VIJAY JANAPA REDDI: Welcome to deploying TinyML, Course 3 of the TinyML edX series.
I'm excited to see you back here and I hope
you had a really good time in Course 2.
As always, I'd like to start off by reminding us
why we are all excited about TinyML.
TinyML is one of the fastest growing fields of machine learning, which
brings together algorithms, hardware, and software so
that we can do on device sensor analytics
right where the data resides in real time.
If we can do this at ultra low power consumption using
small little microcontrollers, then we unlock
the ability to run machine learning models at all times.
And if so, we can run them on battery powered devices,
and this would unlock a whole new range of applications which I've been psyched about in both Course 1 in Course 2.
And of Course 3, we're going to touch on a whole new range
of different applications.
Now that said, let's take a quick recap on where we came from
and where we're going.
In Course 1, we focused on the language of machine learning.
We exposed you to the fundamental building blocks of machine learning.
We taught you how to program in TensorFlow.
We taught you about Colab and how you can use Colab as a programming environment free of your own resources.
In doing so we expose you to things like stochastic gradient descent,
we expose you to loss, we expose you to cost functions,
and so forth, all the building blocks you
need to have in order to build machine learning models.
Then in Course 2, we got excited about different kinds
of applications of TinyML.
And then we use the knowledge that we got from Course 1 in Course two,
specifically we learned how to take the models that we trained in TensorFlow
and shrink them down using the TF Lite converter so that we
can have quantized models that are more easier
to run on small embedded devices.
In doing that, we also learned the art of training
neural networks faster for instance using things like transfer learning.
We did all of this in the context of three different applications ranging
from keyword spotting, to visual wake words,
to anomaly detection, which was all about unsupervised learning.
So if you take a step back and you kind of
reflect on what you have covered just between Course 1 and Course 2,
it's a wealth of knowledge surrounding machine learning.
Now finally in Course 3, this is where the pedal meets the metal my friend.
This is where we're going to take those small compressed
models that we had trained and learn how to deploy them onto a microcontroller.
One of the key things that you will learn here
is that simply having a small little neural network
is not enough in terms of deploying it.
You actually have to understand the art of packaging up
the entire application as a whole binary and being able to push that down.
And when you're running it, you need to really think
about what it means to invoke a model.
So we'll talk about things like pre processing, post processing,
and of course we'll also learn how to write code in TensorFlow Lite
Micro, which is specifically designed for microcontrollers
so that you can actually take a model, a TF Lite model,
and then be able to run it effectively on a microcontroller.
So one of the core things that we really want
to get through by the end of this course is teach you
how to write an end to end ML application.
As I've said before in Course 2, I do not
care if you know how to write a model.

A model alone is a part of the bigger picture.
It really is about orchestrating the entire flow, which
includes the neural network computation as well as all the non neural network computation.
That is what a real application is, and that
is really going to be the big takeaway that I'm going to push hard with you in this particular course.
And that my friend, is exactly why when you finish this course,
you can call yourself a TinyML engineer.
Because you will actually have the hands on knowledge
of working with the physical microcontroller
and learning how to program it with a production grade inference framework, which is TensorFlow Lite Micro.
It has got new APIs, and it's about the state of the art inference
engine that is out there.
So you will be way ahead of the curve.
So stick with me and let's get through Course 3.