# Homework 1

# Analysis of Media and Semantic Forensics in Scientific Literature

## Due: Saturday, July 27, 2023, 11:59:59 p.m. PT

## 1. Overview

In this first homework, we will explore several of the topics discussed in the early portion of class – Big Data – MIME types and their taxonomy – Data Similarity – and so forth. To do this, we will leverage the dataset highlighted in Figure 1 – a set of 200 papers (out of 800) from biomedical and scientific literature identified in the Bik et al paper referenced below that has an "endpoint" conclusion. Endpoint papers are papers from the Bik et al 2017 dataset that resulted in either a retraction, a correction, or a no-action (a statement by the journal that they will not investigate because the paper is either too old, or they do not see any problem). Some examples of media manipulation in papers are summarized graphically shown in Figure 1.

Media manipulation is an increasing concern in scientific and open literature since researchers like Bik et al have seen a tremendous uptick in papers with media manipulations and potential problems over the last decade. Figure 2 plots this growth.

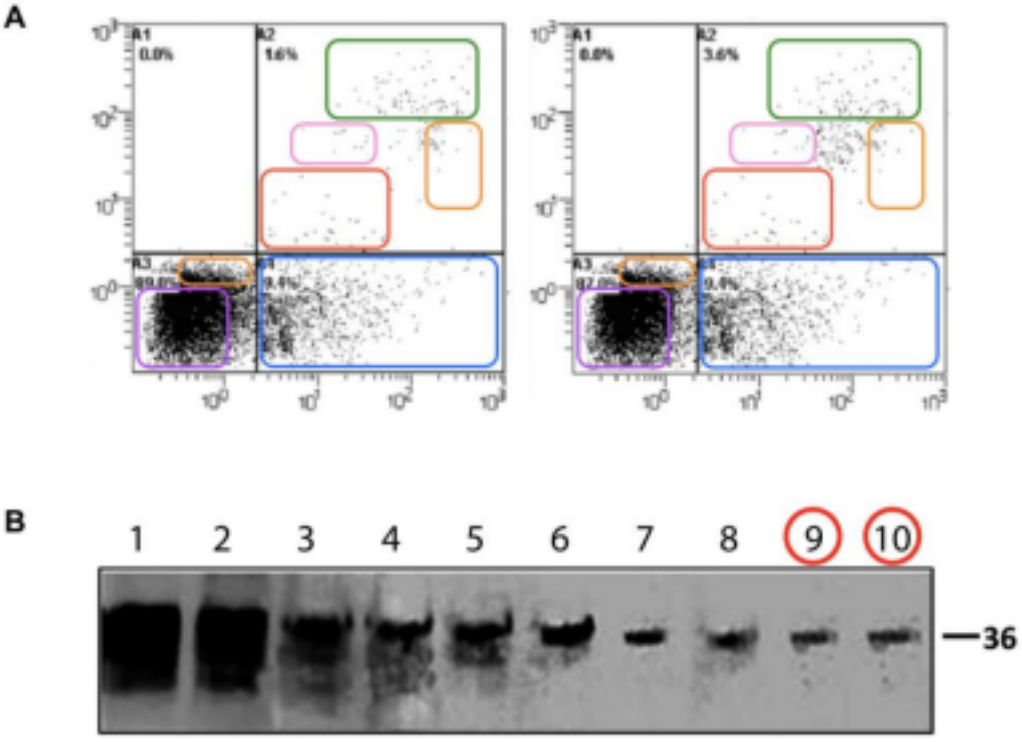



**Figure 1: Potentially problematic Media manipulations present in biomedical research papers. A and B show areas of duplicity in included media in papers.**
**https://journals.asm.org/doi/epub/10.1128/mBio.00809-16**

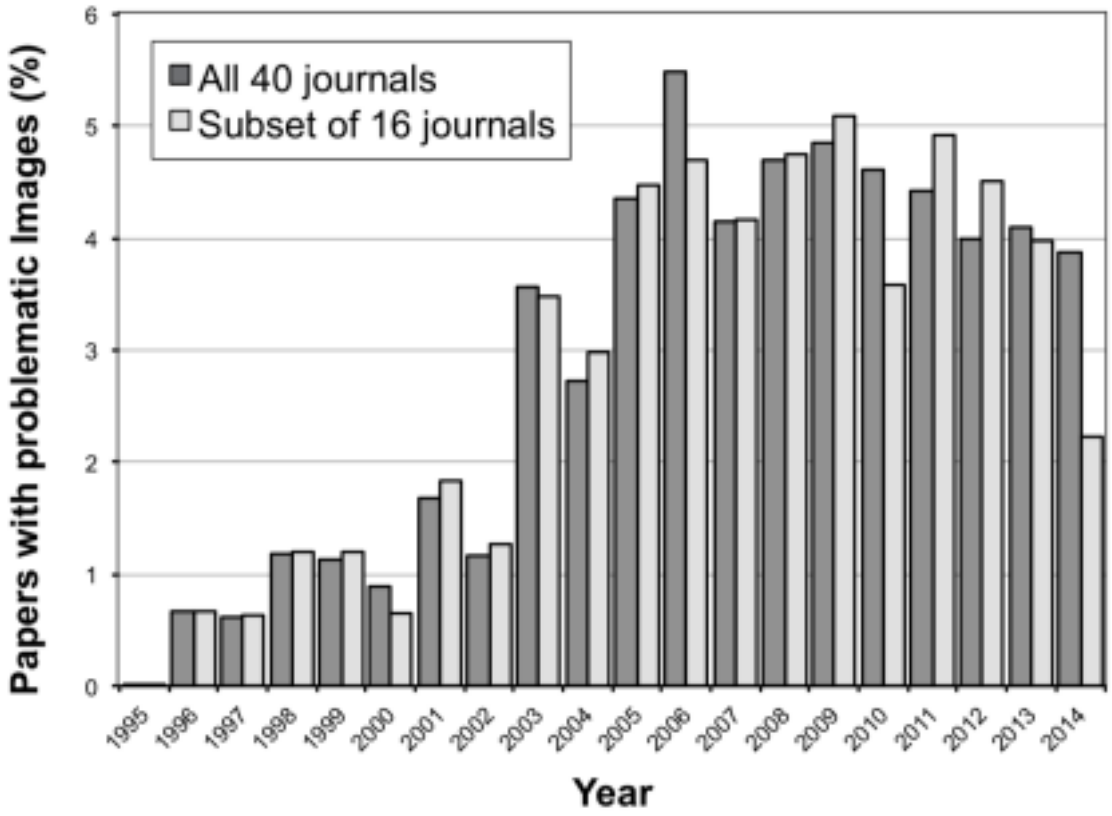
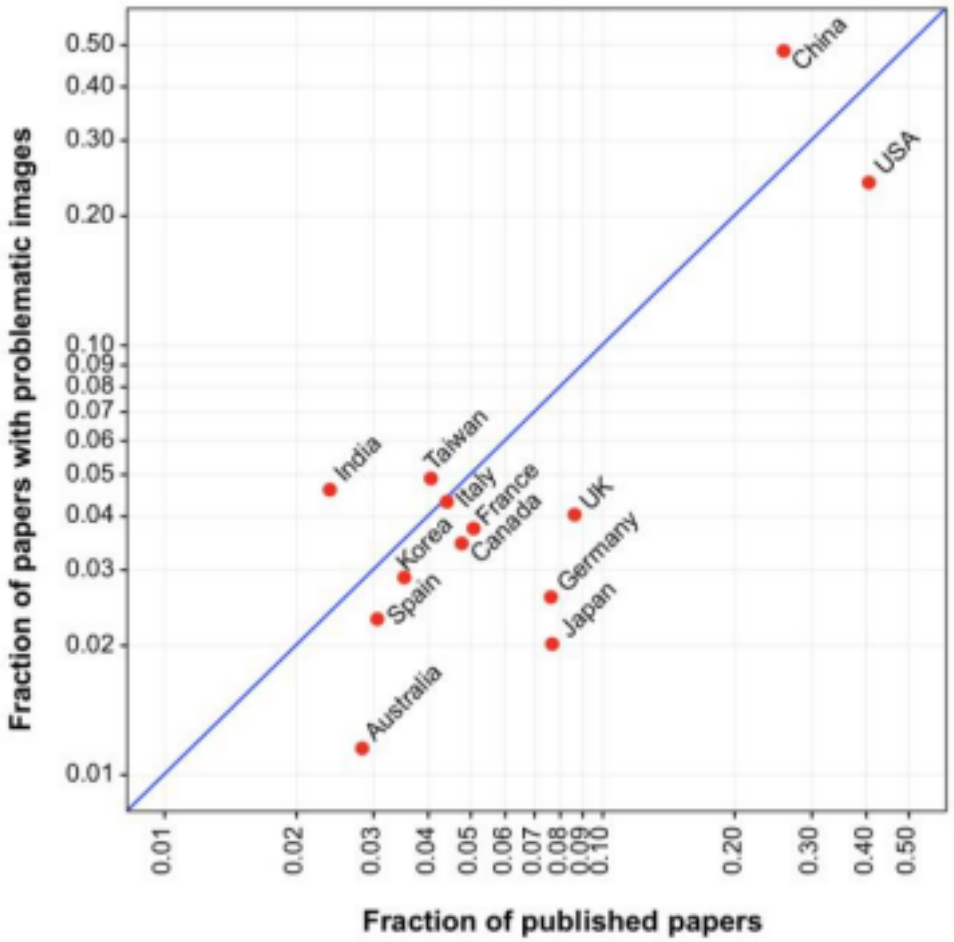



**Figure 2: Growth of papers with problematic or manipulated images 1995-2014**

Additionally, one of the challenges too if you look deeper at the data is that there has been a geographic uptick in the author institutions and affiliations associated with particular countries that seem to be responsible for the uptick when you plot author affiliation / country against fraction of published papers and fraction of media manipulations. Figure 3 has the associated data and presents it visually.

**Figure 3:**




As can be gleaned from the data, the manipulations occur most by authors from The United States and China and then there is a set of countries with similar properties with respect to number of high impact papers and possible manipulations. Bik et al have captured a dataset that records information about papers including Authors, Paper Title, Citation, Digital Object Identifier (DOI), Year, Month, Classification Label (0-3) referring to three major categories: simple duplications, duplications with repositioning, and duplications with alteration and also two potentially problematic areas that may not be manipulations but cause issues (Cuts & Beautification). The dataset also includes a text description of what was found (Findings), if the paper was reported to the journal or not, and whether it was retracted, or a correction issued or not, and finally whether no action was taken and what date (if any) any action was completed on. The data is formatted according to the referenced schema which will be provided to you and is in TSV (TSV) format (MIME type: text/tab-separated-values).

As you can imagine, there are plenty of things we could do with this data. Some of them fall into the realm of the paper we discussed from IEEE Computer in 2013 regarding "Big Data's Unintended Consequences". If you recall, one of the key points of this paper was that companies and governments were moving away from data silos, and instead focused on how data could be *joined* together to form even more comprehensive, statistically relevant, and accurate data on *everyone* – in short, increasing its value, and potentially its veracity, in addition to its volume, and potentially its variety.

Another topic frequently discussed in class thus far has been the framework of the "Five V's" – volume, velocity, variety, veracity and value. We have done several class/group activities so far and thought about many datasets in the real world and how to classify and leverage them along this framework. From judicial data about inmates and prisoners, to health data, to Twitter data, to restaurant review data, there have been extremely useful discussions and points made. You will leverage those discussions, class lectures, and material discussed in this assignment.

## 2. Objective

Looking at the Bik et al media manipulation data, you may ask yourselves: "what other data is available that could be joined with this information" to affect its Five V's – intentionally, or unintentionally. For example, consider if you could automatically pull down a lot more information about a particular author of a paper and her co-authors – their rate of publication, number of students in their lab, information about what other journals they have published in, etc. This is an interesting question, and will form the first part of your assignment which will involve finding additional information about each author for each of the provided publications and collecting and joining it to the Bik dataset. You will add new features to the dataset of "Lab Size (number of students)", "Publication Rate", "Other Journals Published In" and some information about "First Author" including "Affiliation University", "Duration of Career (Years)", highest degree obtained (e.g., "PhD", "MS") and "Degree Area" (e.g., Computer Science). Perhaps there is a pattern that will emerge, for example you could posit that those with a

Masters in Computer Science, with 50 years of experience and 100 students in the lab, may not be critically reviewing papers published in biomedical journals. Or not. These will be the interesting research questions that you will investigate!

What other datasets could you join the Bik et al data to? For example, what about census population per closest city to the first author's institutional affiliation? What about restaurant data indicating open hours and close times for nearby coffee shops - maybe the students in the lab are up late at night and not paying attention when they paste in graphics in the papers? Finally, what about animal population data in the closest county, city, metropolitan area to the author's lab, to give you an idea of the mice population or rabbit or experimental animal population – perhaps there is duplication because all of the animals are cousins, and their cellular data is in fact the same. You could also look at associated conference or journal submission times for the journals or conferences in question, see if the submission times are all late at night... etc.

You will choose at least three publicly accessible datasets along these lines to join the Bik et al data to, and you must add at least three new features per dataset that you join. The datasets you select may not all belong to the same MIME top level type – that is – you must pick a different MIME top level type for each of the three datasets you are joining to this Bik et al problematic papers dataset.

Once the data is joined properly, you will explore the combined dataset using Apache Tika and an associated Python library called Tika Similarity. Using Tika Similarity, you can evaluate data *similarity* (as discussed during the Deduplication lecture in class; and also during data forensics discussions). Tika similarity will allow you to explore and test different distance metrics (Edit-Distance; Jaccard similarity; Cosine similarity, etc.). So, you can figure out how similar papers with problem areas are within the data and ask questions of your new augmented Bik et al dataset. For example, you may ask, how many papers with similar media manipulations came from first authors with more than 50 students in the lab, with small mice populations who were all cousins, and in which the students in the lab have access to coffee shops that keep them up all night and thus they are tired when clicking submit on the papers?

The assignment specific tasks will be specified in the following section.

## 3. Tasks

1. Download and install Apache Tika
   - The lecture on Tika covers some of the basics of building the code, and additionally, see http://tika.apache.org/1.23/gettingstarted.html
   - Install Tika-Python, you can pip install tika to get started.

     Additionally, you can read up on Tika Python here:

     http://github.com/chrismattmann/tika-python
2. Download and install D3.js
   - Visit http://d3js.org/
   - Review Mike Bostock's Visual Gallery Wiki
     i. https://github.com/mbostock/d3/wiki/Tutorials
3. Download the Bik et al dataset:
   - https://drive.google.com/file/d/1FG7Ze6qmI06zNmzBjGwH4ns2bwrBZyiv/view?usp=sharing
   - Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
4. Begin by adding additional features to the data about the authors' or the paper's information from ResearchGate:

• Find additional features (at least one more feature such as that authors' university name, etc.) about at least one author of each paper. Or find additional features about each paper (at least one more feature such as citation count, etc.)

• Write a Python program to collect the above information for each of the 200 papers in the dataset. Example code: https://github.com/kevinyu0506/researchgate-crawler

(**EXTRA CREDIT:** Use BeautifulSoup python package to collect this data https://www.crummy.com/software/BeautifulSoup/bs4/doc/ )

• Note that the purpose of this task is to teach Beautiful Soup scraping skills.

5. Identify **at least one other datasets**, of different top level MIME type (can be e.g., text/*, json, csv, etc) and **extract at least one useful feature** (to be added as columns of your tsv file, same as in part 4) that can contributing to your research questions and similarity score.

a. Check out places including:
https://catalog.data.gov/dataset (Data.gov)
b. For the new dataset, develop a Python program to join the data to your Bik et al papers dataset
i. For each non tsv dataset, be prepared to describe how you featurized the dataset
c. Each dataset that you join must contribute at least one feature (in addition to the features you are adding described in part 4)
d. For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature
f. Calculate similarities (clustering) between papers of initial TSV file (with existing features), then Calculate similarities (clustering) between papers of modified TSV file (with newly found features). Discuss how the new features contribute to the similarity scores and clusters. See details in rubrics.

6. Download and install Tika-Similarity

• Read the documentation

• You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)

• Compare Jaccard similarity, edit-distance, and cosine similarity

## 4. Assignment Setup

### 4.1 Group Formation

You can work on this assignment in groups sized at **maximum 5**. You may reuse your existing groups from discussion in class.

### 4.2 Bik et al papers dataset

Access to the data is provided in the Google Drive. The dataset itself is 96k. You may want to distribute the data between your team-mates since the data is small (for now).

### 4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the tika-app.jar from: http://tika.apache.org/download.html. You should obtain a jar file called tika-app-1.23.jar. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/1.23/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://wiki.apache.org/tika/TikaJAXRS. The Tika Python library is a robust REST client to the Java-side REST API.

You can also get more information about Tika by checking out the book written by Professor Mattmann called “Tika in Action”, available from: http://manning.com/mattmann/.

## 5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the Bik et al papers previously unanswered? What similarity metrics produced more (in your opinion) accurate measurements? Why? What did the additional datasets suggest about "unintended consequences" related to media forensics data? You should also clearly explain which datasets you used to join the problematic papers data and how you extracted the new features from each dataset.

Thinking more broadly, do you have enough information to answer the following:

1. Does staying up late at night matter?
2. Does the animal population for available specimens in the area influence the type of manipulations?
3. What do animal population demographics tell us about the institutions in which media manipulations occur?
    a. Densely populated? Sparsely populated?
4. What insights do the "indirect" features you extracted tell us about the data?
5. Include your thoughts about Apache Tika – what was easy about using it? What wasn't?

## 6. Submission Guidelines

This assignment is to be submitted ***electronically, by 11:59pm PT*** on the specified due date, via D2L > My Tools > Assignments. A team can submit multiple times, but only the last submission counts. Anyone from a team can submit. However, we suggest designating one person to submit.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts that you used to join three other datasets, and what you used to extract additional features.
- Include your updated dataset TSV. We will provide a Dropbox location for you to upload to.

- Also prepare a readme.txt containing any notes you'd like to submit.

- If you used external libraries other than Tika Python and Tika Similarity, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.

- Save your report as a PDF file (TEAM_XX_BIGDATA.pdf) and include it in your submission.

- Compress all of the above into a single zip archive and name it according to the following filename convention:
  **TEAM_NAME_DSCI550_HW_BIGDATA.zip**
  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.

- ***Important Note:***

- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.

- Again, please note, a team can submit multiple times, but only the last submission counts. **To avoid confusion: designate someone to submit.**

## Late Assignment Policy -

10% for every day or part thereof