Application: Photo OCR

主要内容：

- 复杂机器学习系统的组成
- 如何构建机器学习流水线
- 一些 ideas：计算机视觉、人工数据合成

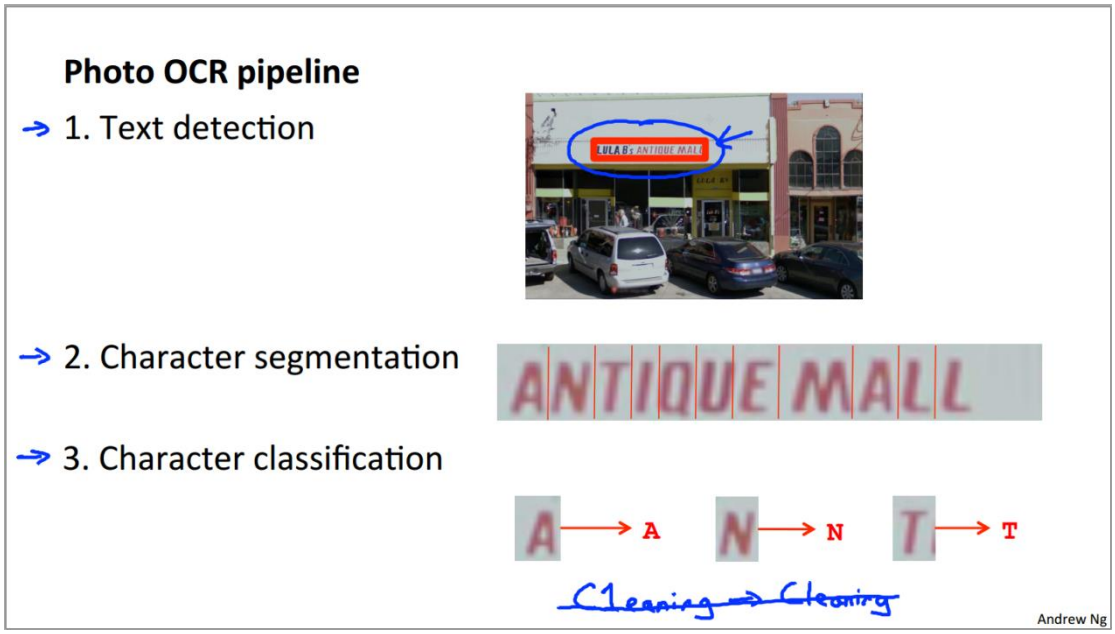
Part 1：照片 OCR

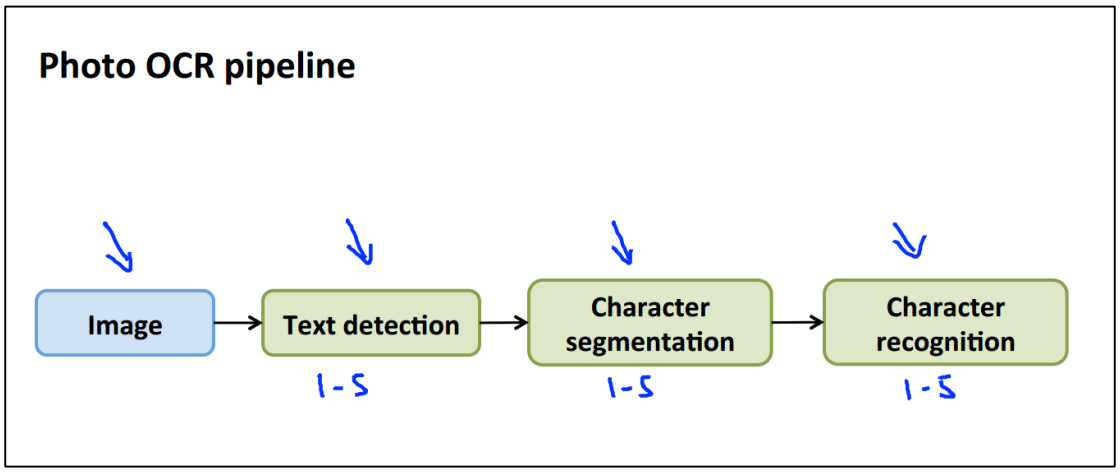
问题介绍 & 机器学习流水线（ML pipeline）

Photo OCR pipeline:

1. Text detection 文字检测
2. Character segmentation 字符分割
3. Character classification 字符识别

(4. Spelling correction 拼写矫正)





滑动窗口 Sliding window

Photo OCR 与行人检测问题类似，但更复杂。

A. Text Detection

1. 先确定需识别的对象的宽高比，据此设置图像块(patch)的标准尺寸，收集带标签的标准尺寸的正负样本，训练分类器；
2. 根据第 1 步中训练样本的宽高比，在需检测的图片上依次用相应尺寸但倍数大小不同的矩形滑动窗口以一定的步长(step size/stride parameter)遍历整个图片，每次的 patch 都 resize 到原训练样本的大小送入分类器中进行检测，最终得到一些可能的区域；
3. 送入展开器(expansion operator)，使对象集合更清晰；
4. 根据识别对象集合的形状特点（如宽高比）筛除一些备选项，将剩下的区域用矩形抠下来；

B. Character segementation

5. 对矩形用 1D 的滑动窗口识别出文字中间的区域，画上线分隔开；

C. Character classification

6. 将分割开的文字逐个送入识别文字的分类器中，得到最终结果。

- 详见课程 PPT。

人工数据合成 & 获取更多数据

人工数据合成的方法：

A. 从零开始创造数据

B. 在已有标记数据的基础上进行变形得到更多数据

- 变形需要考虑实际情况中可能出现的，随机噪声没有用。

获取更多数据时的注意事项：

1. 在绘制 learning curve 确保了模型是 low bias 的的基础上，才增加更多数据；
2. 常常问这个问题："为了得到比现有的多 10 倍的数据，需要付出多少劳动？"经常会发现，花不了多长时间就可以得到更多数据。
   - 人工数据合成
   - 自行收集、标注数据
   - 众包

上限分析 Ceiling analysis

改进机器学习流水线系统时，对每个模块的误差进行数值化的分析评价，对提高效率避免无用功至关重要。

具体方法是，分别对每个模块 plug in ground truth labels，即人为地将其中各个模块的准确度依次变成百分之百，看看最终的准确率有何变化。找出对结果影响最大的环节，再在这个环节上投入精力。