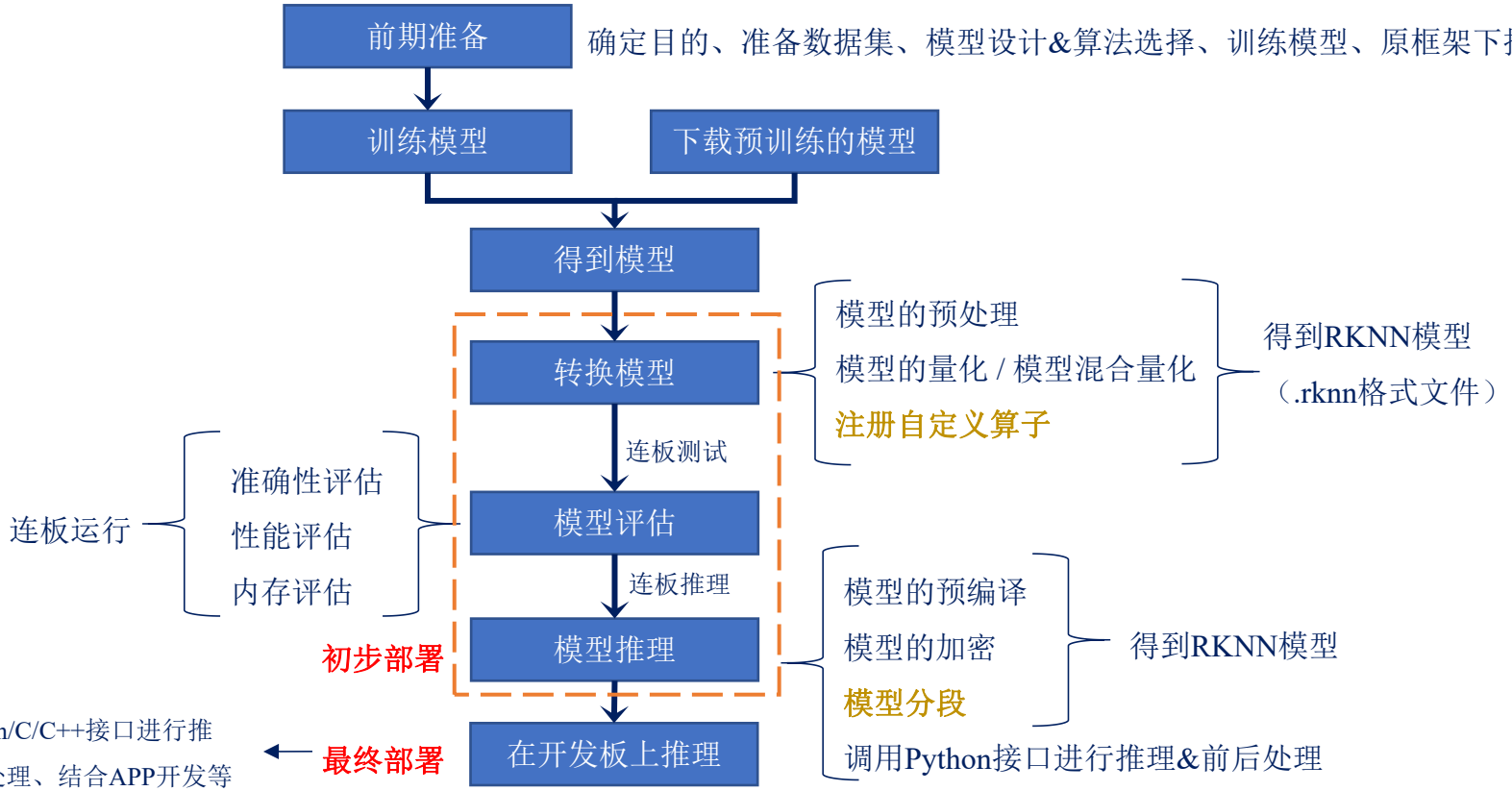
主讲人 ：正点原子团队
硬件平台：正点原子ATK-DLRV1126开发板

资料下载：www.openedv.com/docs/index.html
教学平台：www.yuanzige.com
天猫店铺：zhengdianyuanzi.tmall.com
技术论坛：www.openedv.com/forum.php
公众平台：正点原子

■ AI模型部署流程

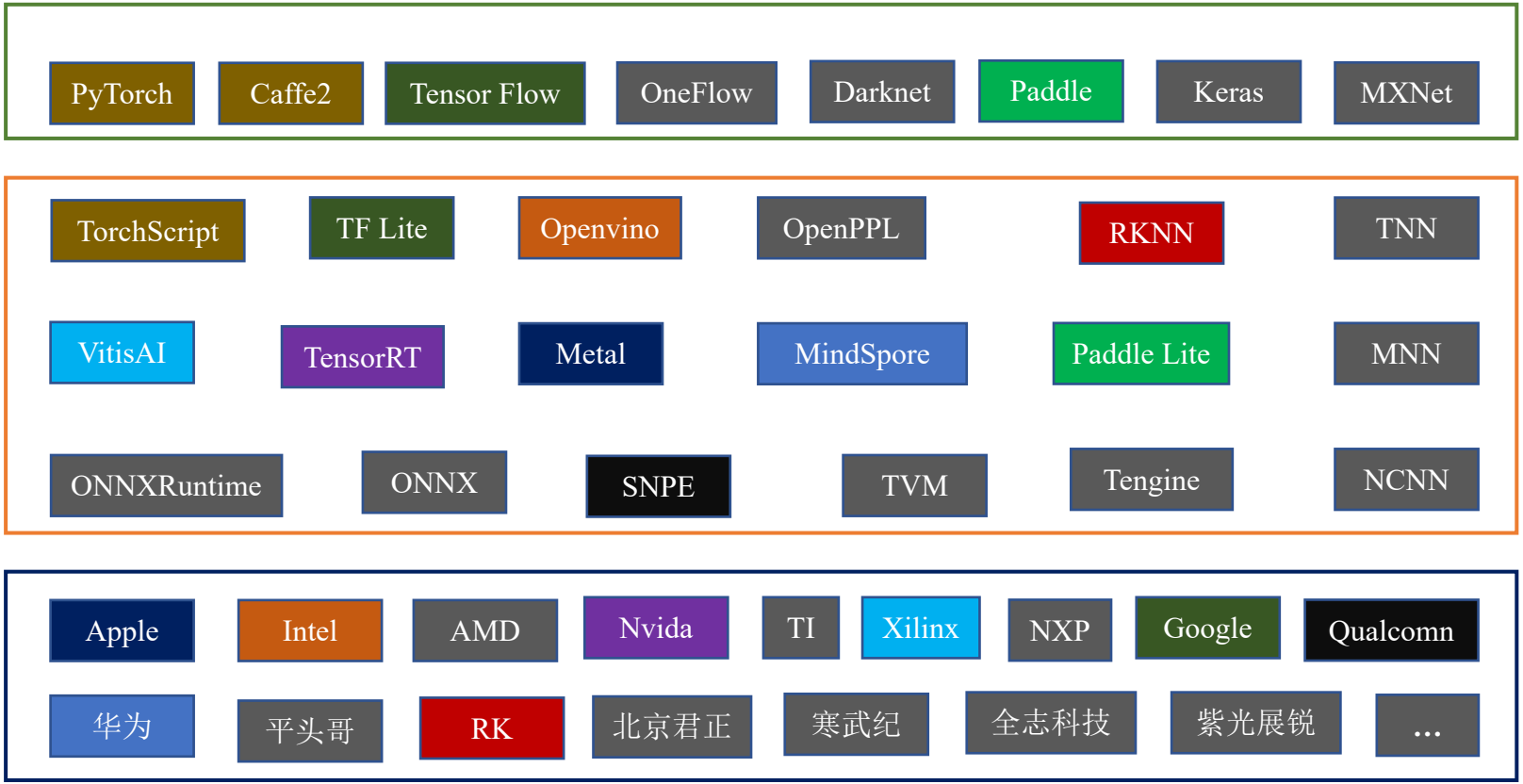
模型的部署，是指将训练好的模型放到运行环境中进行推理的过程。



模型的部署，是指将训练好的模型放到运行环境中进行推理的过程。一般需要经过以下操作：

- 将训练模型转换推理模型→需要转换框架的转换器/转换工具，转换为符合硬件要求的数据结构，模型的预处理等
- 部署阶段的一些性能优化→如算子融合、算子替换、自定义算子、模型的预编译（加快加载模型的时间）等
- 模型压缩→量化，以减少模型精度，剪枝神经网络的稀疏，知识蒸馏等
- 安全保护→模型的加密
- 模型推理→涉及前处理、执行推理、后处理

AI模型训练框架、部署/推理框架、底层硬件



天猫店铺：https://zhengdianyuanzi.tmall.com