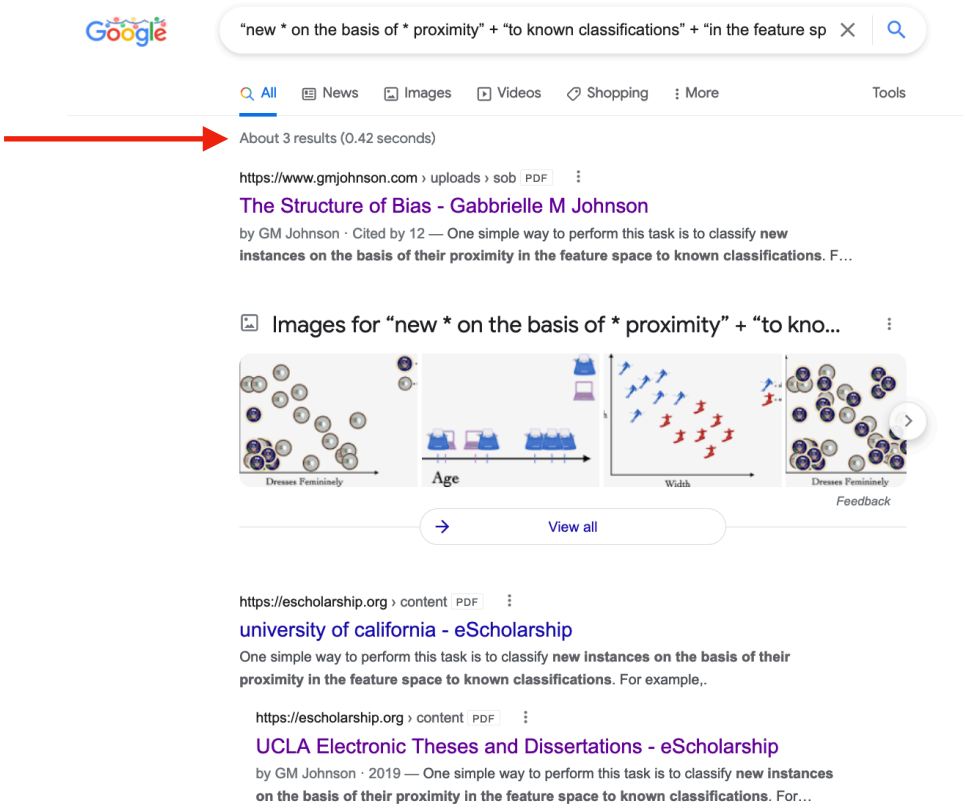
Analysis of ML Operation Overlaps

Here I analyze a series of just partial strings of sentences that show up in Professor Schellenberg and my papers.

First, Professor Schellenberg writes "…classifies new data on the basis of its proximity to known classification in the feature space", which I took to be uncomfortably similar to my passage "…classify new instances on the basis of their proximity in the feature space to known classifications."

To see how much these two descriptions overlap with how "many authors present those same facts", I again used Schellenberg's methodology of Googling the ideas. I put the relevant quotations into Google, this time grouping the prepositions through quotations, and including an * for the synonyms and pronouns:

Ex. 1: **"new * on the basis of * proximity" + "to known classifications" + "in the feature space"**

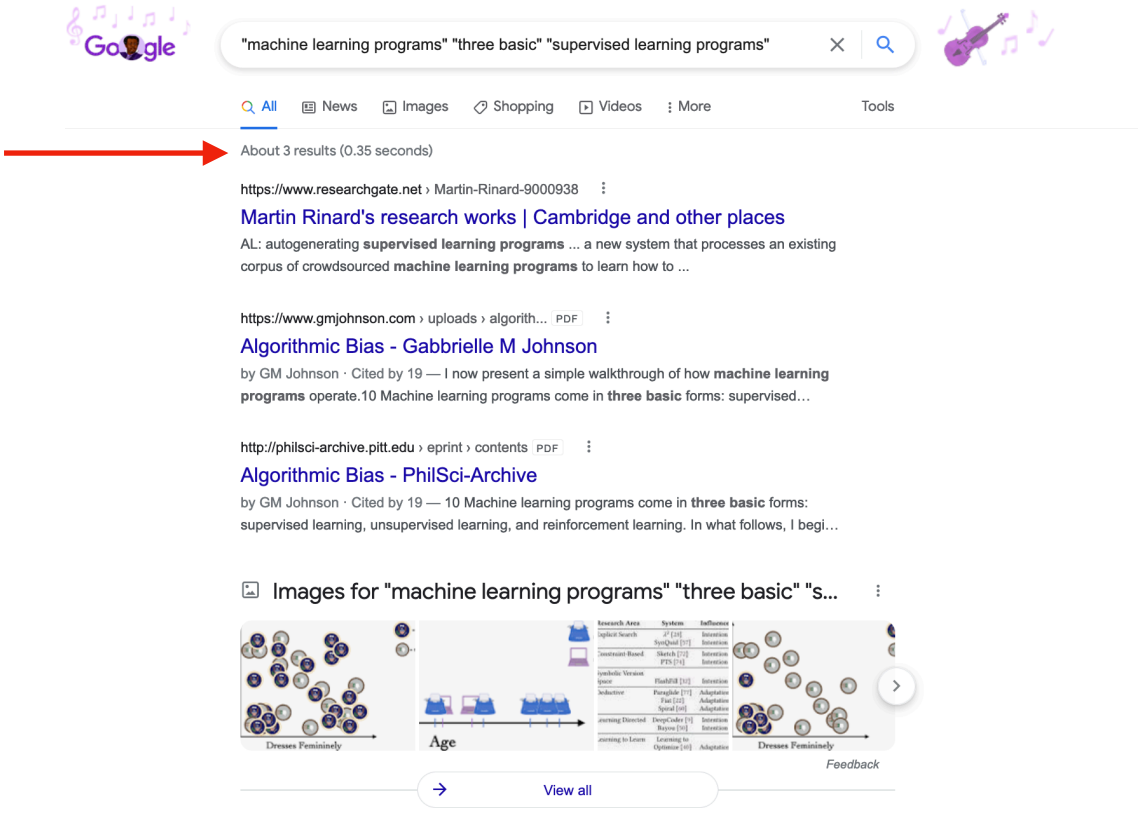


As you can see from the screenshot (or from your own Googling of the same strings), there are just 3 results that come up, all of which are my work (either my "Structure of Bias" paper or my dissertation).

(continue to page 2)

Now compare when Professor Schellenberg writes "There are three basic types of machine learning programs: supervised learning, unsupervised learning, reinforcement learning. For the sake of concreteness, I will focus on supervised learning programs", which I took to be uncomfortably similar to my passage "Machine learning programs come in three basic forms: supervised learning, unsupervised learning, and reinforcement learning. In what follows, I focus on the simpler case of supervised learning programs":

Ex. 2: **"machine learning programs" + "three basic" + "supervised learning programs"**

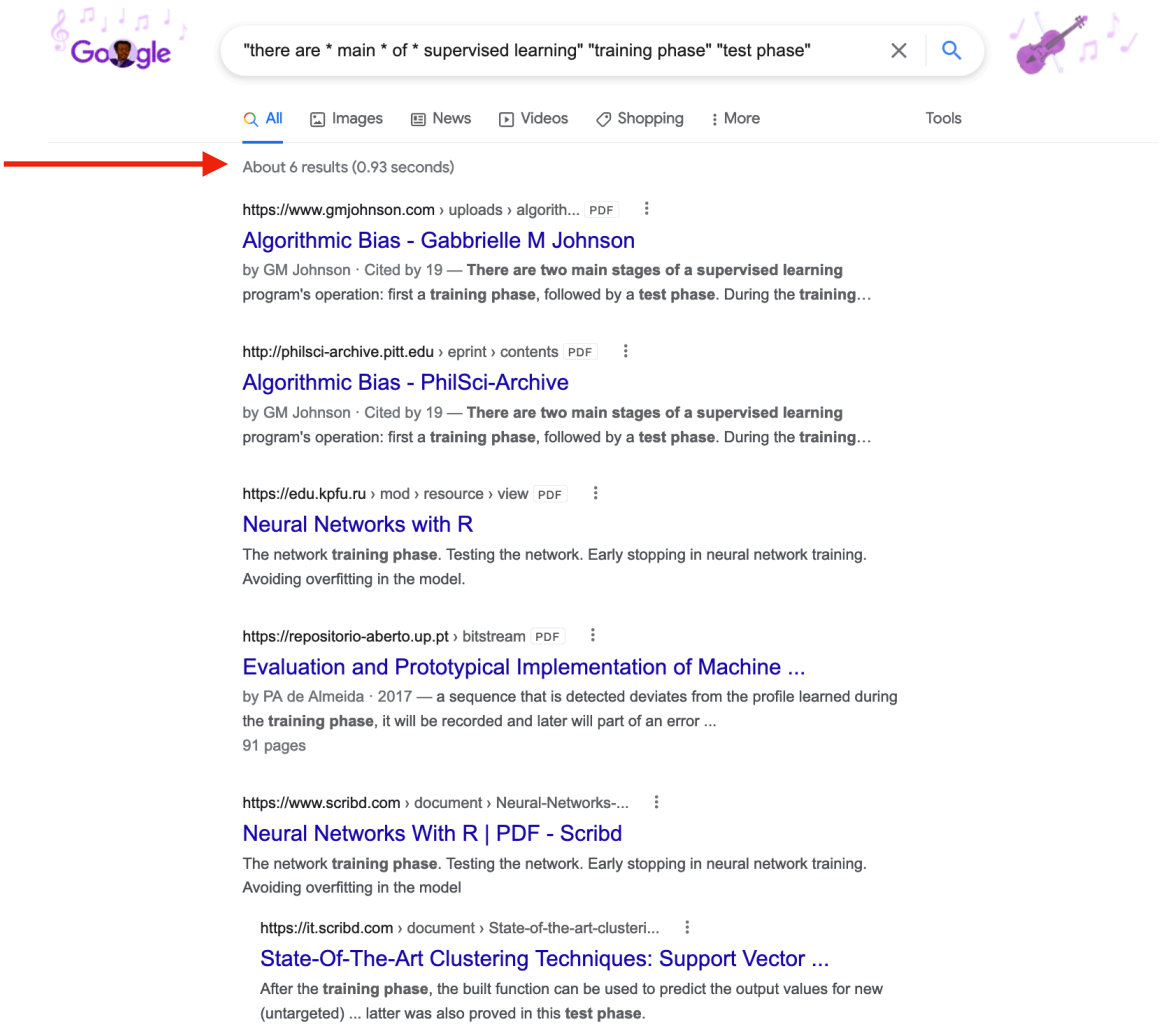


As you can see from the screenshot (or from your own Googling of the same strings), there are again just 3 results that come up, two of which are my work (different links to my "Algorithmic Bias" paper).

(continue to page 3)

Finally compare when Professor Schellenberg writes "There are three main phases of supervised learning: training phase and test phase", which I took to be uncomfortably similar to my passage "there are two main stages of a supervised learning program's operation: first a training phase, followed by a test phase."

Ex. 3: **"there are * main * of * supervised learning" + "training phase" + "test phase"**



As you can see from the screenshot (or from your own Googling of the same strings), there are just 6 results that come up, two of which are my work (different links to my "Algorithmic Bias" paper).

Now keep in mind that this search only checks a subset of the phrases that overlap, and takes each phrase individually. These searches are searching for these phrases throughout entire documents – they do not take into account the fact that in both our papers, the phrases were found in close proximity or within the same paragraph.

Given this proximity, the other overlaps I haven't searched for, and the fact that the paper includes the combination of all these searches together, I would suggest that the probability of this level of syntactic overlap being accidental is extremely low. It is surely enough to deny that Schellenberg and my papers have the level of overlap that her paper would have with any paper in the field. None of the papers she mentions come up in these searches, as their similarity is on the word but not the phrase level.