Authorship Analysis and Obfuscation

Matthias Hagen and Martin Potthast

MLU Halle-Wittenberg and Leipzig University

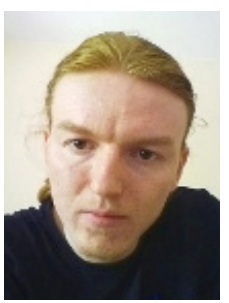
webis.de

Outline

- Introduction
- Technology Basics
- Author Identification
- Constrained Paraphrasing
- Author Obfuscation

Joint work with

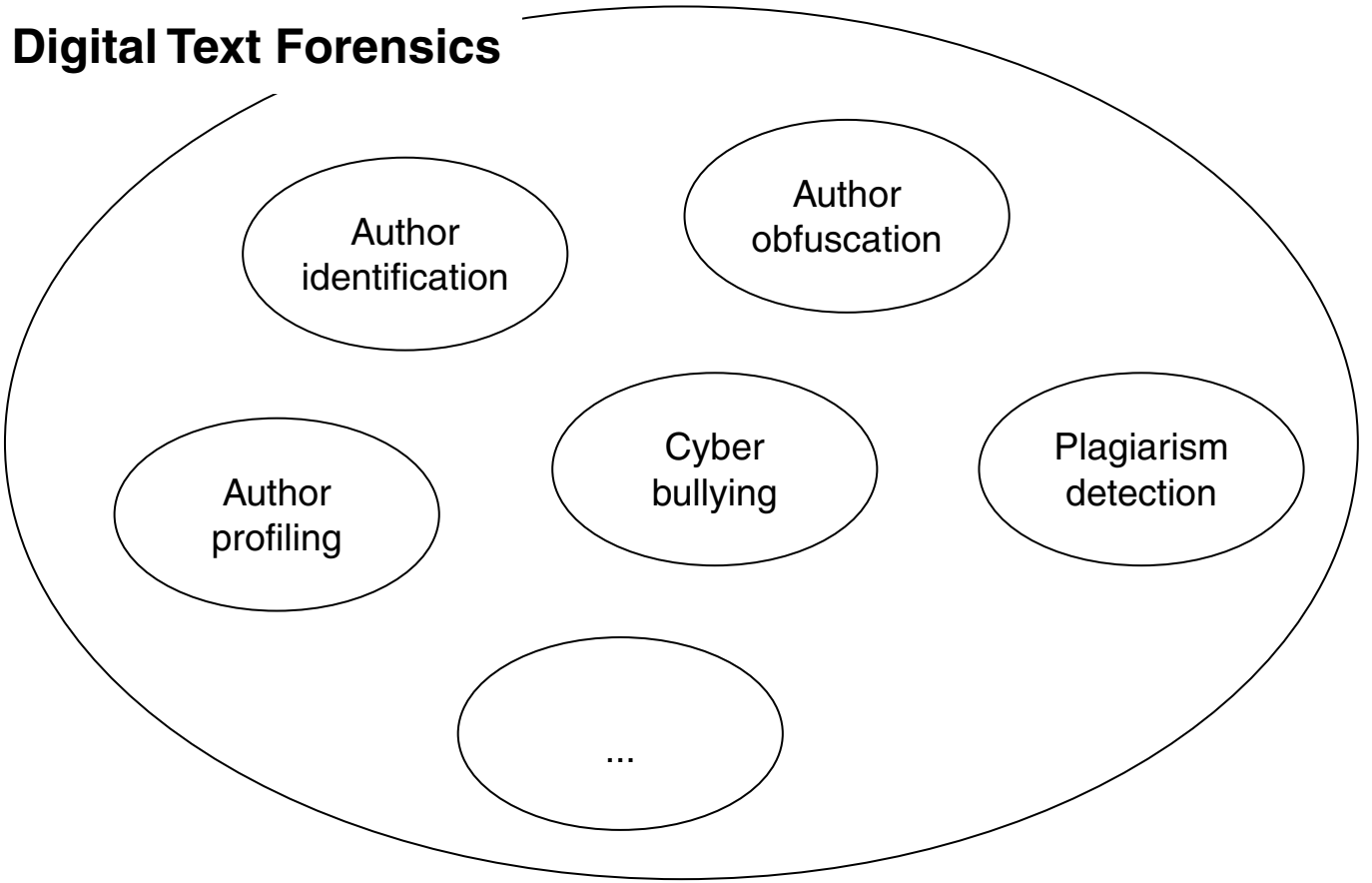
Benno Stein

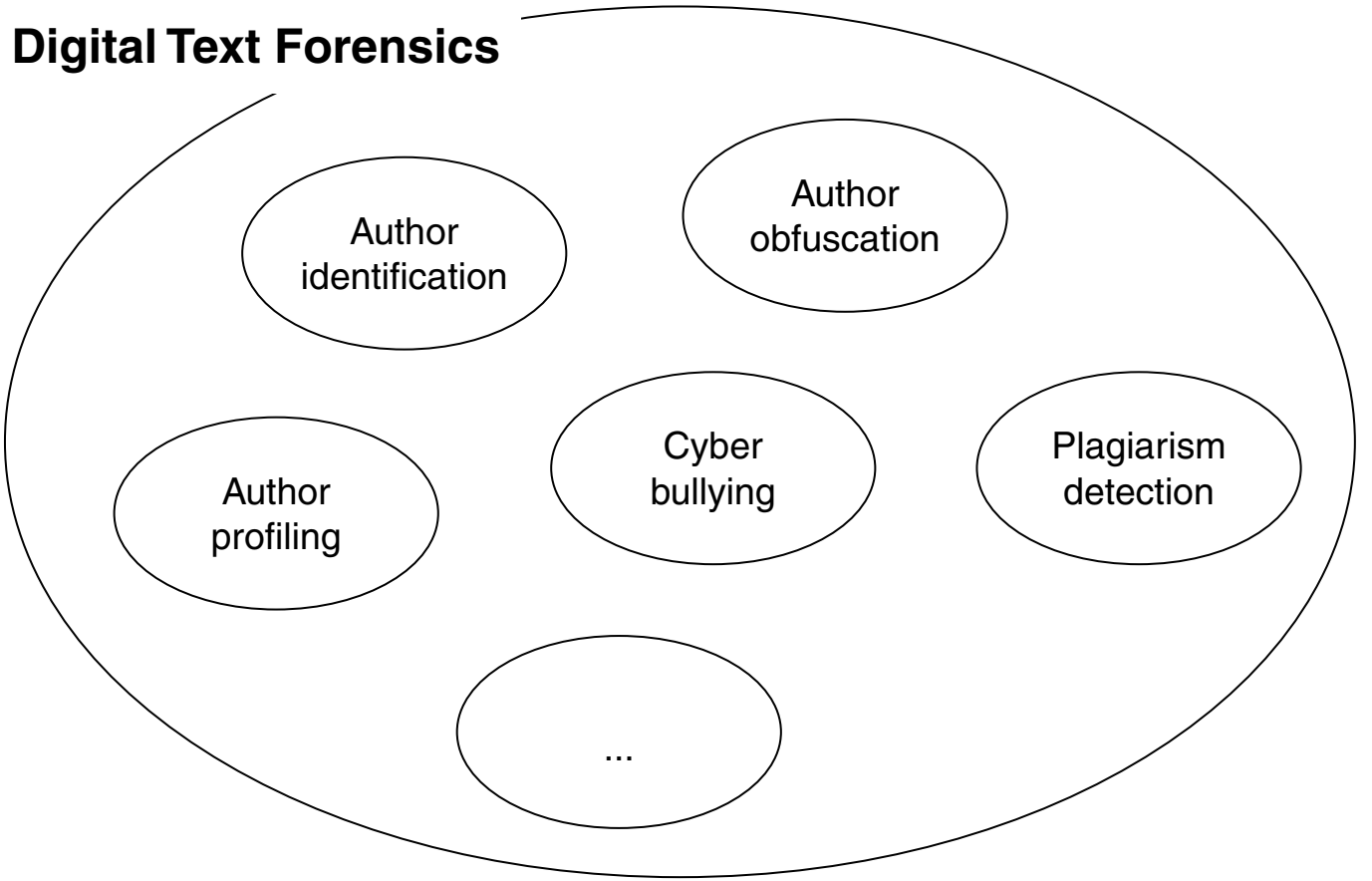


Janek Bevendorff

Christof Bräutigam

Digital Text Forensics





Retrieval models

Algorithms

Corpora

Some Technology Basics

Text with markup:

```
<TEXT> <TITLE>CHRYSLER> DEAL LEAVES UNCERTAINTY FOR AMC
WORKERS</TITLE> <AUTHOR> By Richard Walker, Reuters</AUTHOR>
<DATELINE> DETROIT, March 11 - </DATELINE><BODY>Chrysler
Corp's 1.5 billion dlr bid to takeover American Motors Corp;
AMO> should help bolster the small automaker's sales, but it
leaves the future of its 19,000 employees in doubt, industry
analysts say. It was "business as usual"yesterday at the
American ...
```

Raw text:

```
chrysler deal leaves uncertainty for amc workers by richard
walker reuters detroit march 11 chrysler corp s 1 5 billion
dlr bid to takeover american motors corp should help bolster
the small automaker s sales but it leaves the future of its
19 000 employees in doubt industry analysts say it was
business as usual yesterday at the american
```

Stop words:

chrysler deal leaves uncertainty **for** amc workers **by** richard walker reuters detroit **march 11** chrysler **corp s 1 5 billion dlr** bid **to** takeover american motors **corp should** help bolster **the small** automaker **s** sales **but it** leaves **the** future **of its 19 000** employees **in** doubt industry analysts **say it was** business **as usual** yesterday **at the** american

After stemming:

```
chrysler deal leav uncertain amc work richard walk reut
detroit takeover american motor help bols automak sal leav
futur employ doubt industr analy business usual yesterday
```

After stemming:

```
chrysler deal leav uncertain amc work richard walk reut
detroit takeover american motor help bols automak sal leav
futur employ doubt industr analy business usual yesterday
```

$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{motor} & w_2 \\ \cdots & \\ \texttt{cat} & w_x \\ \texttt{dog} & w_y \\ \texttt{mouse} & w_z \end{pmatrix}$$

After stemming:

```
chrysler deal leav uncertain amc work richard walk reut
detroit takeover american motor help bols automak sal leav
futur employ doubt industr analy business usual yesterday
```

$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{motor} & w_2 \\ \cdots & \\ \texttt{cat} & w_x \\ \texttt{dog} & w_y \\ \texttt{mouse} & w_z \end{pmatrix} \quad \rightsquigarrow \quad \begin{pmatrix} \texttt{chrysler} & 0.2 \\ \texttt{motor} & 0.3 \\ \cdots & \\ \texttt{cat} & 0.0 \\ \texttt{dog} & 0.1 \\ \texttt{mouse} & 0.1 \end{pmatrix}$$

Weight computation:

term frequency (tf), inverse document frequency (idf), divergence from randomness

After stemming:

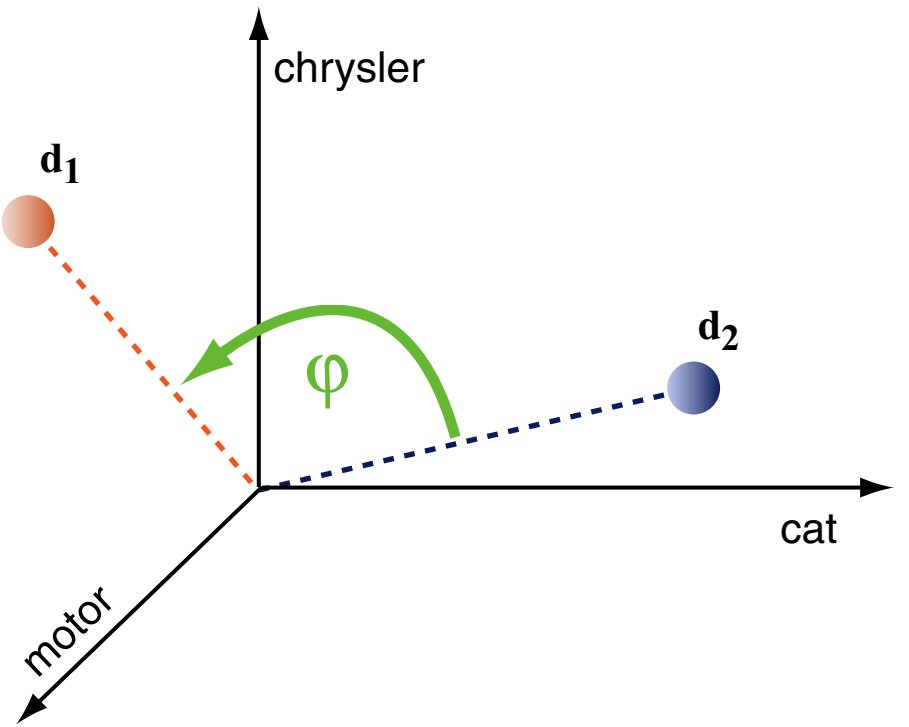
```
chrysler deal leav uncertain amc work richard walk reut
detroit takeover american motor help bols automak sal leav
futur employ doubt industr analy business usual yesterday
```

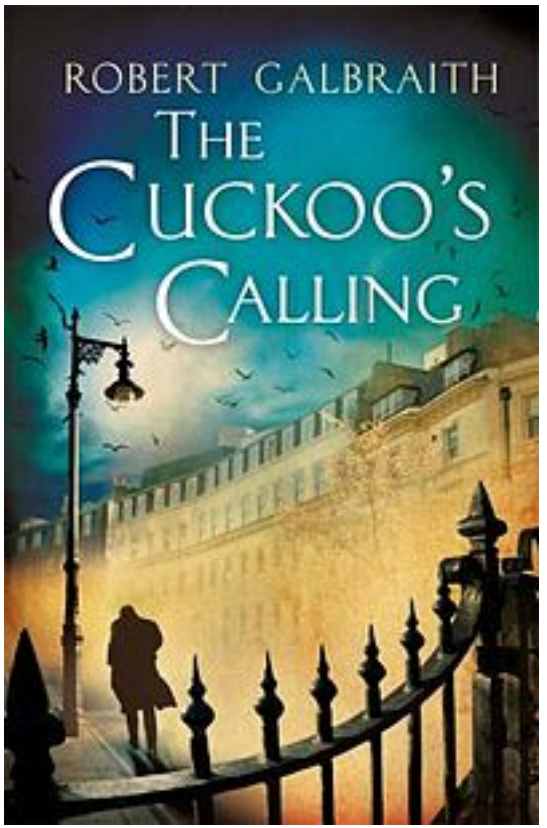
$$\mathbf{d} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{motor} & w_2 \\ \cdots & \\ \texttt{cat} & w_x \\ \texttt{dog} & w_y \\ \texttt{mouse} & w_z \end{pmatrix} \quad \rightsquigarrow \quad \left\langle \begin{pmatrix} \texttt{chrysler} & 0.2 \\ \texttt{motor} & 0.3 \\ \cdots & \\ \texttt{cat} & 0.0 \\ \texttt{dog} & 0.1 \\ \texttt{mouse} & 0.1 \end{pmatrix}, \begin{pmatrix} \texttt{chrysler} & 0.1 \\ \texttt{motor} & 0.2 \\ \cdots & \\ \texttt{cat} & 0.2 \\ \texttt{dog} & 0.0 \\ \texttt{mouse} & 0.0 \end{pmatrix} \right\rangle$$

Weight computation:

term frequency (tf), inverse document frequency (idf), divergence from randomness

Vector space:





Fake likes

Fake news

Fake clicks

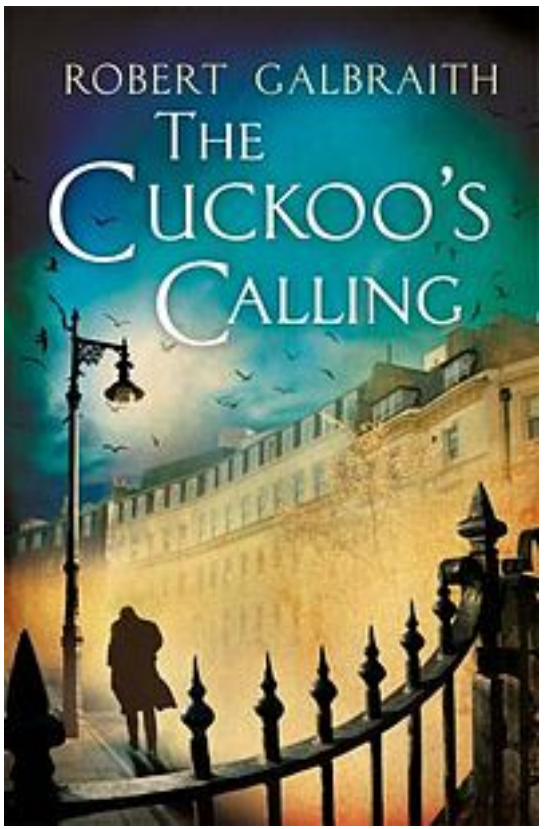
Fake users

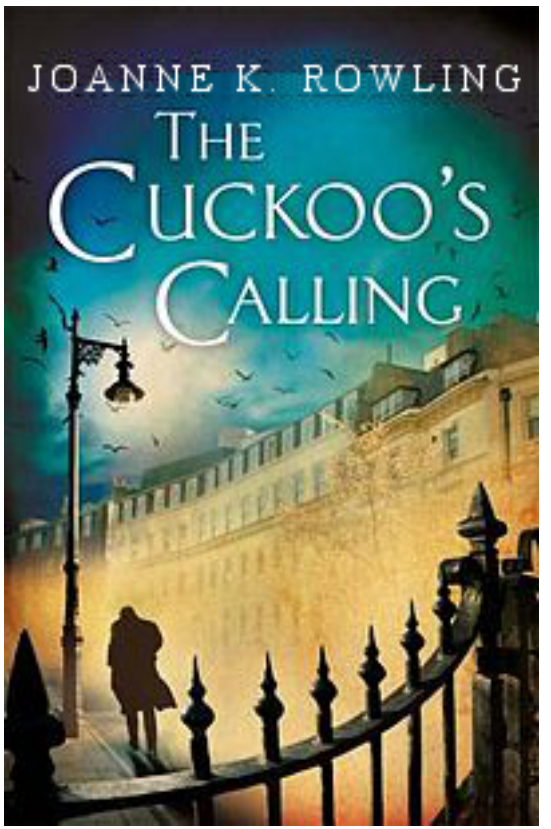
Fake reviews

Fake comments

⋮

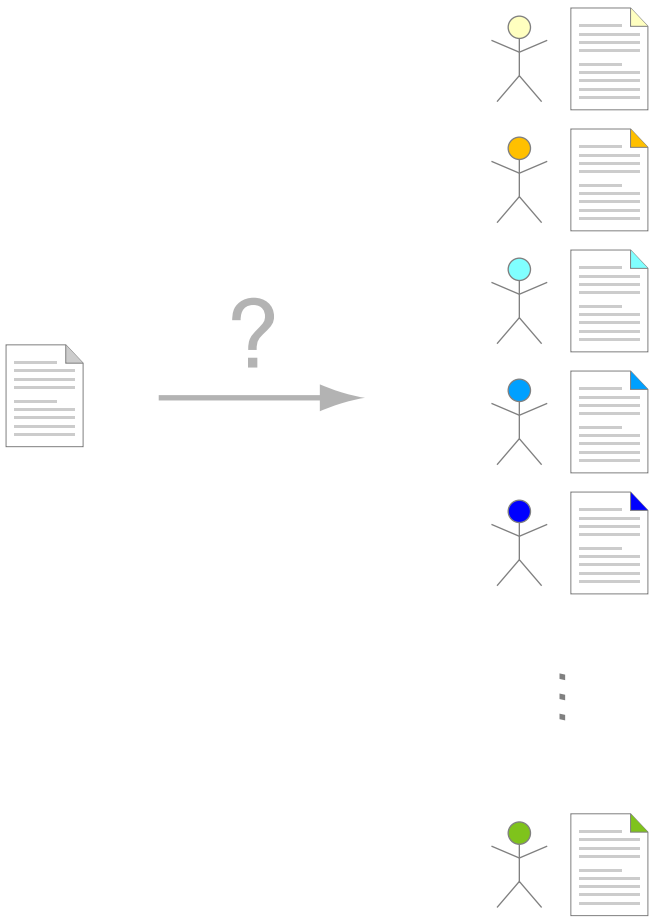
Fake identities (pseudonyms)


ROBERT GALBRAITH
THE CUCKOO'S CALLING


JOANNE K. ROWLING
THE CUCKOO'S CALLING

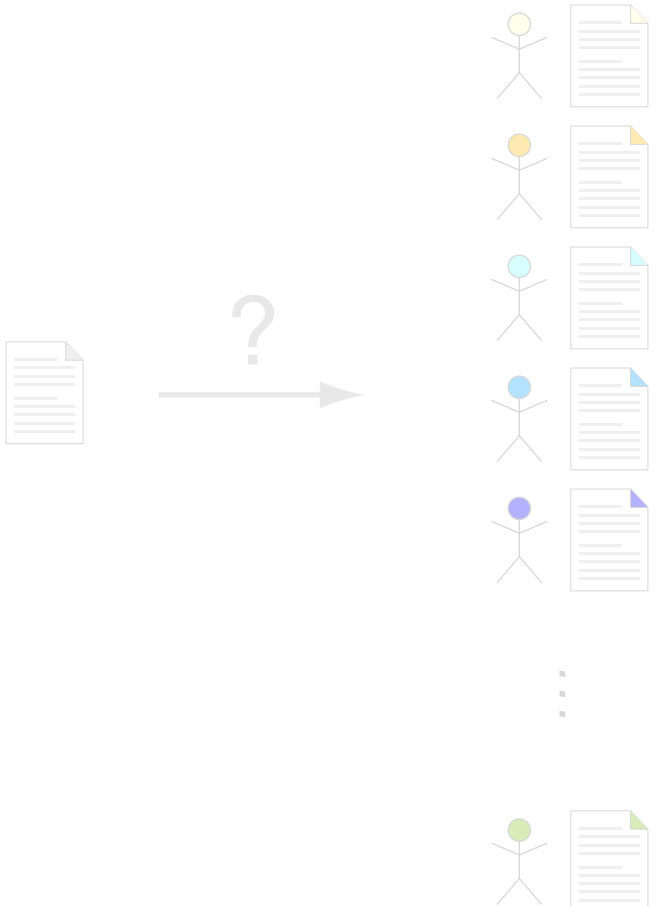
Author Identification

Authorship Attribution



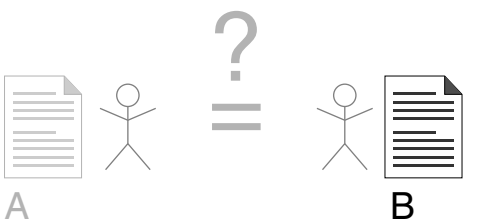
To which author does a text belong?

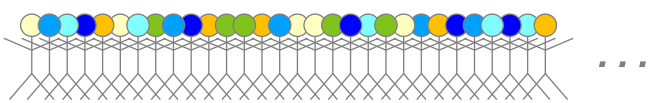
Authorship Attribution



To which author does a text belong?

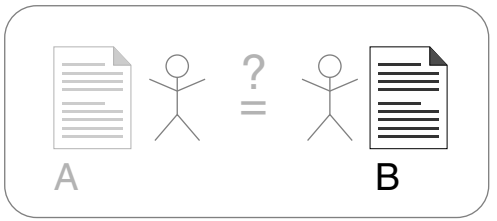
Authorship Verification



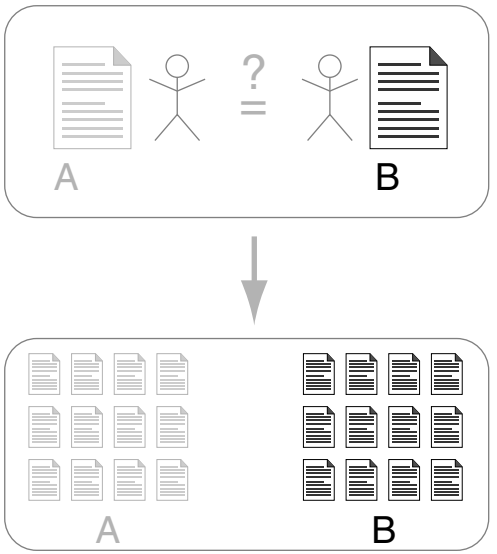


Originate two texts from the same author?

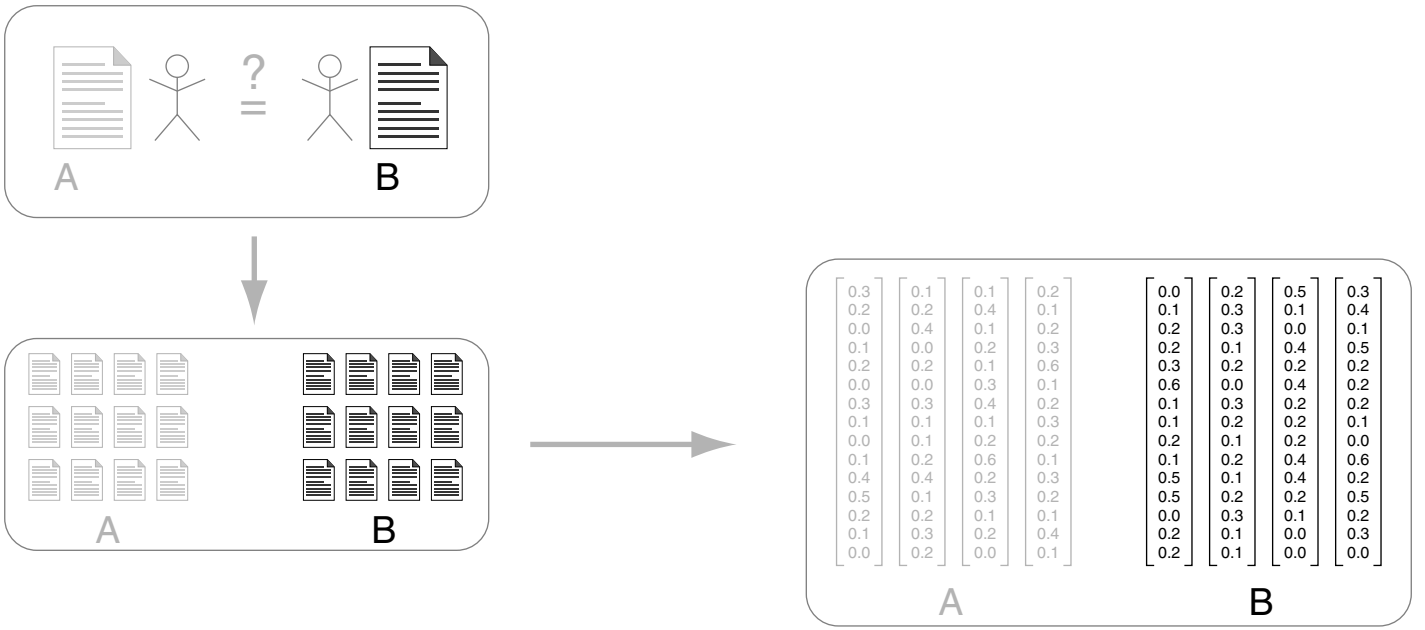
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



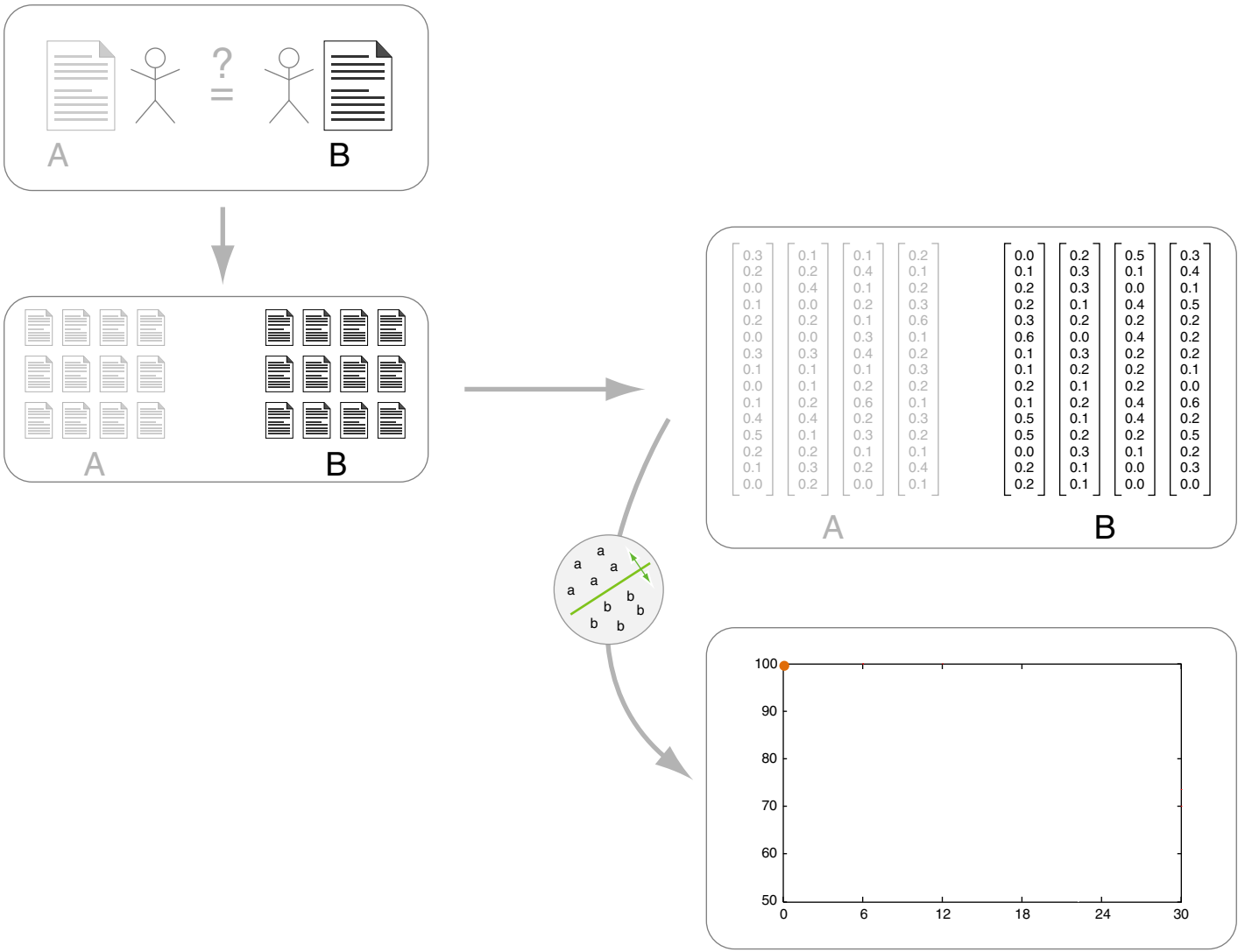
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



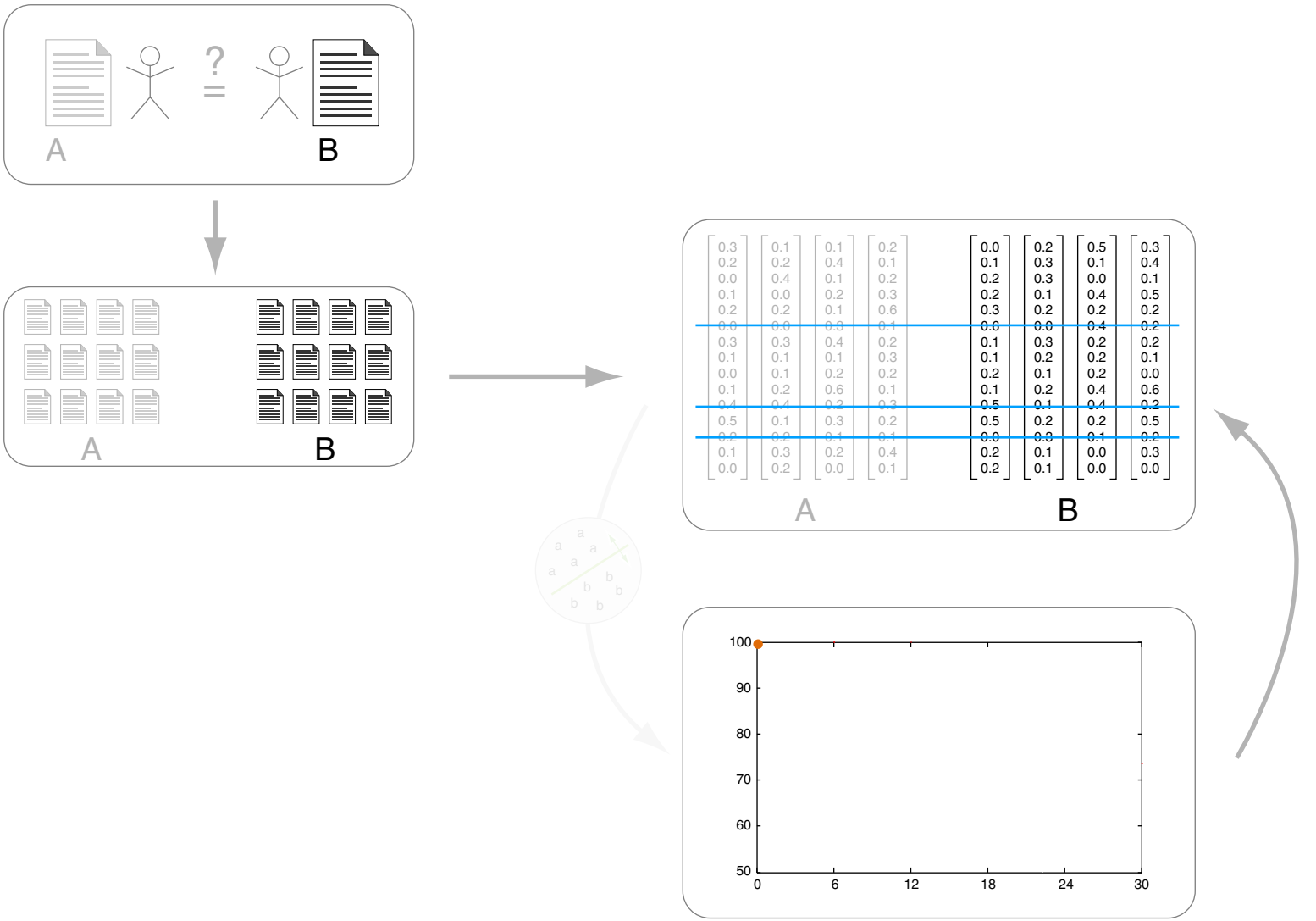
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



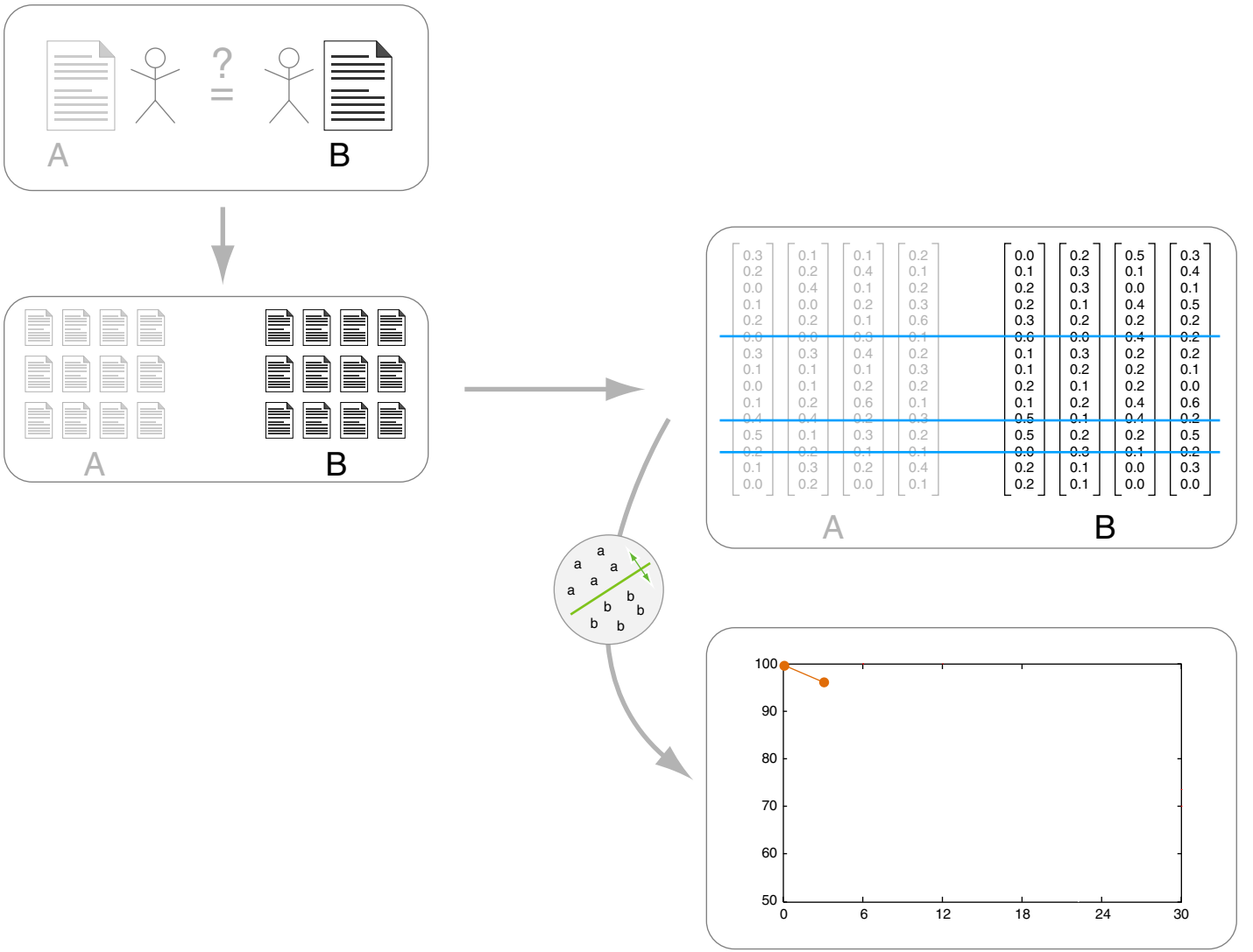
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



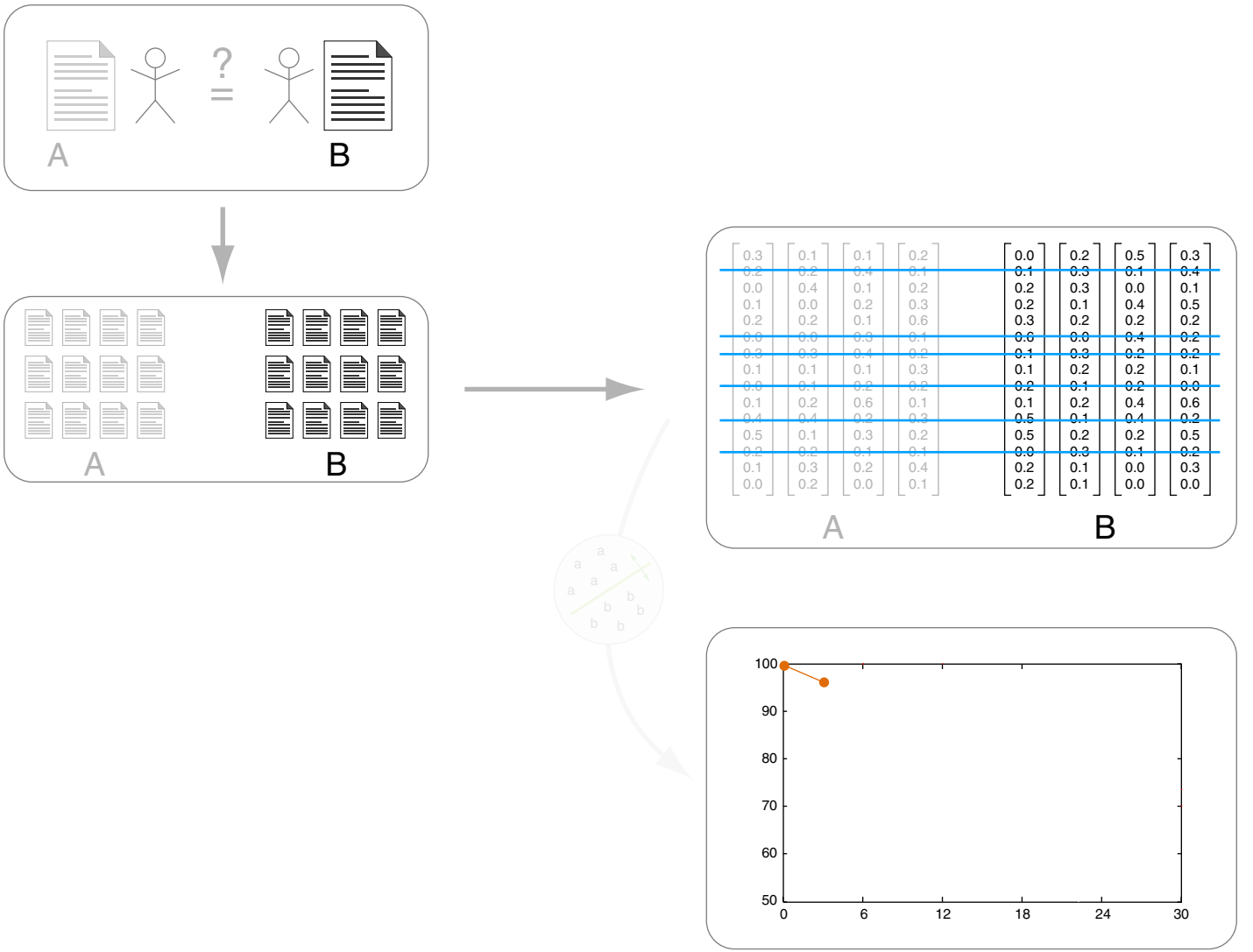
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



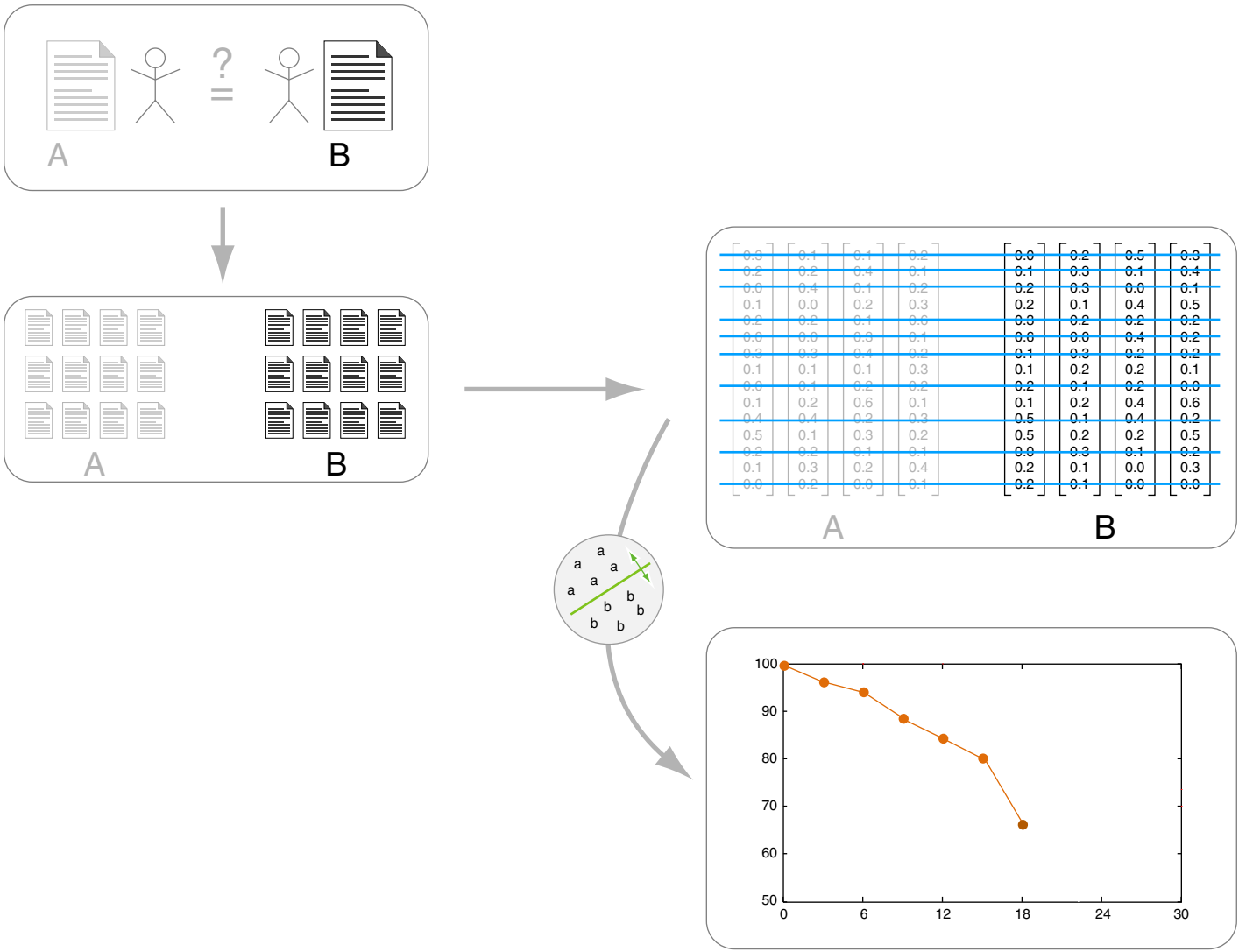
Authorship Verification via “Unmasking” [Koppel/Schler 2004]



Authorship Verification via “Unmasking” [Koppel/Schler 2004]

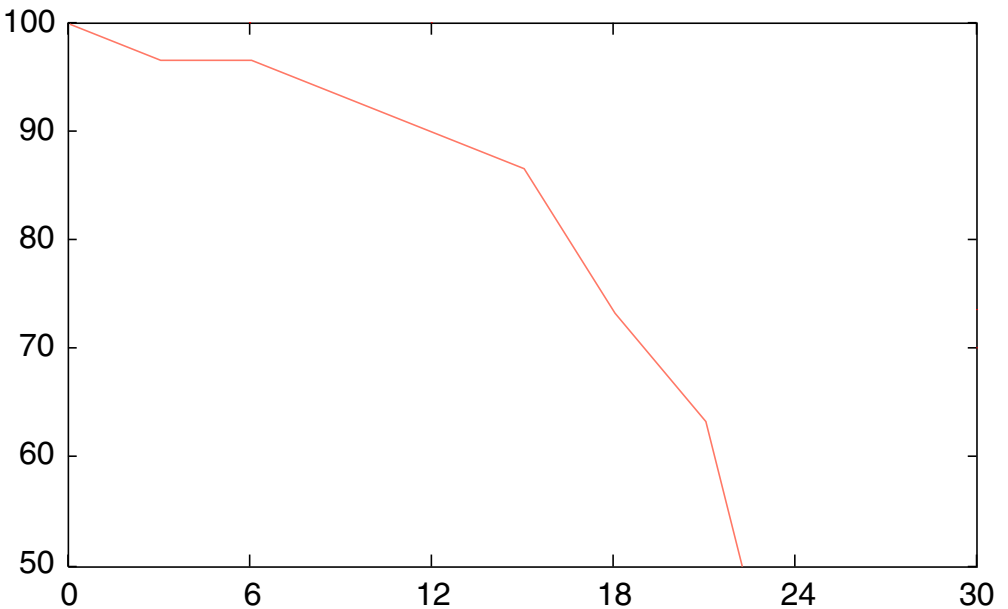


Authorship Verification via “Unmasking” [Koppel/Schler 2004]



Authorship Verification via “Unmasking”

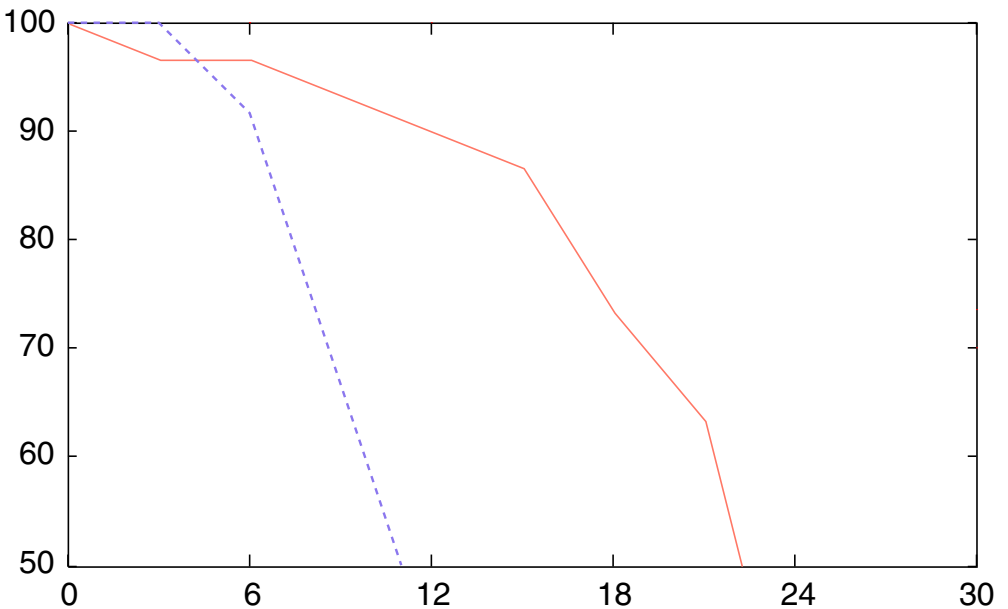
Typical learning characteristic for . . .



different authors ($A \neq B$)

Authorship Verification via “Unmasking”

Typical learning characteristic for . . .

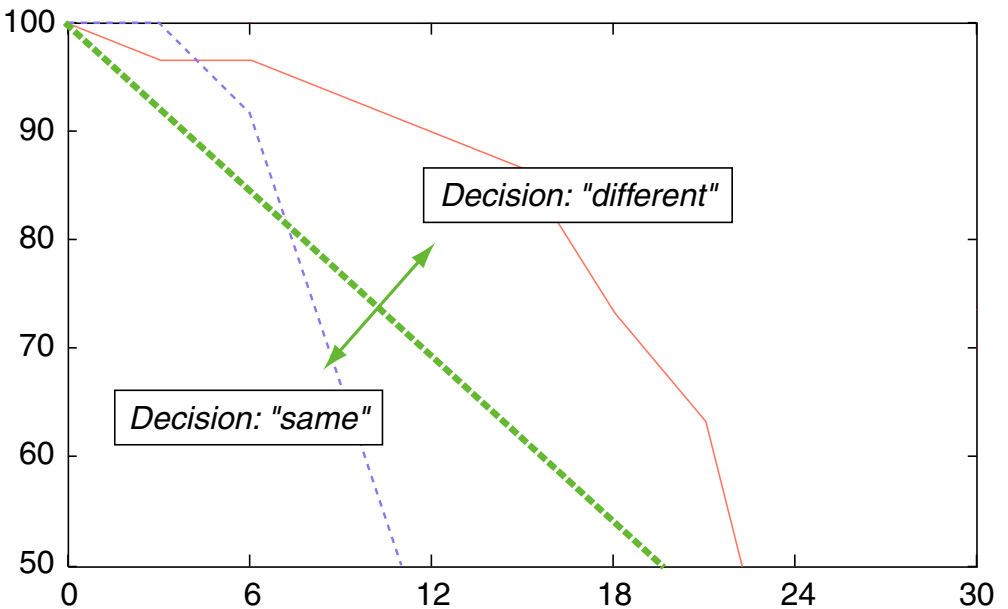


different authors $(A \neq B)$

same author $(A = B)$

Authorship Verification via “Unmasking”

Typical learning characteristic for . . .

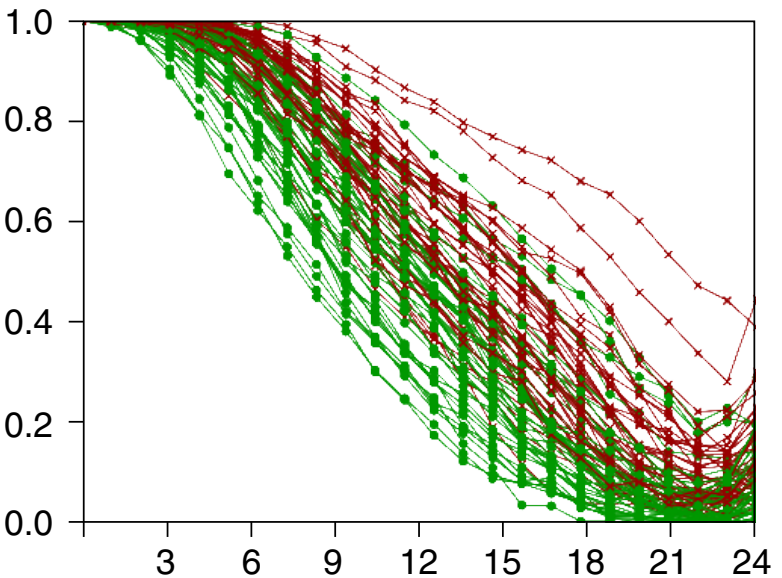


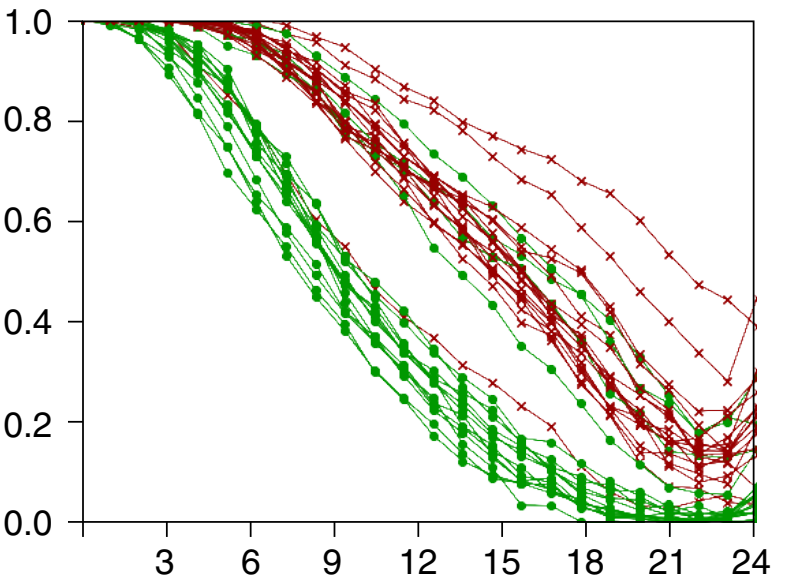
different authors $(A \neq B)$

same author $(A = B)$

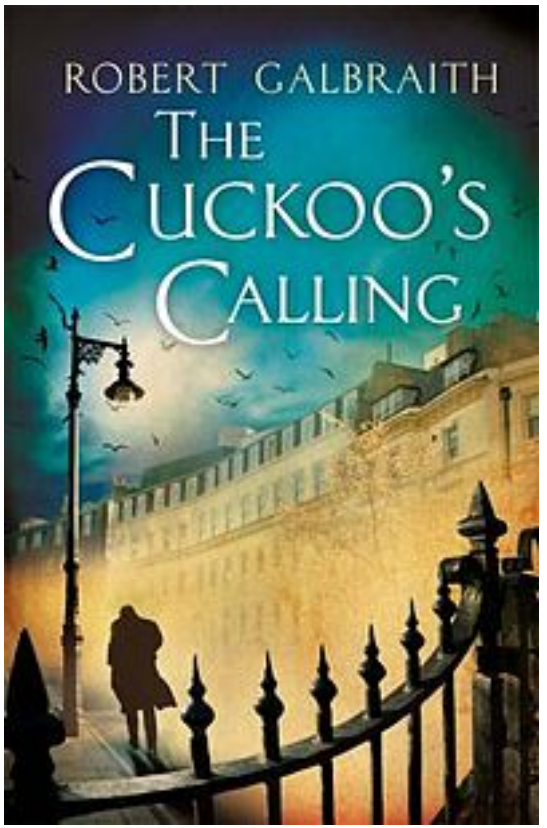
The typical learning characteristic can be learned.

Authorship Verification via “Unmasking”





Applied to 78 pairs of texts, 4,000 words each ➜ 26% of decisions are “safe”



Fake likes

Fake news

Fake clicks

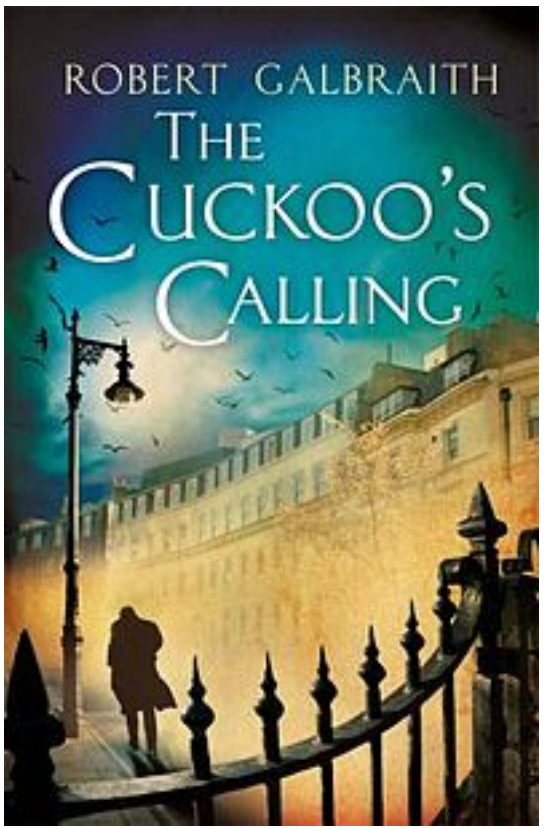
Fake users

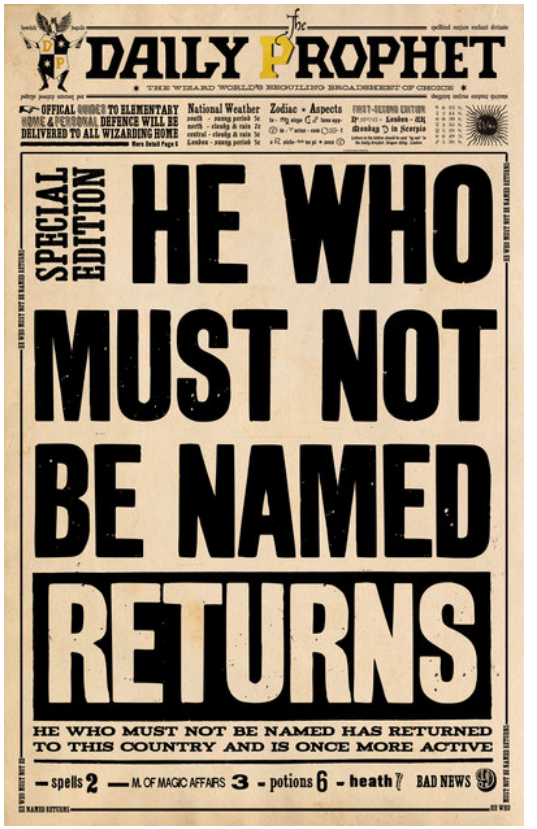
Fake reviews

Fake comments

⋮

Fake identities (pseudonyms)


ROBERT GALBRAITH
THE CUCKOO'S CALLING


The DAILY PROPHET
THE WIZARD WORLD'S BEGUILING BROADSHEET OF CHOICE
OFFICAL GUIDES TO ELEMENTARY HOME & PERSONAL DEFENCE WILL BE DELIVERED TO ALL WIZARDING HOME
National Weather
Zodiac • Aspects
FIRST-SECOND EDITION
SPECIAL EDITION
HE WHO MUST NOT BE NAMED RETURNS
HE WHO MUST NOT BE NAMED HAS RETURNED TO THIS COUNTRY AND IS ONCE MORE ACTIVE
– spells 2 — M. OF MAGIC AFFAIRS 3 - potions 6 - heath 7 BAD NEWS

Constrained Paraphrasing

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

[Veto message for the Shipyard project, Port of San Francisco. Oct. 12th, 2009]

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

[Veto message for the Shipyard project, Port of San Francisco. Oct. 12th, 2009]

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another [illegible] Californians overwhelmingly [illegible] it is unnecessary [illegible]

Sincerely,

Arnold Schwarzenegger

"*My goodness. What a coincidence* [. . .] "

[Aaron McLear, Schwarzenegger spokesman, Oct. 2009]

On Acrostics

An acrostic is a poem or other form of writing in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out a word or a message. [Wikipedia]

A poem [Kuperavage 2000] :

```
H  He broke my heart
E  Every piece, shattered
A  All I wanted was his love
R  Real, as he promised
T  True, as mine for him
...
```

On Acrostics

An acrostic is a poem or other form of writing in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out a word or a message. [Wikipedia]

A poem [Kuperavage 2000] :

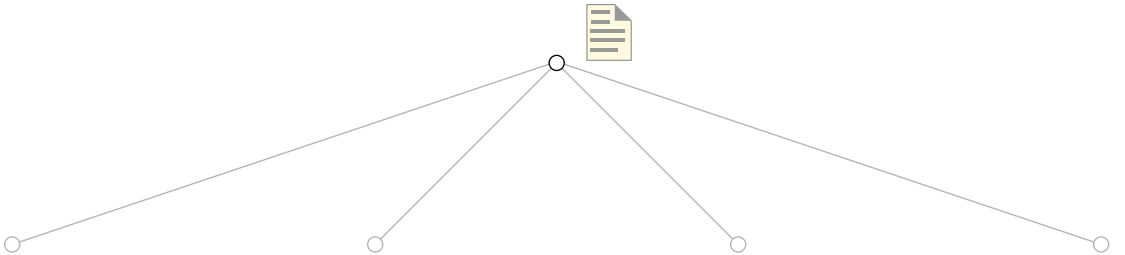
```
H  He broke my heart
E  Every piece, shattered
A  All I wanted was his love
R  Real, as he promised
T  True, as mine for him
...
```

Task [Stein/Hagen/Bräutigam 2014]

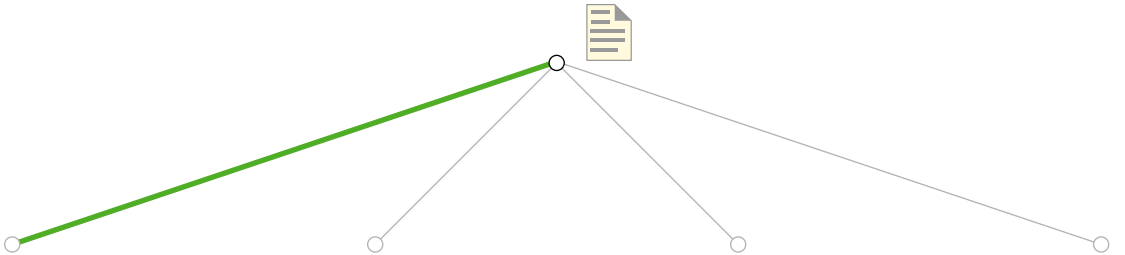
Given: (1) A text T and an acrostic x.

(2) Lower and upper bounds on the desired line lengths.

Task: Find a paraphrased version T^* of T in monospaced font that encodes x in some consecutive lines, if possible. Each line of T^* has to meet the length constraints.



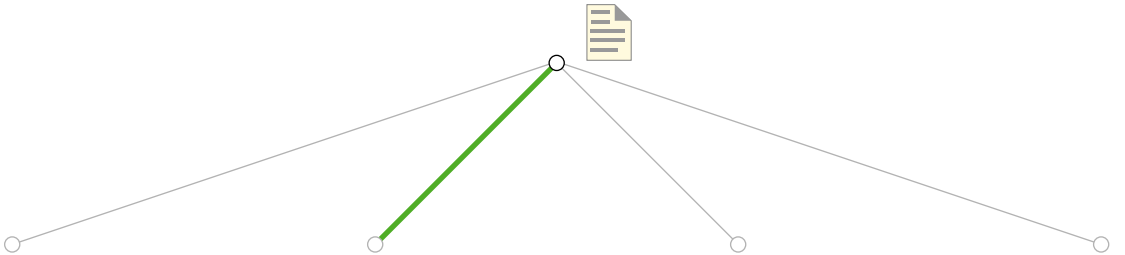
Subtask: Create the character bauhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

Subtask: Create the character **b**auhaus



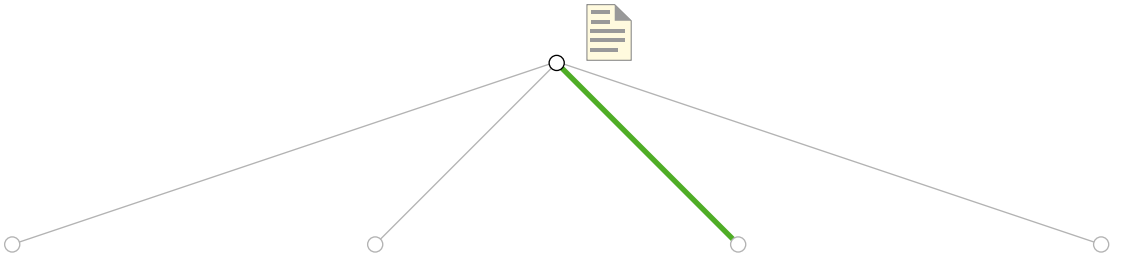
~~Before~~ ~~some time now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

For some time now I have lamented **but** the fact that major issues are overlooked while many bills

«Add Connective»

Subtask: Create the character **b**auhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

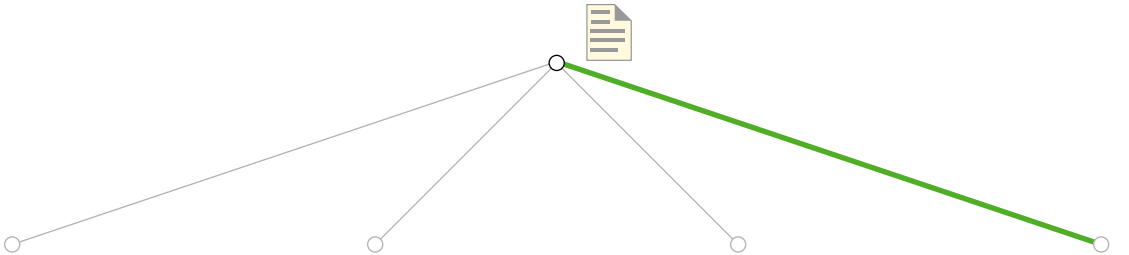
For some time now I have lamented **but** the fact that major issues are overlooked while many bills

«Add Connective»

Been for some time now I have lamented the fact that major issues are overlooked while many bills come to

«Change Tense»

Subtask: Create the character **b**auhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

For some time now I have lamented | **but** the fact that major issues are overlooked while many bills

«Add Connective»

Been for some time now I have lamented the fact that major issues are overlooked while many bills come to

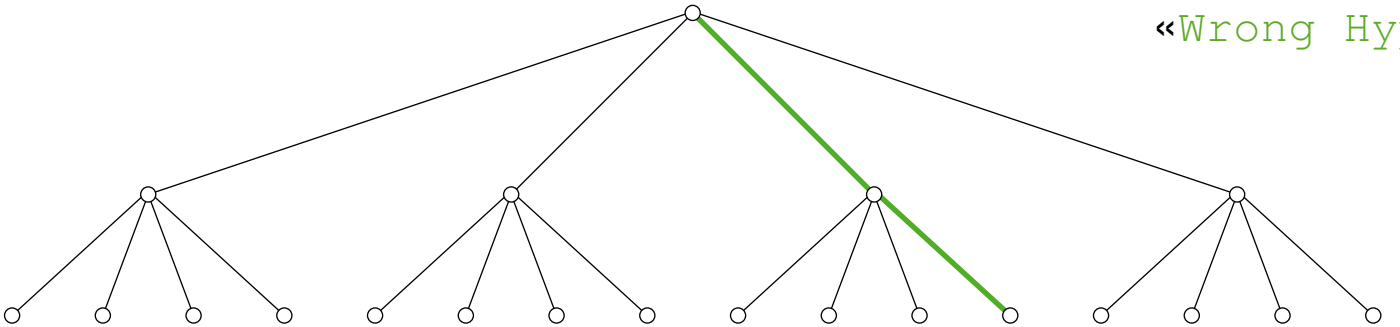
«Change Tense»

For some time now I have lamented the fact that major issues are overlooked while many |
bills come to

«Linebreak»

Subtask: Create the character **b**auhaus

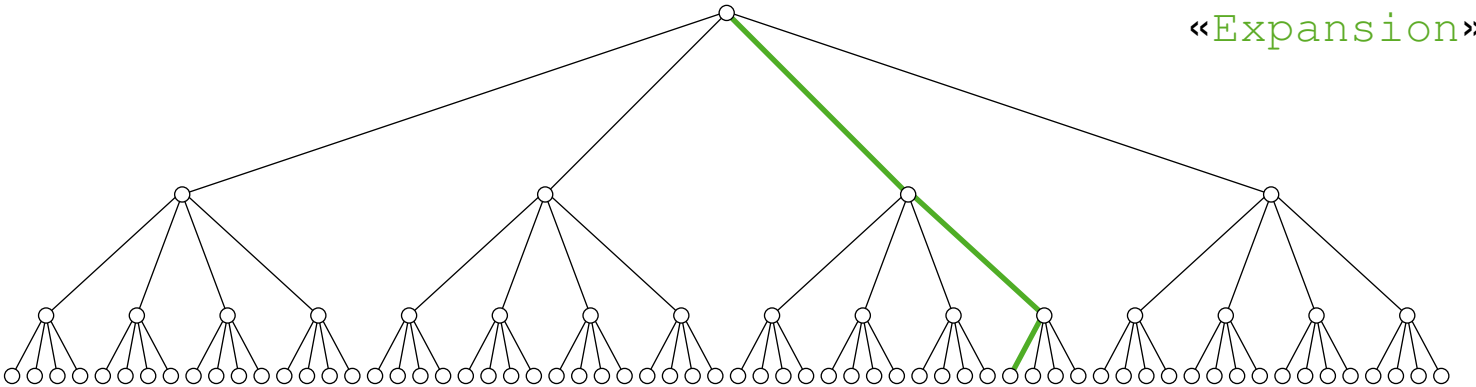
«Wrong Hyphen»



Subtask: Create the character bauhaus

Decision 3 of 64

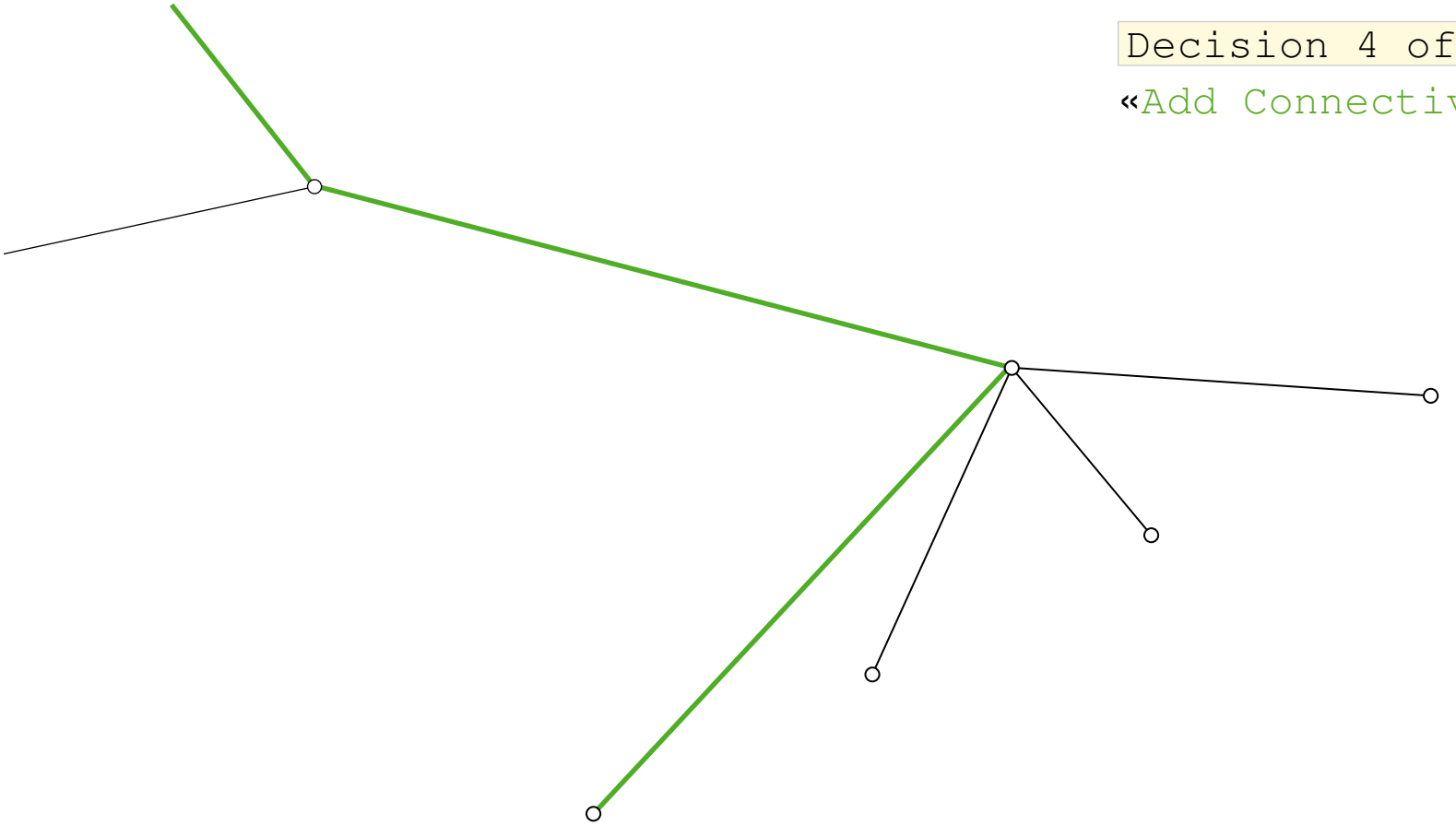
«Expansion»



Subtask: Create the character ba**u**haus

Decision 4 of 256

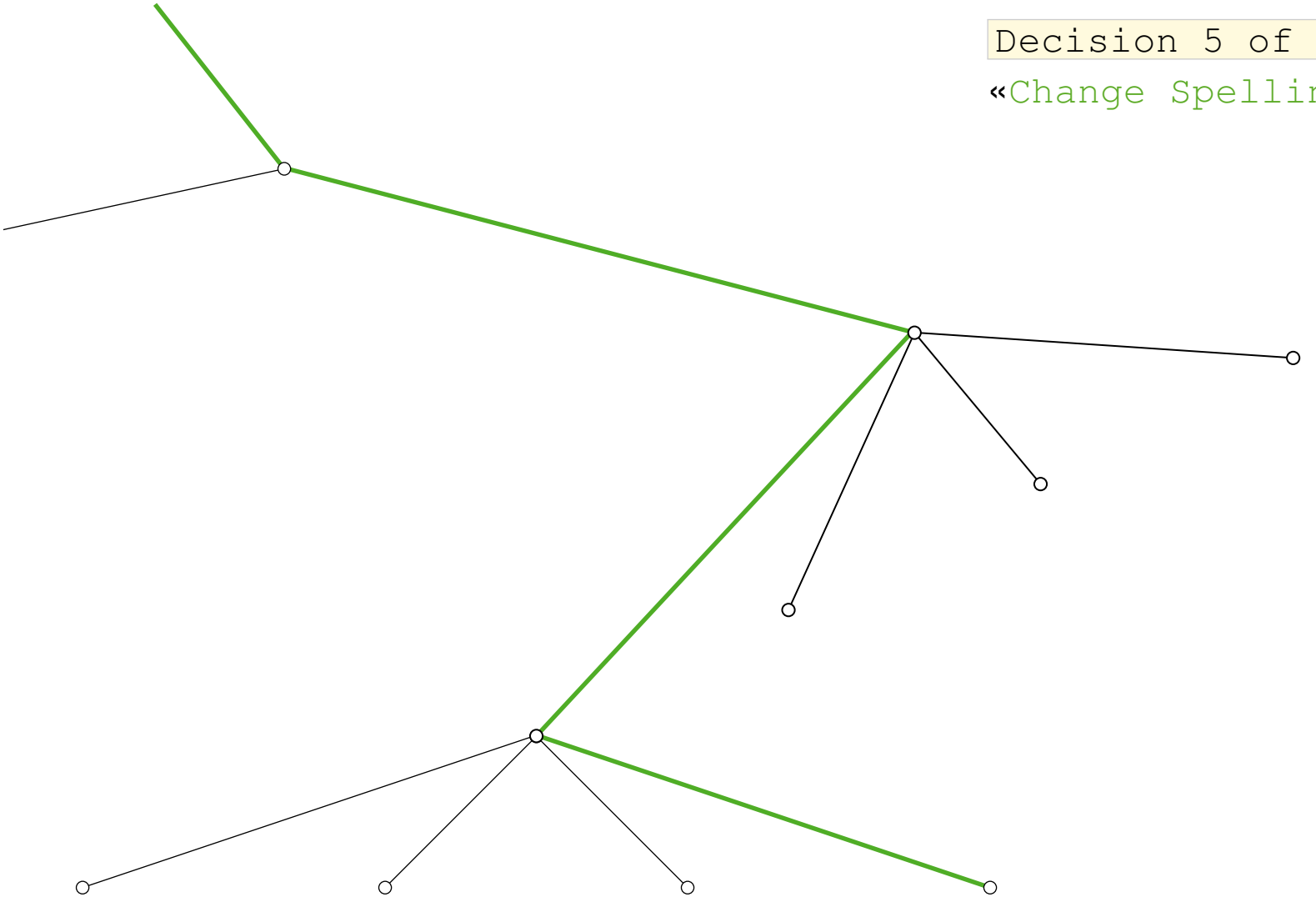
«Add Connective»



Subtask: Create the character bau**h**aus

Decision 5 of 1024

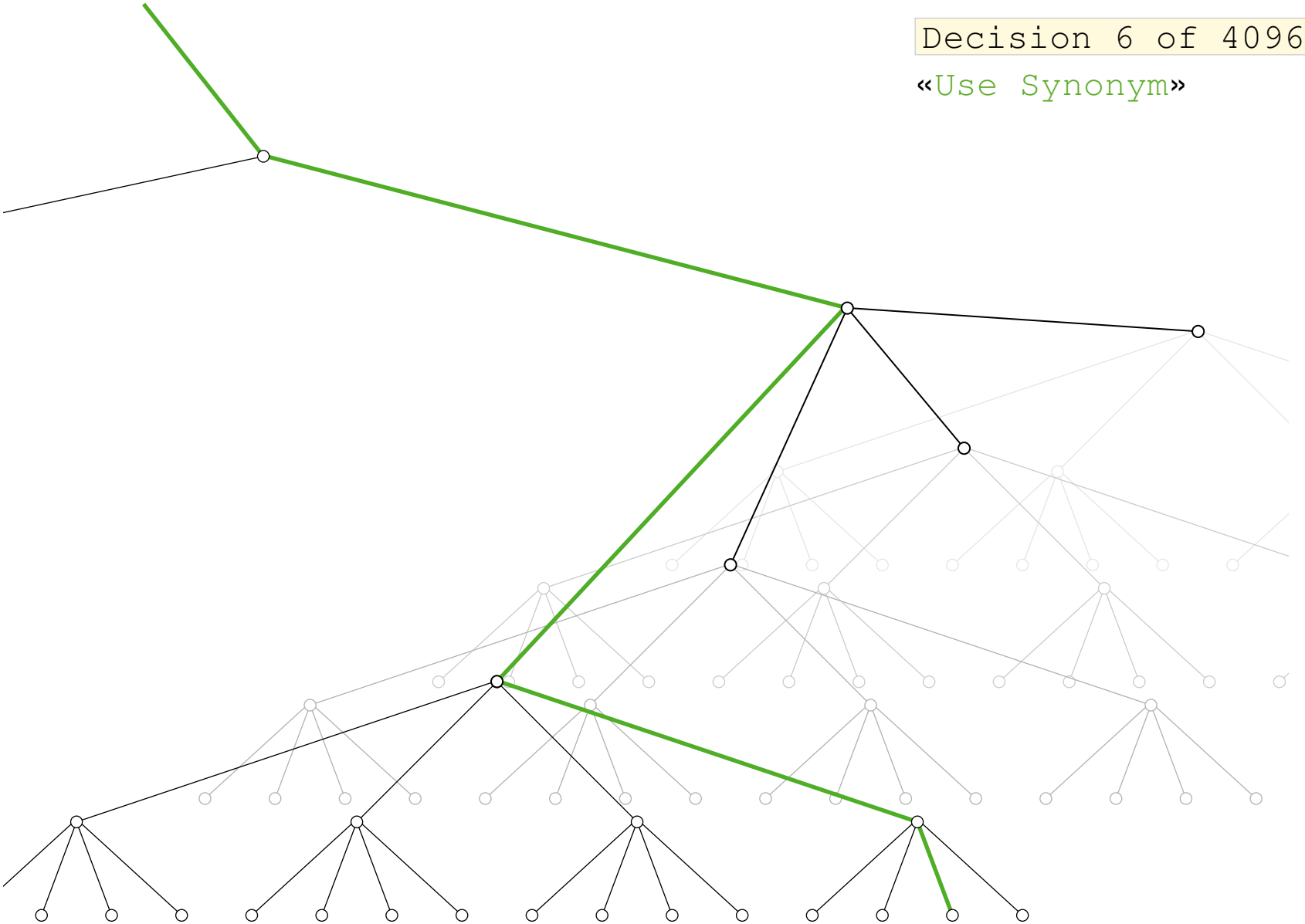
«Change Spelling»



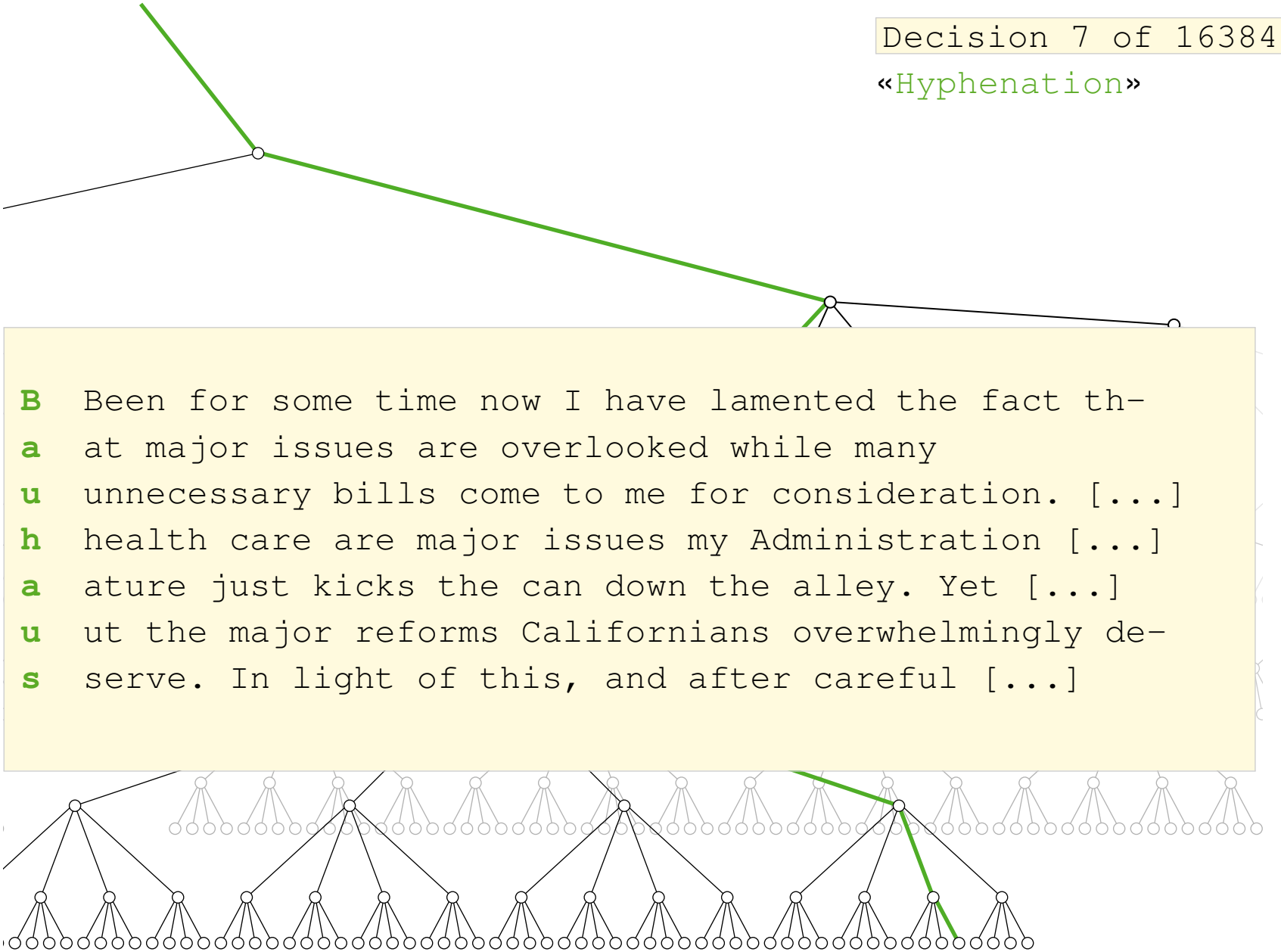
Subtask: Create the character bauh**a**us

Decision 6 of 4096

«Use Synonym»



«Hyphenation»



B Been for some time now I have lamented the fact th-
a at major issues are overlooked while many
u unnecessary bills come to me for consideration. [...]
h health care are major issues my Administration [...]
a ature just kicks the can down the alley. Yet [...]
u ut the major reforms Californians overwhelmingly de-
s serve. In light of this, and after careful [...]

Searchspace Facts

Searchspace Facts

Consider a text with a length of 100 words (the Schwarzenegger Letter) . . .

- $\approx$ $10 \cdot 3$ possibilities to change tense
- $\approx$ 100 possibilities to break a line
- $\approx$ $100 \cdot 3$ possibilities to introduce a synonym
- $\approx$ $100 \cdot 3$ possibilities to introduce filler words
- $\approx$ $100 \cdot 5$ possibilities to hyphenate a word
- $\gg$ 100 possibilities to introduce tautologies

. . .

Searchspace Facts

Consider a text with a length of 100 words (the Schwarzenegger Letter) . . .

- $\approx$ 10 · 3 possibilities to change tense
- $\approx$ 100 possibilities to break a line
- $\approx$ 100 · 3 possibilities to introduce a synonym
- $\approx$ 100 · 3 possibilities to introduce filler words
- $\approx$ 100 · 5 possibilities to hyphenate a word
- $\gg$ 100 possibilities to introduce tautologies

  . . .

➜ $>$ 1 000 possible operations to generate a single letter of an acrostic

➜ $\mathcal{O}(10^{3n})$ possibilities to synthesize an n = 7 letter word like ‘`Bauhaus`’

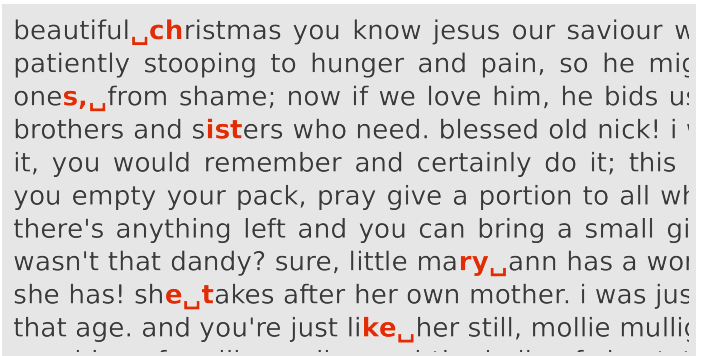
Compare the following numbers:

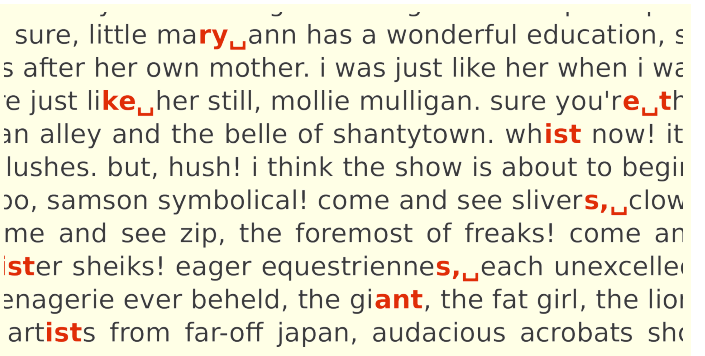
10^{80} atoms in the observable universe

10^{123} game-tree complexity of chess

Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

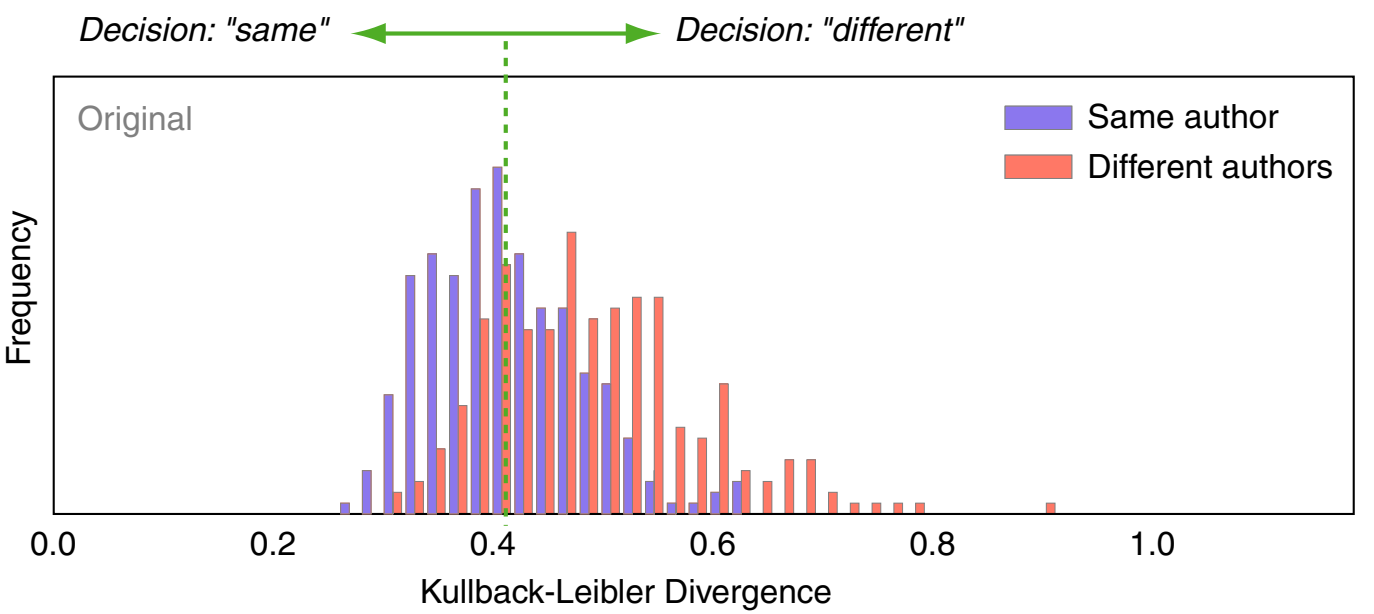




Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

beautiful␣**ch**ristmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
one**s,␣**from shame; now if we love him, he bids u
brothers and s**ist**ers who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little ma**ry␣**ann has a wo
she has! sh**e␣t**akes after her own mother. i was jus
that age. and you're just li**ke␣**her still, mollie mulli

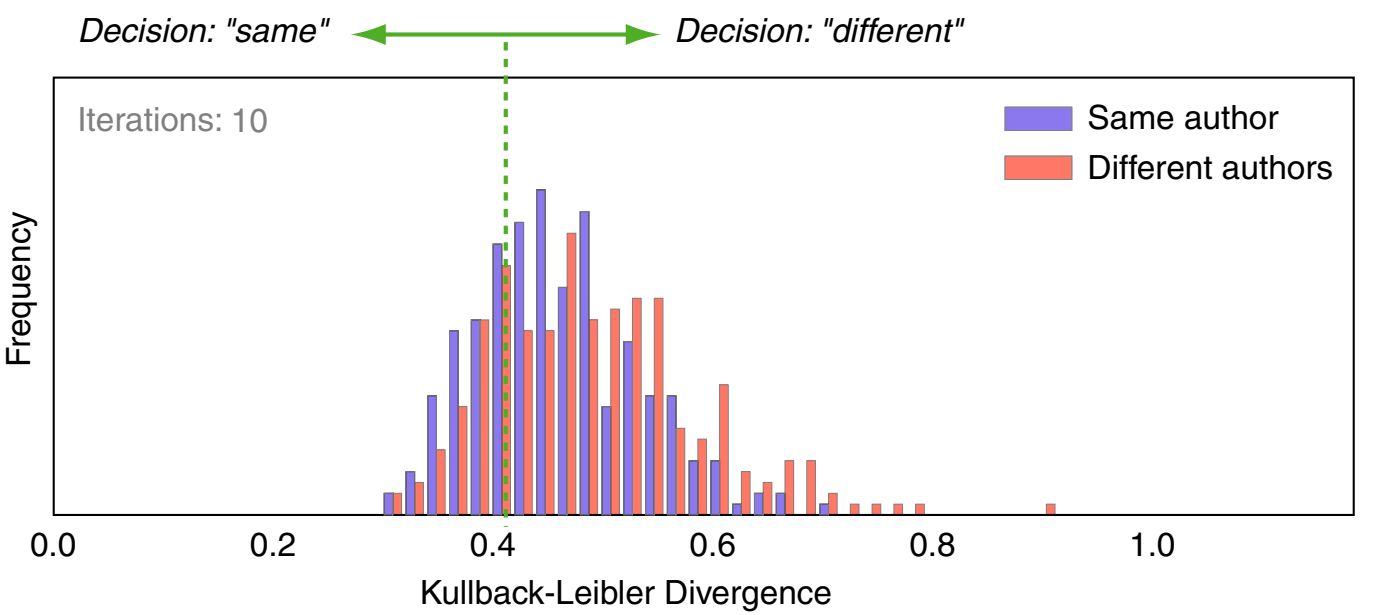
sure, little ma**ry␣**ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just li**ke␣**her still, mollie mulligan. sure you're**␣t**h
an alley and the belle of shantytown. wh**ist** now! it
lushes. but, hush! i think the show is about to begi
oo, samson symbolical! come and see sliver**s,␣**clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestrienne**s,␣**each unexcelle
enagerie ever beheld, the gi**ant**, the fat girl, the lio
art**ist**s from far-off japan, audacious acrobats sh



Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

beautiful␣**ch**ristmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
one**s,**␣from shame; now if we love him, he bids us
brothers and s**ist**ers who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little ma**ry**␣ann has a wo
she has! sh**e**␣**t**akes after her own mother. i was jus
that age. and you're just li**ke**␣her still, mollie mulli

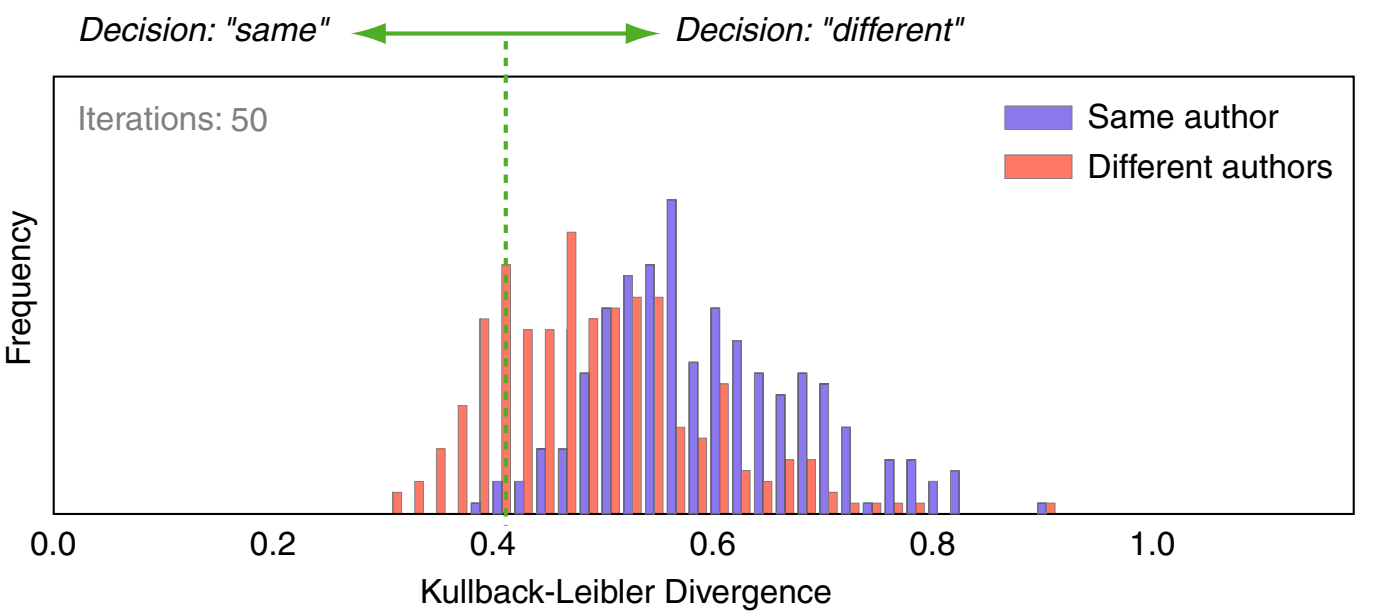
sure, little ma**ry**␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just li**ke**␣her still, mollie mulligan. sure you're**e**␣**t**h
an alley and the belle of shantytown. wh**ist** now! it
lushes. but, hush! i think the show is about to begi
oo, samson symbolical! come and see sliver**s,**␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestrienne**s,**␣each unexcelle
enagerie ever beheld, the gi**ant**, the fat girl, the lio
art**ist**s from far-off japan, audacious acrobats sh



Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

beautiful␣**ch**ristmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
one**s,**␣from shame; now if we love him, he bids u
brothers and s**ist**ers who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little ma**ry**␣ann has a wo
she has! sh**e**␣**t**akes after her own mother. i was jus
that age. and you're just li**ke**␣her still, mollie mulli

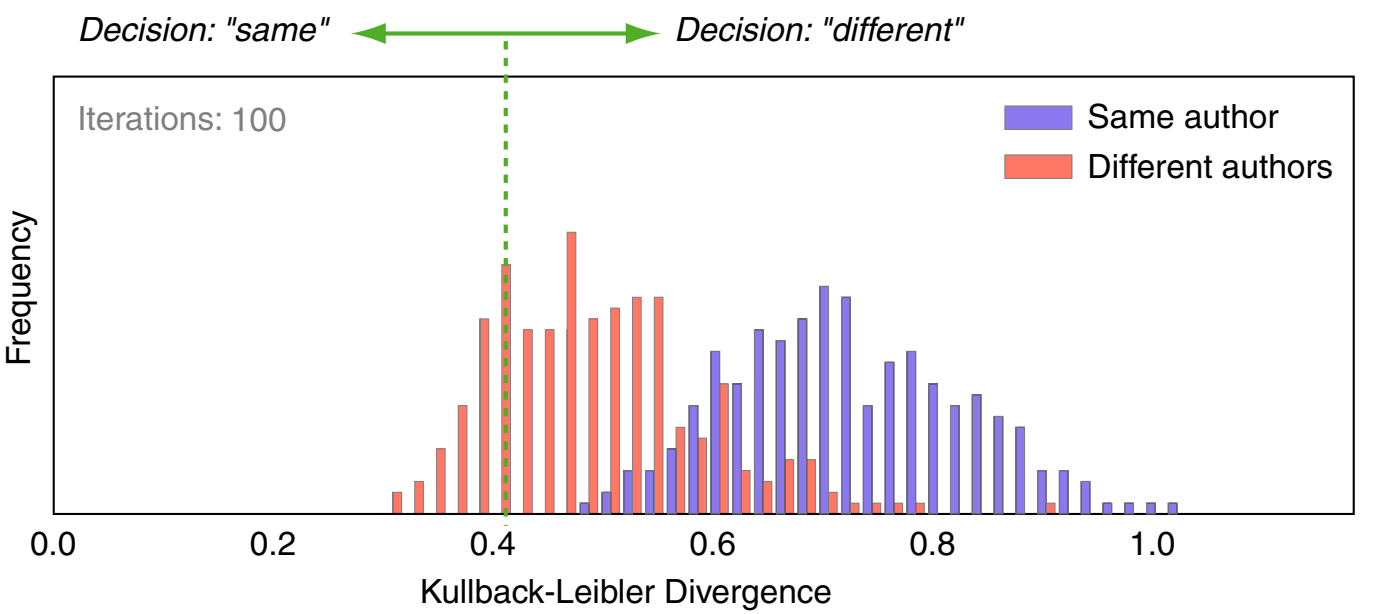
sure, little ma**ry**␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just li**ke**␣her still, mollie mulligan. sure you're**e**␣**t**h
an alley and the belle of shantytown. wh**ist** now! it
lushes. but, hush! i think the show is about to begi
oo, samson symbolical! come and see sliver**s,**␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestrienne**s,**␣each unexcelle
enagerie ever beheld, the gi**ant**, the fat girl, the lio
art**ist**s from far-off japan, audacious acrobats sh



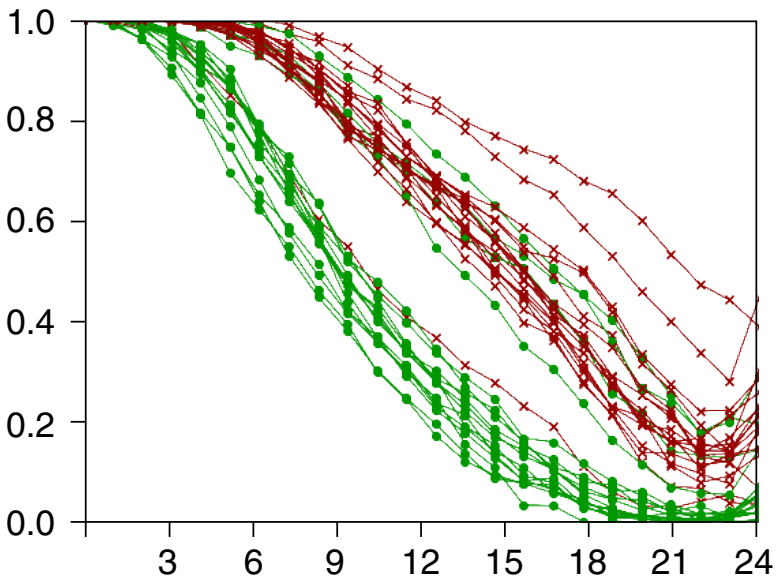
Toward Author Obfuscation [Stein/Potthast/Hagen/Bevendorff 2016]

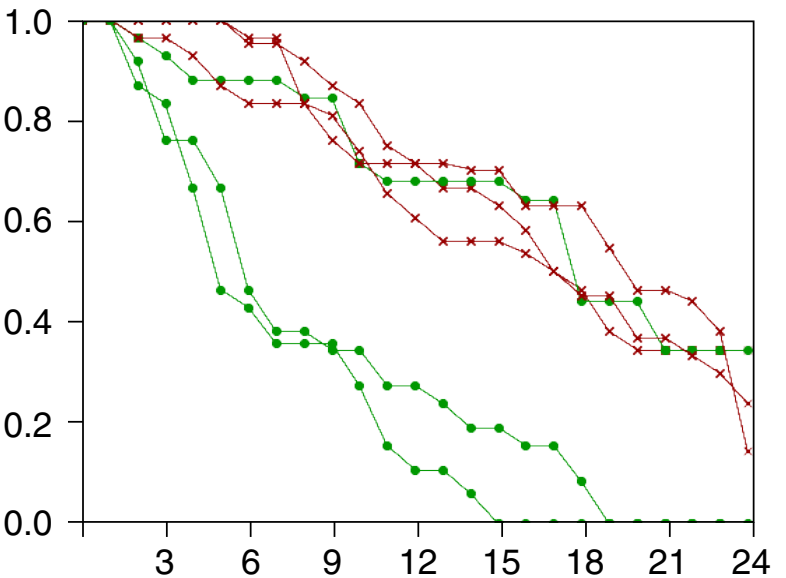
beautiful␣**ch**ristmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
one**s,**␣from shame; now if we love him, he bids u
brothers and s**ist**ers who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little ma**ry**␣ann has a wo
she has! sh**e**␣**t**akes after her own mother. i was jus
that age. and you're just li**ke**␣her still, mollie mulli

sure, little ma**ry**␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just li**ke**␣her still, mollie mulligan. sure you're**e**␣**t**h
an alley and the belle of shantytown. wh**ist** now! it
lushes. but, hush! i think the show is about to begi
oo, samson symbolical! come and see sliver**s,**␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestrienne**s,**␣each unexcelle
enagerie ever beheld, the gi**ant**, the fat girl, the lio
art**ist**s from far-off japan, audacious acrobats sh



Authorship Verification via “Unmasking”





26% of decisions are “safe” ➜ 10% remain safe when obfuscating 1% of the text

Conclusion

Summary

- Constrained paraphrasing via heuristic search in style space
- 5 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

Conclusion

Summary

- Constrained paraphrasing via heuristic search in style space
- 5 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

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly anyone considers obfuscation a threat

Conclusion

Summary

- Constrained paraphrasing via heuristic search in style space
- 5 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

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly anyone considers obfuscation a threat

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

Neither can live while the other survives

Conclusion

Summary

- Constrained paraphrasing via heuristic search in style space
- 5 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

Take-away messages

- State of the art in authorship verification vulnerable to obfuscation
- Automatic obfuscation is feasible, yet far from perfection
- Hardly anyone considers obfuscation a threat

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

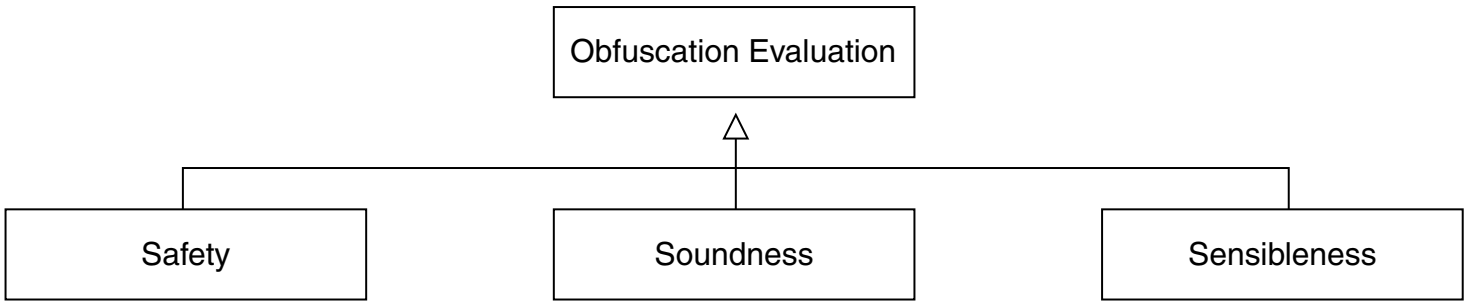
Neither can live while the other survives

Thank you for your attention!

Evaluating Author Obfuscation

Obfuscation Evaluation

Taxonomy of Evaluation Dimensions

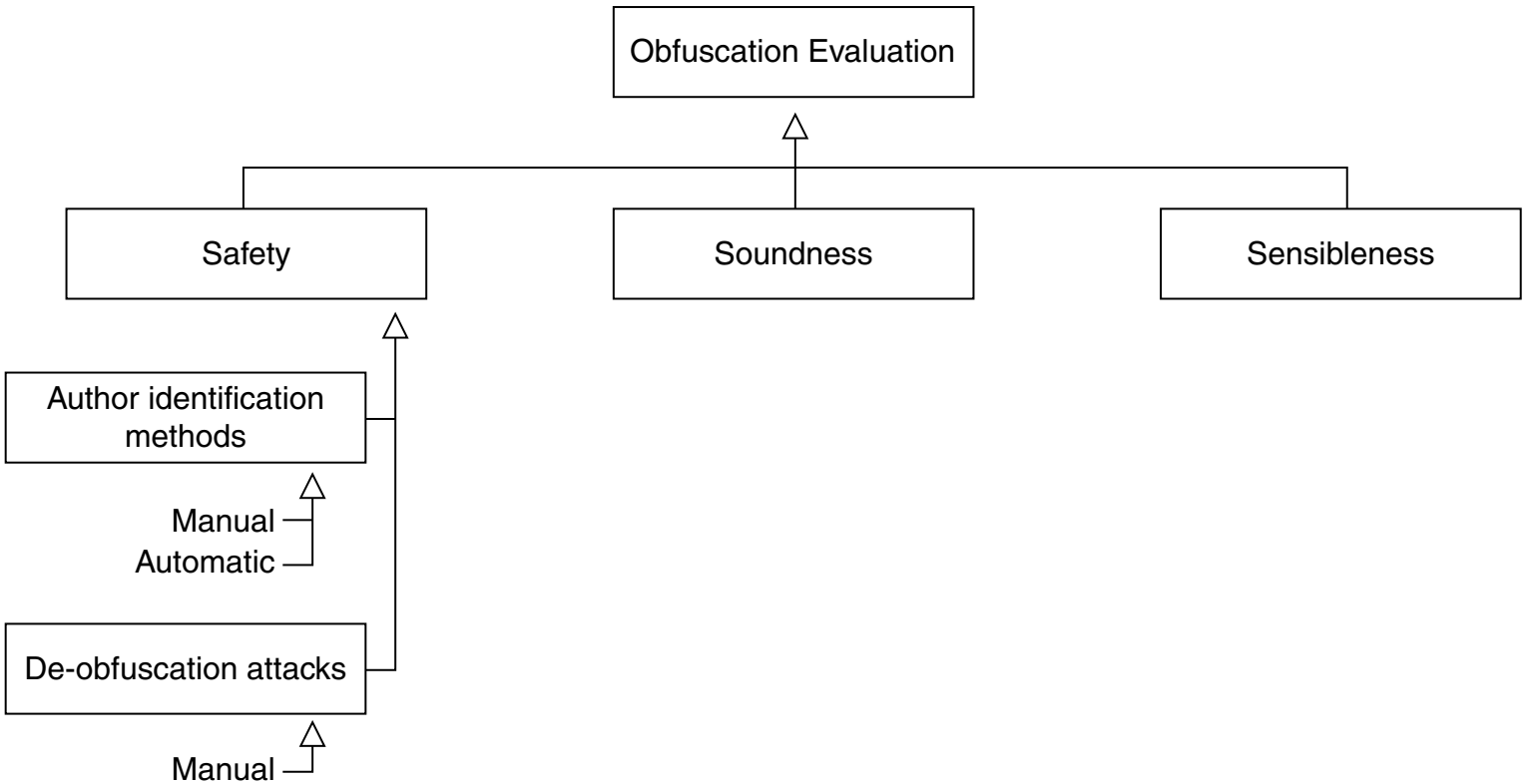


We call an obfuscation software

- **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

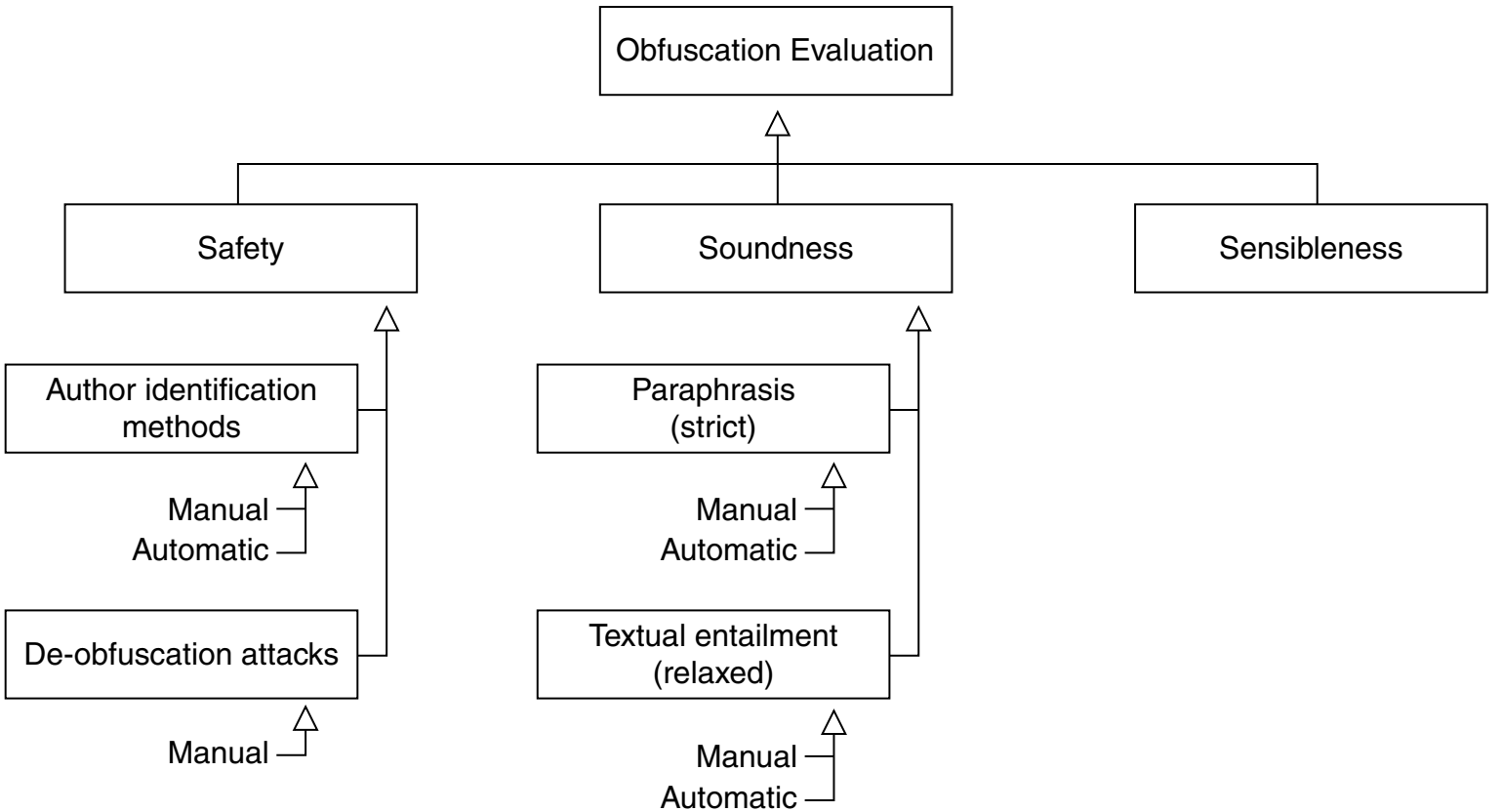
Taxonomy of Evaluation Dimensions



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

Obfuscation Evaluation

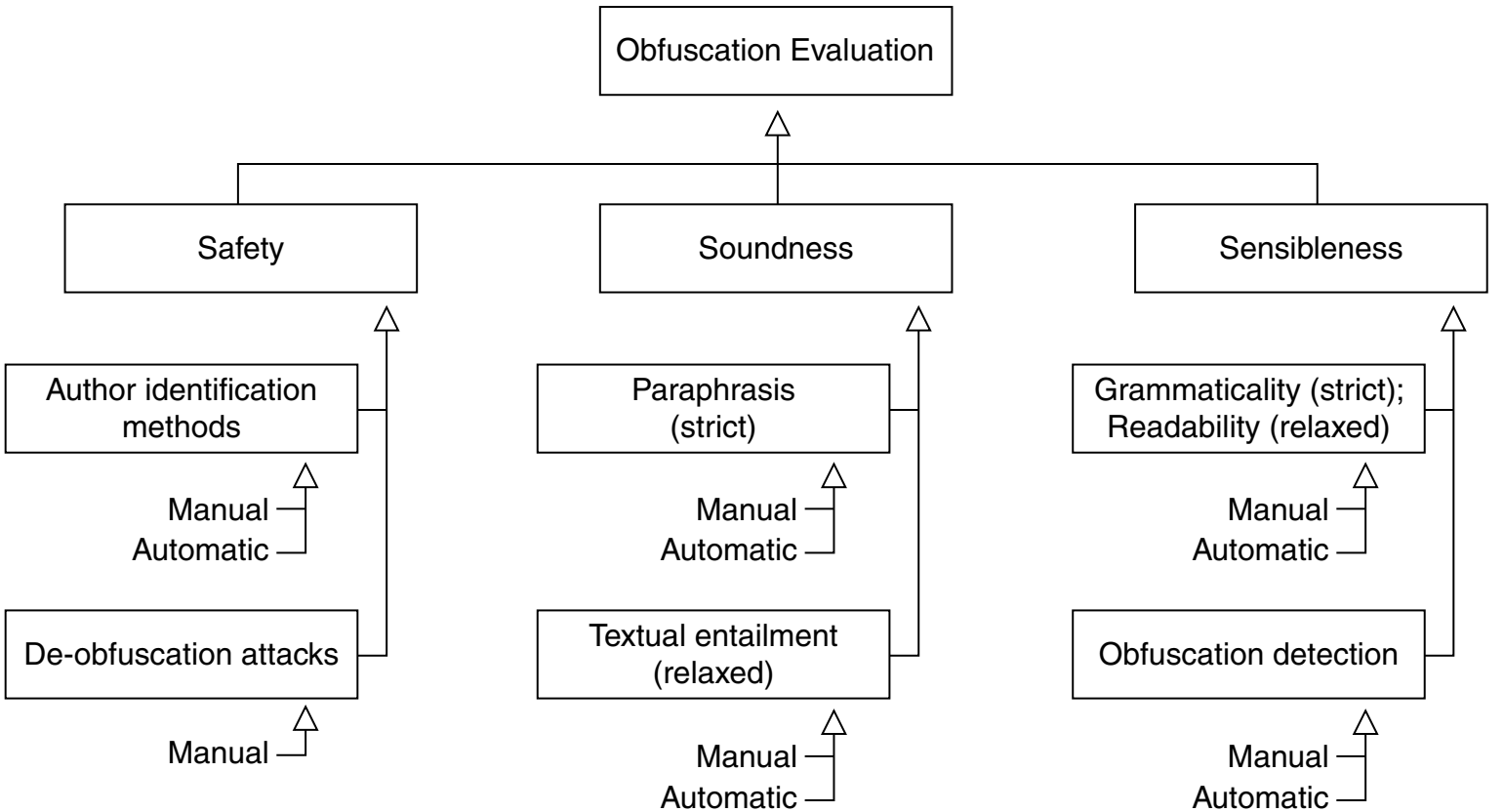
Taxonomy of Evaluation Dimensions



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

Obfuscation Evaluation

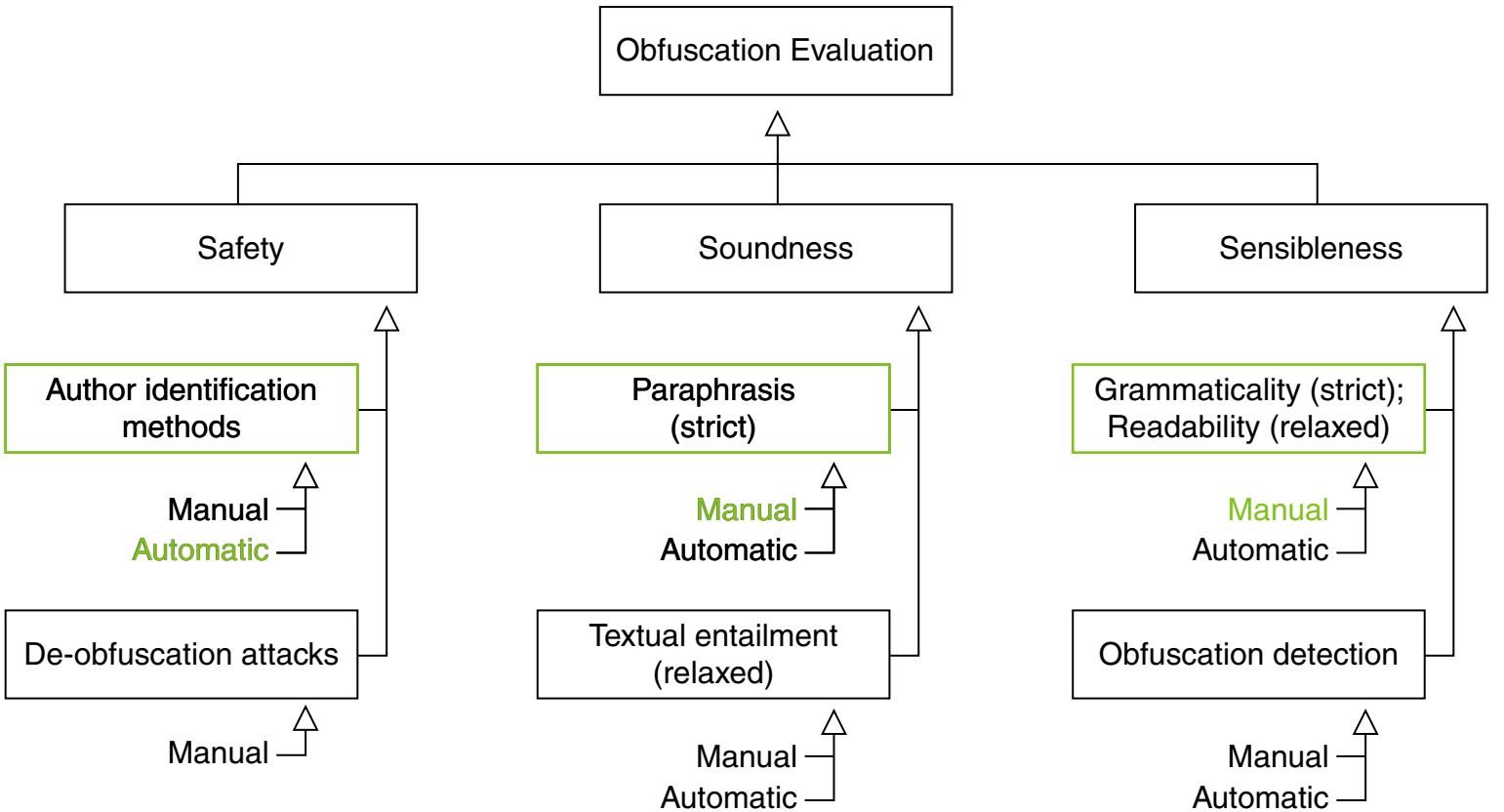
Taxonomy of Evaluation Dimensions



- 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

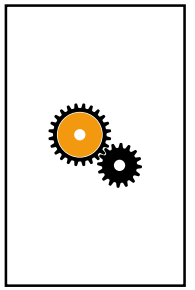
Taxonomy of Evaluation Dimensions

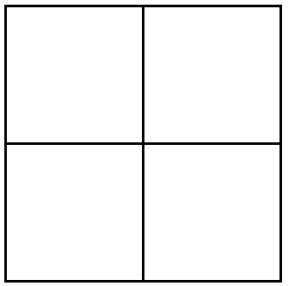


- Evaluations conducted in our shared task

Obfuscation Evaluation

Shared Task Setup

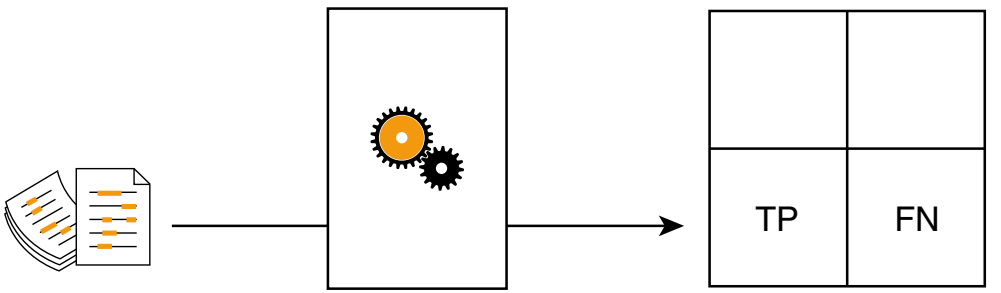




Obfuscation Evaluation

Shared Task Setup

PAN 13/14/15: Authorship Verification | Evaluation

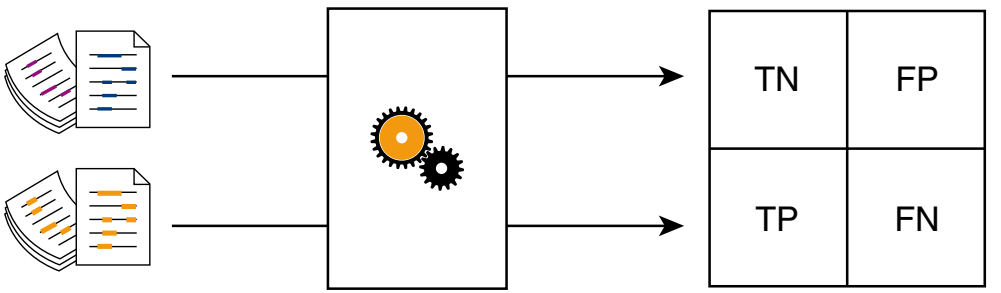


- TP = true positive
- FN = false negative

Obfuscation Evaluation

Shared Task Setup

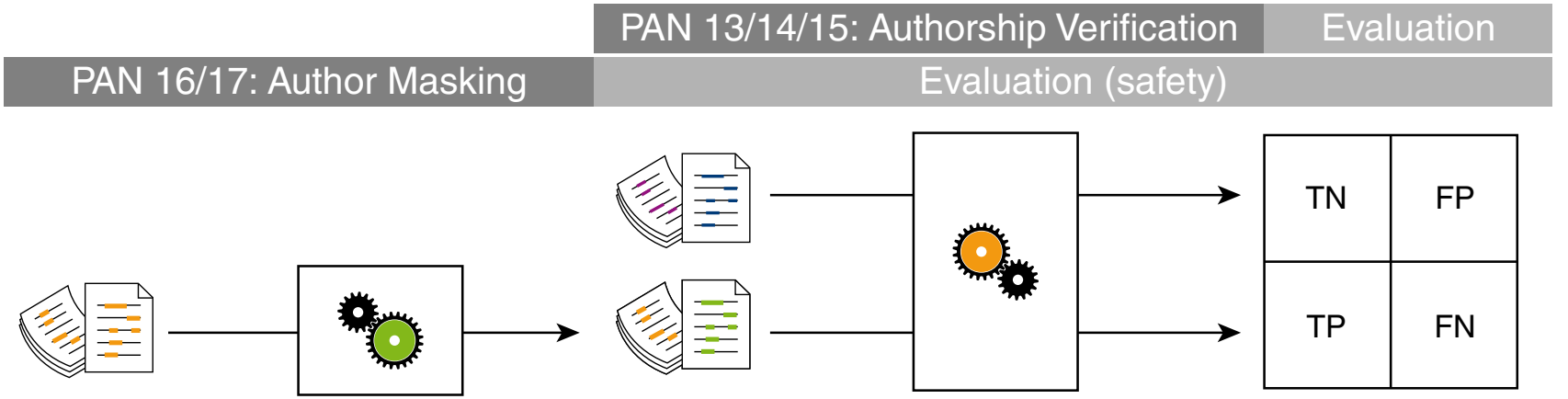
PAN 13/14/15: Authorship Verification | Evaluation



- TP = true positive
- FN = false negative
- TN = true negative
- FP = false positive

Obfuscation Evaluation

Shared Task Setup

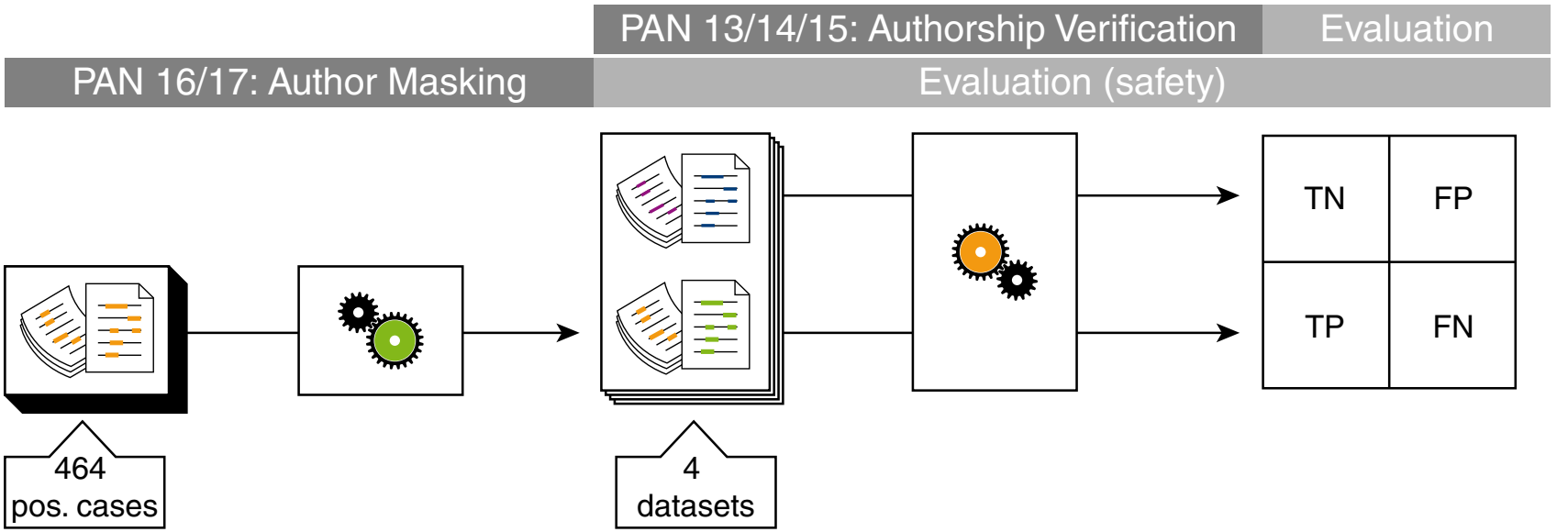


This setup tells us

- whether an obfuscator can defeat a verifier

Obfuscation Evaluation

Shared Task Setup

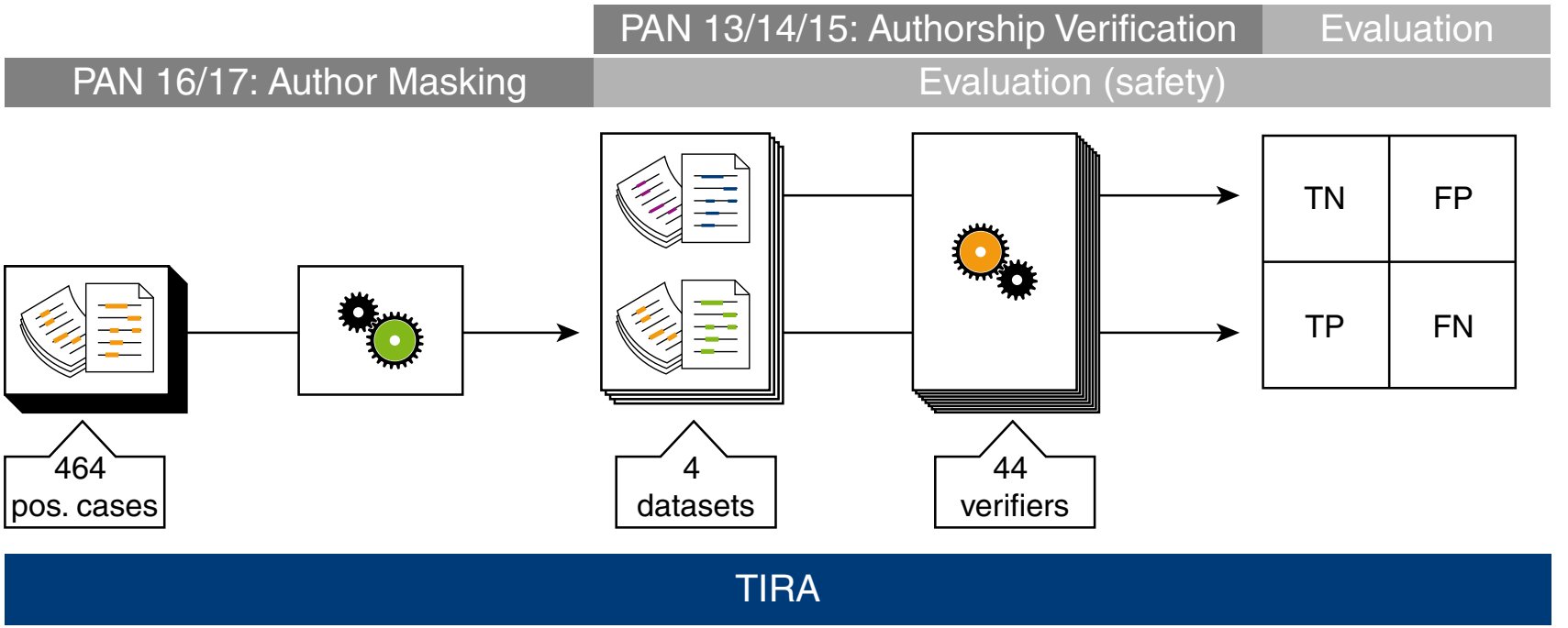


This setup tells us

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

Obfuscation Evaluation

Shared Task Setup

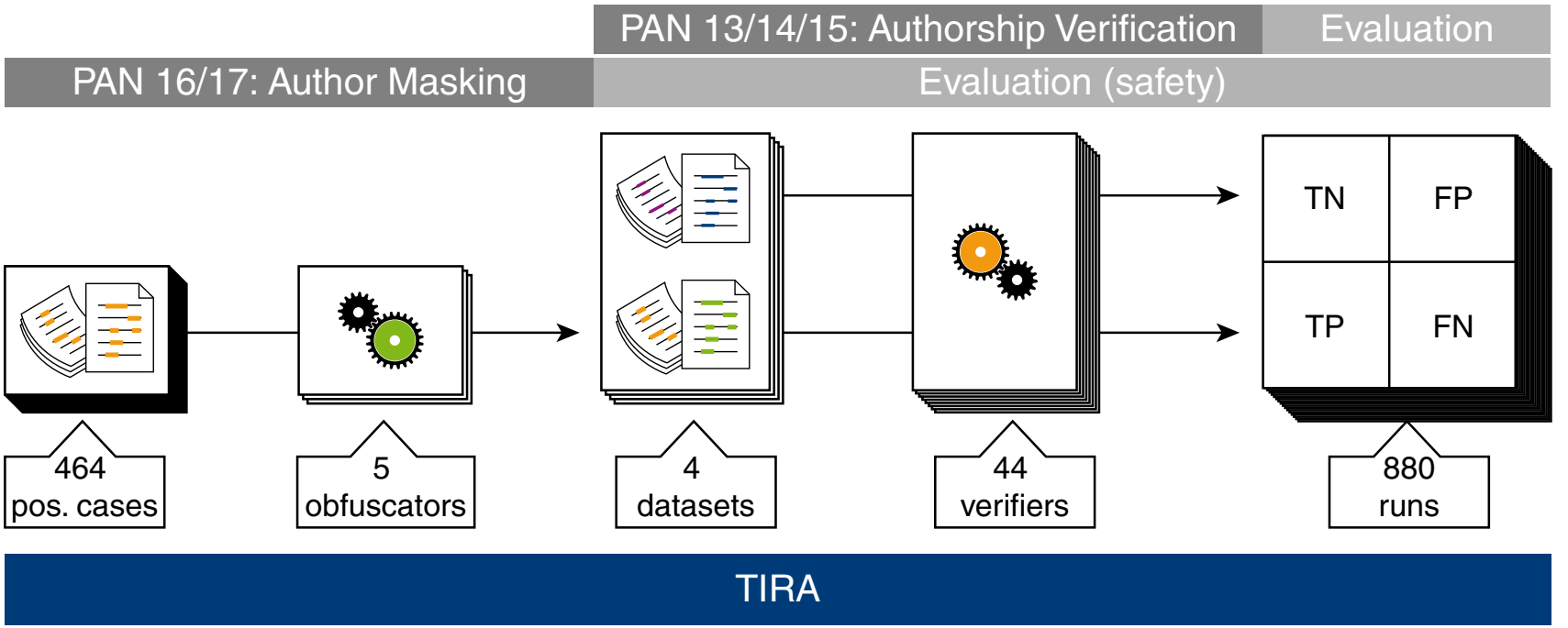


This setup tells us

- 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

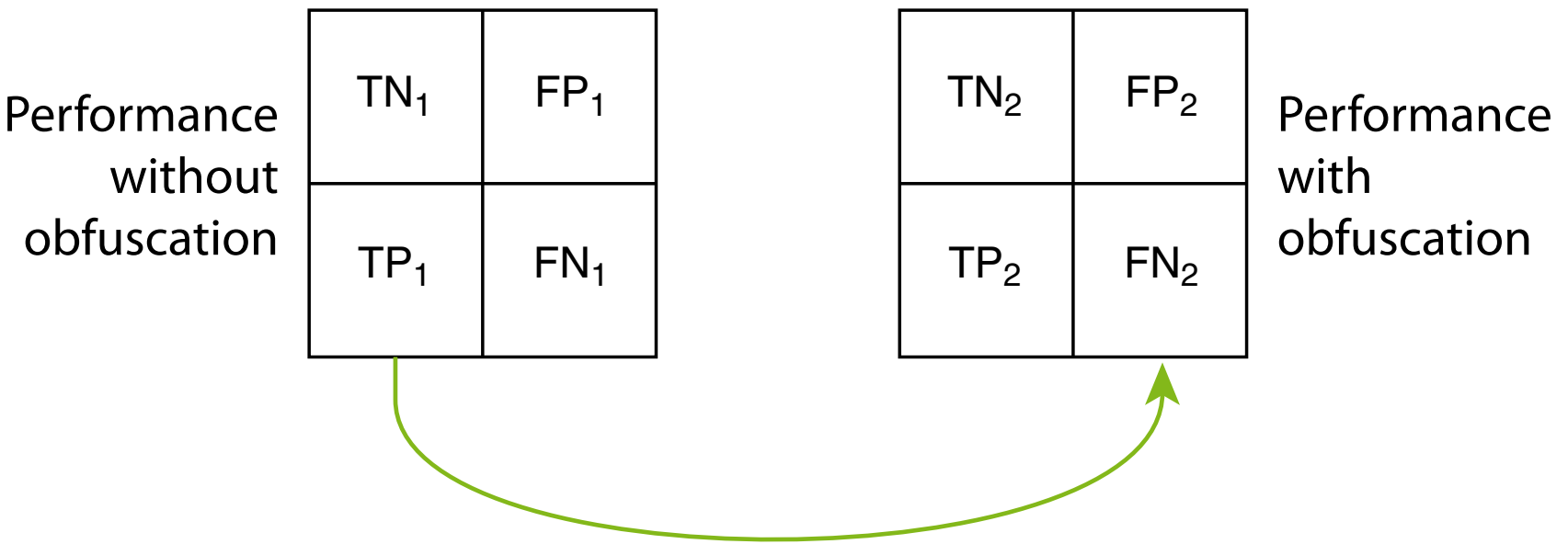
TN_1	FP_1
TP_1	FN_1

TN_2	FP_2
TP_2	FN_2

Performance with obfuscation

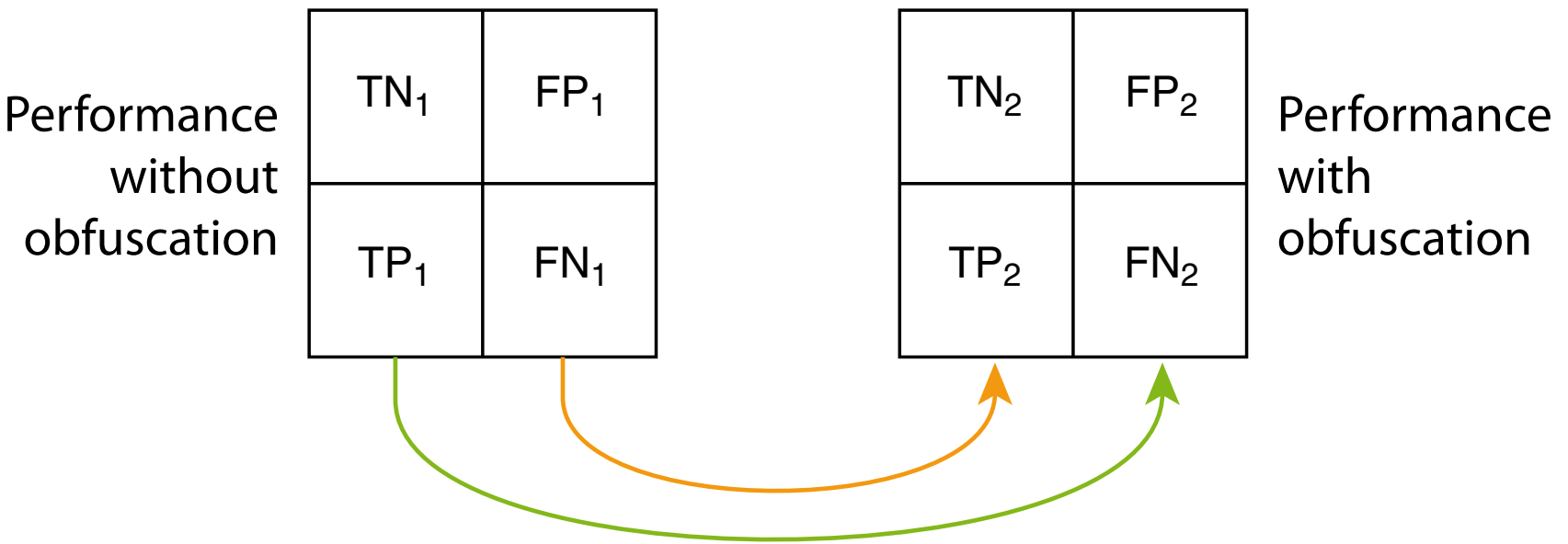
Obfuscation Evaluation

Measuring Obfuscation Impact



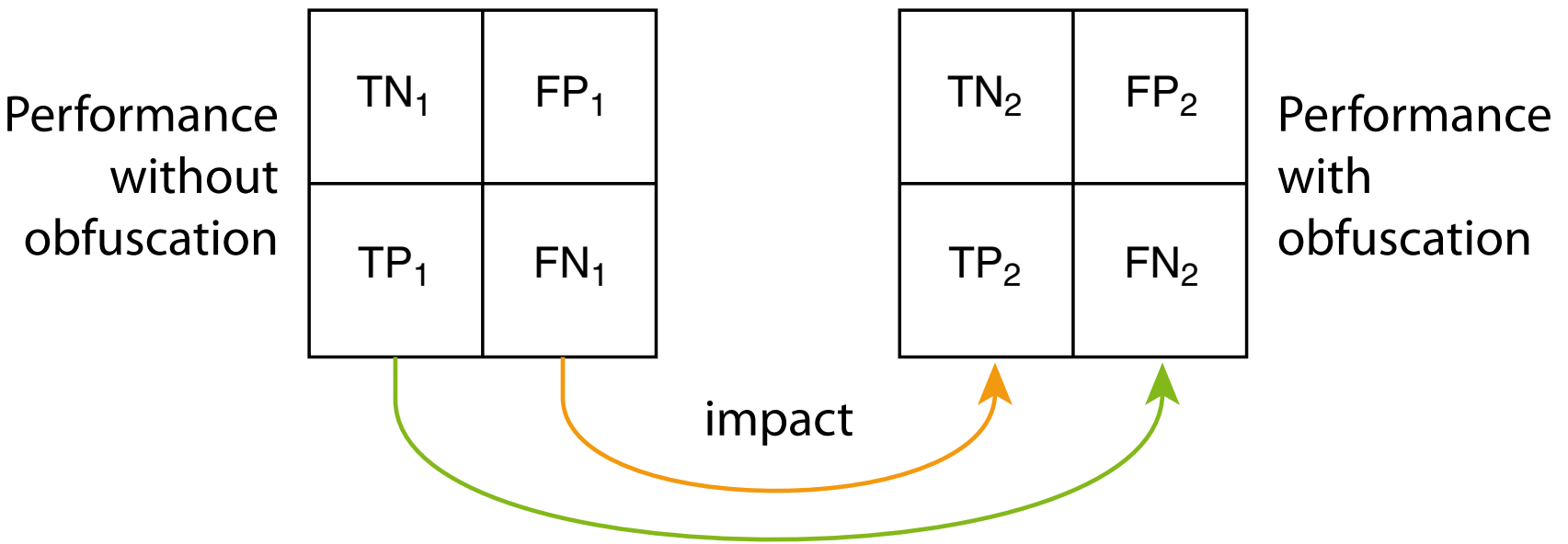
Obfuscation Evaluation

Measuring Obfuscation Impact



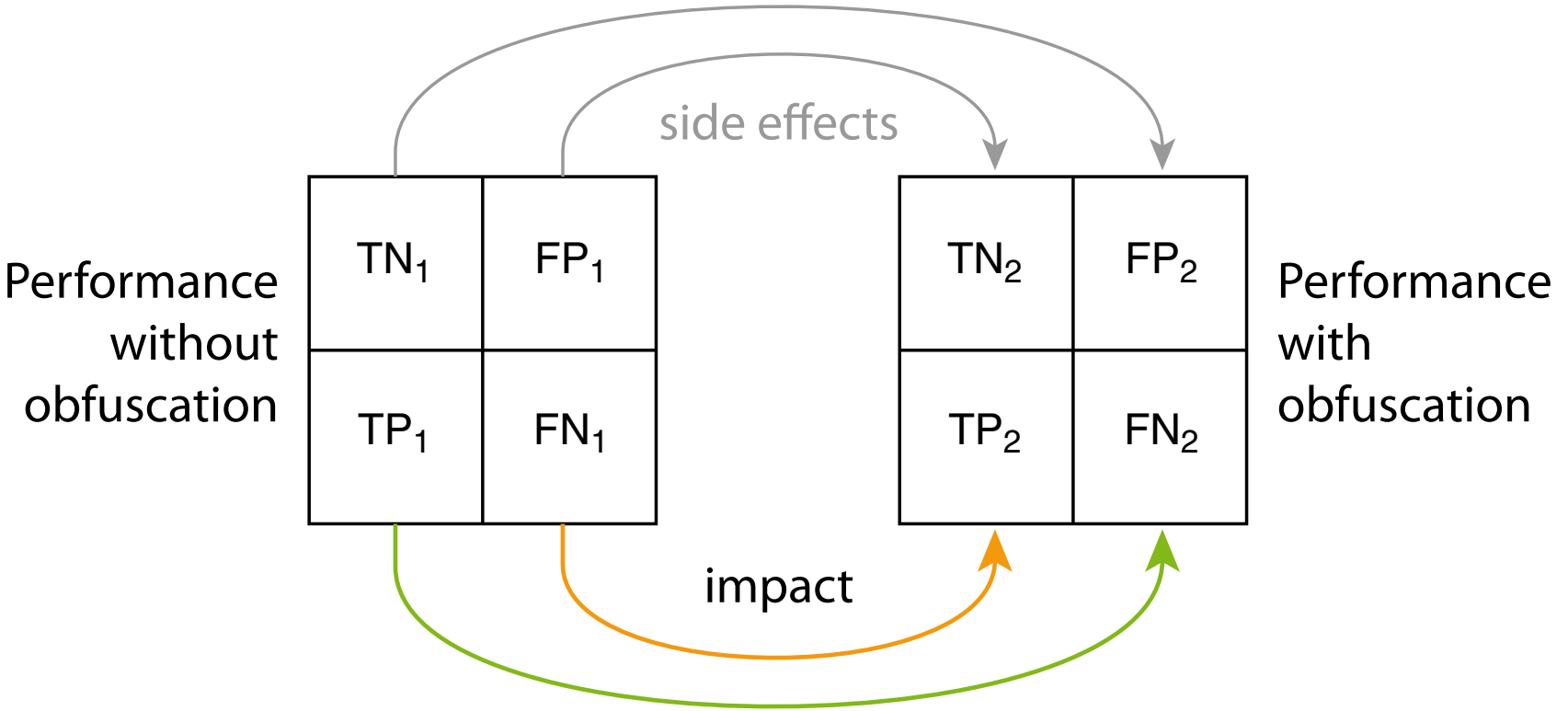
Obfuscation Evaluation

Measuring Obfuscation Impact



Obfuscation Evaluation

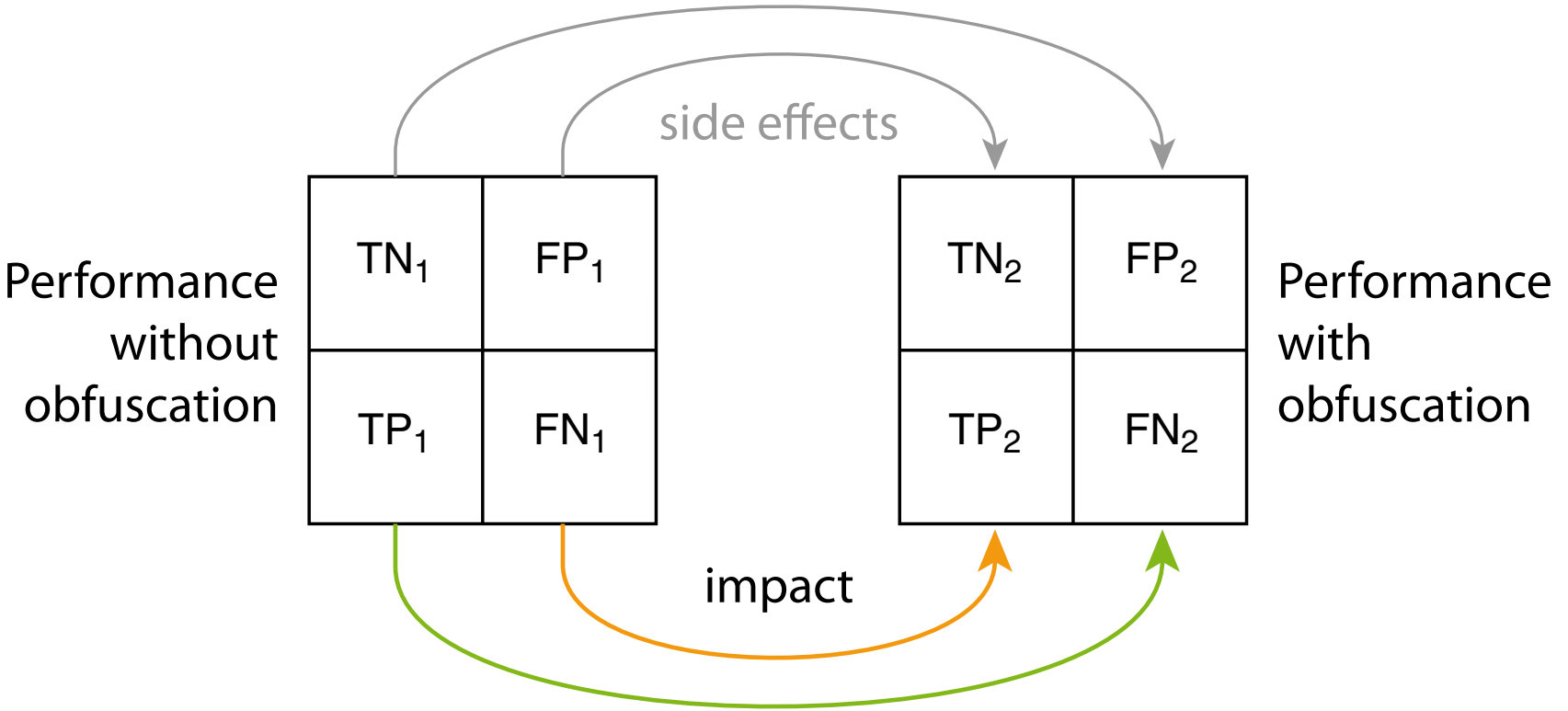
Measuring Obfuscation Impact



- 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

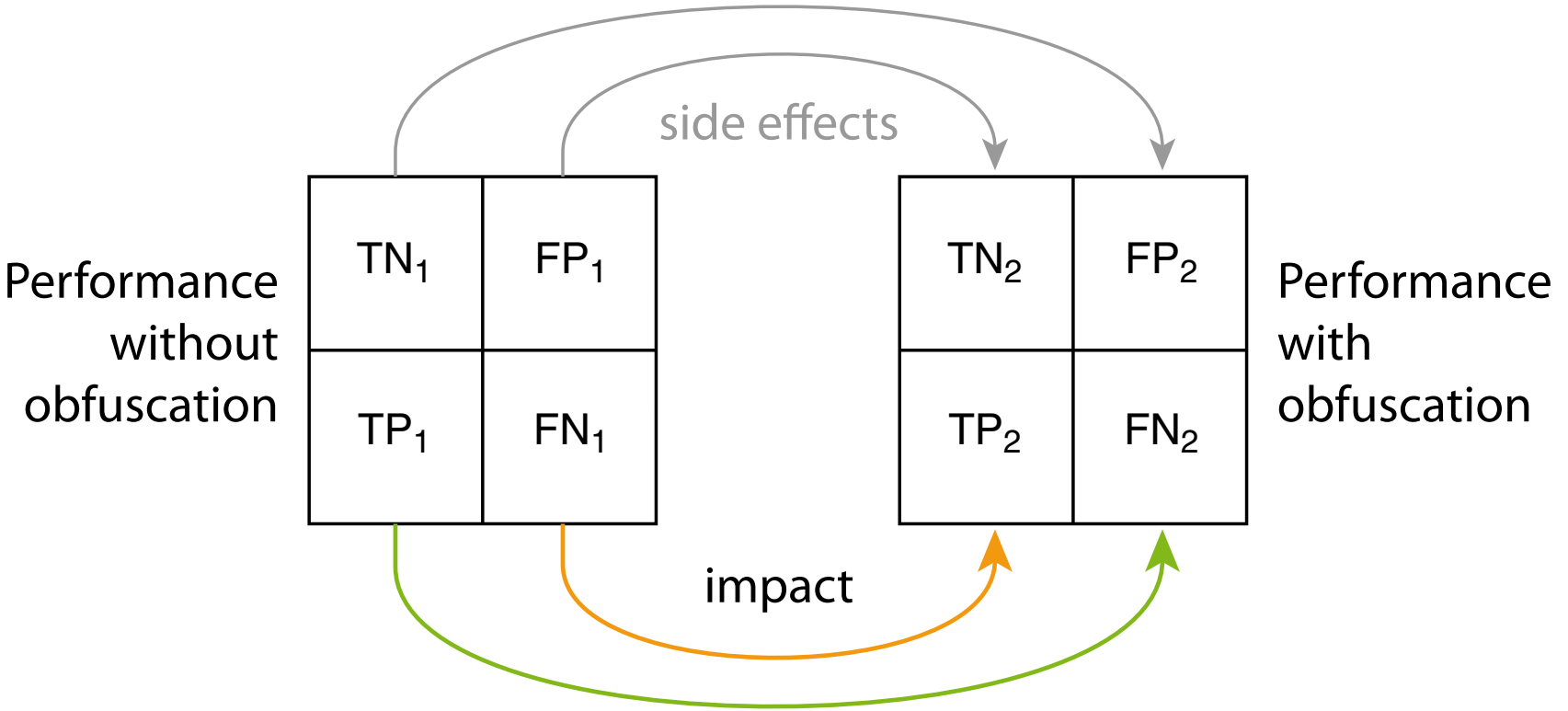


$$\text{rec}_1 = \frac{\text{TP}_1}{\text{TP}_1 + \text{FN}_1} \qquad \text{rec}_2 = \frac{\text{TP}_2}{\text{TP}_2 + \text{FN}_2}$$

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

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**
Castro et al.	PAN13	14	-0.2449	0.4175
Keswani et al.	PAN13	14	-0.2361	0.4245
Bakhteev et al.	PAN13	14	-0.1667	0.2881
Mansoorizadeh et al.	PAN13	14	-0.0933	0.1442
Mihaylova et al.	PAN14 EE	100	**-0.2304**	**0.4891**
Castro et al.	PAN14 EE	100	-0.2273	0.4328
Keswani et al.	PAN14 EE	100	-0.1873	0.4058
Bakhteev et al.	PAN14 EE	100	-0.1177	0.2558
Mansoorizadeh et al.	PAN14 EE	100	-0.1038	0.2512
Mihaylova et al.	PAN14 EN	100	**-0.2456**	**0.4750**
Castro et al.	PAN14 EN	100	-0.1900	0.3811
Keswani et al.	PAN14 EN	100	-0.1783	0.3769
Bakhteev et al.	PAN14 EN	100	-0.1129	0.2354
Mansoorizadeh et al.	PAN14 EN	100	-0.0958	0.2345
Mihaylova et al.	PAN15	250	**-0.2009**	**0.3649**
Castro et al.	PAN15	250	-0.1973	0.3087
Keswani et al.	PAN15	250	-0.1298	0.2543
Bakhteev et al.	PAN15	250	-0.1314	0.2172
Mansoorizadeh et al.	PAN15	250	-0.0994	0.1952