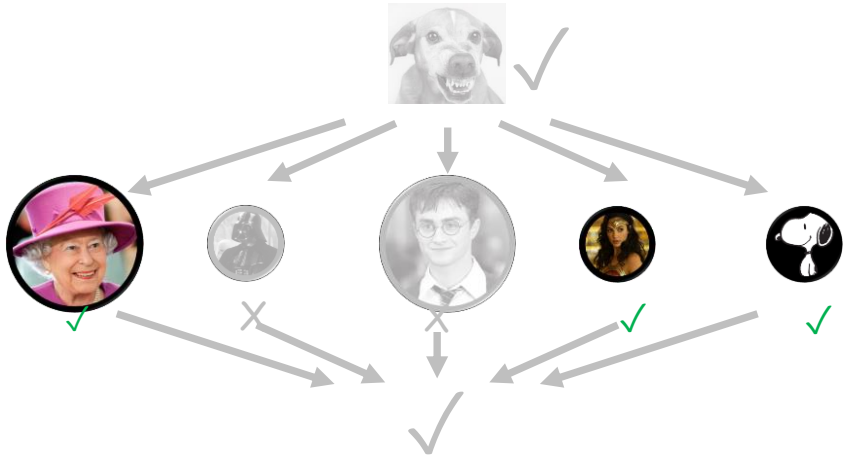
And make our choice.

Neural Networks



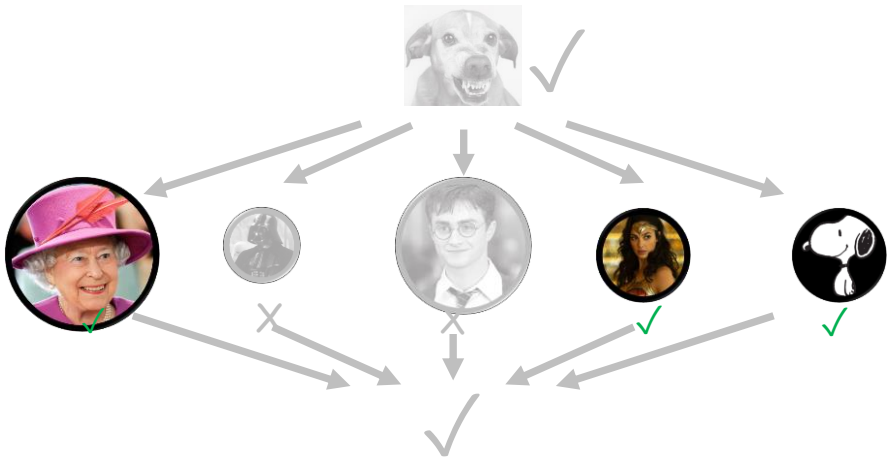
Now this time, we got it right.

Neural Networks



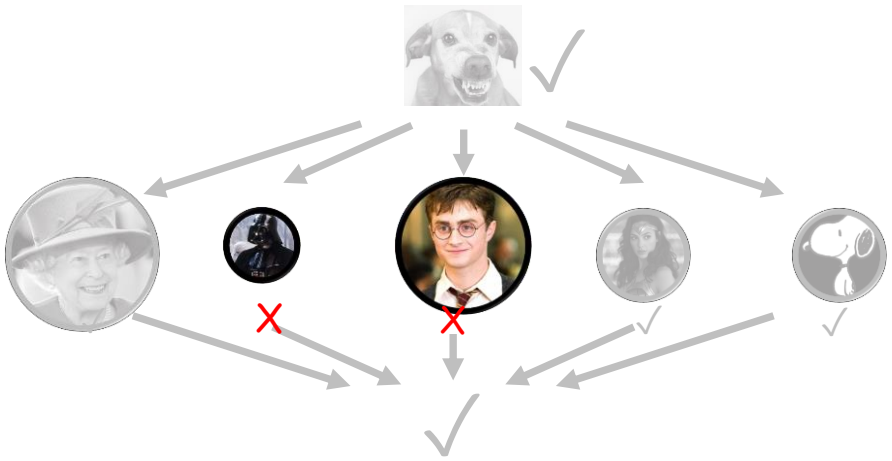
So we take the experts who got it right.

Neural Networks



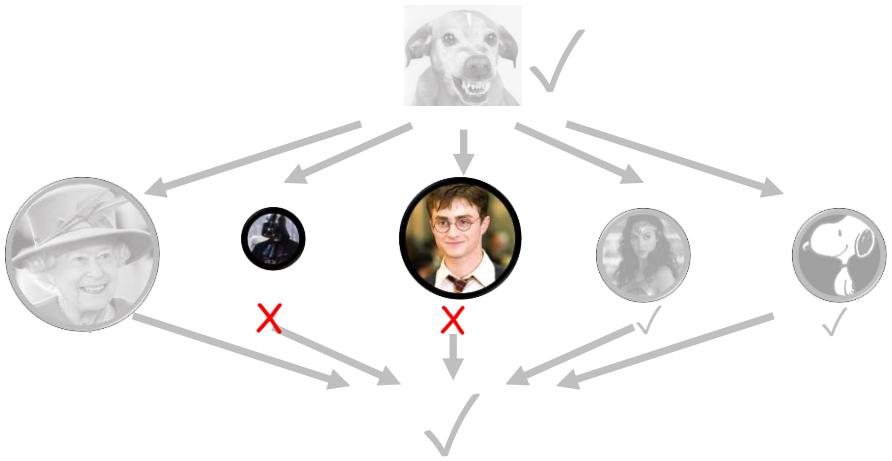
And amplify them.

Neural Networks



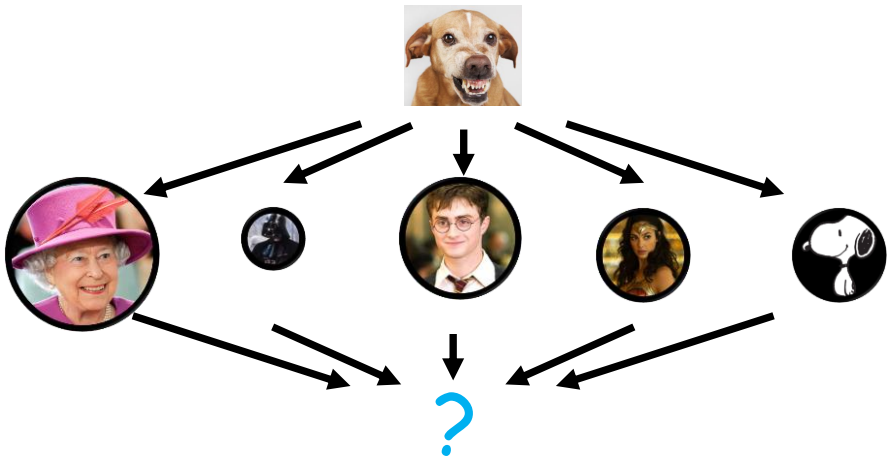
And take the ones who got it wrong.

Neural Networks



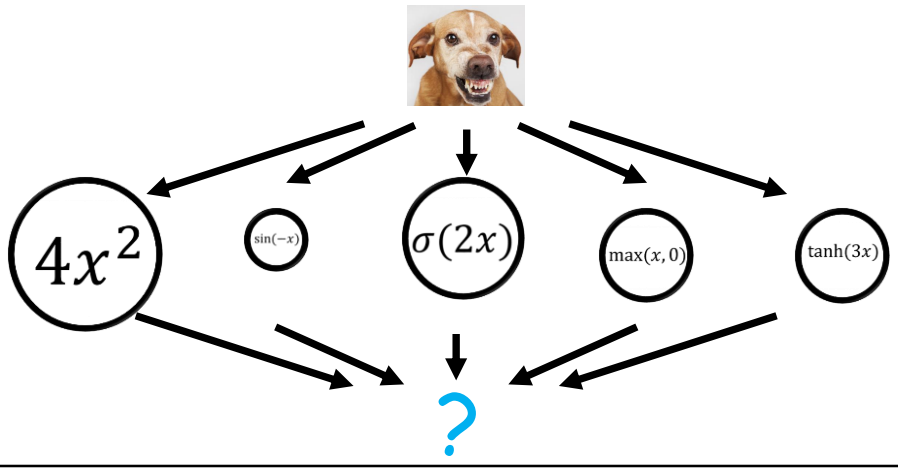
And reduce their effect

Neural Networks



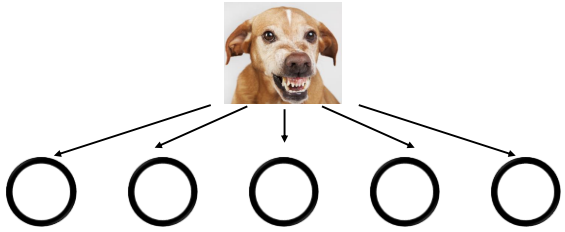
And in doing so, we slowly learn our neural network. Now obviously, we don't actually ask people. Instead, we use random mathematical functions that have the same effect. And we modify these functions as we learn.

Neural Networks



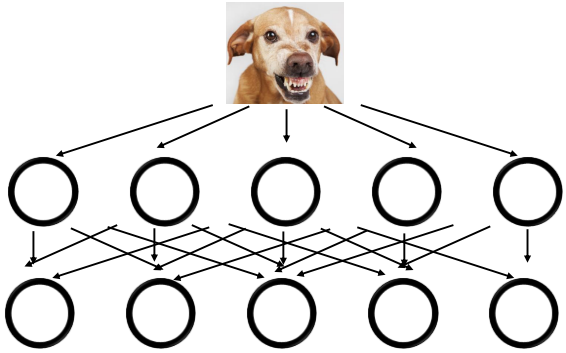
Now, there's one more thing. In practice, we don't just consult one set of experts.

Neural Networks



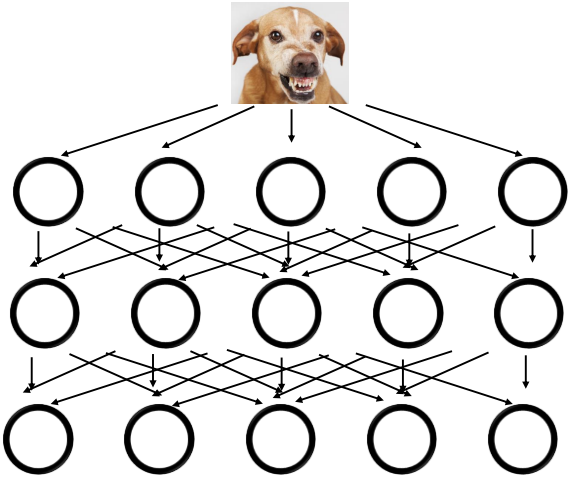
We first talk to one set of experts. Then we take their opinions to another set of experts.

Neural Networks



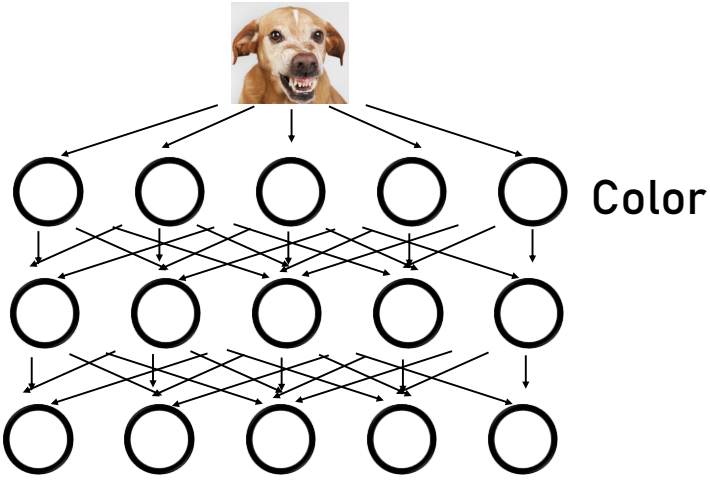
And then another

Neural Networks



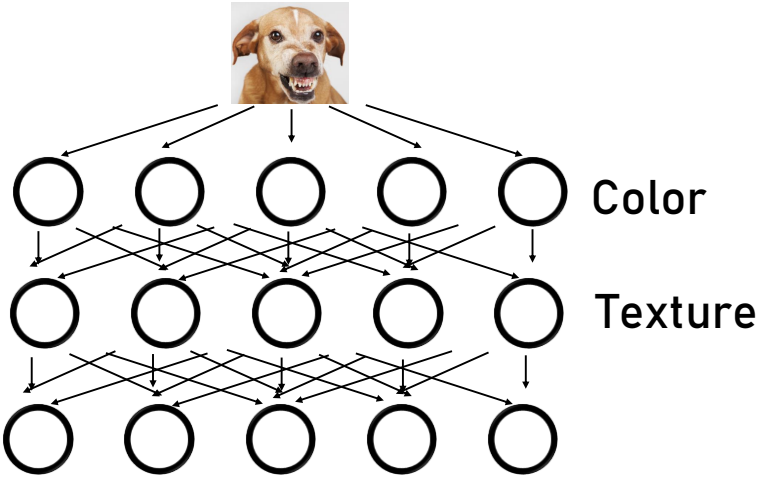
And we can repeat this for many many “layers” of experts. This is where deep learning gets its name – these networks are deep.
And you might ask the question, why do I need multiple layers of experts? What can the lower layers do with the information from the higher layers?
The answer is that, in practice, we find that higher layers can learn bigger ideas. For example, the first layer might learn color

Neural Networks



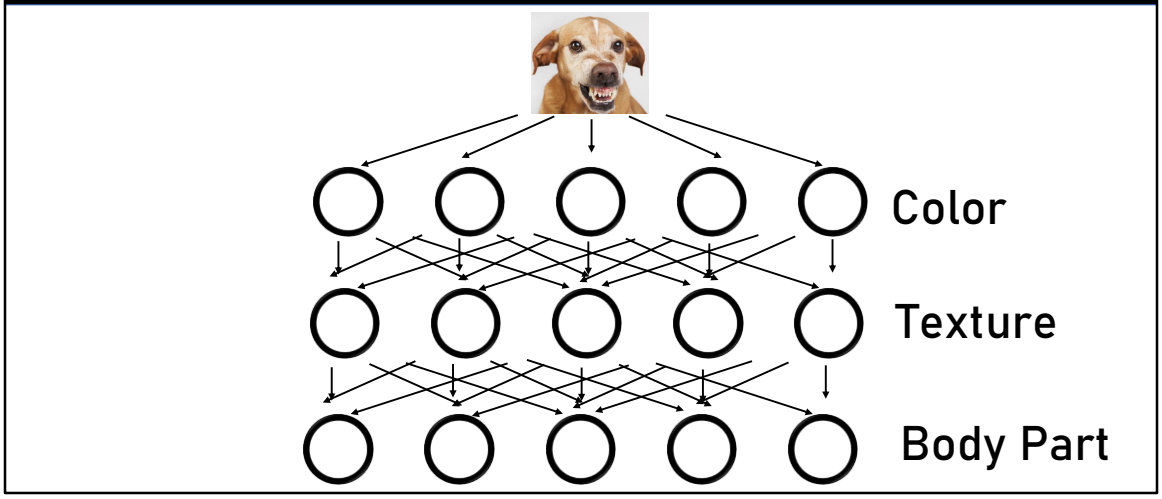
And the next layer texture

Neural Networks



And we can repeat this for many many “layers” of experts. This is where deep learning gets its name – these networks are deep.
And you might ask the question, why do I need multiple layers of experts? What can the lower layers do with the information from the higher layers?

Neural Networks



And the next layer body part. And gradually these higher level features are useful for making our final decision.

Neural Networks

Now in practice, these neural networks are big. How big?

Neural Networks

If we want to identify small pictures of hand-written digits…

Neural Networks

99% Accuracy

We can get to 99% accuracy

Neural Networks

99% Accuracy

3 Layers

With a three layer neural network

Neural Networks

99% Accuracy

3 Layers

300K Connections

That has 300,000 connections between experts.

Neural Networks

99% Accuracy

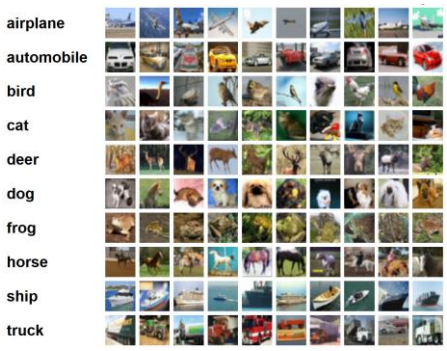
3 Layers

300K Connections

$0.10

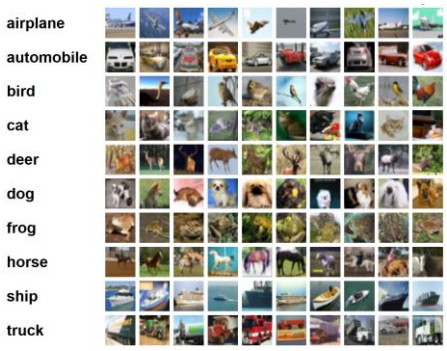
For very little cost. You could train this network on your laptop in a few minutes.

Neural Networks



Now let's consider a more challenging task with small, color images of objects. Ten different kinds of objects, to be specific.

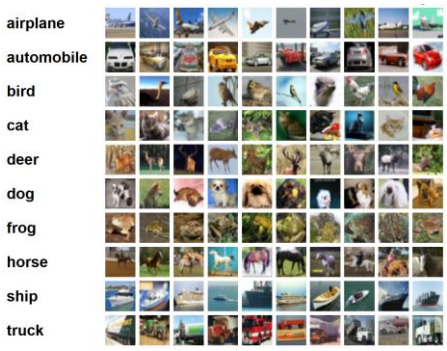
Neural Networks



93% Accuracy

We can get to 93% accuracy

Neural Networks

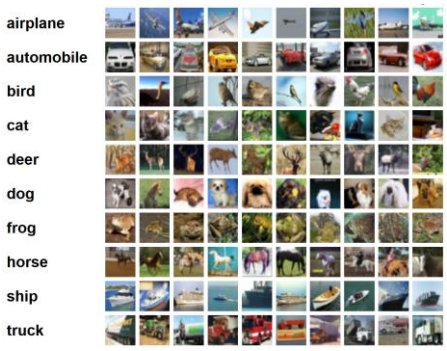


93% Accuracy

20 Layers

With a network with 20 layers

Neural Networks



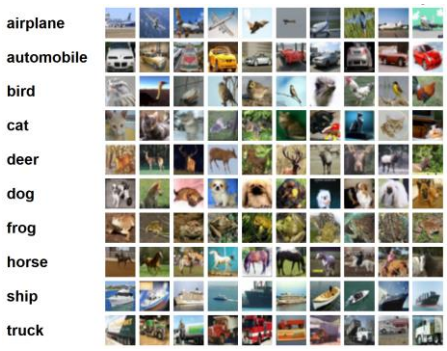
93% Accuracy

20 Layers

1M Connections

And about 1 million connections

Neural Networks



93% Accuracy

20 Layers

1M Connections

$1

For about a dollar of computation time. Okay, one more.

Neural Networks

If we look at bigger pictures with 1000 different categories of images.

Neural Networks

76% Accuracy

We can get to about 76% accuracy.

Neural Networks

76% Accuracy

50 Layers

With 50 layers

Neural Networks

76% Accuracy

50 Layers

25M Connections

25 million connections

Neural Networks

76% Accuracy

50 Layers

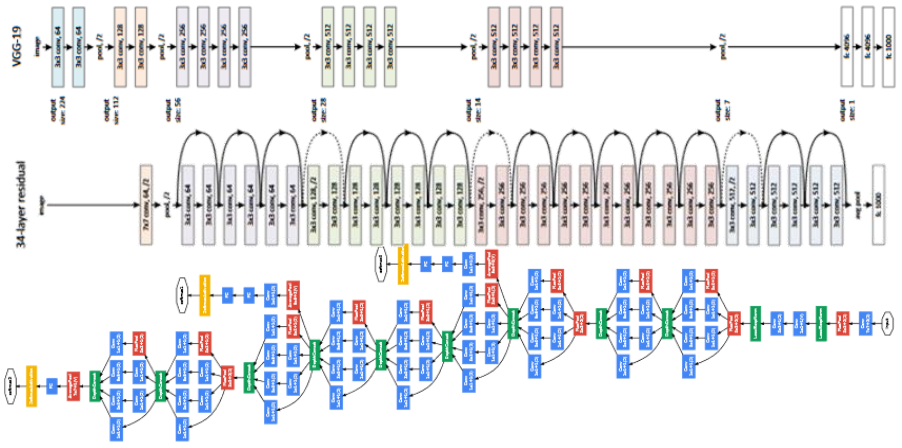
25M Connections

$400

And about $400 for the computation necessary to train it. And these are pretty simple tasks – just categorizing small images.

My point here is to communicate that the neural networks we use in practice are really big and really complicated. And, as far as we know right now, they have to be. These are hard problems we're trying to solve, and the models we need are complicated.

Neural Networks



For the task I just described, here are sample architectures that people have proposed in the past few years. I had to put them sideways to get them to fit on the page.

Neural Networks

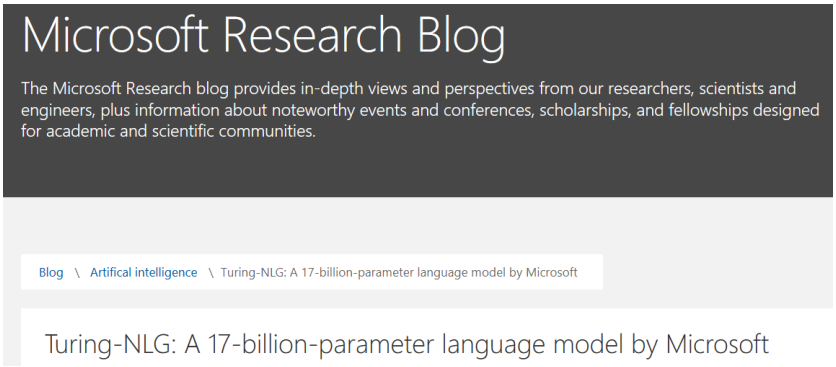
Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting.

And they're getting bigger. Did anyone hear about this "GPT-2" model that generates really realistic text? It has 1.5 billion connections.

Neural Networks



And Microsoft just released one with 17 billion connections.

What Can Go Wrong

And with these giant, complicated statistical machines, we can run into a lot of unexpected problems. I'm going to go through a few of them quickly.

What Can Go Wrong

Artificial intelligence > OECD Principles on AI

What are the OECD Principles on AI?

The OECD Principles on Artificial Intelligence promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values. They were adopted in May 2019 by OECD member countries when they approved the **OECD Council Recommendation on Artificial Intelligence**. The OECD AI Principles are the first such principles signed up to by governments. Beyond OECD members, other countries including Argentina, Brazil, Colombia, Costa Rica, Peru and Romania have already adhered to the AI Principles, with further adherents welcomed.

The OECD AI Principles set standards for AI that are practical and flexible enough to stand the test of time in a rapidly evolving field. They complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct.

In June 2019, the **G20 adopted human-centred AI Principles** that draw from the OECD AI Principles.

The OECD AI Principles

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

What can governments do?

Consistent with these value-based principles, the OECD also provides five recommendations to governments:

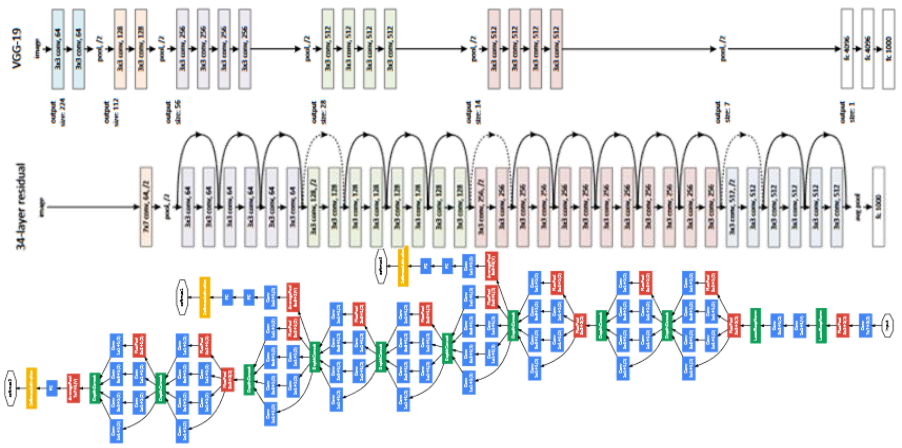
- Facilitate public and private investment in research & development to spur innovation in trustworthy AI.
- Foster accessible AI ecosystems with digital infrastructure and technologies and mechanisms to share data and knowledge.
- Ensure a policy environment that will open the way to deployment of trustworthy AI systems.
- Empower people with the skills for AI and support workers for a fair transition.
- Co-operate across borders and sectors to progress on responsible stewardship of trustworthy AI.

And I'll say that most of what I'm going to discuss is reflected already in the OECD principles on artificial intelligence.

1. Explainability

So the first big problem is explainability. We'd like to understand *why* the model made a particular decision. Why did it recommend someone to get a loan or go to jail? Can it tell us what it was thinking? Or can it at least give us a good reason that the decision it made is the right one?

1. Explainability



So the problem with explainability, as you can already imagine, is that these models are wildly complicated. In a model with 25 million connections, or even 1.5 billion, it's very hard to understand what any particular connection is doing or how it contributes to the overall decision. And it's unclear what logic the model is using to process the data. The logic may not be easy for a human to understand, and it could rely on things we can't even see or conceptualize as a human.

Now, I do want to say that neural networks aren't necessarily "black boxes." They're just very very complicated, and we don't yet know how to make it easy to understand that complexity.

1. Explainability

Simoyan et al., 2013.

Now, we're making progress on this question.

One popular way to try to explain a model is to try to show what parts of the input had an influence on the output. But this is still more art than science at the moment, and I don't think we really have robust explainability solutions that work in general.

There's

1. Explainability

Tradeoff between completeness and comprehensibility

Now there's one last thing I want to say about explainability. It's that there's a tradeoff between what we call "completeness" and "comprehensibility." I can tell you every mathematical detail of the model. And that would be a complete explanation. But it's not very comprehensible. Or I can give you a simplified picture of what the model might have done, but it will necessarily leave out a lot of detail.

One thing we can do to reduce this tradeoff is to use simpler models. Maybe, instead of 25 million connections, we use something simple that only has ten or twenty. The tradeoff here is that

This is a hard problem, and we don't have clear answers for how to address it yet.

1. Explainability

Tradeoff between simplicity and expressive power.

One thing we can do to reduce this tradeoff is to use simpler models. Maybe, instead of 25 million connections, we use something simple that only has ten or twenty. The tradeoff here is that, by using a simpler model that might be easier to completely understand, we lose some expressive power. We can't model more complicated tasks as effectively, and our model won't be as accurate.

Explainability is a hard problem, and we don't have clear answers for how to address it yet. If we want to demand explainability as a policy choice, it is likely that we won't be able to use the most advanced neural network models.

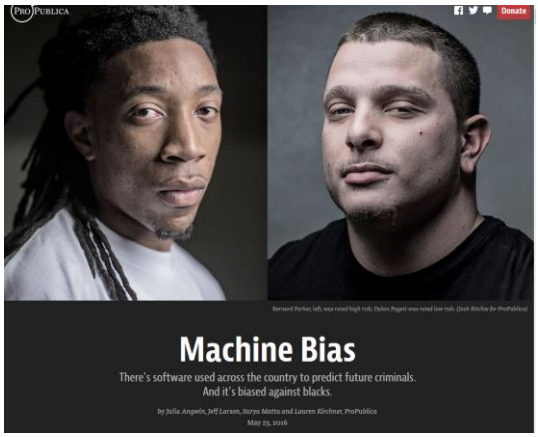
2. Bias and Fairness

Now let's talk about another problem. Bias and fairness. I think we're probably pretty familiar with this problem at this point. It has received a lot of attention lately.

2. Bias and Fairness

I showed you an example before, when Google's image recognition algorithm called black people gorillas.

2. Bias and Fairness



But there are also examples of criminal risk scoring algorithms that were found to be biased in the united states.

2. Bias and Fairness

Bias: the model performs differently on different groups.

Now I usually define bias as when the model performs differently on different groups.

2. Bias and Fairness

Bias: the model performs differently on different groups.

Fairness: frameworks for measuring bias.

Now, where does bias come from? Remember before, we had that four step machine learning process.

2. Bias and Fairness

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation
4. Deployment

Bias can enter the system in a number of places.

2. Bias and Fairness

(Decide on the problem)

1. Data collection

2. Training

3. Evaluation

4. Deployment

In data collection, we might collect data that is unrepresentative. Google studied this problem, and here is what they found.

2. Bias and Fairness

(Decide on the problem)

1. Data collection

Here are three images of weddings from a dataset. These are all western-style weddings.

2. Bias and Fairness

(Decide on the problem)

1. Data collection

This wedding on the right may not be a part of the dataset. Our dataset may not be representative. Now, there's another problem. Remember, we don't just need to collect the data, we also need to label it.

2. Bias and Fairness

(Decide on the problem)

1. Data collection

And our labels can reflect our biases. The western style weddings were labeled as “wedding,” “bride,” “groom,” etc., and the non-western wedding was labeled as “person,” and “people.”

2. Bias and Fairness

(Decide on the problem)

1. Data collection

2. Training

3. Evaluation

4. Deployment

Now it can also slip in during training. When we train a model, we're trying to choose the model that gives us the best performance. But what do we mean by performance?

2. Bias and Fairness

(Decide on the problem)

1. Data collection

2. Training

3. Evaluation

4. Deployment

Accuracy?

Performance could mean Accuracy.

2. Bias and Fairness

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation
4. Deployment

Accuracy?

Fairness?

But maybe we should pick the model that is the fairest.

2. Bias and Fairness

(Decide on the problem)

1. Data collection
2. Training
3. Evaluation
4. Deployment

Accuracy?

Fairness?

A mix?

Or maybe some mix of the two. The point is that we can train a model to meet whatever goals we want, and accuracy is only one goal. So the training process can introduce (or be used to improve) fairness.

2. Bias and Fairness

(Decide on the problem)

1. Data collection

2. Training

3. Evaluation

4. Deployment

And finally, we need to ask how we're evaluating the model. Again, our evaluation dataset could be biased. But we need to imagine ways that bias could show up and try to look for them. If we aren't a very diverse group, we're likely to miss important sources of bias that will show up during deployment.

3. Robustness

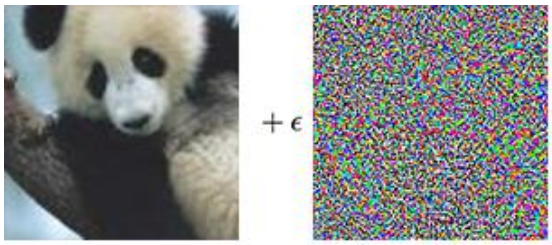
Let’s imagine I give your model a Panda.

3. Robustness

"panda"
57.7% confidence

And your model is pretty good. It thinks it's a Panda with 57.7% confidence.

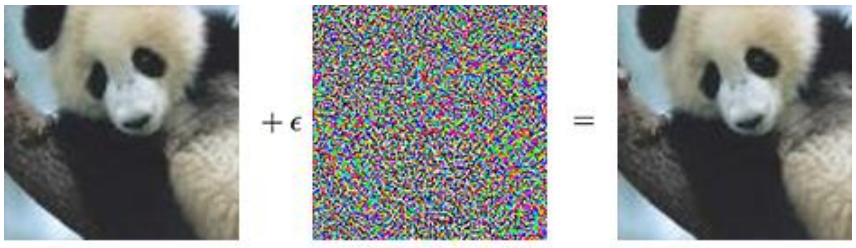
3. Robustness



"panda"
57.7% confidence

Now what if I change the image a little bit by adding some carefully-chosen static.

3. Robustness



"panda"
57.7% confidence

The image still looks like a Panda to you and me.

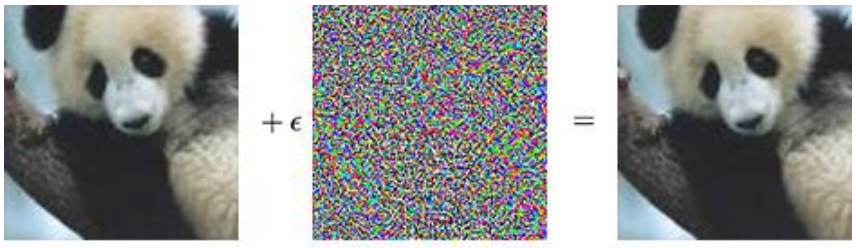
3. Robustness

"panda"
57.7% confidence

"gibbon"
99.3% confidence

But the model thinks it's a gibbon with 99% confidence. This is known as an adversarial example.

3. Robustness



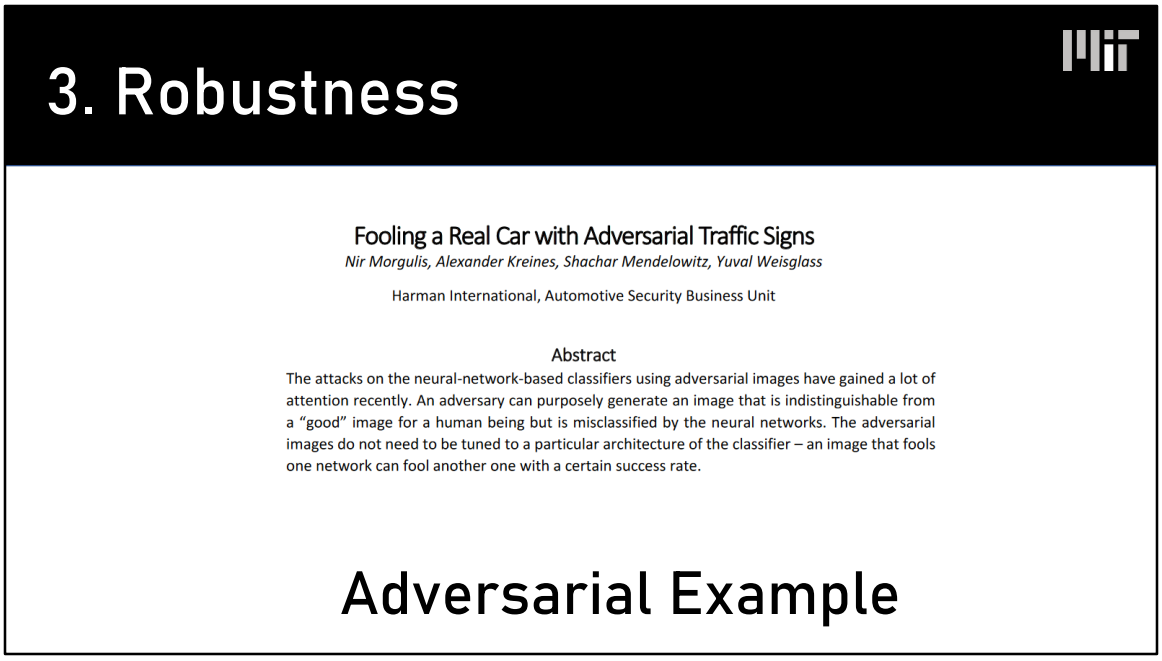
"panda"
57.7% confidence

"gibbon"
99.3% confidence

Adversarial Example

And neural networks are generally vulnerable. If I have access to your model or something similar to it, I can easily create these examples that look normal to a human but completely change the model.

3. Robustness



Fooling a Real Car with Adversarial Traffic Signs

Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, Yuval Weisglass

Harman International, Automotive Security Business Unit

Abstract

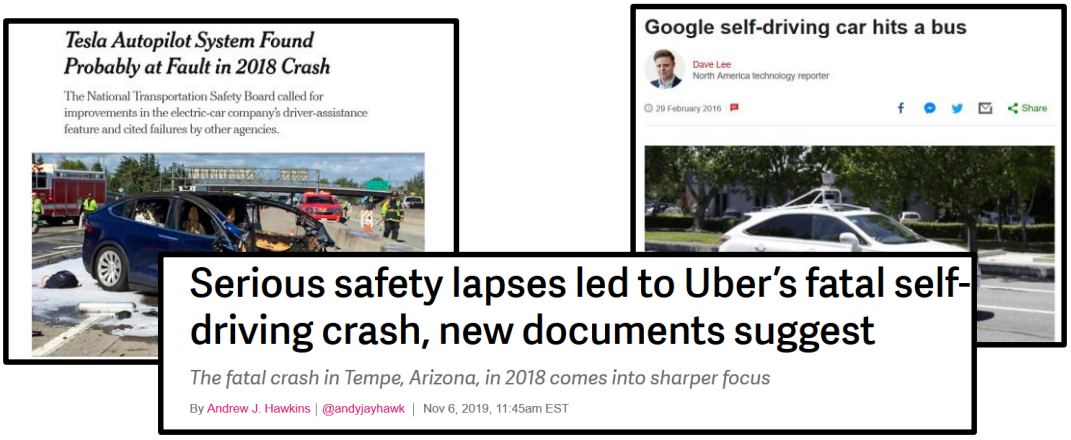
The attacks on the neural-network-based classifiers using adversarial images have gained a lot of attention recently. An adversary can purposely generate an image that is indistinguishable from a "good" image for a human being but is misclassified by the neural networks. The adversarial images do not need to be tuned to a particular architecture of the classifier – an image that fools one network can fool another one with a certain success rate.

Adversarial Example

You can use this to try to fool a self-driving car and get it to make a mistake or an unsafe decision.

Now we've been talking about adversarial examples, where someone has deliberately tried to fool a model. Now, that I have your attention with adversarial examples, I want to remind you that these machine learning models often fail even in seemingly normal circumstances.

3. Robustness



Mistakes happen. Accidents happen. We don't expect AI to be perfect. Humans make mistakes, and tens of thousands of people die in car accidents every year. But making models robust to new circumstances or unexpected inputs is important even when we don't have an adversary. How will a self-driving car trained in California do in the snow, for example?

4. Privacy

Now the last thing I want to mention is privacy. I feel like privacy often gets left out of conversations about AI. We're so concerned about fairness or bias or explainability that we forget about this part. But AI models are trained on millions or billions of examples, and this data has to come from somewhere. This means that a lot of people's data is getting swept up into these models. And that's necessary. If we want to do an algorithm for healthcare, we're going to need people's medical information. If we want to do facial recognition, we need pictures of people's faces. So privacy is a huge challenge in AI. I think it often gets forgotten, though.

I looked through many principles frameworks from countries, organizations, and companies, and few mentioned privacy. Those that did often put it last. So I implore you not to forget it.

What Can Go Wrong

1. Explainability
2. Bias and Fairness
3. Robustness
4. Privacy

Just to summarize, I've discussed four things that can go wrong: explainability, bias and fairness, robustness, and privacy. These aren't the only things that can go wrong, to be certain, but they're some of the most prominent ones as we understand it today.

Policy Suggestions

Now, as we wrap up, I want to leave you with three concrete policy suggestions that I hope you'll take home as you think about these issues.

Policy Suggestions

1. Proportionality

There are three of them. First, remember to be proportional in your demands for explainability, fairness, robustness, and privacy. There are cases where these are overriding concerns, and cases where they're less important and will get in the way of innovation. We have to balance these situations.

Policy Suggestions

1. Proportionality
2. Address the workflow, not the specific technology.

Second, this is something that I mentioned much earlier and promised I would come back to. I think it's silly to try to make policy about AI. I think we should instead address workflows or situations, not the system itself. As an example, I think we should create policy for "automated decisionmaking" or "credit scoring" rather than for "AI" – we should be clear about the goals we have for any system used for these workflows, whether human, computer program, or AI.

Policy Suggestions

1. Proportionality
2. Address the workflow, not the specific technology.
3. Look to the future.

And the last thing I'll say is to keep your eyes on the future. It's easy to get caught up in the AI that's already here. But innovation is happening rapidly, and I think existing policy frameworks are short-sighted. I'll explain more about that in a moment.

1. Proportionality

Let's start with proportionality. To discuss it, I want to bring up an example use of AI: facial recognition.

1. Proportionality

Facial recognition involves using a picture of someone's face.

1. Proportionality

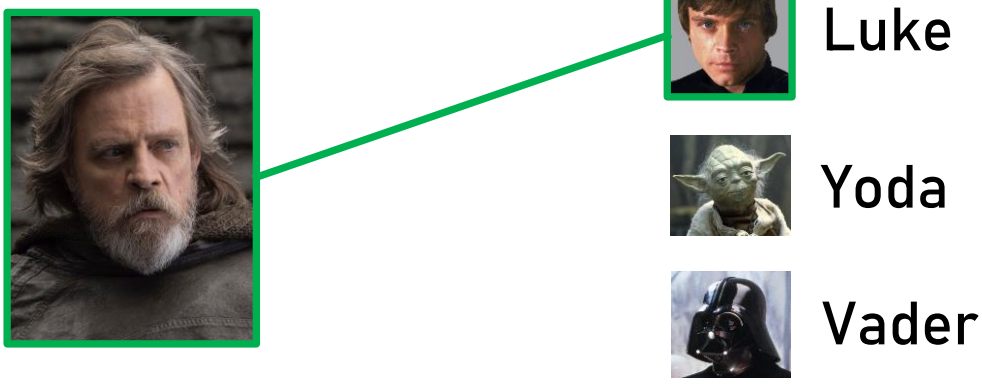
Luke

Yoda

Vader

And trying to match it to one of several other faces whose identities you know.

1. Proportionality



If we're successful, we match it to the right person.

Now we have excellent facial recognition algorithms. The technology tends to work very well in practice. Well enough that it has begun to pose privacy and civil liberty concerns

1. Proportionality

The New York Times

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

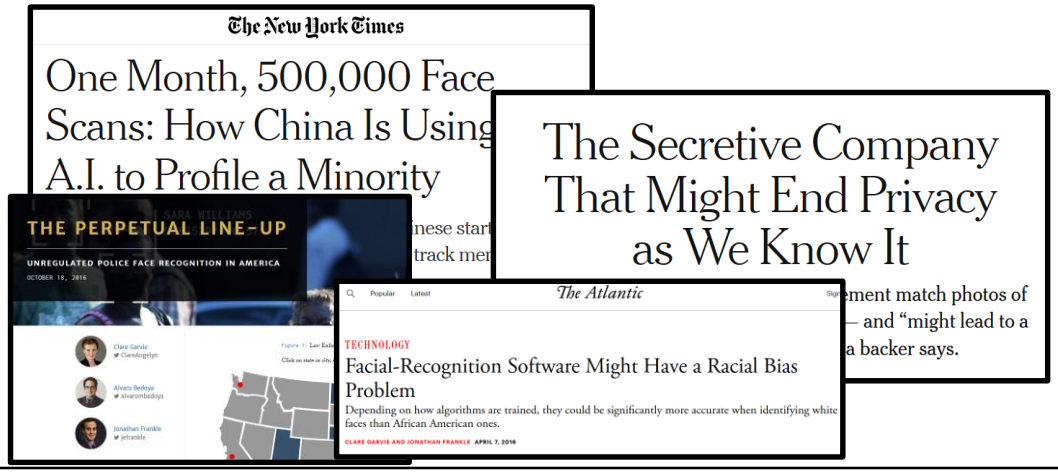
In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.

It's a powerful surveillance tool.

1. Proportionality

It's a powerful surveillance tool.

1. Proportionality



And it's something that I've personally been very critical of. Here's why: It's not very explainable.

1. Proportionality

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

It raises privacy concerns

1. Proportionality

And it has problems with bias.

1. Proportionality

What Happens When Employers Can Read Your Facial Expressions?

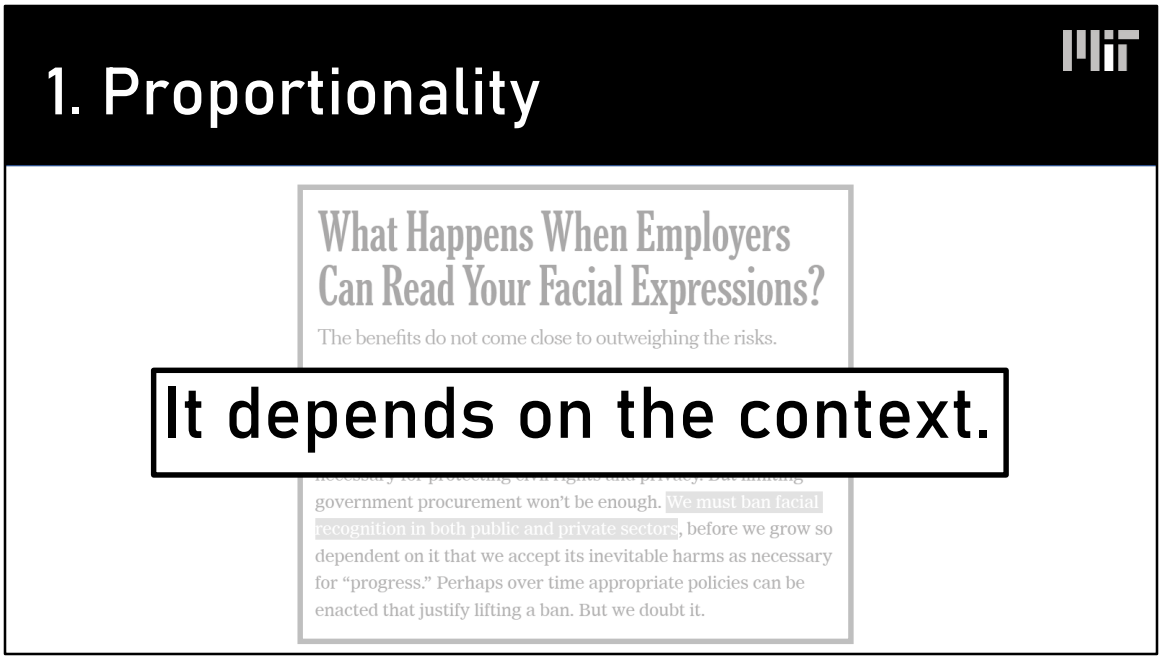
The benefits do not come close to outweighing the risks.

By Evan Selinger and Woodrow Hartzog

We think the senator is right: Stopping this technology from being procured — and its attendant databases from being created — is necessary for protecting civil rights and privacy. But limiting government procurement won't be enough. We must ban facial recognition in both public and private sectors, before we grow so dependent on it that we accept its inevitable harms as necessary for "progress." Perhaps over time appropriate policies can be enacted that justify lifting a ban. But we doubt it.

So should we just ban it?

1. Proportionality



And my answer is: It depends on the context.

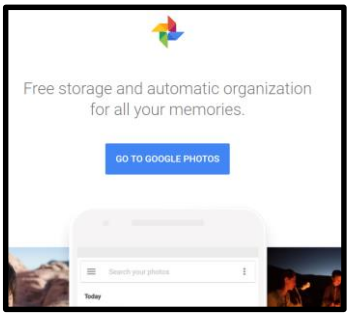
1. Proportionality

Let’s imagine two different contexts.

1. Proportionality

The first context is facial recognition for surveillance. Whenever we are searching for a criminal, we turn on facial recognition and try to find a match with the criminal. If we detect them, we arrest them.

1. Proportionality



Our other context is Google Photos. Google Photos also has facial recognition. You upload your photos, and it will try to identify the face of the same person in many photos. If you label that photo as “grandma,” then it will label all of the photos that match as grandma as well. It’s a convenient way to tag people in lots of photos.

1. Proportionality

1. Explainability
2. Bias and Fairness
3. Robustness
4. Privacy

Now let's go down our list of risks.

1. Proportionality

1. Explainability
2. Bias and Fairness
3. Robustness
4. Privacy

When it comes to surveillance, we should demand high standards of explainability. If you're going to arrest someone – to deprive them of their freedom, you better have a good explanation for why the photo is a match. When it comes to google photos, the risk of a mismatch is small. You might accidentally label aunt jenny as grandma, but the consequences are small.

1. Proportionality

1. Explainability
2. Bias and Fairness
3. Robustness
4. Privacy

When it comes to bias and fairness, we should have high standards for surveillance. Surveillance already has risks of bias, and it's important that we exercise policing powers fairly. Again, the risk of a mistake is high. In Google photos, bias and unfairness is certainly a concern to dignity, and it may disproportionately affect the experience of certain users, but it doesn't put anyone's freedom at risk.

1. Proportionality

1. Explainability
2. Bias and Fairness
3. Robustness
4. Privacy

The privacy risks are also very different. One system can identify anyone without their knowing. The other can only identify the photos you give it.

1. Proportionality

1. [illegible]
2. [illegible]
3. [illegible]
4. Privacy

Our demands should be proportional to the risk that an application poses.

And so I advocate for proportionality. We don't need to demand rigorous explainability everywhere. We should demand explainability in proportion to the risk associated with the particular application

2. Address Workflows

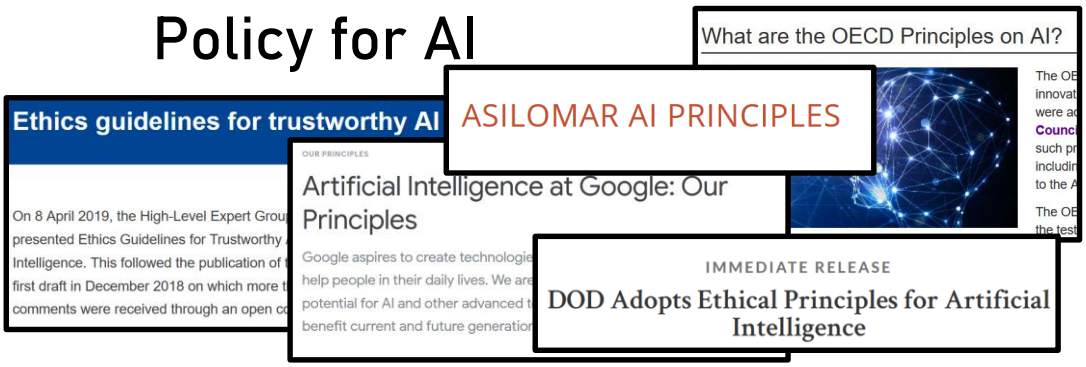
My second recommendation is that we should address workflows rather than specific technologies. Let me give you an example.

2. Address Workflows

Policy for AI

One way to handle AI policy is to try to create a blanket policy for AI. This is the approach that most organizations have taken.

2. Address Workflows



2. Address Workflows

Policy for workflows

My proposed alternative is to create policy for workflows.

2. Address Workflows

Policy for workflows:

- Automated decisionmaking.
- Content generation.
- Planning (robotics).

Examples of workflows would be things like “automated decisoinmaking,” “content generation,” and “planning.”

2. Address Workflows

Policy for workflows:

- Automated decisionmaking.
- Content generation.
- Planning (robotics).

I would even argue that all of the AI policy frameworks so far have focused on automated decisionmaking, not on AI more broadly, despite their names. Concepts like fairness, bias, and explainability make much more sense in the context of automated decisionmaking than they do in the context of content generation.

2. Address Workflows

Policy for workflows:

- Credit scoring
- Hiring
- Criminal justice

I would even argue that we should break this down even further. The criteria that will be important to us are different in each of these areas. And the policies we create won't be specific to AI.

2. Address Workflows

Policy for workflows:

- Credit scoring
- Hiring
- Criminal justice

Not all policy must be AI specific

And the policies we create don't need to be specific to AI. In many cases, AI is just a natural evolution of existing automated processes. And in those settings where it isn't, we should create policy specific to the workflow, not to Ai in general.

2. Address Workflows

Policy for workflows:

- Automated decisionmaking.
- Content generation.
- Planning (robotics).

Now remember those three areas I mentioned before?

2. Address Workflows

Policy for workflows:

- Automated decisionmaking.
- Content generation.
- Planning (robotics).

The other two are coming. Content generation might include deepfakes. Planning and robotics includes autonomous vehicles. I don't think our existing AI policy frameworks were designed with these workflows in mind.

3. Look to the Future

Policy for workflows:

- Automated decisionmaking.
- Content generation.
- Planning (robotics).

So we should look to the future. Which AI applications are coming and how can we address them proactively. My believe is that, with the existing OECD principles, we have only taken the first step toward building the policy frameworks necessary to address AI.

Policy Suggestions

1. Proportionality
2. Address the workflow, not the specific technology.
3. Look to the future.

And the last thing I'll say is to keep your eyes on the future. It's easy to get caught up in the AI that's already here. But innovation is happening rapidly, and I think existing policy frameworks are short-sighted. I'll explain more about that in a moment.

Summary

What is artificial intelligence?

How are AI systems built?

What should we worry about?

Our goals were

Summary

1. Definitions
2. How to build a model
3. Deep learning
4. How AI systems fail
5. What we should do about it (policy)

We discussed…

Artificial Intelligence for Policymakers

Jonathan Frankle
jfrankle@mit.edu