Author Obfuscation

Attacking the State of the Art in Authorship Verification

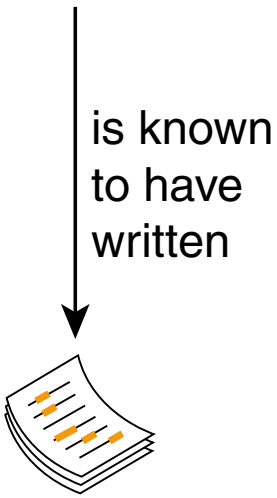
Martin Potthast, Matthias Hagen, and Benno Stein
Bauhaus-Universität Weimar

www.webis.de

Obfuscation vs. Identification

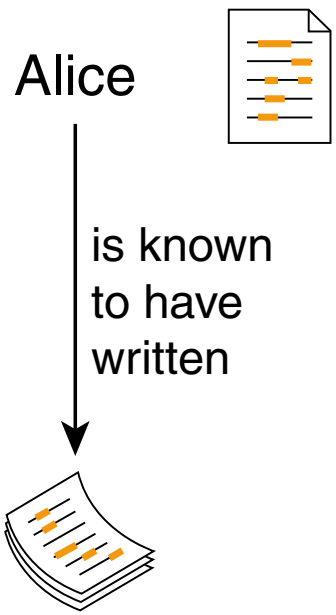
Introduction

Alice



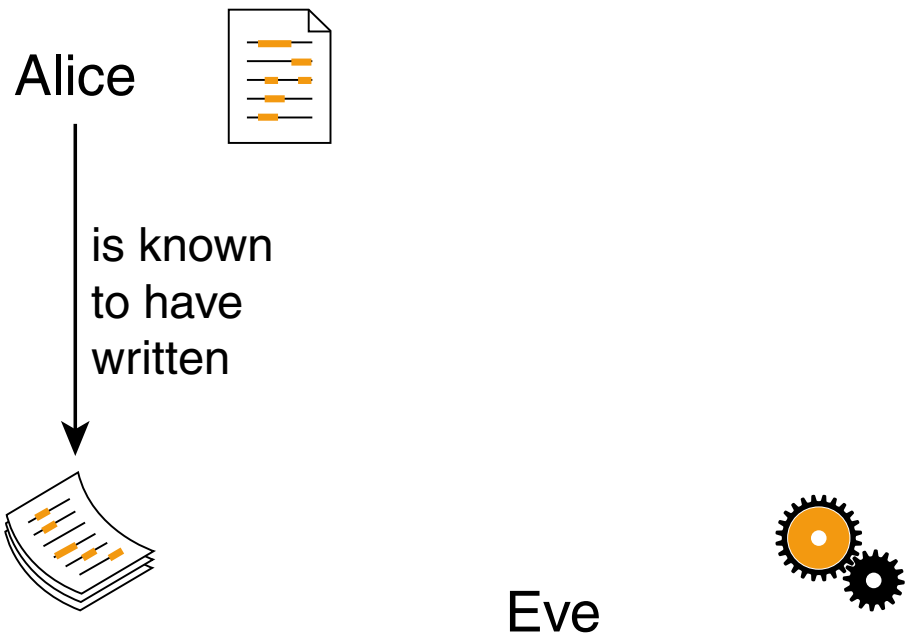
Obfuscation vs. Identification

Introduction



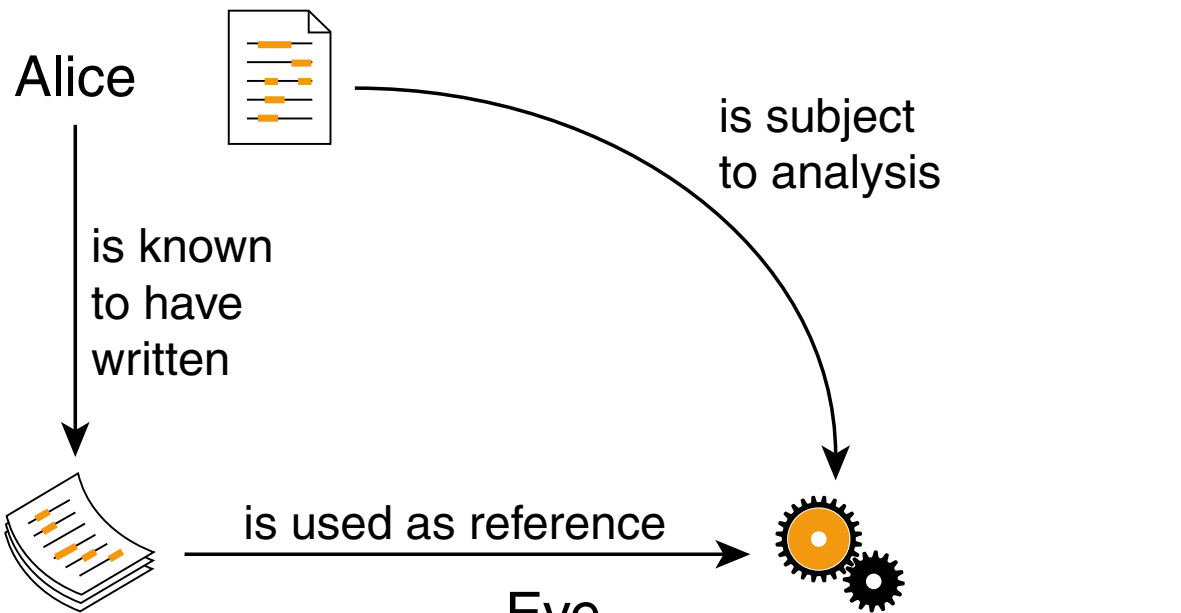
Obfuscation vs. Identification

Introduction



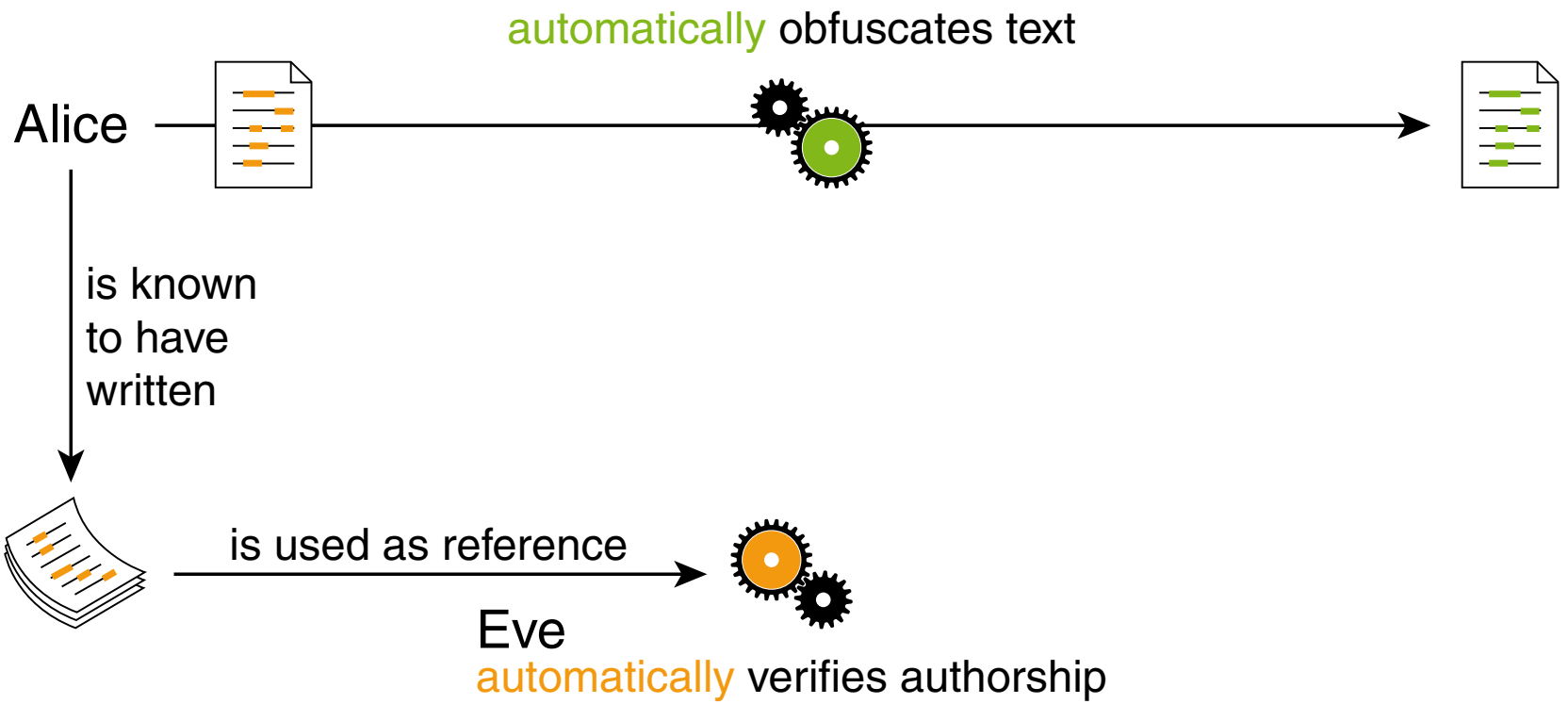
Obfuscation vs. Identification

Introduction



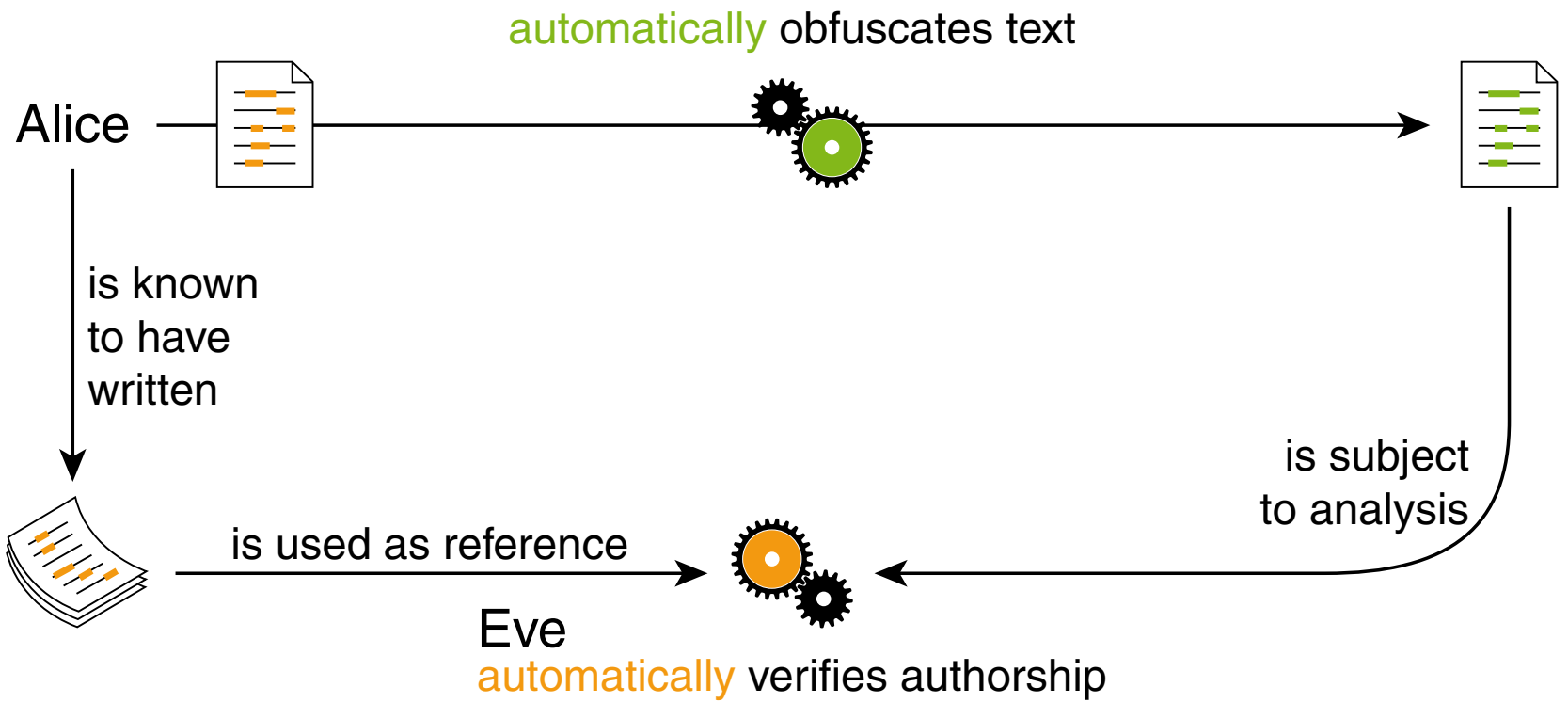
Obfuscation vs. Identification

Introduction



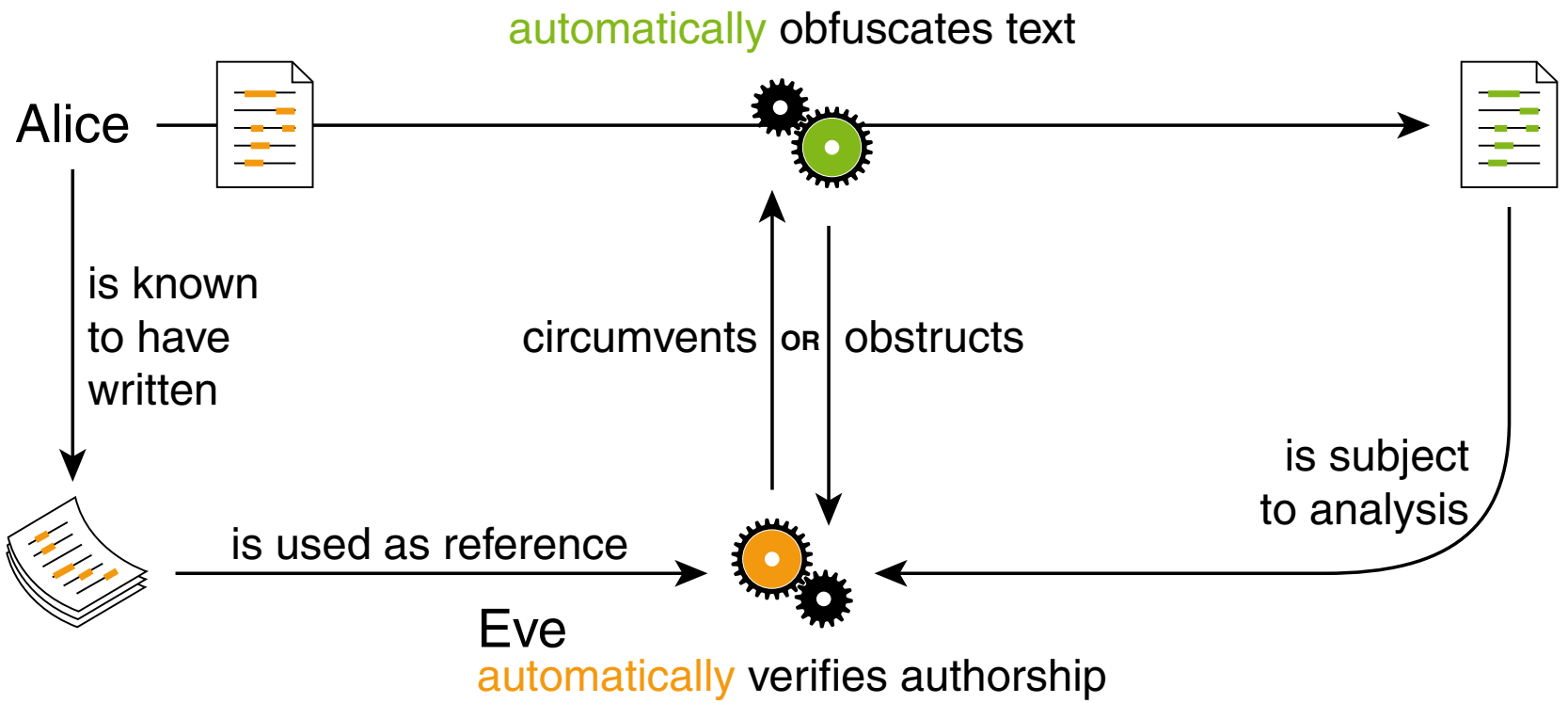
Obfuscation vs. Identification

Introduction



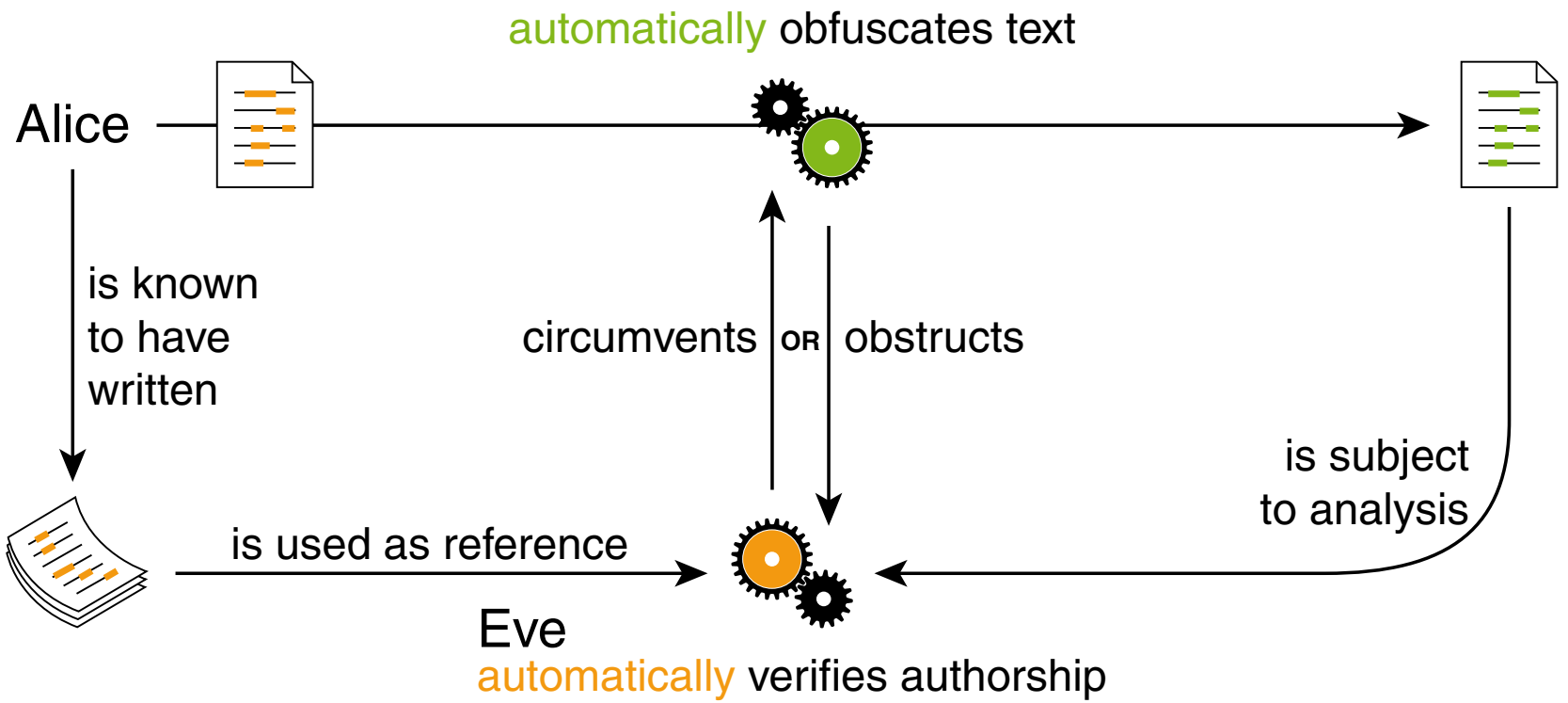
Obfuscation vs. Identification

Introduction



Obfuscation vs. Identification

Introduction



Author masking:

Given two documents by the same author, paraphrase the designated one so that the author cannot be verified anymore.

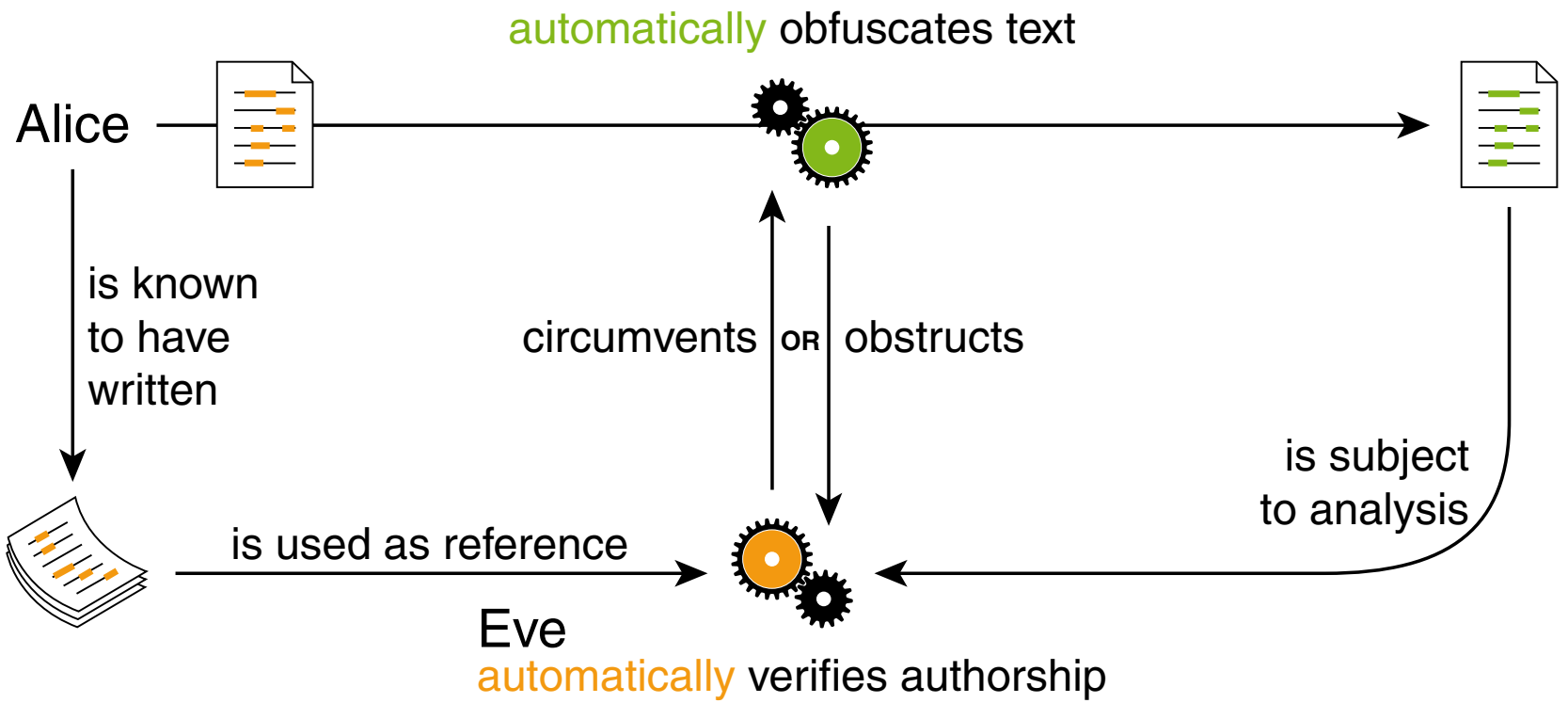
vs.

Authorship verification:

Given two documents, decide whether they have been written by the same author.

Obfuscation vs. Identification

Introduction



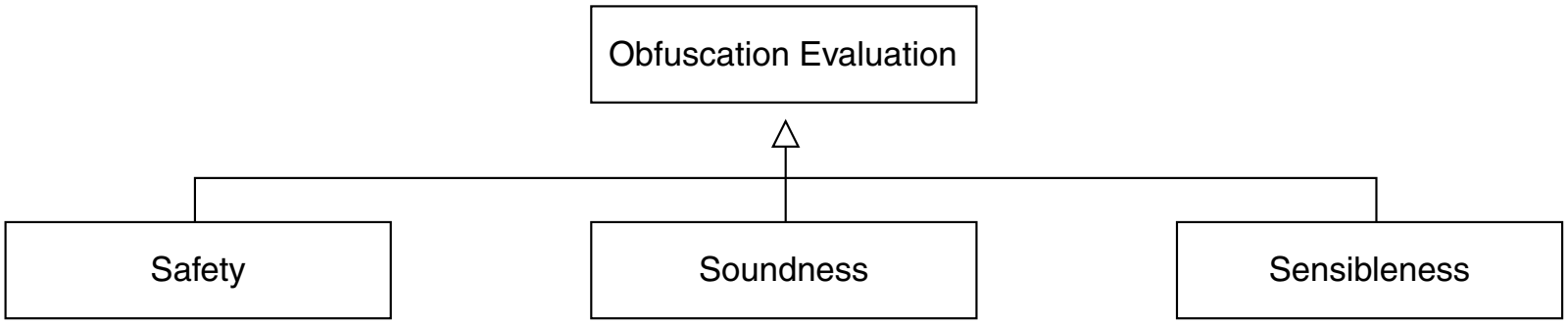
Key questions

- How vulnerable are verification approaches to obfuscation?
- How successful are obfuscation approaches against verification?
- Which technology will ultimately dominate the other?

Obfuscation Evaluation

Obfuscation Evaluation

Taxonomy of Evaluation Dimensions

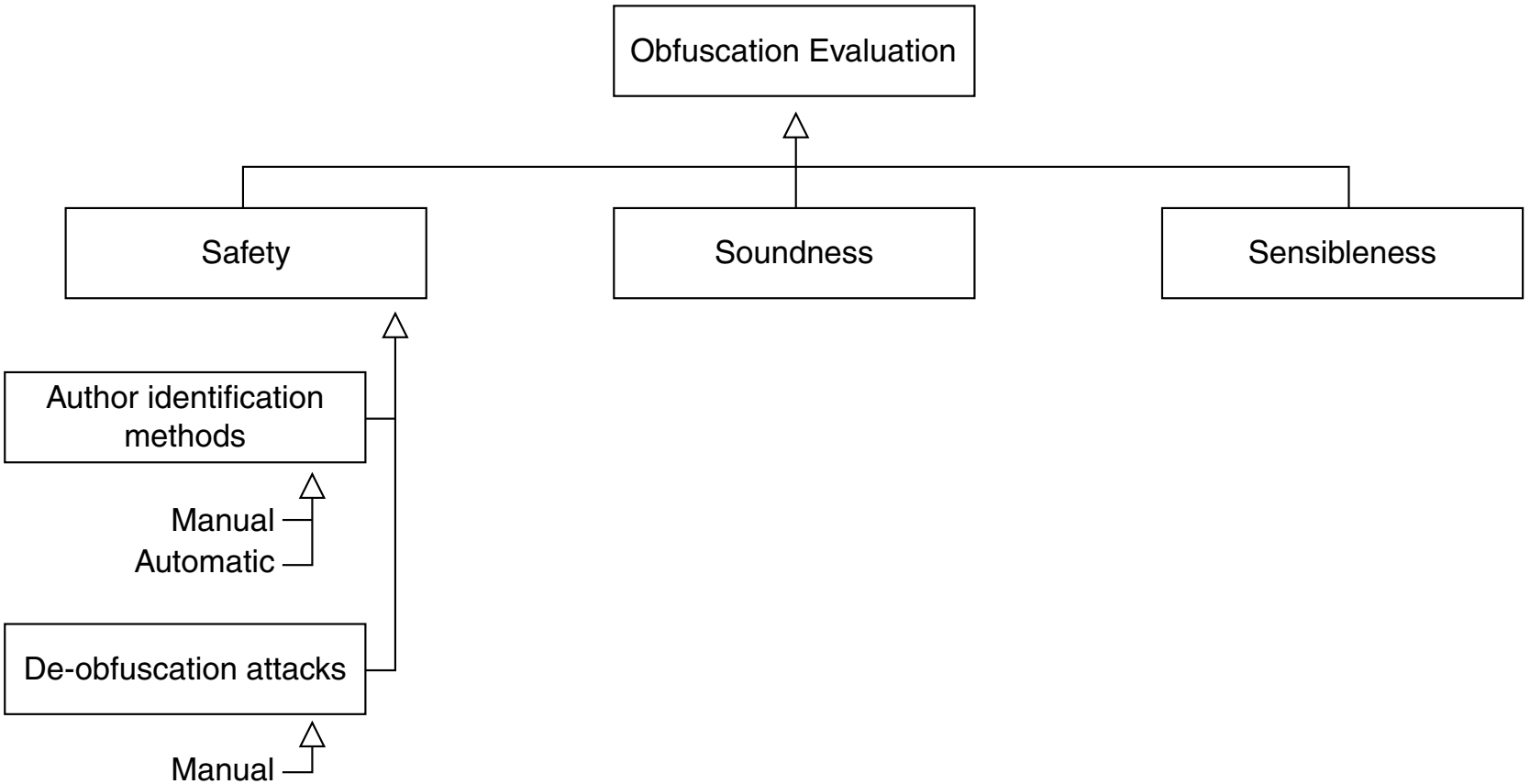


We call an obfuscation software

- **safe**, if its obfuscated texts can not be attributed to their original authors,
- **sound**, if its obfuscated texts are textually entailed by their originals, and
- **sensible**, if its obfuscated texts are well-formed and inconspicuous.

Obfuscation Evaluation

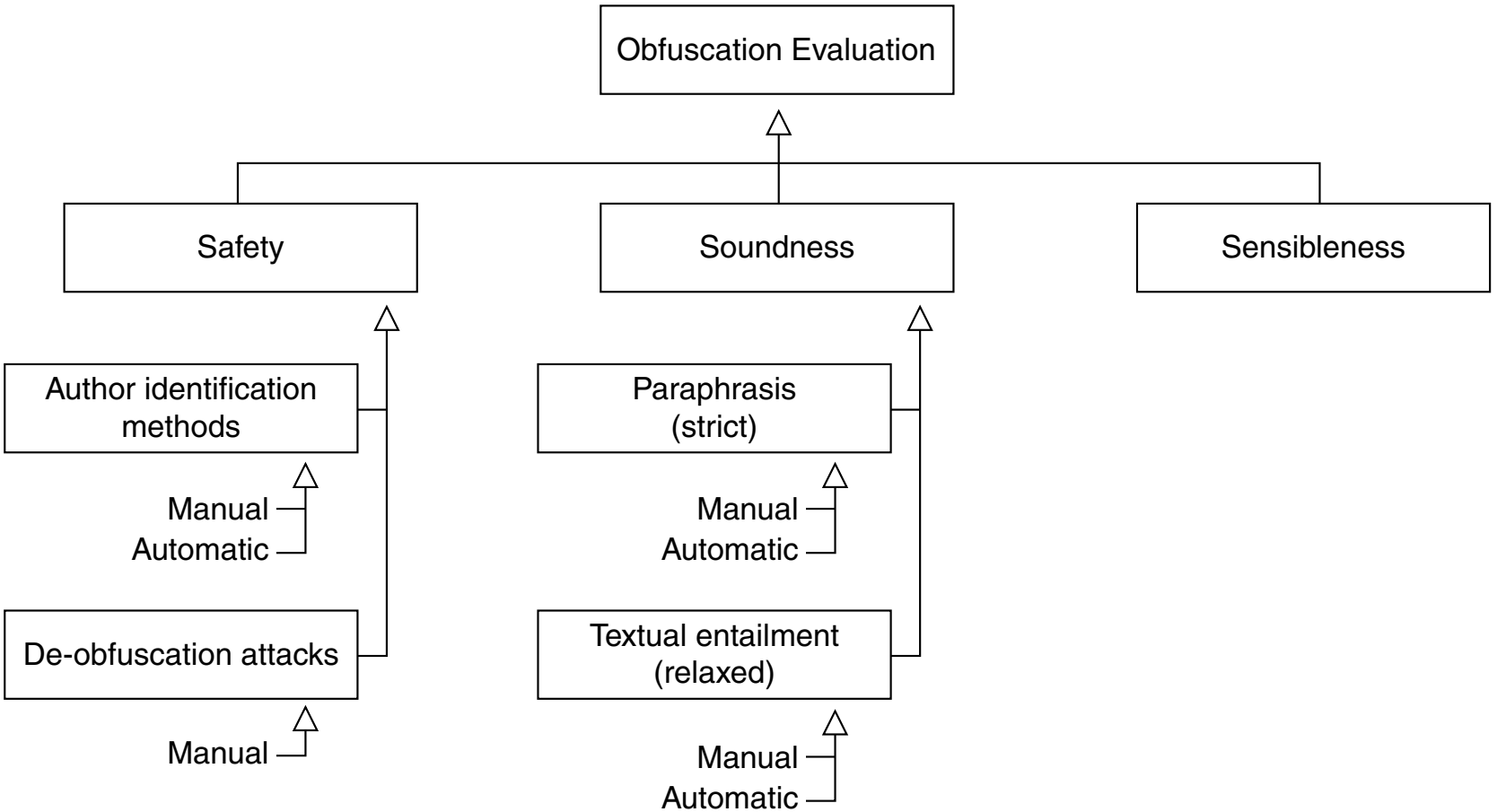
Taxonomy of Evaluation Dimensions



- Manual safety evaluation against forensic linguists not scalable
- Automatic safety evaluation requires large amount of implementations
- Several obfsucation approaches can be undone

Obfuscation Evaluation

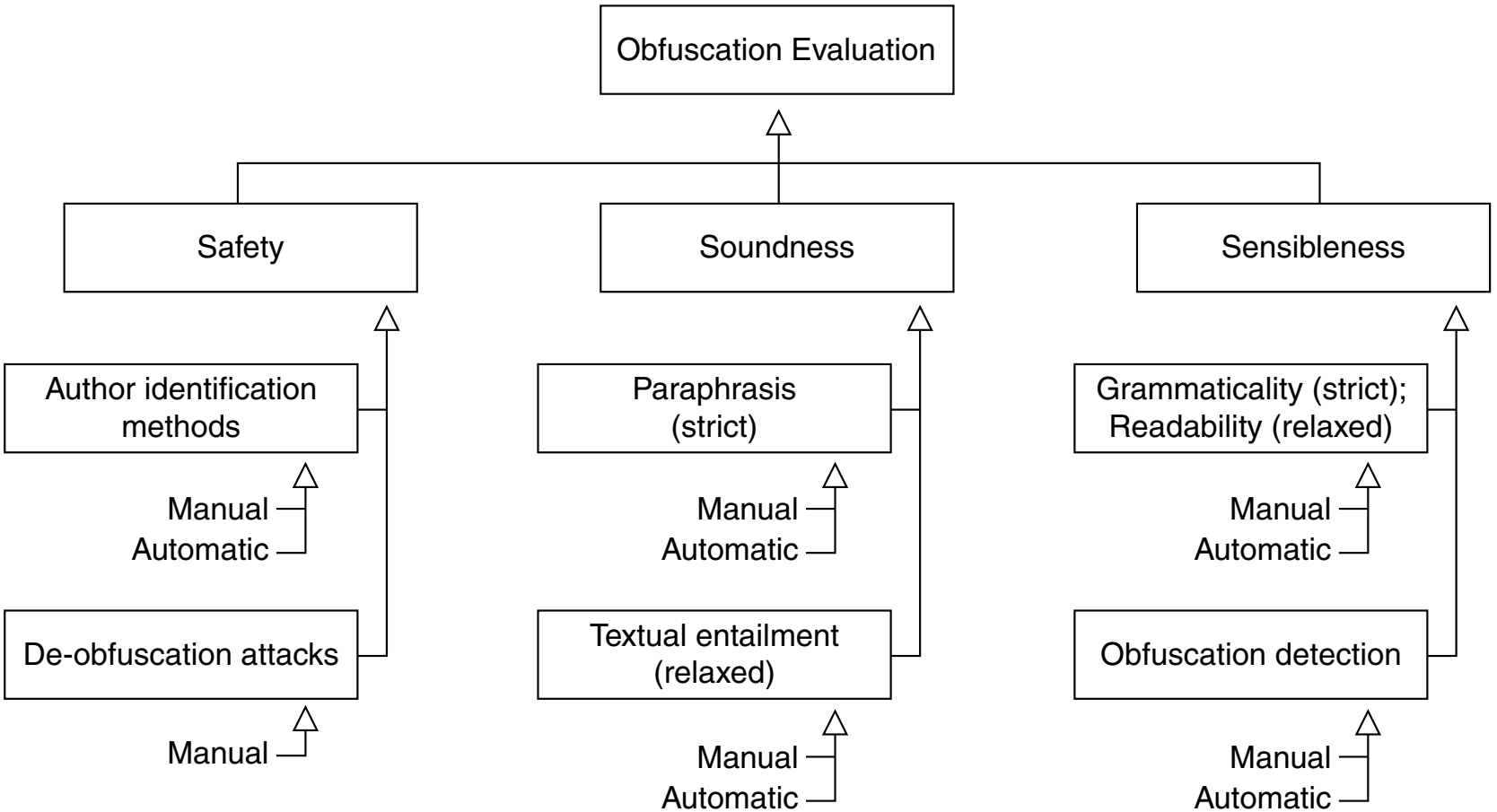
Taxonomy of Evaluation Dimensions



- Paraphrase: obfuscation restates the original with different words
- Textual entailment: obfuscation follows logically from original
- Support manual review with visual text comparison

Obfuscation Evaluation

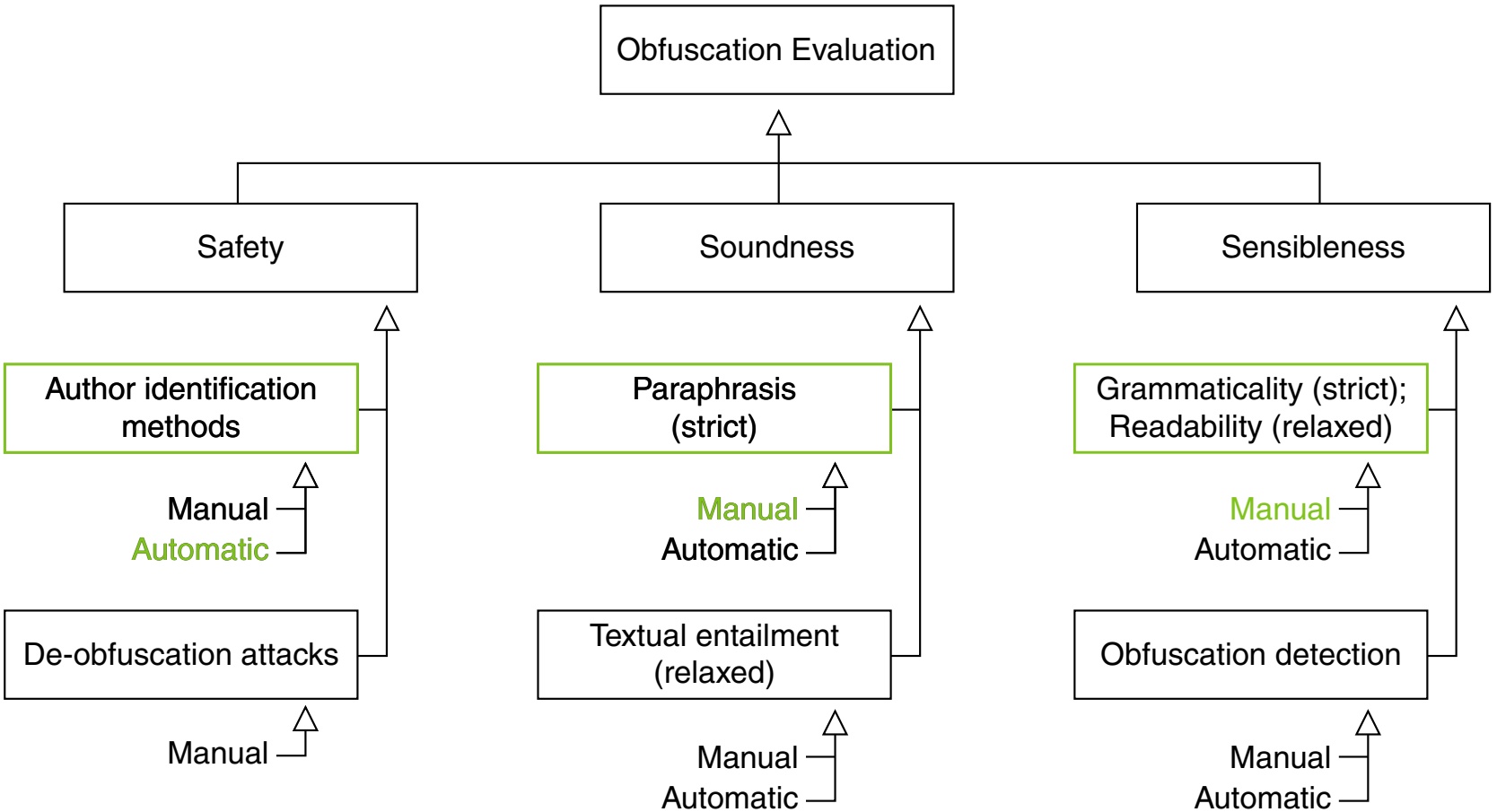
Taxonomy of Evaluation Dimensions



- Relax grammaticality: machine translation also not perfect, yet useful
- Hiding obfuscation useful to avoid in-depth (manual) forensic analysis
- Automatic evaluation involves cutting edge research

Obfuscation Evaluation

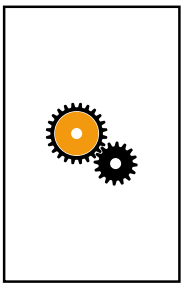
Taxonomy of Evaluation Dimensions

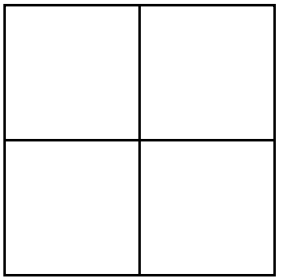


- Evaluations conducted in our shared task

Obfuscation Evaluation

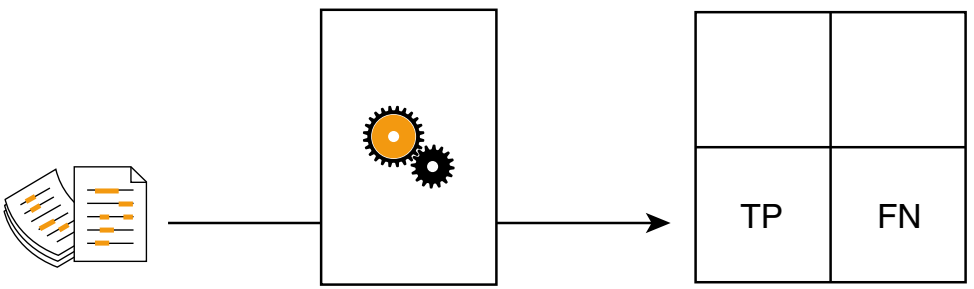
Shared Task Setup





Obfuscation Evaluation

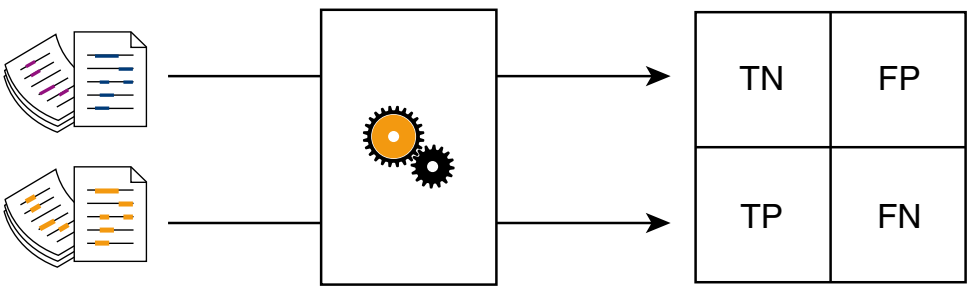
Shared Task Setup



Obfuscation Evaluation

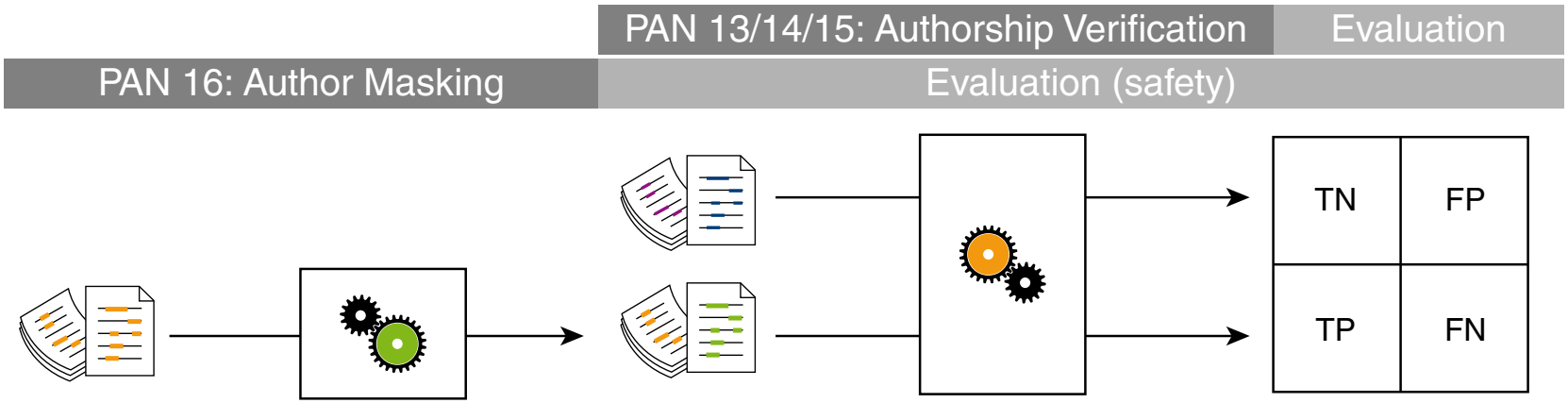
Shared Task Setup

PAN 13/14/15: Authorship Verification | Evaluation



Obfuscation Evaluation

Shared Task Setup

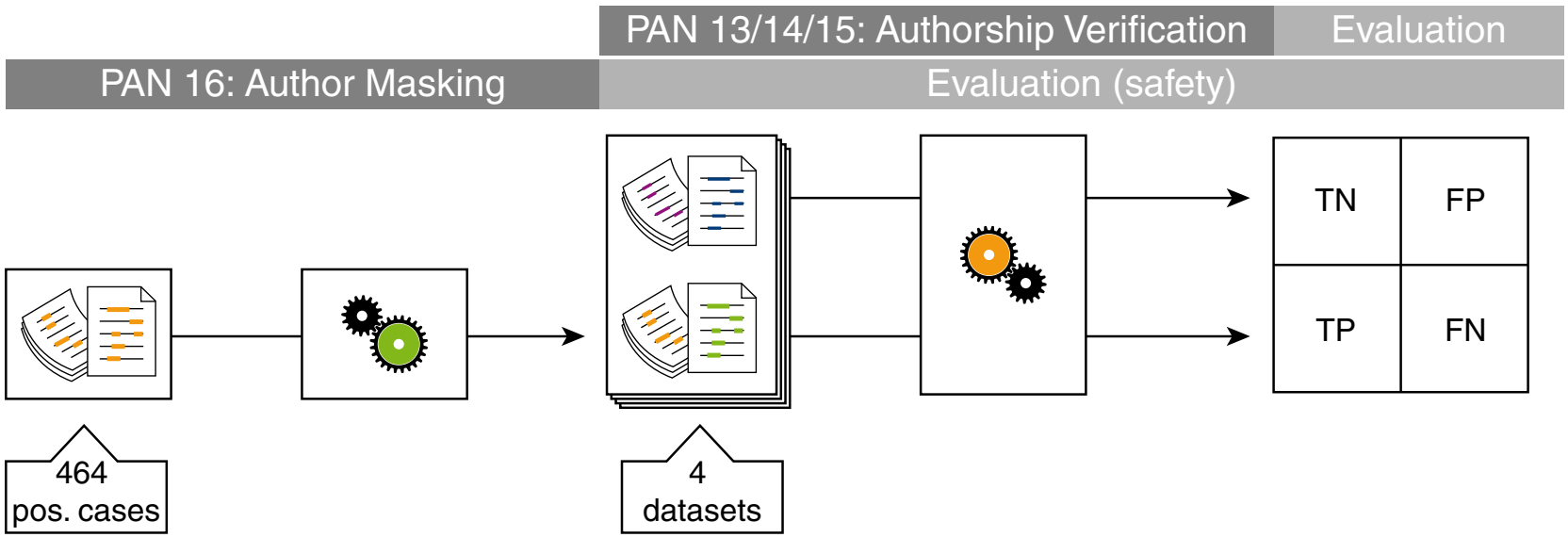


This setup tells us

- whether an obfuscator can defeat a verifier

Obfuscation Evaluation

Shared Task Setup

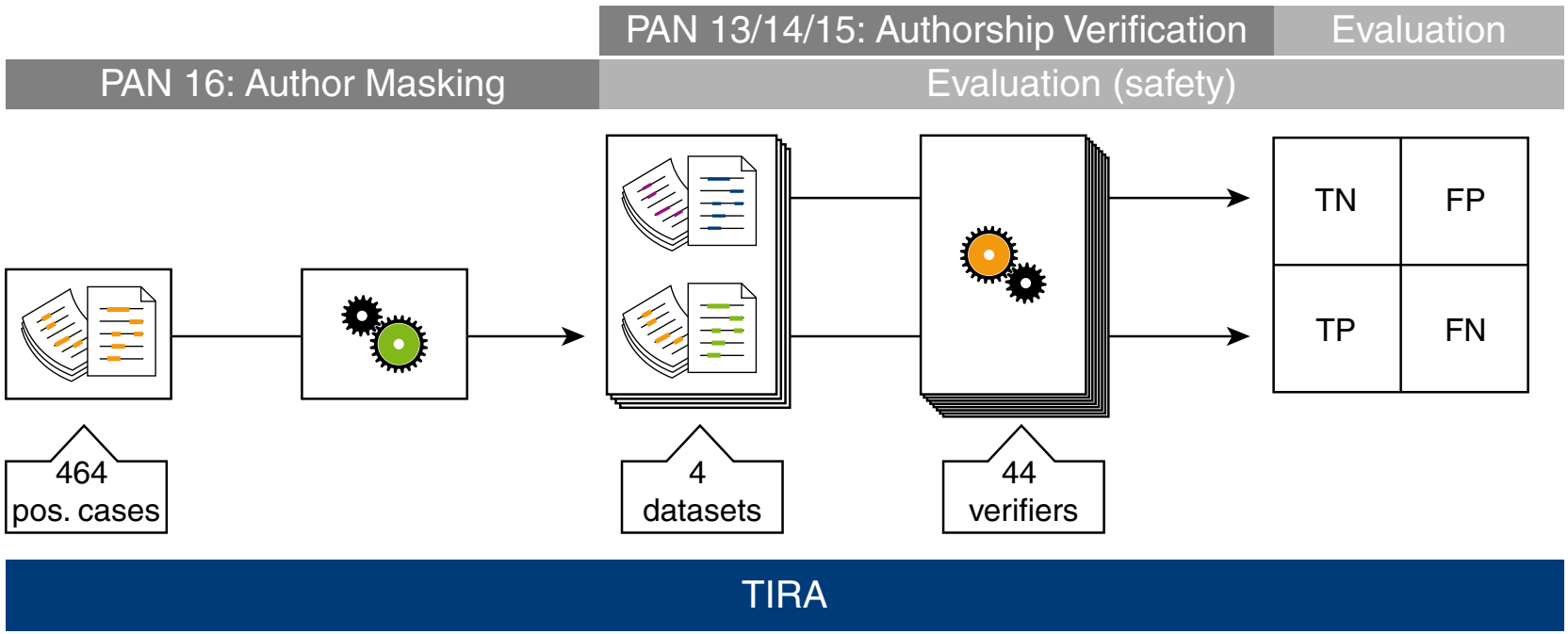


This setup tells us

- whether an obfuscator can defeat a verifier
- whether an obfuscator can defeat a verifier in general

Obfuscation Evaluation

Shared Task Setup

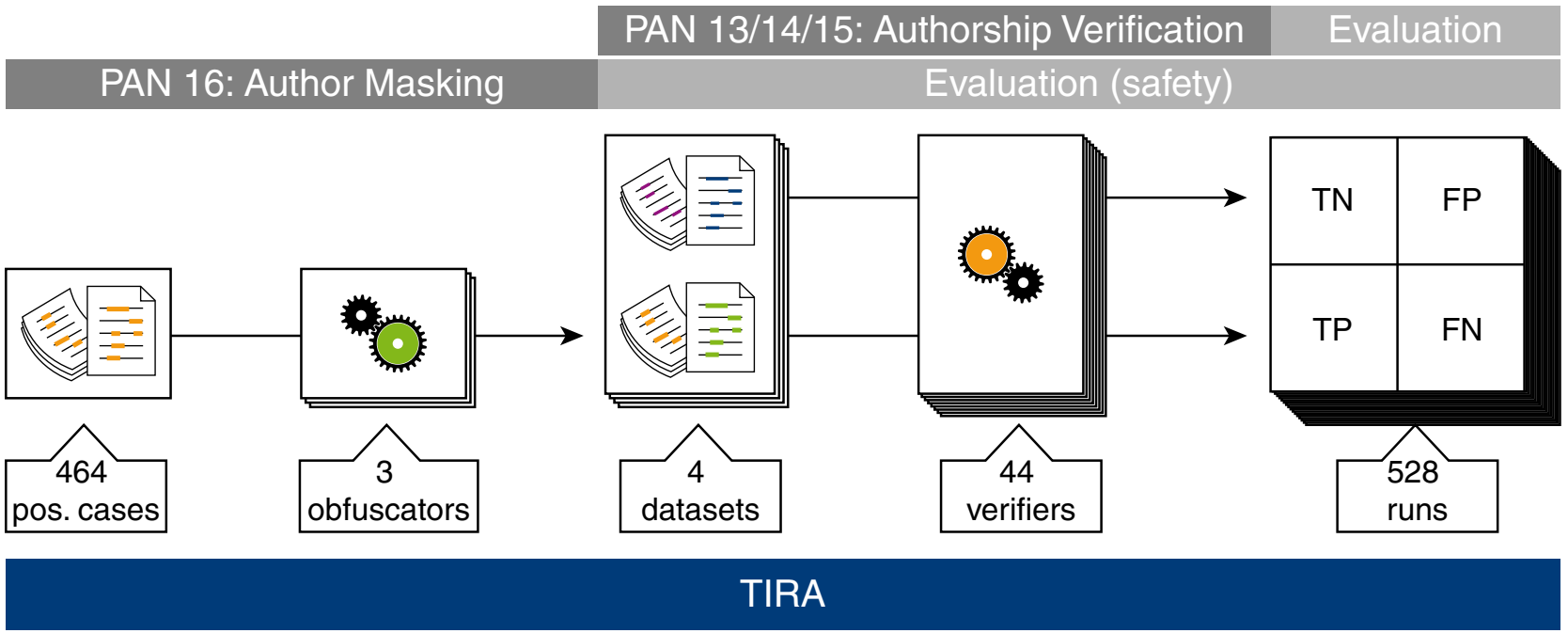


This setup tells us

- whether an obfuscator can defeat a verifier
- whether an obfuscator can defeat a verifier in general
- whether an obfuscator can defeat verifiers in general

Obfuscation Evaluation

Shared Task Setup



This setup tells us

- whether an obfuscator can defeat a verifier
- whether an obfuscator can defeat a verifier in general
- whether an obfuscator can defeat verifiers in general
- whether obfuscators can defeat verifiers in general

Obfuscation Evaluation

Measuring Obfuscation Impact

Performance without obfuscation

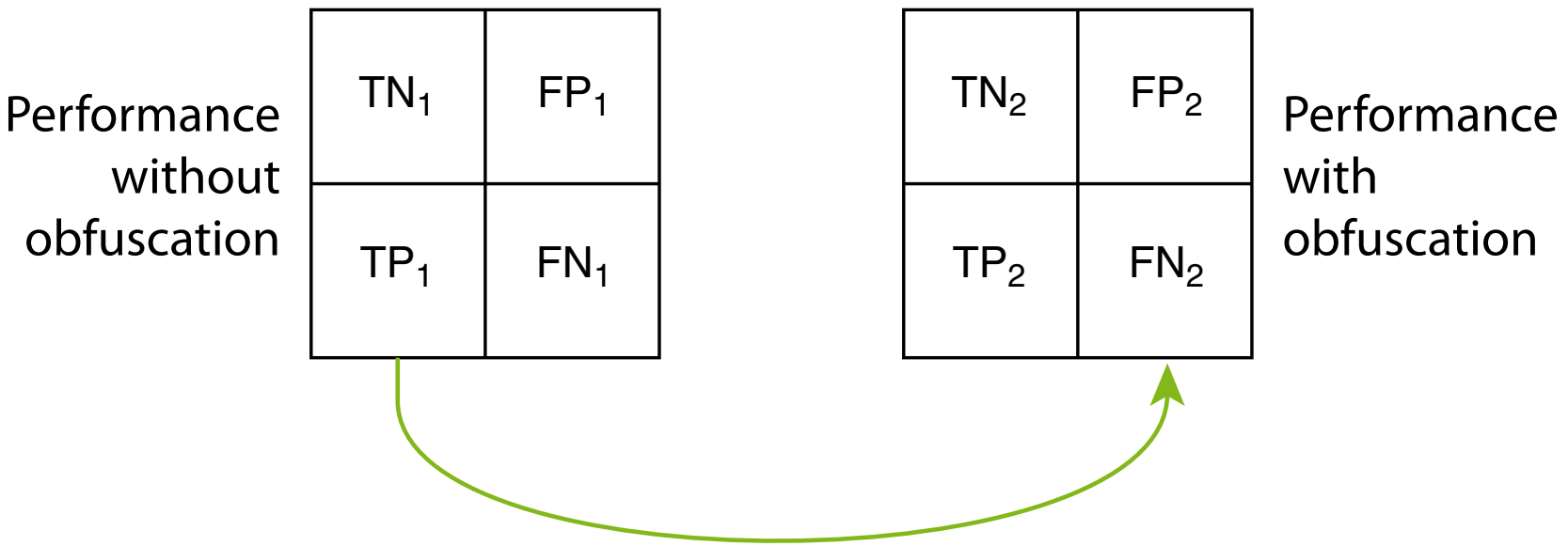
TN_1	FP_1
TP_1	FN_1

TN_2	FP_2
TP_2	FN_2

Performance with obfuscation

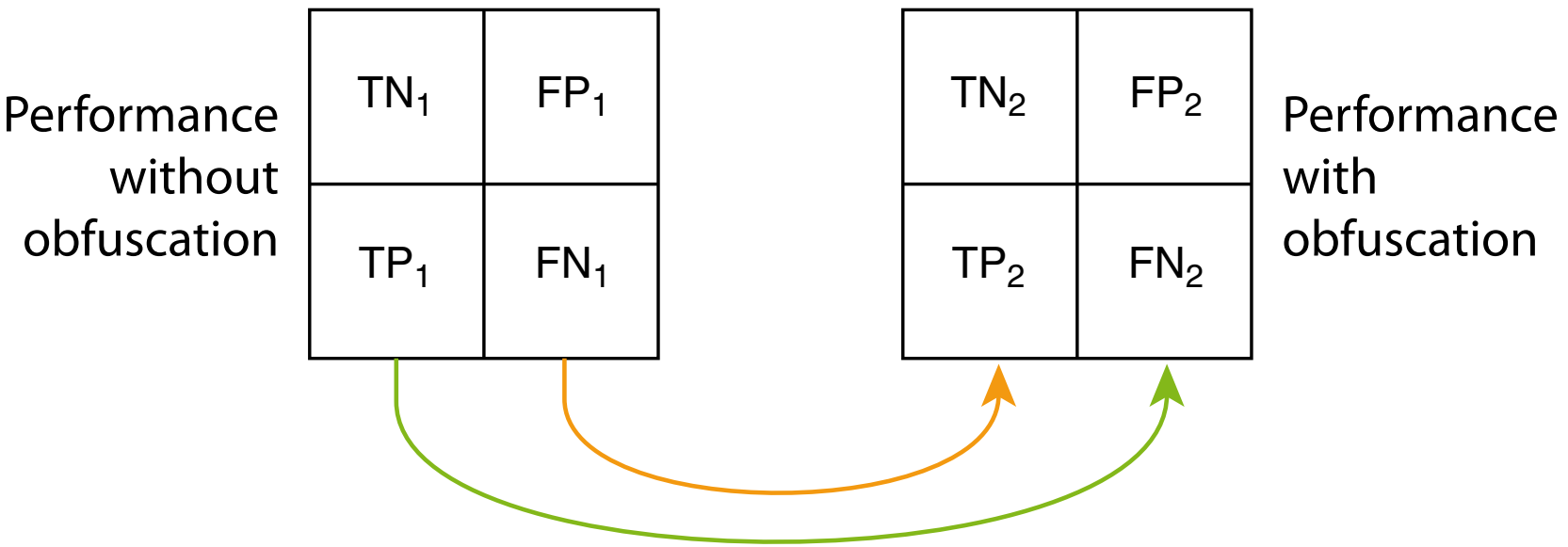
Obfuscation Evaluation

Measuring Obfuscation Impact



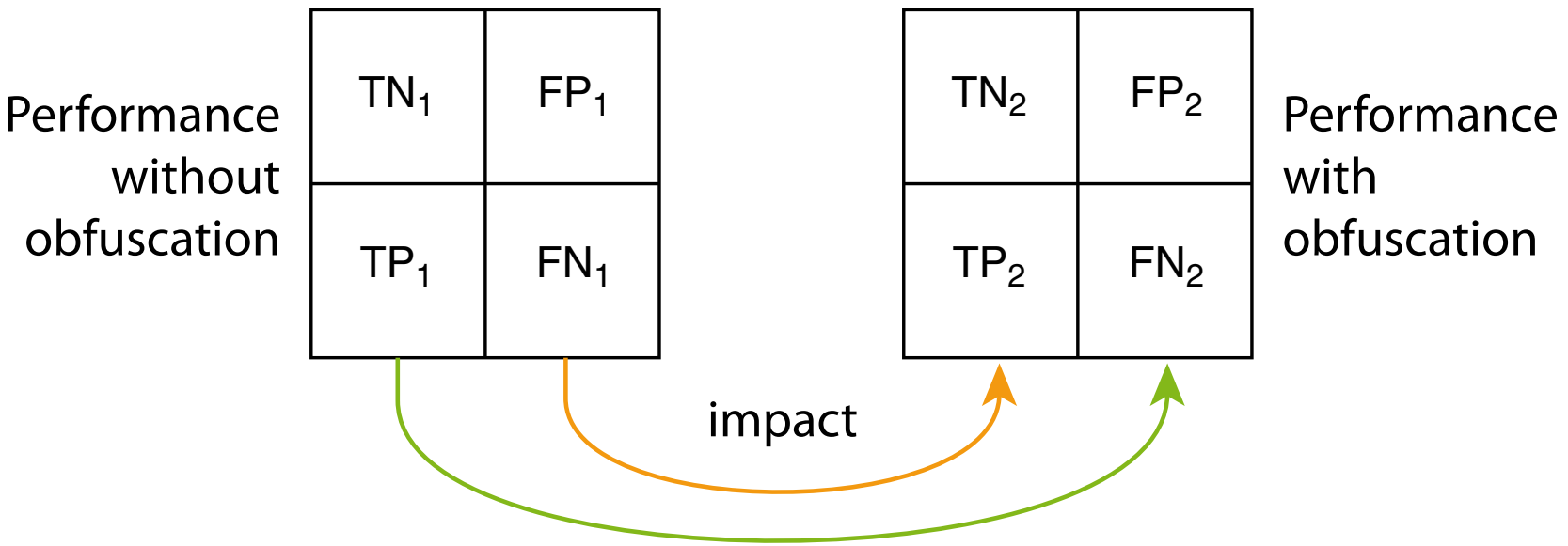
Obfuscation Evaluation

Measuring Obfuscation Impact



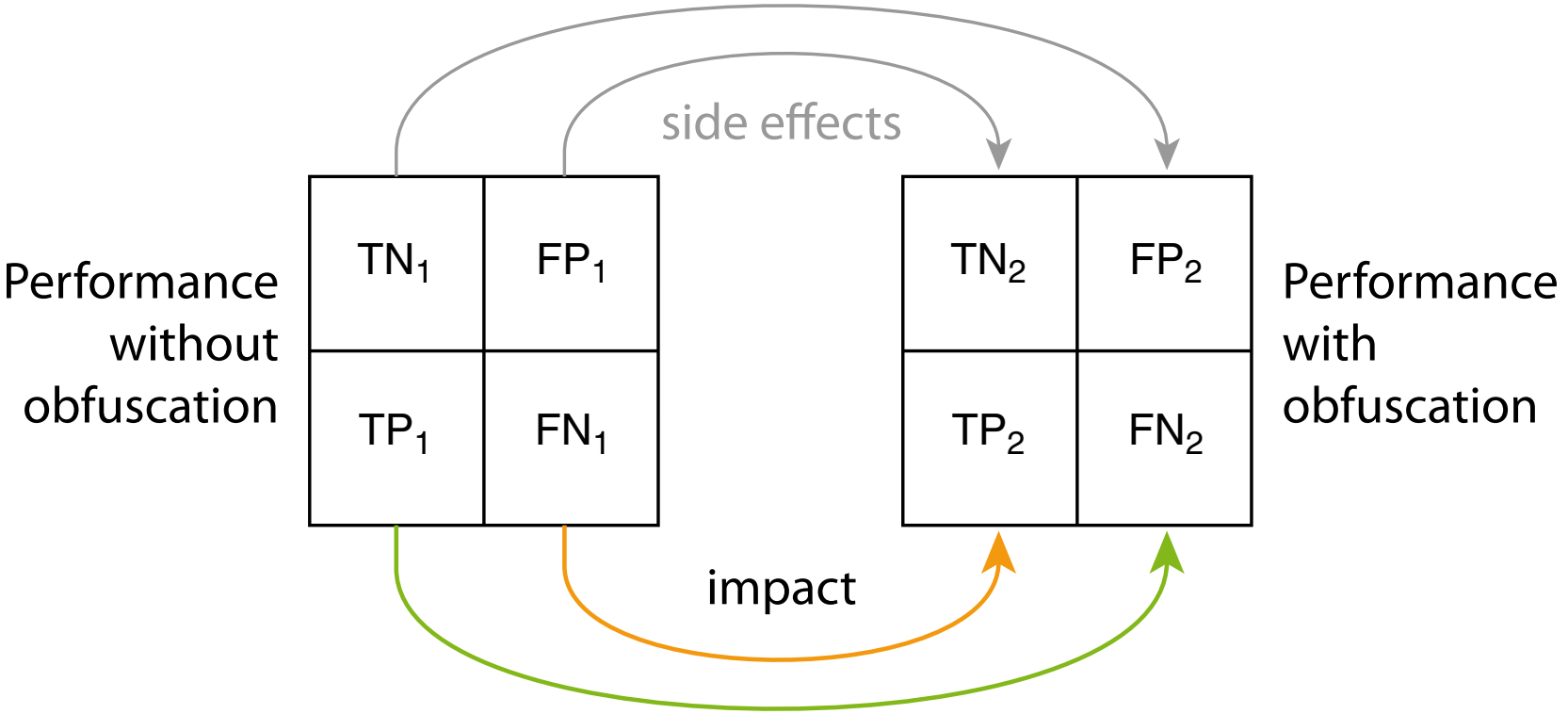
Obfuscation Evaluation

Measuring Obfuscation Impact



Obfuscation Evaluation

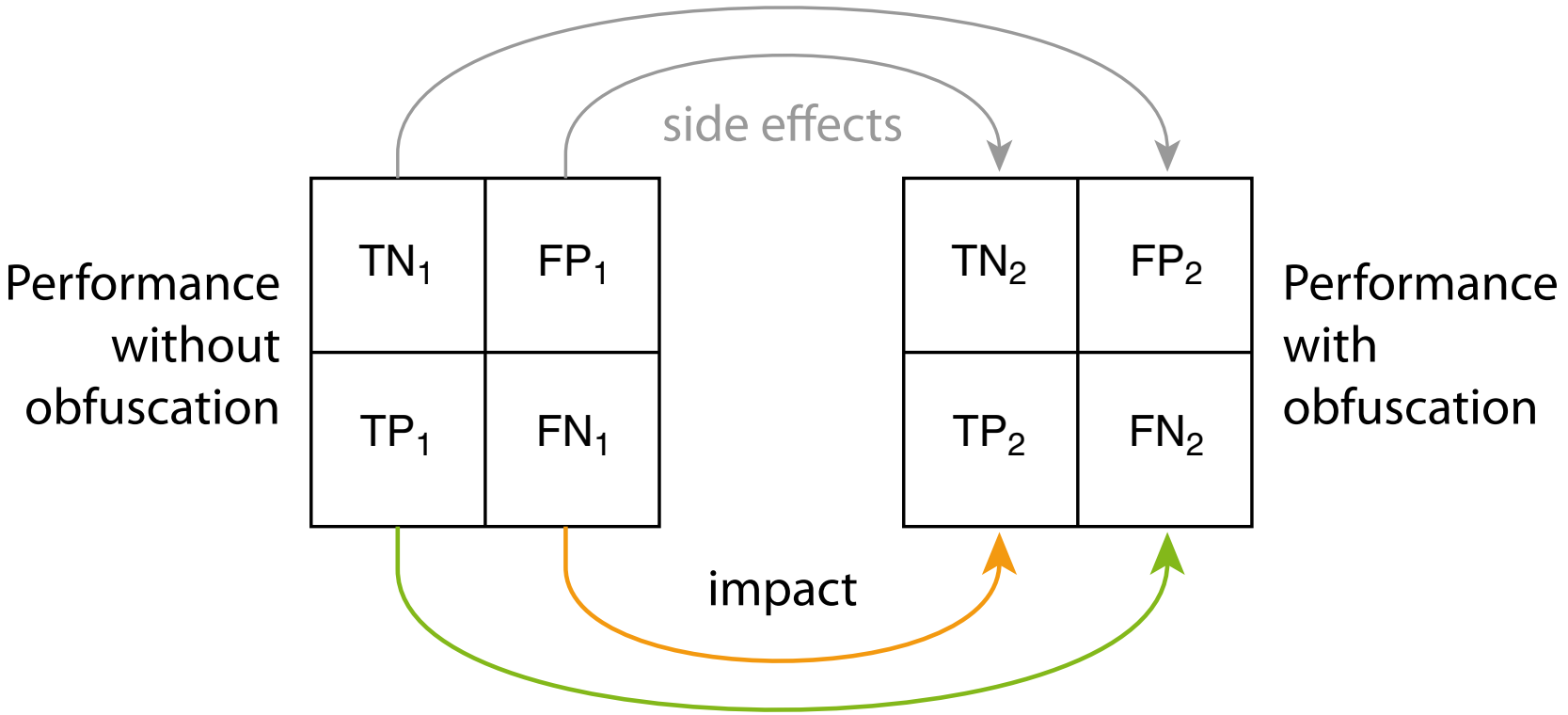
Measuring Obfuscation Impact



- Side effects indicate that the verifier employs corpus-relative features
- Corpus-relative features are an anti-pattern since verification cases do not come in groups

Obfuscation Evaluation

Measuring Obfuscation Impact

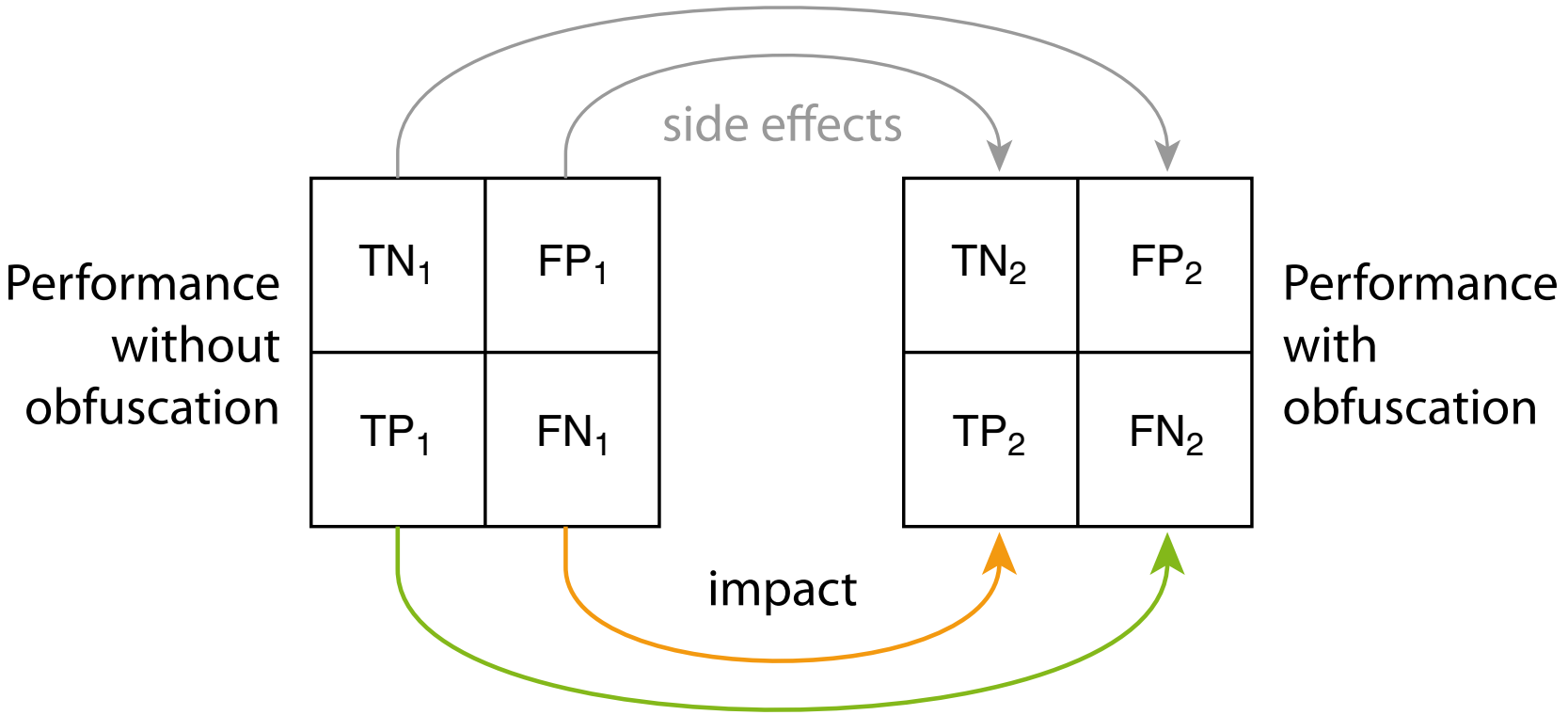


$$\mathrm{rec}_1 = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FN}_1} \qquad \mathrm{rec}_2 = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FN}_2}$$

$$\Delta_{\mathrm{rec}} = \mathrm{rec}_2 - \mathrm{rec}_1$$

Obfuscation Evaluation

Measuring Obfuscation Impact



$$\text{imp} = \begin{cases} -\frac{\Delta_{\text{rec}}}{\text{rec}_1} & \text{if } \Delta_{\text{rec}} < 0, \\ -\frac{\Delta_{\text{rec}}}{1-\text{rec}_1} & \text{else.} \end{cases}$$

Obfuscation Evaluation

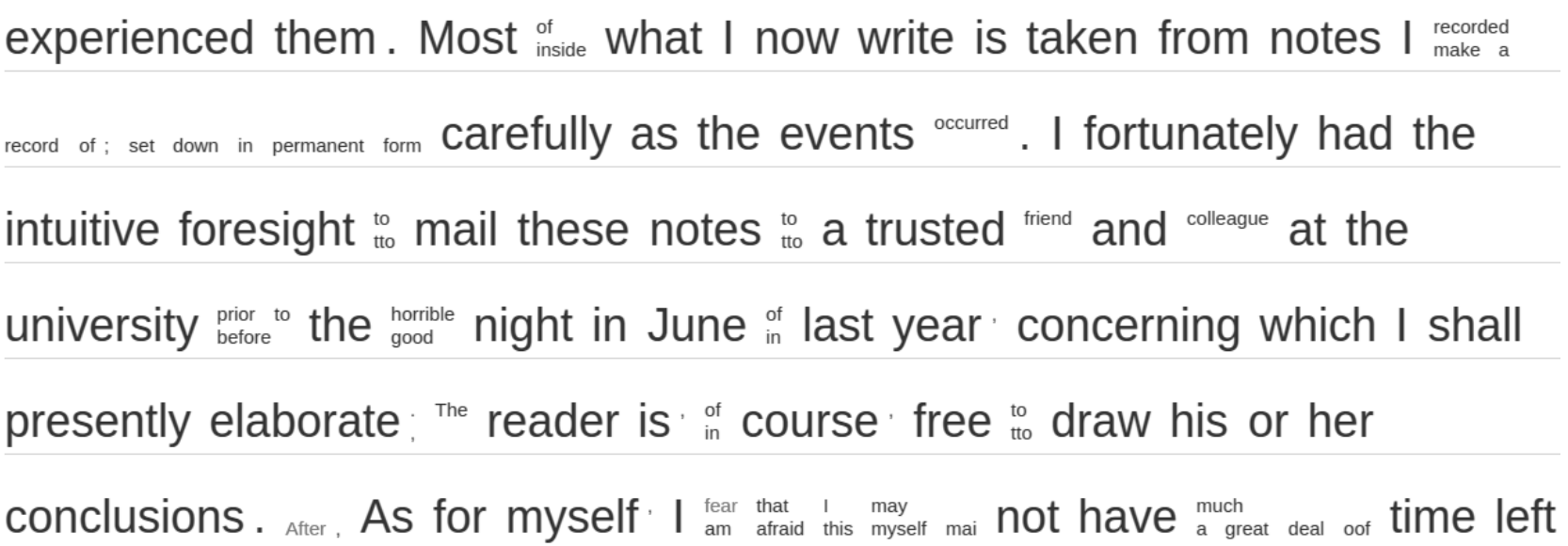
Safety Evaluation Results

Obfuscator	**Dataset**	**Pos. cases**	avg $\triangle_{rec}$	avg imp
Mihaylova *et al.*	PAN13	14	**-0.2778**	**0.4690**
Keswani *et al.*	PAN13	14	-0.2361	0.4245
Mansoorizadeh *et al.*	PAN13	14	-0.0933	0.1442
Mihaylova *et al.*	PAN14 EE	100	**-0.2304**	**0.4891**
Keswani *et al.*	PAN14 EE	100	-0.1873	0.4058
Mansoorizadeh *et al.*	PAN14 EE	100	-0.1038	0.2512
Mihaylova *et al.*	PAN14 EN	100	**-0.2456**	**0.4750**
Keswani *et al.*	PAN14 EN	100	-0.1783	0.3769
Mansoorizadeh *et al.*	PAN14 EN	100	-0.0958	0.2345
Mihaylova *et al.*	PAN15	250	**-0.2009**	**0.3649**
Keswani *et al.*	PAN15	250	-0.1298	0.2543
Mansoorizadeh *et al.*	PAN15	250	-0.0994	0.1952

Obfuscation Approaches

Obfuscation Approaches

Mihaylova et al.



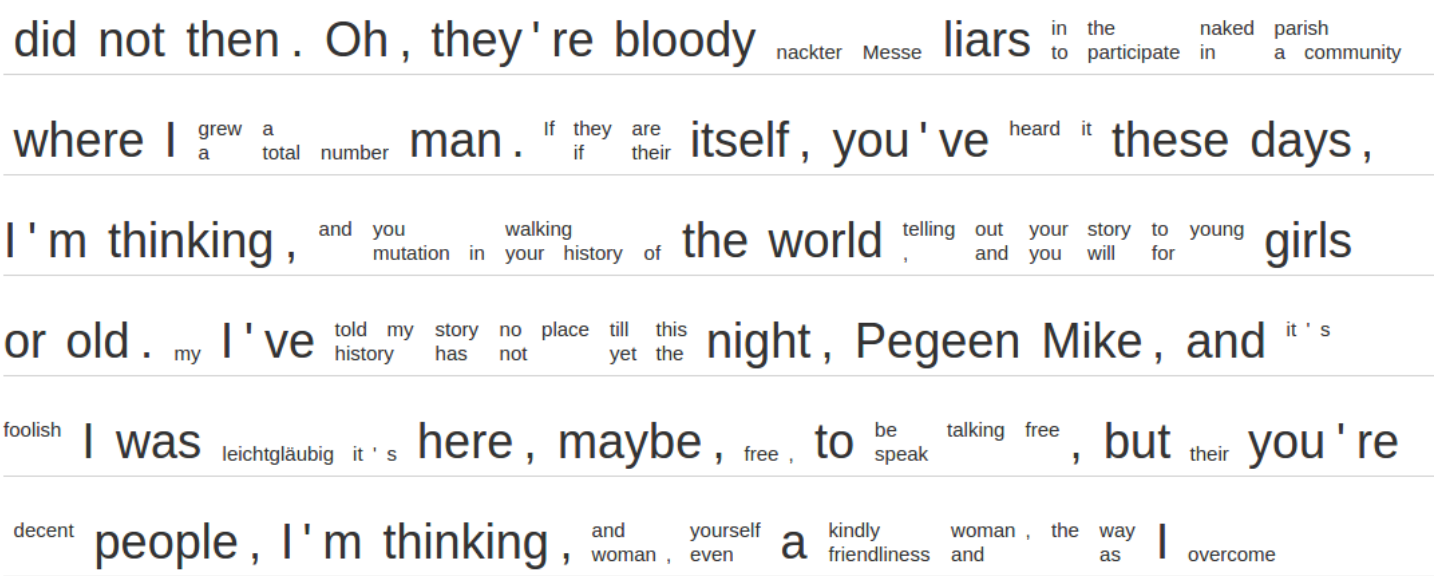
- Targets 6 style features: sentence length, punctuation, stop words, parts of speech, all caps, word frequencies
- Computation of expected values based on background corpus
- Obfuscation towards the average using rule-based text operations; 14 rules

Observations

- Unfitting replacements, semantic distortions, overdone error insertion

Obfuscation Approaches

Keswani et al.



- Round-trip translation English – German – French – English
- Based on Moses SMT toolkit, trained on Europarl corpus

Observations

- Fragments of non-English text remain from translation
- Europarl corpus ill-suited for the genres of the test datasets
- Text unsound and unreadable

Obfuscation Approaches

Mansoorizadeh et al.

run - time system organization called the JRE . This approach has some advantages and disadvantages and it is worth comparing these three options in order to appreciate the implications for the Java coffee developer . The compiler translates the source code into machine car code for the relevant hardware / OS combination . Strictly speaking there are two stages : compilation of program units (usually files) , followed by ‘ linking ’ when the complete accomplished

- Conservative paraphrasing: at most 1 word per sentence replaced
- Replacement candidates chosen among 200 most frequent words
- Replacements obtained from WordNet, based on word similarity
- Selection based on commonness under 4-gram language model

Observations

- Lots of problems with genre-specific terminology

Obfuscation Evaluation

Peer Evaluation

- Given training runs of each obfuscator, evaluate their performance
- Participants anonymized; yet, self-identification likely
- 3 participating teams, 1 independent (i.e., without obfuscator submission): Mihaylova et al., Mansoorizadeh et al., and Liebeck et al.

Obfuscation Evaluation

Peer Evaluation

- Given training runs of each obfuscator, evaluate their performance
- Participants anonymized; yet, self-identification likely
- 3 participating teams, 1 independent (i.e., without obfuscator submission): Mihaylova et al., Mansoorizadeh et al., and Liebeck et al.

Safety

- In all cases, based on GLAD verification system
- 2 rank Mihaylova $\succ$ Keswani $\succ$ Mansoorizadeh (in agreement with us)
- Mansoorizadeh ranks Keswani $\succ$ Mansoorizadeh $\succ$ Mihaylova

Obfuscation Evaluation

Peer Evaluation

- Given training runs of each obfuscator, evaluate their performance
- Participants anonymized; yet, self-identification likely
- 3 participating teams, 1 independent (i.e., without obfuscator submission): Mihaylova et al., Mansoorizadeh et al., and Liebeck et al.

Safety

- In all cases, based on GLAD verification system
- 2 rank Mihaylova $\succ$ Keswani $\succ$ Mansoorizadeh (in agreement with us)
- Mansoorizadeh ranks Keswani $\succ$ Mansoorizadeh $\succ$ Mihaylova

Soundness

- Liebeck employ SemEval STS measure, scoring from 5 (best) to 0 (worst)
- All obfuscators score >4
- Mihaylova and Mansoorizadeh assign scores manually
- 2 rank Mansoorizadeh $\succ$ Mihaylova $\succ$ Keswani (in agreement with us)
- Mihaylova ranks Mihaylova $\succ$ Mansoorizadeh $\succ$ Keswani

Obfuscation Evaluation

Peer Evaluation (cont'd)

Sensibleness

- All assessors resort to manual review
- Scales include 0-5 (Mansoorizadeh), 0-2 (Liebeck), 0,1 (Mihaylova)
- All rank Mansoorizadeh $\succ$ Mihaylova $\succ$ Keswani (in agreement with us)

Obfuscation Evaluation

Peer Evaluation (cont'd)

Sensibleness

- All assessors resort to manual review
- Scales include 0-5 (Mansoorizadeh), 0-2 (Liebeck), 0,1 (Mihaylova)
- All rank Mansoorizadeh $\succ$ Mihaylova $\succ$ Keswani (in agreement with us)

Organizer Evaluation wrt. Soundness and Sensibleness

Sensibleness

- School grading 1-5 (fail) of 1 text per obfuscator per dataset
- Our ranking: Mansoorizadeh (2) $\succ$ Mihaylova (4-5) $\succ$ Keswani (5)

Soundness

- After grading, reading the originals, and comparison with Diff visualization
- Judgment of whole texts on scale 0 ("incorrect"), 1 ("passable"), 2 ("correct")
- Our ranking: Mansoorizadeh (1+) $\succ$ Mihaylova (0+) $\succ$ Keswani (0)

Conclusion

Summary

- 3 author obfuscators vs. 44 authorship verifiers in 4 settings
- Authorship verifiers represent the state of the art as per PAN'13/14/15
- Obfuscators flip on average from 20% up to 49% of true positive decisions
- Even conservative obfuscation has significant impact

Conclusion

Summary

- 3 author obfuscators vs. 44 authorship verifiers in 4 settings
- Authorship verifiers represent the state of the art as per PAN'13/14/15
- Obfuscators flip on average from 20% up to 49% of true positive decisions
- Even conservative obfuscation has significant impact

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly any authorship researcher considers obfuscation a threat
- Better ways of assessing soundness and sensibleness at scale needed

Conclusion

Summary

- 3 author obfuscators vs. 44 authorship verifiers in 4 settings
- Authorship verifiers represent the state of the art as per PAN'13/14/15
- Obfuscators flip on average from 20% up to 49% of true positive decisions
- Even conservative obfuscation has significant impact

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly any authorship researcher considers obfuscation a threat
- Better ways of assessing soundness and sensibleness at scale needed

- Author obfuscation and author identification are locked in an instance of the "Potter-Voldemort Conundrum:"

Neither can live while the other survives

Conclusion

Summary

- 3 author obfuscators vs. 44 authorship verifiers in 4 settings
- Authorship verifiers represent the state of the art as per PAN'13/14/15
- Obfuscators flip on average from 20% up to 49% of true positive decisions
- Even conservative obfuscation has significant impact

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly any authorship researcher considers obfuscation a threat
- Better ways of assessing soundness and sensibleness at scale needed

- Author obfuscation and author identification are locked in an instance of the "Potter-Voldemort Conundrum:"

Neither can live while the other survives

Thank you for your attention!