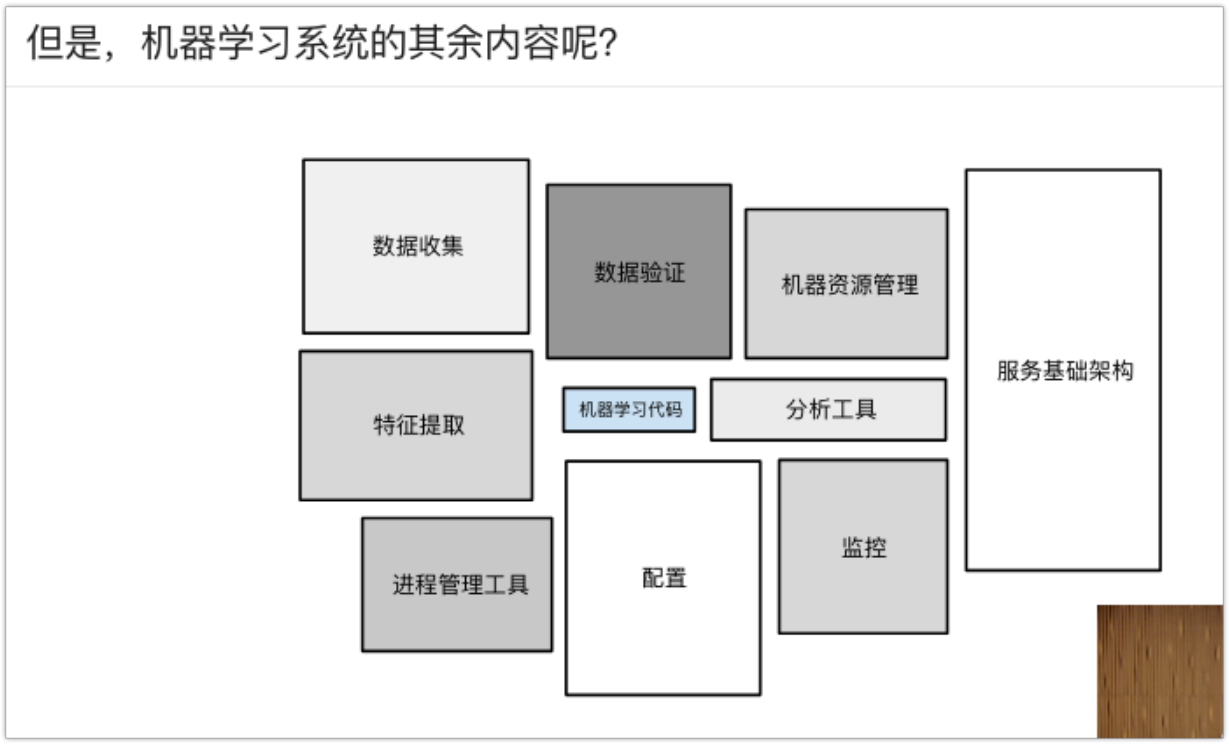
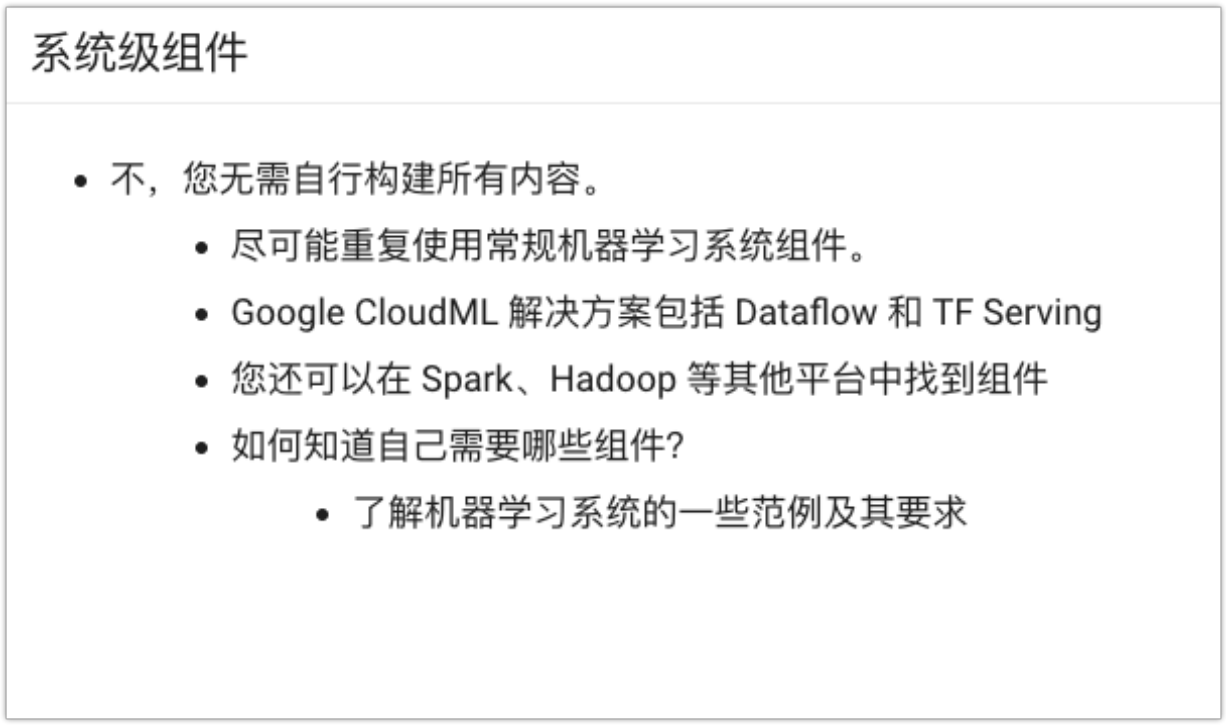
# 机器学习工程




## 静态训练与动态训练(Static vs. Dynamic Training)

从广义上讲，训练模型的方式有两种：

- **静态模型**采用离线训练方式。也就是说，我们只训练模型一次，然后使用训练后的模型一段时间
- **动态模型**采用在线训练方式。也就是说，数据会不断进入系统，我们通过不断地更新系统将这些数据整

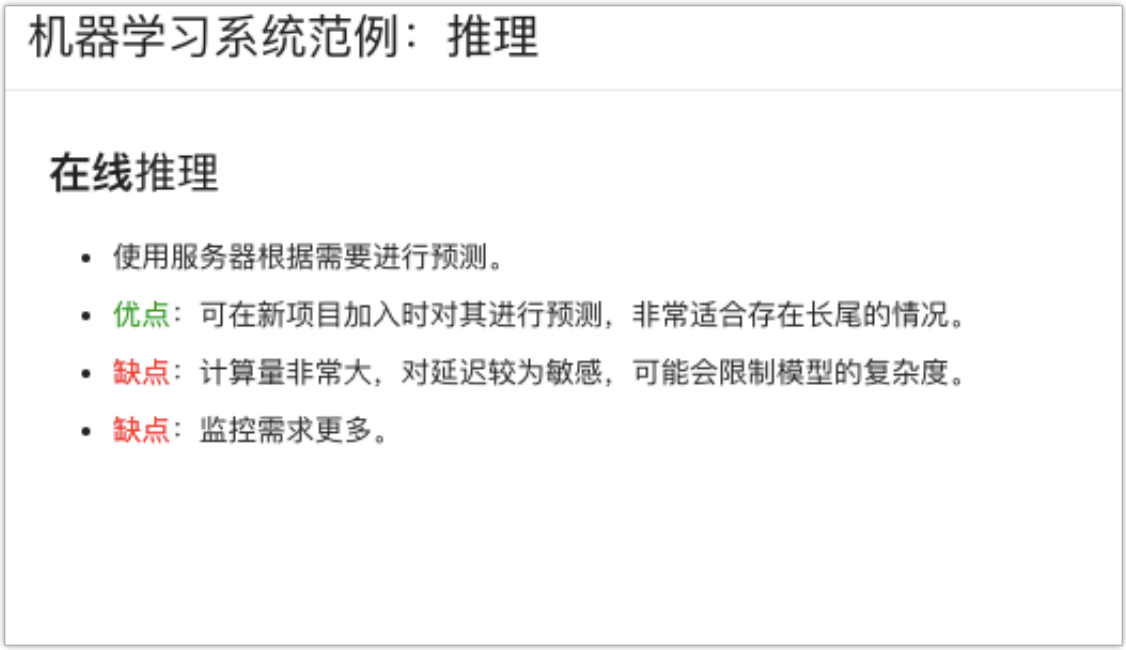
合到模型中。


## 静态推理与动态推理(Static vs. Dynamic Inference)

你可以选择以下任一推理策略:

- **离线推理**，指的是使用MapReduce或类似方法批量进行所有可能的预测。然后，将预测记录到SSTable或Bigtable中，并将它们提供给一个缓存/查询表
- **在线推理**，指的是使用服务器根据需要进行预测


机器学习系统范例：推理
离线推理
使用 MapReduce 或类似方法批量进行所有可能的预测。
记录到表格中，然后提供给缓存/查询表。
优点：不需要过多担心推理成本。
优点：可以使用批量方法。
优点：可以在推送之前对数据预测执行后期验证。
缺点：只能对我们知晓的数据进行预测，不适用于存在长尾的情况。
缺点：更新可能延迟数小时或数天。


机器学习系统范例：推理
在线推理
使用服务器根据需要进行预测。
优点：可在新项目加入时对其进行预测，非常适合存在长尾的情况。
缺点：计算量非常大，对延迟较为敏感，可能会限制模型的复杂度。
缺点：监控需求更多。

## 数据依赖关系

## 特征管理

- 输入数据（特征）决定机器学习系统的行为。
    - 我们可以针对软件库编写单元测试，但数据呢？
- 选择输入信号时要谨慎。
    - 甚至比决定要依赖哪个软件库时更谨慎吗？

## 对输入数据提出的问题

- 可靠性
    - 信号不可用时会出现什么情况？您知道吗？
- 版本控制
    - 计算此信号的系统是否发生过变化？多久一次？会出现什么情况？
- 必要性
    - 信号的实用性是否能证明值得添加此信号？

## 针对输入数据询问的问题（续）

- 相关性
    - 是否有任何输入信号密不可分，以至于需要采取额外策略来梳理它们？
- 反馈环
    - 哪个输入信号可能会受到我的模型输出的影响？

以下哪个模型容易受到反馈环的影响?

✓ 大学排名模型 - 将选择率（即申请某所学校并被录取的学生所占百分比）作为一项学校评分依据。

此模型的排名可能会提高学生对高评分学校的兴趣，从而使这些学校收到的申请增加。如果这些学校录取的学生人数继续保持不变，则选择率会增大（录取的学生所占百分比会下降）。这样会提升这些学校的排名，从而进一步提高未来有意申请这些学校的学生的兴趣，如此循环下去...

正确答案共有 3 个，您目前选中了 1 个。

✓ 交通状况预测模型 - 使用海滩上的人群规模作为特征之一预测海滩附近各个高速公路出口的拥堵情况。

有些准备前往海滩的游客可能会根据交通状况预测结果来制定出行计划。如果海滩上人群规模很大且交通预计会拥堵，则许多人可能会另做打算。这样一来，海滩上游客的数量就会减少，进而使模型作出交通畅通的预测，然后这又会导致前往海滩的游客增加，这样，这个循环就会反复下去。

正确答案共有 3 个，您目前选中了 2 个。

选举结果预测模型 - 在投票结束后对 2% 的投票者进行问卷调查，以预测市长竞选的获胜者。

人脸检测模型：检测照片中的人是否在微笑（根据每月自动更新的照片数据库定期进行训练）。

✓ 图书推荐模型 - 根据小说的受欢迎程度（即图书的购买量）向用户推荐其可能喜欢的小说。

图书推荐有可能吸引用户购买，而且这些额外销量将作为输入项反馈回模型，从而使该模型更有可能在将来推荐同样的图书。

正确答案共有 3 个，您目前选中了 3 个。

住宅价值预测模型 - 使用建筑面积（以平方米为单位计算的面积）、卧室数量和地理位置作为特征预测房价。