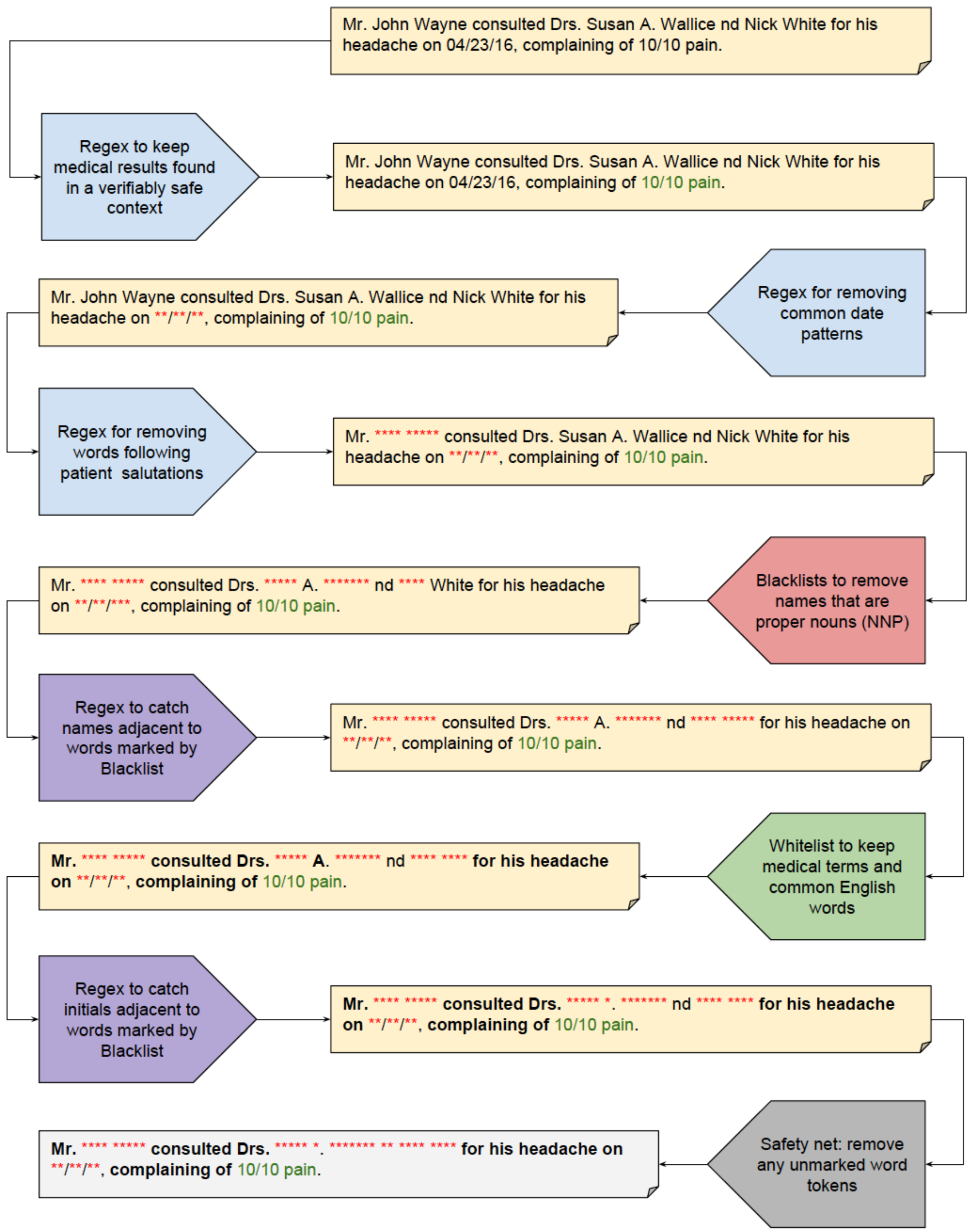
UCSFPhilter: User-Friendly De-Identification of Clinical Text

UCSF Bakar Computational Health Sciences Institute has created UCSFPhilter, a tool that removes Protected Health Information (PHI) from clinical text. The tool removes 18 HIPAA-defined types of PHI from the user-provided text by implementing an extremely effective privacy-centric algorithm, that combines rule-based and statistical NLP approaches. The algorithm was originally developed by Beau Norgeot, PhD from Atul Butte's Lab at UCSF Bakar Computational Health Sciences Institute. UCSF Philter's PHI removal pipeline was evaluated against 2 strongest text de-identification tools, Physionet and Scrubber, on both UCSF and i2b2 (Integrating Biology and Bedside) annotated corpora of clinical notes. UCSFPhilter by far outperformed both comparators based on the overall recall (rate of PHI removal) and for each category of PHI, yielding respective 99.46% and 99.92% overall PHI removal rate on UCSF and i2b2 test corpora.

How It Works

UCSFPhilter PHI removal algorithm is based on identifying parts of text that are highly unlikely to be PHI using a combination of state-of-the-art rule-based and statistical methods. For additional security and precision, the algorithm incorporates approaches for identifying words and phrases that are likely to be PHI. UCSFPhilter PHI removal pipeline includes the following steps:

1. Tokenization (separation of individual words)
2. Identifying safe (likely non-PHI) phrases by matching with a pre-defined library of "PHI-safe" regular expressions patterns and applying a pre-compiled "whitelist" of non-PHI words, such as medical terms, codes, concepts and abbreviations, and most common English words.
3. Identifying unsafe (likely PHI) phrases by matching with a pre-defined library of PHI patterns
4. Part-of-speech (POS) tagging of all tokens in the text document
5. Applying pre-configured "blacklist" to all tokens, along with the POS tags and the surrounding context to detect PHI
6. Removing patient and provider initials and additional removal of names using predefined patterns

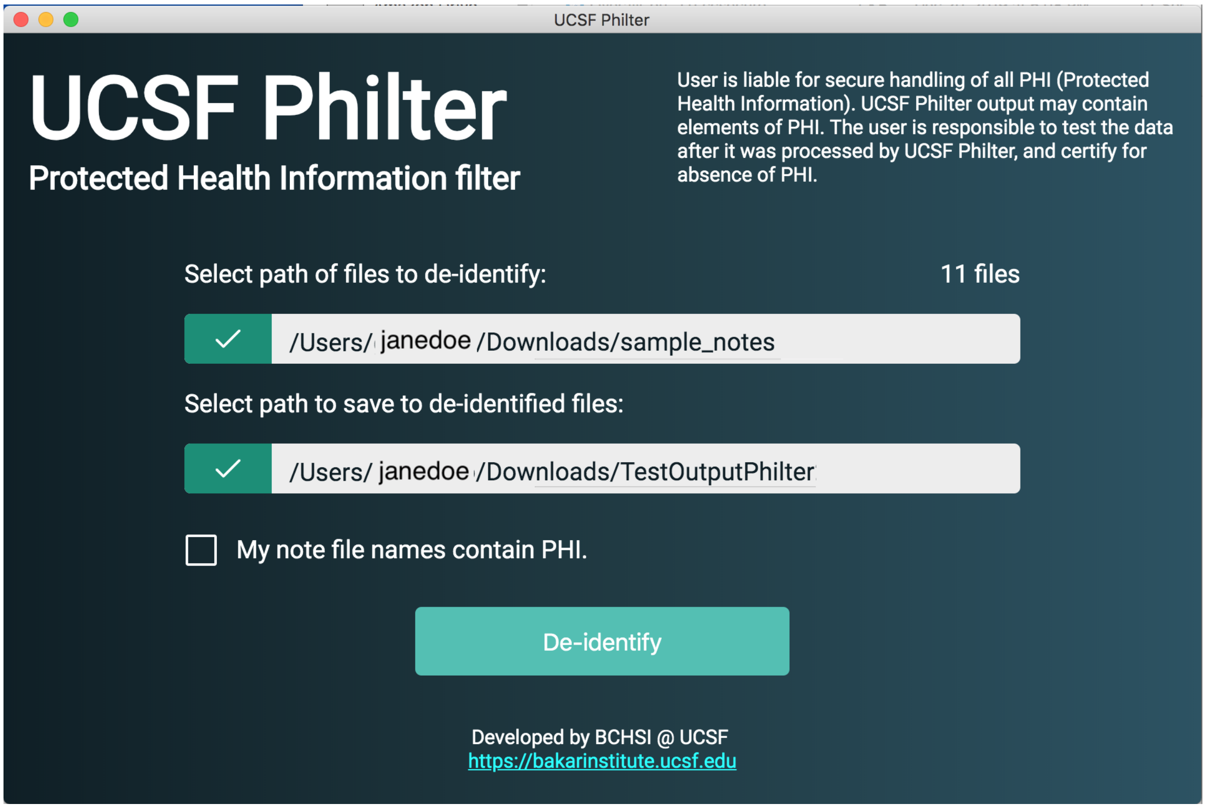


To learn more about the UCSFPhilter underlying architecture, please see Beau Norgeot's dissertation (Chapter 8) and the Open Source repository on GitHub (note that the open source code represents the command line version of UCSFPhilter).

Using UCSFPhilter At Your Institution

UCSFPhilter is now available for other medical academic institutions! View license

The user-friendly version of UCSFPhilter is a one-screen desktop application (available for Windows and Mac OS) that uses a generalized pre-configured pipeline and requires minimal set of inputs: the location of the text documents user wants to de-identify, the output location and an additional switch for de-identifying file names.



If you are interested in using UCSFPhilter to de-identify clinical text at your institution, please contact us at info.commons@ucsf.edu. We will provide you the software executables and user documentation. Although the software is very easy to install and use, we gladly offer technical help as needed to ensure its successful use. Please note, even though UCSFPhilter is a top performing PHI removal tool, some PHI may still remain in its output, and it is each user's responsibility to handle this output safely, according with the HIPAA guidelines and your institution's privacy policies. Please credit UCSF Bakar Computational Health Sciences Institute in any work enabled by UCSFPhilter.