This document displays the outcomes of the in-class workshop from 8 September, 2014. Together with the PowerPoint presentation from the lecture, this document can be used as support when working with your own texts.

The task was to analyze the introductions of two texts; Rossow et al, *Prudent Practices for Designing Malware Experiments: Staus Quo and Outlook* and Riloff, *Little Words can Make a Big Difference for Text Classification*, based on the following seven questions. A summary of the comments from students and teachers are provided in connection with the text below.

1. Does the Introduction provide information about context, indicate motivation for the paper, define focus, explanation of document structure? Please indicate where in the text

2. What referencing system is used? Comment on the number of in-text references?

3. Have a look at the organization of each paragraph. Do the paragraphs hold one idea starting with a topic sentence? How does the rest of the rest of the paragraph match the topic sentence?

4. Have a look at the structure of sentences. Are sentences generally short or long? Any fuzzy sentences?

5. Does the author use linking devices as glue between sentences (such as “in addition”, “however” etc)?

6. What about the language: is the style formal / informal? Give examples
Does the author use the passive or active voice? Give examples

7. Any additional comments:

A. Rossow et al.	
I. INTRODUCTION Observing the host- or network-level behavior of malware as it executes constitutes an essential technique for researchers seeking to understand malicious code. Dynamic malware analysis systems like Anubis [8], CWSandbox [50] and others [16, 22, 27, 36, 42] have proven invaluable in generating ground truth characterizations of malware behavior. The anti-malware community regularly applies these ground truths in scientific experiments, for example to evaluate malware detection technologies [2, 10, 17, 19, 24, 26, 30, 33, 44, 48, 52–54], to disseminate the results of large-scale malware experiments [6, 11, 42], to identify new groups of malware [2, 5, 38, 41], or as training datasets for machine learning approaches [20, 34, 35, 38, 40, 41, 47, 55]. However, while analysis of malware execution clearly holds importance for the community, the data collection and subsequent analysis processes face numerous potential pitfalls. In this paper we explore issues relating to prudent experimental evaluation for projects that use malware-execution datasets. Our interest in the topic arose while analyzing malware and researching detection approaches ourselves, during which we discovered	First paragraph present the background. The context is provided in the first three sentences. Motivation introduced Small motivation introduction Small focus presentation The main part of the motivation starts here

that well-working lab experiments could perform much worse in real-world evaluations. Investigating these difficulties led us to identify and explore the pitfalls that caused them. For example, we observed that even a slight artifact in a malware dataset can inadvertently lead to unforeseen performance degradation in practice.

Thus, we highlight that performing prudent experiments involving such malware analysis is harder than it seems. Related to this, we have found that the research community's efforts (including ours) frequently fall short of fully addressing existing pitfalls. Some of the shortcomings have to do with presentation of scientific work, i.e., authors remaining silent about information that they could likely add with ease. Other problems, however, go more deeply, and bring into question the basic representativeness of experimental results.

As in any science, it is desirable for our community to ensure we undertake prudent experimental evaluations. We define experiments reported in our paper as prudent if they are are correct, realistic, transparent, and do not harm others. Such prudence provides a foundation for the reader to objectively judge an experiment's results, and only wellframed experiments enable comparison with related work. As we will see, however, experiments in our community's publications could oftentimes be improved in terms of transparency, e.g., by adding and explaining simple but important aspects of the experiment setup. These additions render the papers more understandable, and enable others to reproduce results. Otherwise, the community finds itself at risk of failing to enable sound confirmation of previous results.

In addition, we find that published work frequently lacks sufficient consideration of experimental design and empirical assessment to enable translation from proposed methodologies to viable, practical solutions. In the worst case, papers can validate techniques with experimental results that suggest the authors have solved a given problem, but

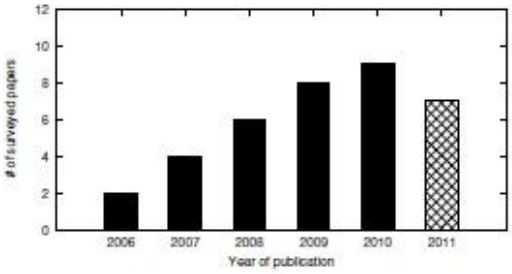

Figure 1: Surveyed papers using malware execution, per year.

the solution will prove inadequate in real use. In contrast, well-designed experiments significantly raise the quality of science. Consequently, we argue that it is important to have guidelines regarding both experimental design and presentation of research results.

We aim in this work to frame a set of guidelines for describing and designing experiments that incorporate such prudence, hoping to provide touchstones not only for authors, but also for reviewers and readers of papers based on analysis of malware execution. To do so, we define goals that we regard as vital for prudent malware experimentation: transparency, realism, correctness, and safety. We then translate these goals to guidelines that researchers in our field can use.	Focus
We apply these guidelines to 36 recent papers that make use of malware execution data, 40% from top-tier venues such as ACM CCS, IEEE S&P, NDSS and USENIX Security, to demonstrate the importance of considering the criteria. Figure 1 shows the number of papers we reviewed by publishing year, indicating that usage of such datasets has steadily increased. Table II (on page 6) lists the full set of papers. We find that almost all of the surveyed papers would have significantly benefited from considering the guidelines we frame, indicating, we argue, a clear need for more emphasis on rigor in methodology and presentation in the subfield. We also back up our assessment of the significance of some of these concerns by a set of conceptually simple experiments performed using publicly available datasets.	Introducing the structure of the text
We acknowledge that fully following the proposed guidelines can be difficult in certain cases, and indeed this paper comes up short in some of these regards itself. For example, we do not fully transparently detail our survey datasets, as we thought that doing so might prove more of a distraction from our overall themes than a benefit. Still, the proposed guidelines can—when applicable—help with working towards scientifically rigorous experiments when using malware datasets.	
To summarize our contributions: • We identify potential pitfalls when designing experiments based on malware execution, and estimate the impact of these pitfalls in a few experiments. • We devise guidelines to help with designing and presenting scientifically rigorous experiments. • Our survey of 36 papers shows that our community could better address a number of shortcomings in typical malware datasets by adhering to these guidelines. • We show that, contrary to our expectations, most of the problems occur equally in publications in top-tier research conferences and in less prominent venues.	

<u>From analysis and discussion</u>

1. Information about context, motivation

· Clear introduction of context and motivation.

· Focus shifted from a specific topic to something written to appeal to a broader audience.

· Examples of context, motivation provided in the text. (see examples in the right column).

· There is no ‘roadmap’ in the introduction. But the text structure is introduced in the beginning of

each section.
· The end of the conclusion is a summary of their contributions, which gives a nice orientation.

2. References
· There are a lot of in-text references, which arguably could speak for the paper's relevance.
· Backing up claims by using many references.
· IEEE is used.
· The grouping of references is useful in describing relations between cited papers. This can also be good in providing the reader with a context to revisit and also as background reading.
· References could be used more optimal in the text. More referencing in some places. Some arguments in the middle and end of the introduction could benefit from supporting references.
· References are used which shows that research has been made.
· It would be an idea to present this large number of references in a separate "Related Work" section instead of the beginning of the introduction.

3. Paragraphs
· Clear topic sentences starting paragraphs. The authors follow their own train of thought.
· Generally paragraphs introduce an idea and stay consistent to the idea throughout the paragraphs. The third paragraph seems to be a continuation of the second one, though.
· The paragraphs often elaborate on the content of the previous paragraphs and this makes the text flow nicely.

4. Structure of sentences
· Some sentences are long but are generally clear.
· Long sentences provide a feeling of formal language, but on the other hand this does create some fuzzy sentences. E.g., "We aim in this work to frame a set of guidelines for describing and designing experiments that incorporate such prudence, hoping to provide touchstones not only for authors, but also for reviewers and readers of papers based on analysis of malware execution."
· The sentence length is appropriate with this type of text using rather long and descriptive sentences and thereby avoiding choppiness and maintaining a good flow.

5. Linking devices
· There are a number of examples: "however, thus, for example, in addition, consequently, still, related to this"

6. Style / formal informal
· The style is quite formal. However, they are using active voice. But most of the text is formal, not written in a 'talking' style.
· Vocabulary used is mostly formal, e.g. "Thus we highlight"
· They use a good vocabulary but use "we, our, us" in several places. Scaling down the use of personal pronouns will make the text more a bit more formal.
· The text is definitely formal, in spite of extensive use of the active form. Certain sentences are very formal, e.g. the first one, but further into the text we find several less formal expressions and constructs.
· The style is adequately formal for the topic. An example for a not very strict formality is the sentence "Our interest in the topic arose while analyzing malware and researching detection approaches ourselves, [...]" in the second paragraph, as we expect very formal author's not to talk about themselves like that.
· Overuse of the word "prudent" in the text. The reason for promoting this word feels like a "gimmick"

B. Riloff

Introduction

Most information retrieval systems use a stopword list to prevent common words from being used as indexing terms. Highly frequent words, such as determiners and prepositions, are not considered to be content words because they appear in virtually every document. Stopword lists are almost universally accepted as a necessary part of an information retrieval system. For example, consider the following quote from a recent information retrieval textbook:

"It has been recognized since the earliest days of information retrieval (Luhn 1957) that many of the most frequently occurring words in English (like "the", "of", "and", "to", etc.) are worthless indexing terms." ([Frakes and Baeza-Yates, 1992], p. 113)

Many information retrieval systems also use a stemming algorithm to conflate morphologically related words into a single indexing term. The motivation behind stemming algorithms is to improve recall by generalizing over morphological variants. Stemming algorithms are commonly used, although experiments to determine their effectiveness have produced mixed results (e.g., see [Harman, 1991; Krovetz, 1993]).

One benefit of stopword lists and stemming algorithms is that they significantly reduce the storage requirements of inverted files. But at what price? We have found that some types of words, which would be removed by stopword lists or merged by stemming algorithms, play an important role in making certain domain discriminations. For example, similar expressions containing different prepositions and auxiliary verbs behave very differently. We have also found that singular and plural nouns produce dramatically different text classification results.

First, we will describe a text classification algorithm that uses linguistic expressions called "relevancy signatures" to classify texts. Next, we will present results from text classification experiments in two domains which show that similar signatures produce substantially different classification results. Finally, we discuss the implications of these results for information retrieval systems.

From analysis and discussion

1. Information about context, motivation

· The paper provides information about the context in the first two sentences in the 1st paragraph. Then the focus continues to be discussed down to the third paragraph After the 2nd sentence. Then

follows the motivation for the text and the last paragraph is the document structure.
· It is nice that the text has a roadmap of the document structure, but it would be better to write "Section 1", "Section 2" etc rather than "First", "Next", "Finally"
· The author doesn't say why automatic text classification is important (but the paper was written for a conference on information retrieval)

2. References
· ACM is used in this paper.
· There are few references. More would be appropriate.

3. Paragraphs
· Yes the paragraphs hold one idea starting with topic sentence. They work.
· Only the last paragraph is different as it explains the document structure.

4. Structure of sentences
· The sentences are neither too long nor too short. There are no fuzzy sentences.
· Some sentence are somewhat short.
· The sentences tend to be short and concise, making the text easy to follow.

5. Linking devices
· There are some linking devices such as, "for example, first, next, finally", etc.
· The authors barely use any linking devices. But as it is a short introduction, the readers do not have a hard time to follow the ideas anyway.

6. Style / formal informal
· In general the text is formal but only 1 or 2 sentences informal way. For example: "But at what price?"
· The author uses active voice. For example,"we have found that .." and "we discuss the implications.." making it less formal.

7. Additional comments
· There is a mix of verb tense in the final paragraph of the introduction, e.g. "First, we will describe" and "Finally, we discuss", which should be avoided.
· Generally it is easy to follow the text structure.
· The background information / related work are introduced precisely where the reader needs to know about it. Given that it is such a short paper, more sections are not needed.