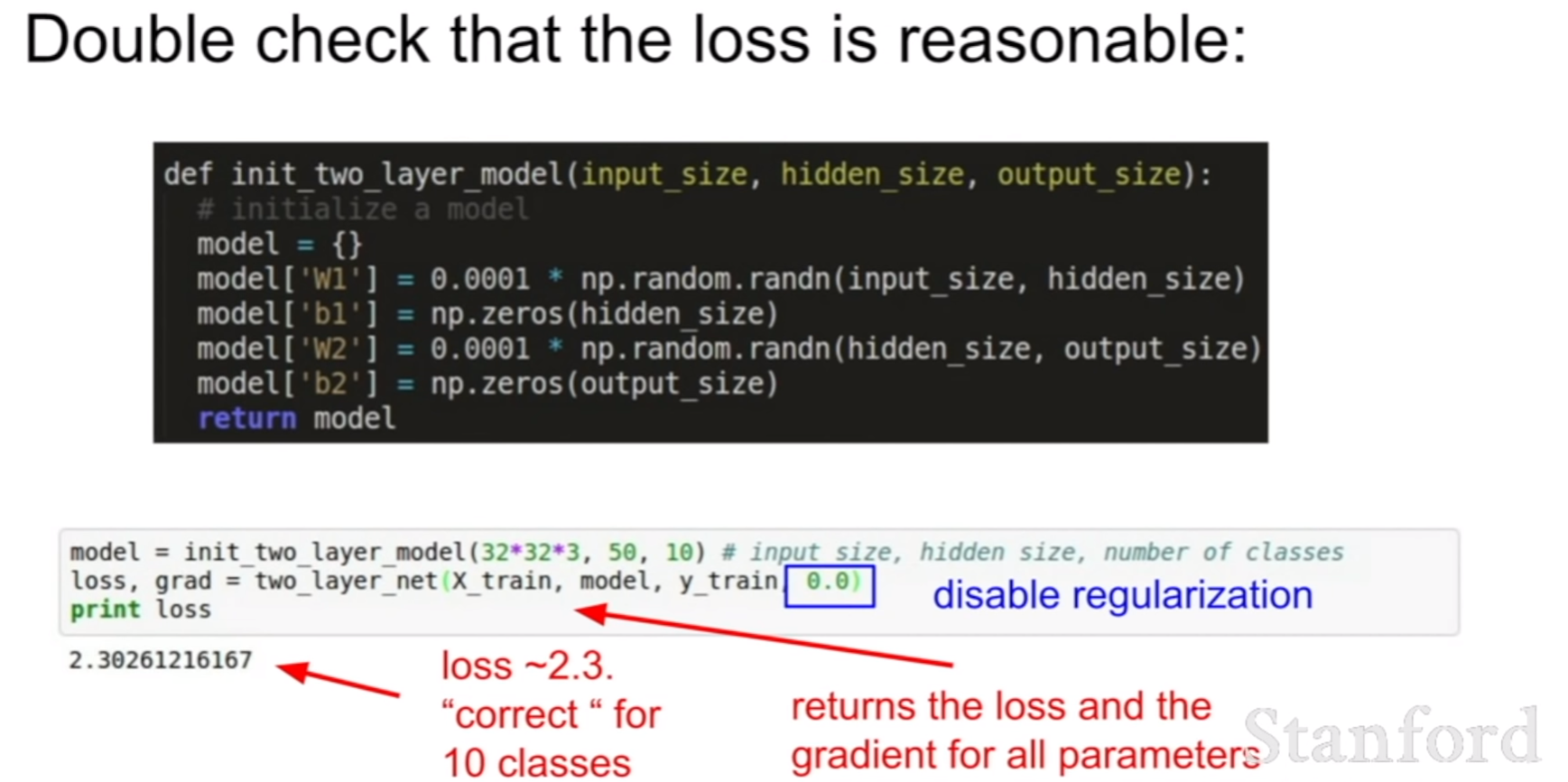
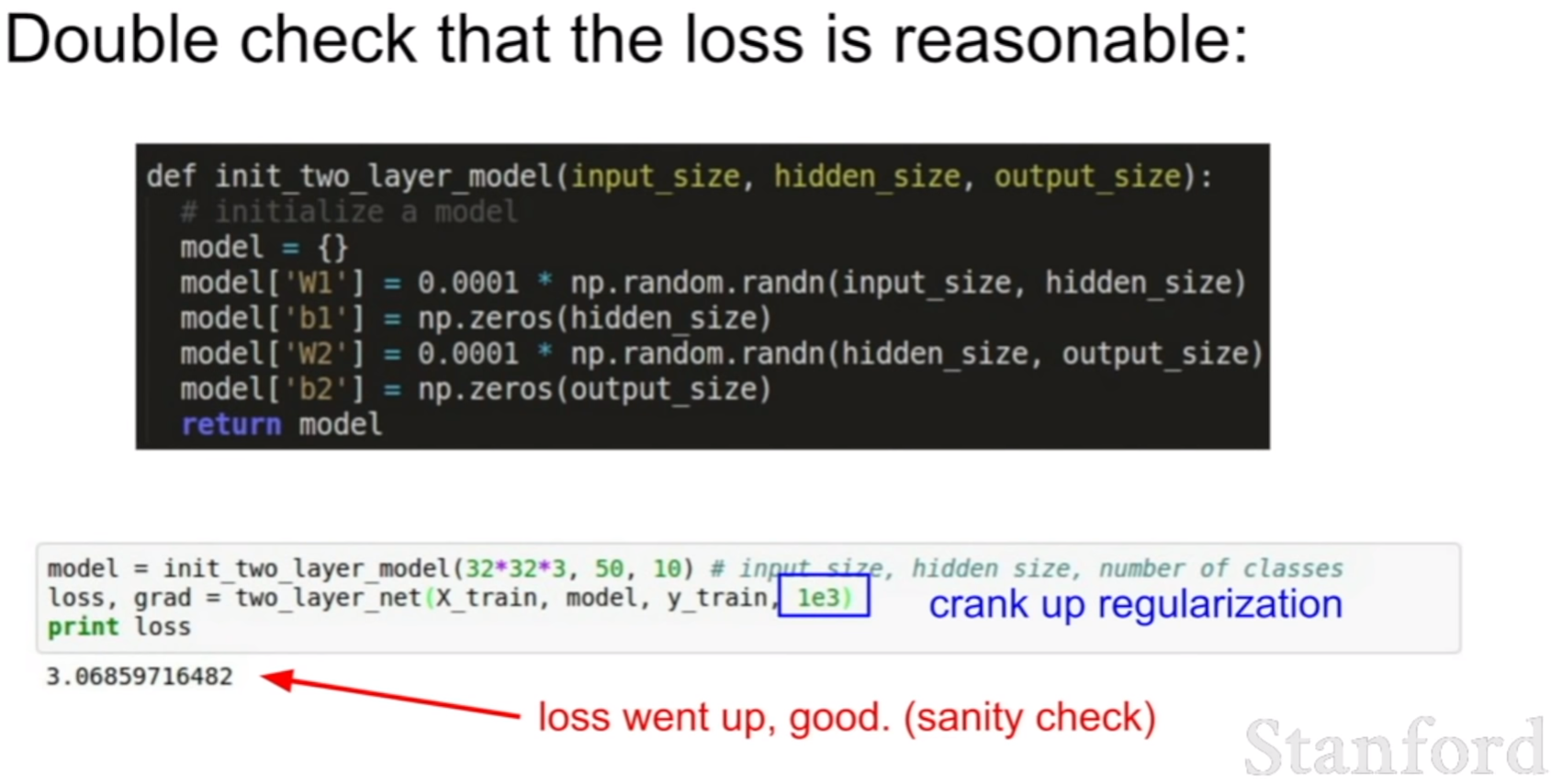
步骤：

预处理数据，如去中心化，PCA等；选择网络结构

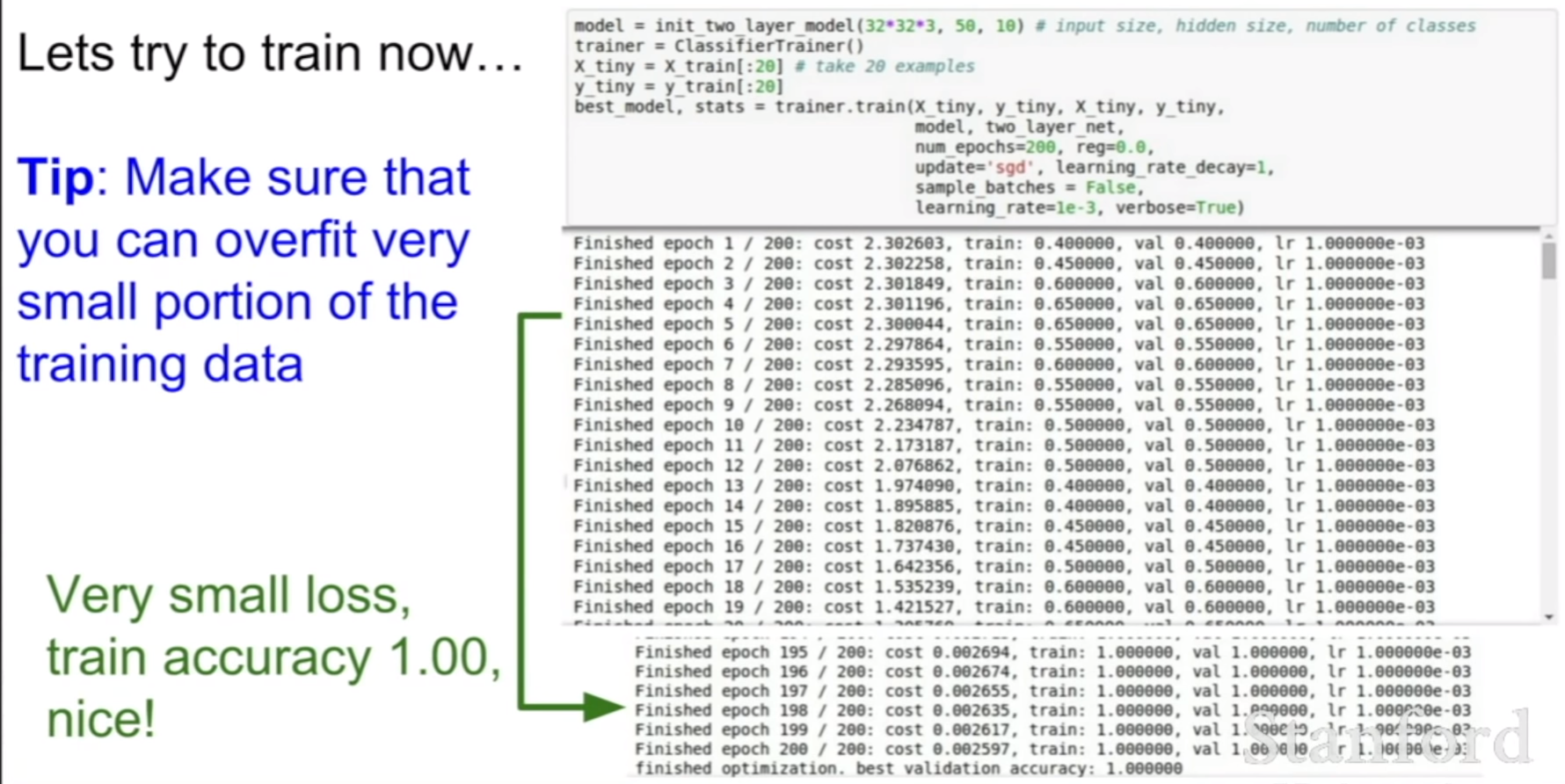
1. 检测第一次的loss是否合理（先把正则化关掉）。



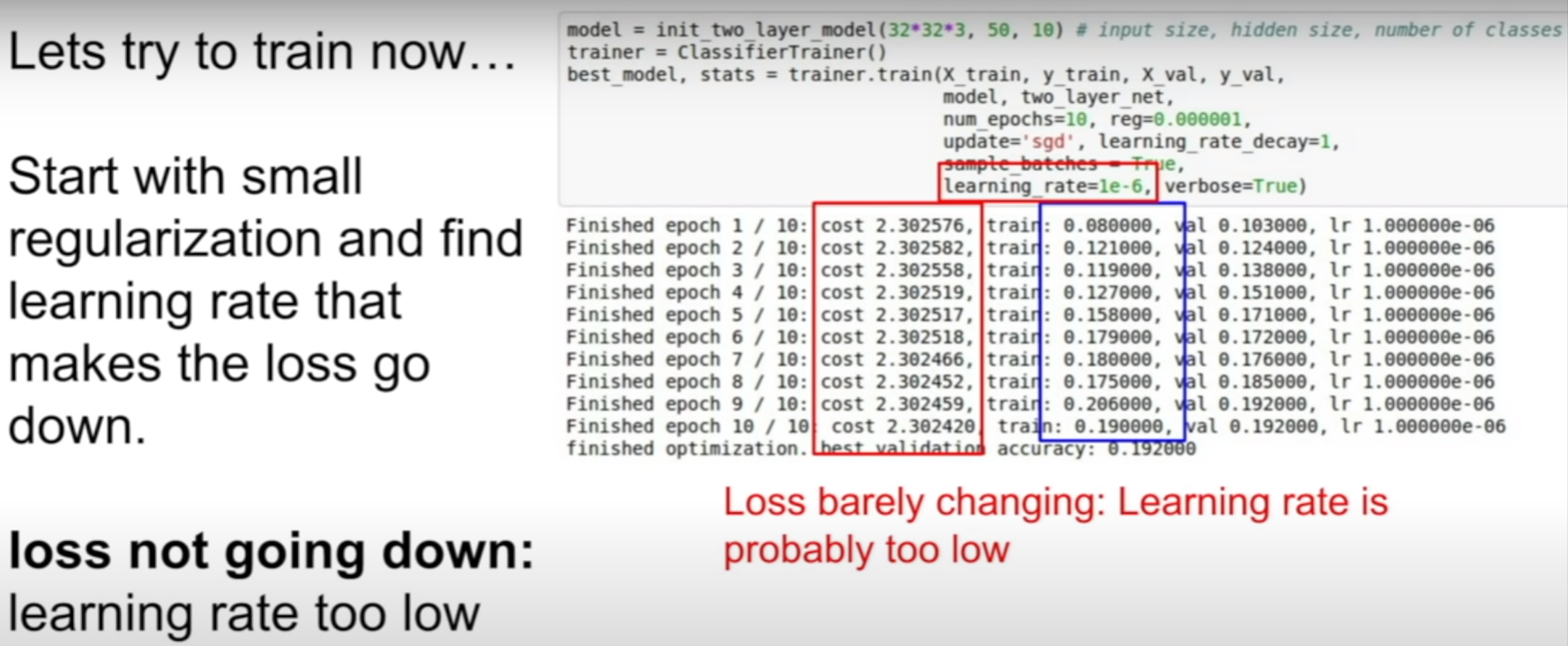
1. 打开正则化，观察loss是否增大。



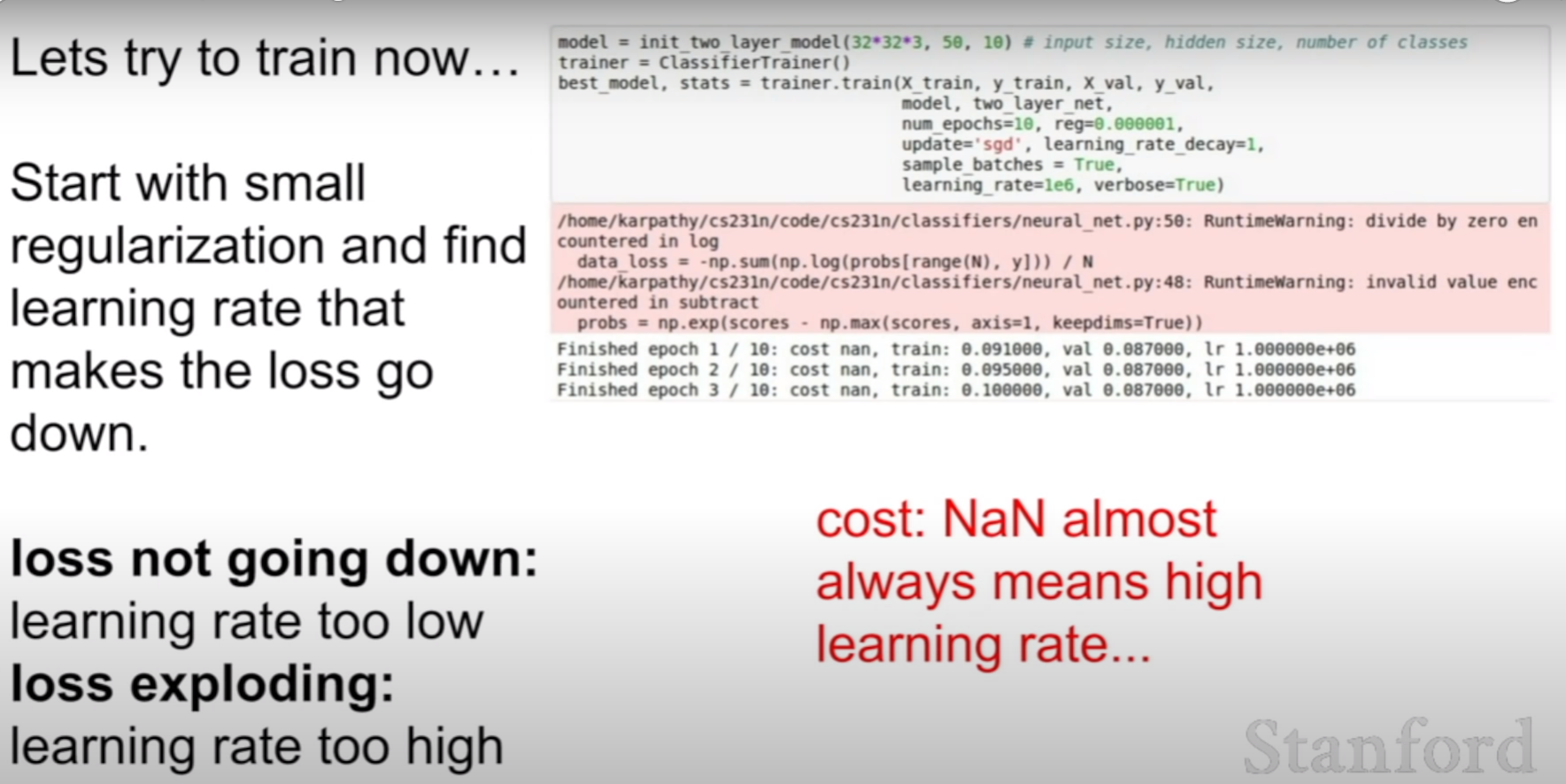
1. 用一小部分数据进行训练使得模型得到预期的过拟合。Loss很小，精度很高。



1. 采用全部的数据进行训练。先采用较小的正则化值，找出哪个学习率使得loss下降。学习率应该是第一个调整并确定的超参数。

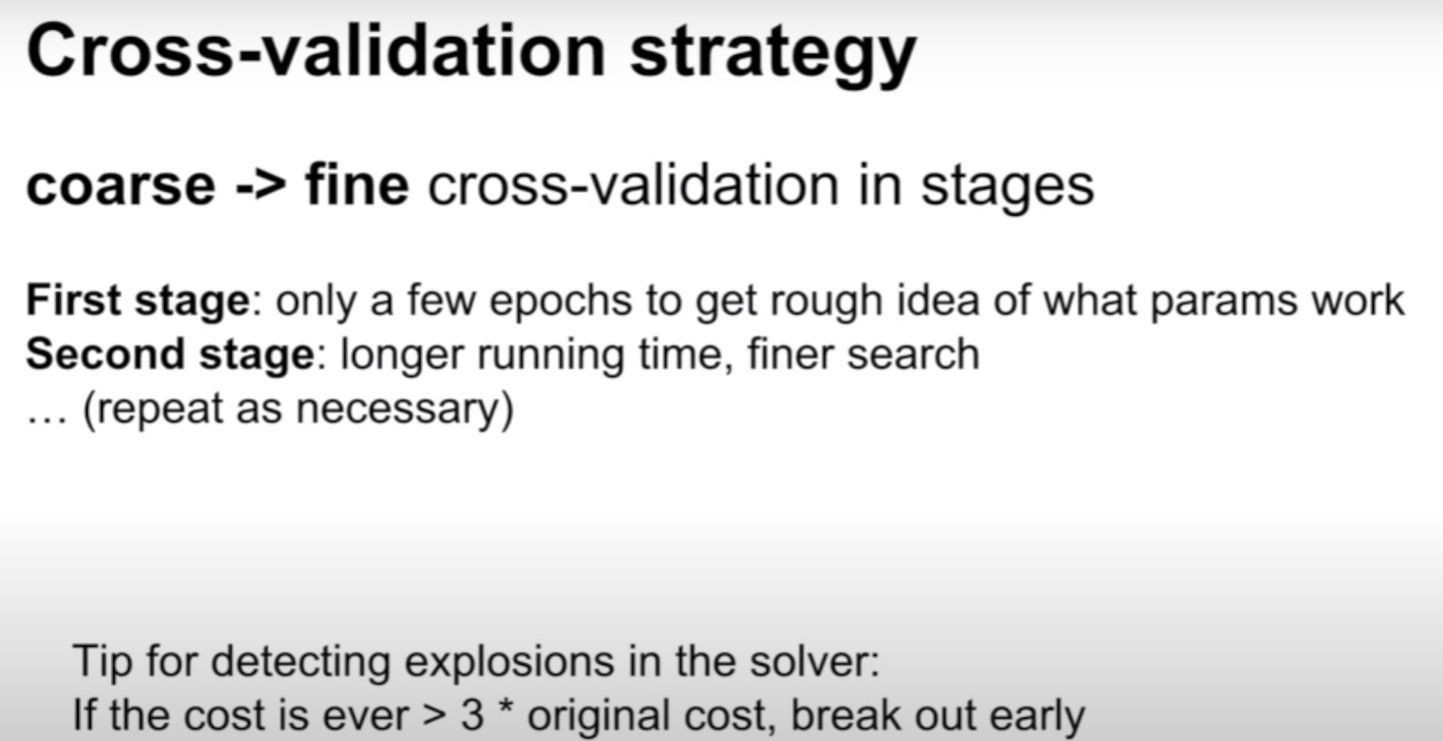


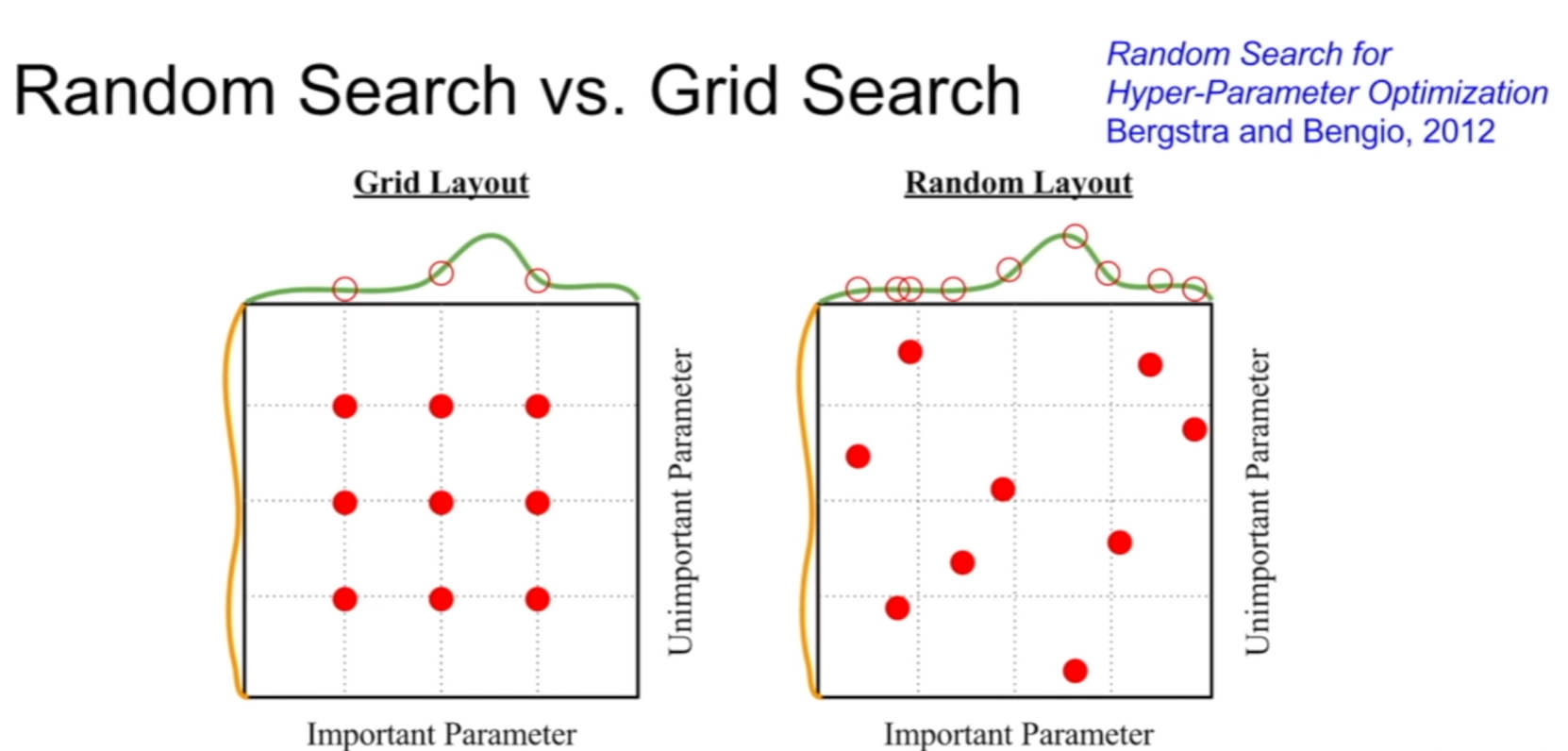
从上面可以观察到，因为学习率太小，所以loss没怎样优化到。但是我们却可以看到精度反而上升了20%，为什么？因为此时，概率分布仍然很分散，所以损失项仍然很相似，但是当我们在正确的方向上稍微shift所有这些概率（即在正常的方向上进行梯度下降修改权重），从而导致精度突然上升。

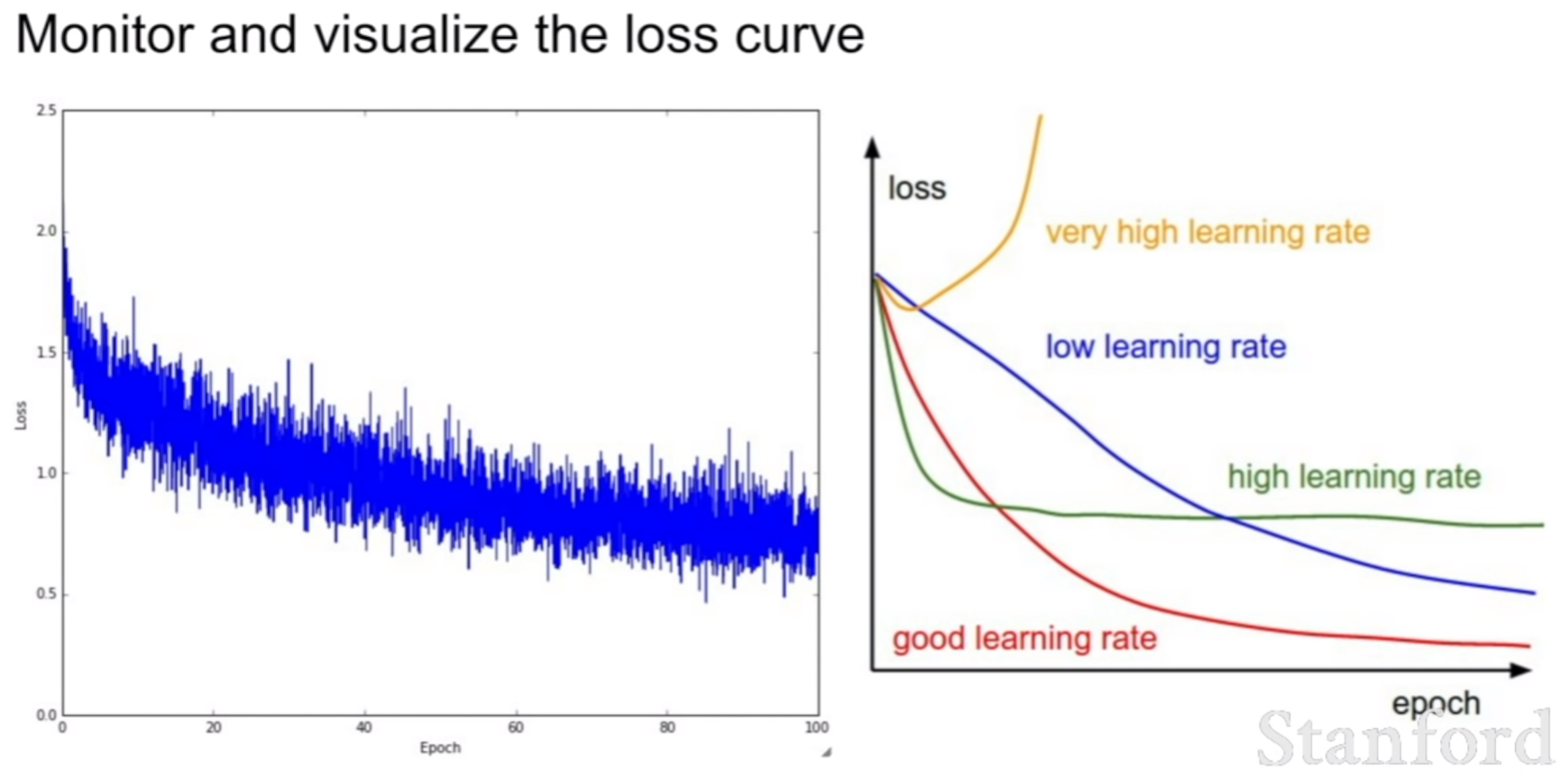


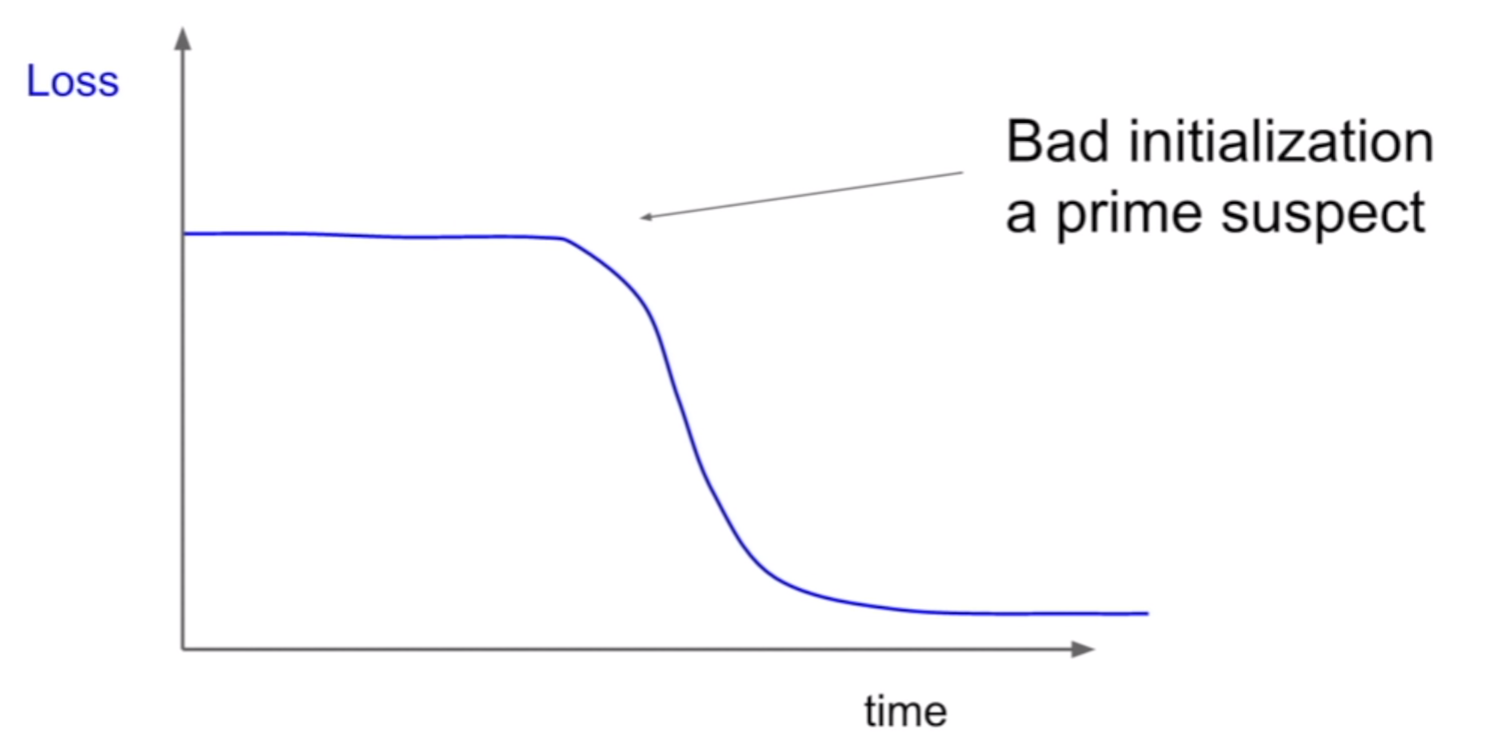
学习率太高，导致损失爆炸。

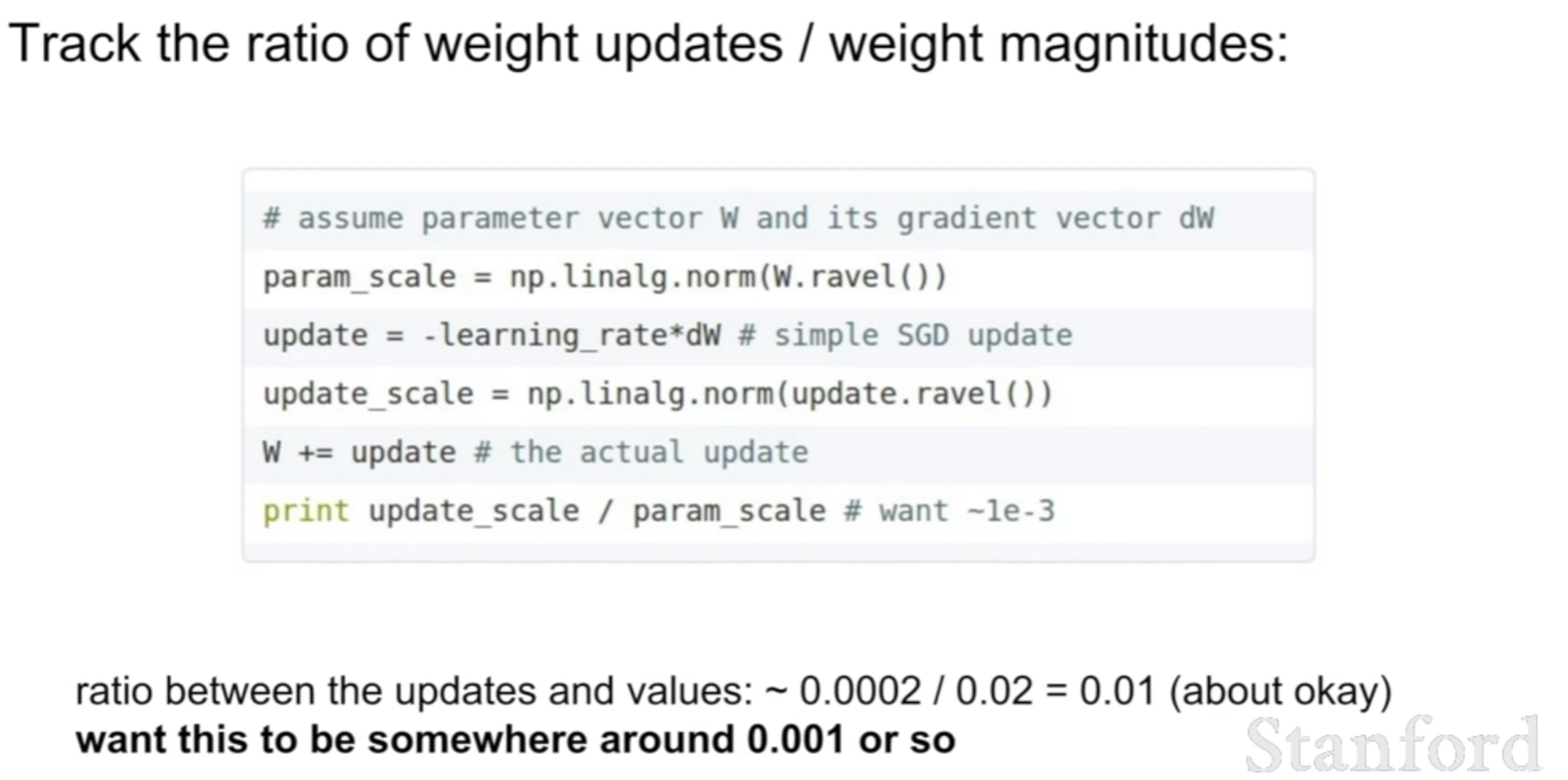
**超参数优化**：











***学习率设置***

随时间衰减。（1）step decay：经过几个epoch后衰减一半；（2）exponential decay：α=α0e-*kt*；（2）1/t decay：α=α0/(1+kt)。学习率衰减结合SGD+momentum的方法比较常见，结合Adam的不太常见。

