Figure 2: Classifying regions within a scholarly document. All text in the document (first panel, page from (Neyshabur and others 2013)) is located and grouped into blocks (second panel). Next the graphical components are isolated and used to determine regions of the page that contain graphics (third panel). To build the final output (fourth panel) these two elements are put together and each text block is classified as body text (filled boxes), image text (box outlines), or caption (box outlines).

gins are labelled as figure text. Left aligned blocks of text that are either small, or tall and narrow, are also classified as figure text because they are usually axis labels or columns within tables that happen to be aligned to the margins. Section headers, titles, and page headers are handled as special cases. To detect pages headers we scan the document to determine whether pages are consistently headed by the same phrase (for example, a running title of the paper might start each page) and if so label those phrases as body text. A similar procedure is used to detect if the pages are numbered and, if so, classify the page numbers found as body text. Section headers are detected by looking for text that is bold, and either column centered or aligned to a margin.

This phase also identifies regions containing graphical elements. To this end, we render the PDF using a customized renderer that ignores all text commands. We then filter out graphics that are contained or nearly contained by a body text region in order to remove lines and symbols that were used as part of a text section. Finally we use the bounding boxes of the connected components of the remaining elements to denote image regions. Figure 2 shows the steps that make up this entire process.

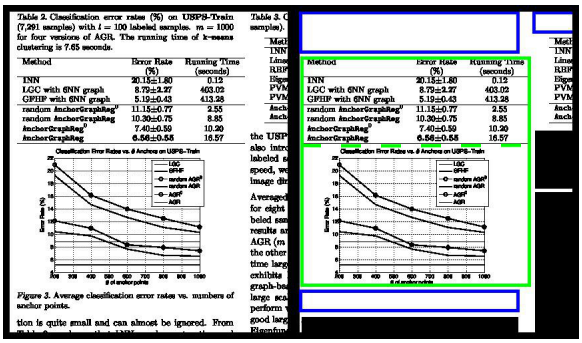

Figure 3: Example of a figure and table being directly adjacent (left panel, from (Liu, He, and Chang 2010)). In this case the proposed figure regions for each caption will by identical and encompass both the plot and table (right, solid lines). We handle this case by detecting that the region is divided in the middle by a section of whitespace, and then splitting the proposed figure region across that whitespace (dashed line).

Figure Assignment

In this final phase we assign each caption to a region of space within the document. Our algorithm is based on the observation that the region of space a figure occupies is almost always both adjacent to one side of its caption and has a box like shape. This algorithm has three parts. Region proposal, which generates, for each caption, potential regions of the document that caption could refer to. Region scoring, which is a function that gives each region a score reflecting how likely it is that the region contains a figure. Finally region selection, which uses the proposed regions and the scoring function to select a proposed region for each caption.

Region proposal is performed by building four regions adjacent to each caption by first expanding that caption's bounding box either to the left, right, up or down as far as possible without running into the body text or the page margin. These regions are then further expanded in the orthogonal directions (for example, boxes would be expanded to the left and right if they were created by expanding the caption's bounding box up or down) as much as possible without running into page margins or body text. We employ one additional heuristic, in two column papers we do not allow the box to cross the center of the page during the second expansion stage unless the caption itself spans both columns. Completing this procedure for each possible expansion direction results in four proposed figure regions for each caption.

Region scoring is a function that gives each region a score