

# Algorithms for Authorship Analytics

---

Benno Stein  
Bauhaus-Universität Weimar  
Intelligent Information Systems Group  
[webis.de](http://webis.de)

Lange Nacht der Wissenschaften · Weimar · November 18, 2022

## Dialogue and co-operation are the only way to cyber security



Graeme Hirst

As one of the greatest inventions in the 20th century, the internet has brought profound changes to our way of thinking, working and living. At the same time, it is prone to security risks and challenges. Wiretapping, attacks and terrorism in cyberspace have become global problems that call for global solutions. This means countries must work together instead of accusing one country for all the problems as some countries recently did against China, not to mention how

a lawless land. No country would tolerate fraud, cheating, stealing, terrorism or incitement of religious extremism.

The Chinese government has no part in stealing commercial secrets, nor do we in any way encourage or support any individual or company to do so. On the contrary, China has been opposing and cracking down on all forms of cyber theft all along.

In recent years, China has strengthened rule of law in cyberspace and kept improving the relevant laws and regulations: the Cyber Security Law and The National Cyberspace Security Strategy were issued in 2016; The first internet court was established in Hangzhou in 2017, followed by the second and third in Beijing and

As a responsible big country, China has been actively pushing for bilateral and multilateral cooperation on cyber security, engaging with the US, the UK and the EU through dialogue mechanisms, and sharing China's wisdom at the UN and the G20 on improving international cooperation in cyberspace. Moreover, China has hosted five sessions of the World Internet Conference since 2014 to promote international cooperation on cyber security and cyber governance.

All these show that the accusations against China on cyber security are unfair, groundless and the opposite of the fact. People of the world need not be reminded who has conducted massive cyber wiretapping against foreign governments – even allies, who has engaged in organised cyber theft

# Outline

- ① Some Technology Basics
- ② Authorship Identification
- ③ Authorship Obfuscation



## Some Technology Basics

# Bag of Words Model

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 ...
```

# Bag of Words Model

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

# Bag of Words Model

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

# Bag of Words Model

After stop word removal and 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

# Bag of Words Model

After stop word removal and 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

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

# Bag of Words Model

After stop word removal and 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

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

# Bag of Words Model

After stop word removal and 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

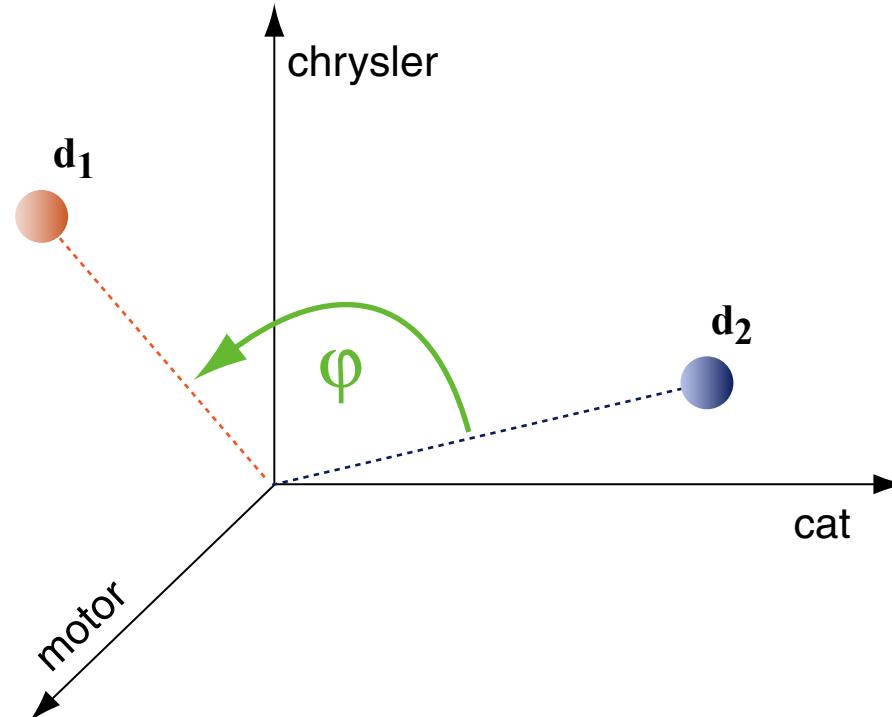
$$\mathbf{d} = \begin{pmatrix} \text{chrysler} & w_1 \\ \text{motor} & w_2 \\ \dots & \\ \text{cat} & w_x \\ \text{dog} & w_y \\ \text{mouse} & w_z \end{pmatrix} \rightsquigarrow \left\langle \begin{pmatrix} \text{chrysler} & 0.2 \\ \text{motor} & 0.3 \\ \dots & \\ \text{cat} & 0.1 \\ \text{dog} & 0.0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

Comparison via normalized dot product.

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

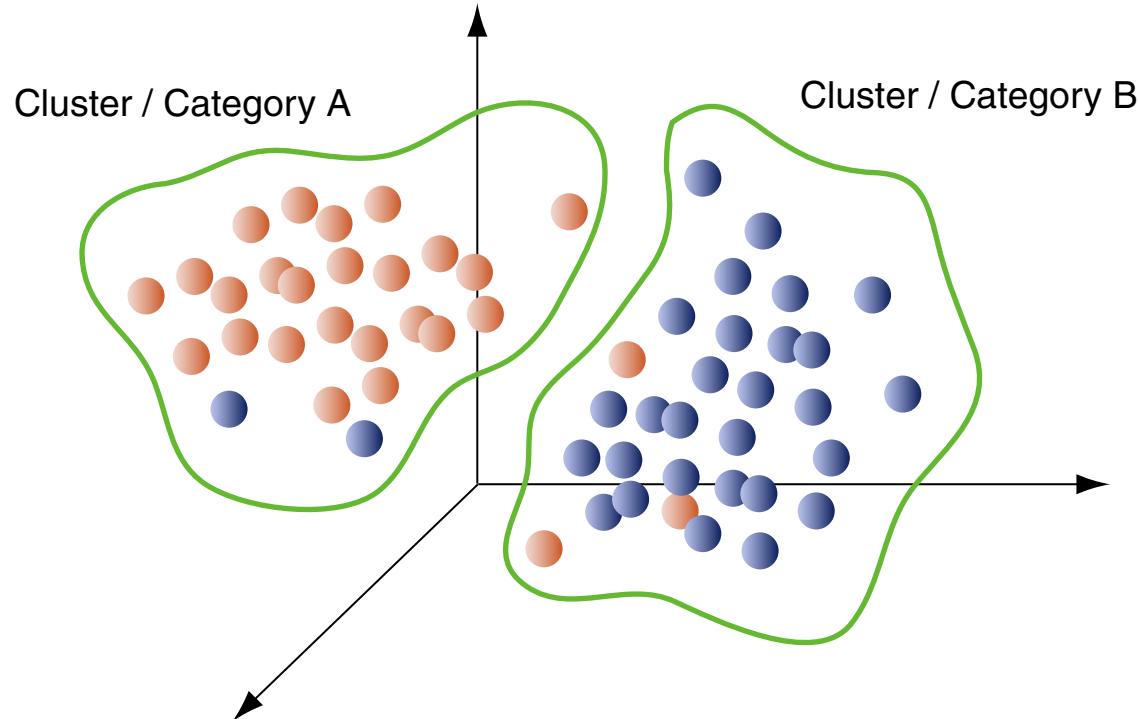
# Vector Space Illustration

Similarities correspond to angles:



# Vector Space Illustration

Document grouping via cluster analysis:



# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The **e** migrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The **m i**grants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The **m i g**igrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The m**i g r**ants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The mi**g r a**nts who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migr**a n t**s who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrant s who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrant**ts** who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a col**ony**, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, and, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **a**nd, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **a n d**, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **and**, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **and**, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

$$\mathbf{d} = \begin{pmatrix} \text{and } c_1 \\ \text{to\_ } c_2 \\ \text{our } c_3 \\ \text{the } c_4 \\ \text{\_sh } c_5 \\ \dots \end{pmatrix}$$

# Char-n-Gram Model

Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **and**, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

$$\mathbf{d} = \begin{pmatrix} \text{and } c_1 \\ \text{to\_ } c_2 \\ \text{our } c_3 \\ \text{the } c_4 \\ \text{\_sh } c_5 \\ \dots \end{pmatrix} \rightsquigarrow \begin{pmatrix} \text{and } 5 \\ \text{to\_ } 3 \\ \text{our } 3 \\ \text{the } 7 \\ \text{\_sh } 2 \\ \dots \end{pmatrix}, \begin{pmatrix} \text{and } 3 \\ \text{to\_ } 6 \\ \text{our } 5 \\ \text{the } 4 \\ \text{\_sh } 1 \\ \dots \end{pmatrix}$$

# Char-n-Gram Model

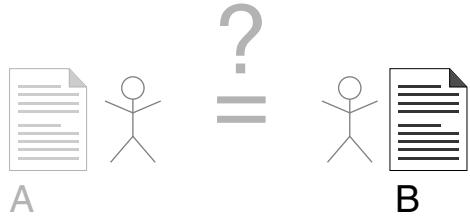
Char-trigrams → sliding window with  $n = 3$ :

The migrants who sailed with Gilbert were better fitted  
for a crusade than a colony, **and**, disappointed at not  
at once finding mines of gold and silver, many deserted;  
and soon there were not enough sailors to man all the four  
ships. Accordingly, the Swallow was sent back...

$$\mathbf{d} = \begin{pmatrix} \text{and } c_1 \\ \text{to\_ } c_2 \\ \text{our } c_3 \\ \text{the } c_4 \\ \text{\_sh } c_5 \\ \dots \end{pmatrix} \rightsquigarrow \begin{pmatrix} \text{and } 5 \\ \text{to\_ } 3 \\ \text{our } 3 \\ \text{the } 7 \\ \text{\_sh } 2 \\ \dots \end{pmatrix}, \begin{pmatrix} \text{and } 3 \\ \text{to\_ } 6 \\ \text{our } 5 \\ \text{the } 4 \\ \text{\_sh } 1 \\ \dots \end{pmatrix} \rightsquigarrow \sum_{i \in \text{trigrams}} P[i] \log \frac{P[i]}{Q[i]} \rightarrow \min$$

Comparison via Kullback-Leibler or Jensen Shannon Divergence.

# Measuring Model Effectiveness



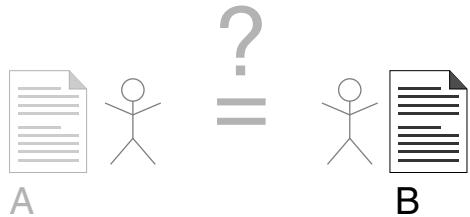
How reliable is our claim of  
same authorship?

~ **Precision**

Did we get all cases of same  
authorship?

~ **Recall**

# Measuring Model Effectiveness

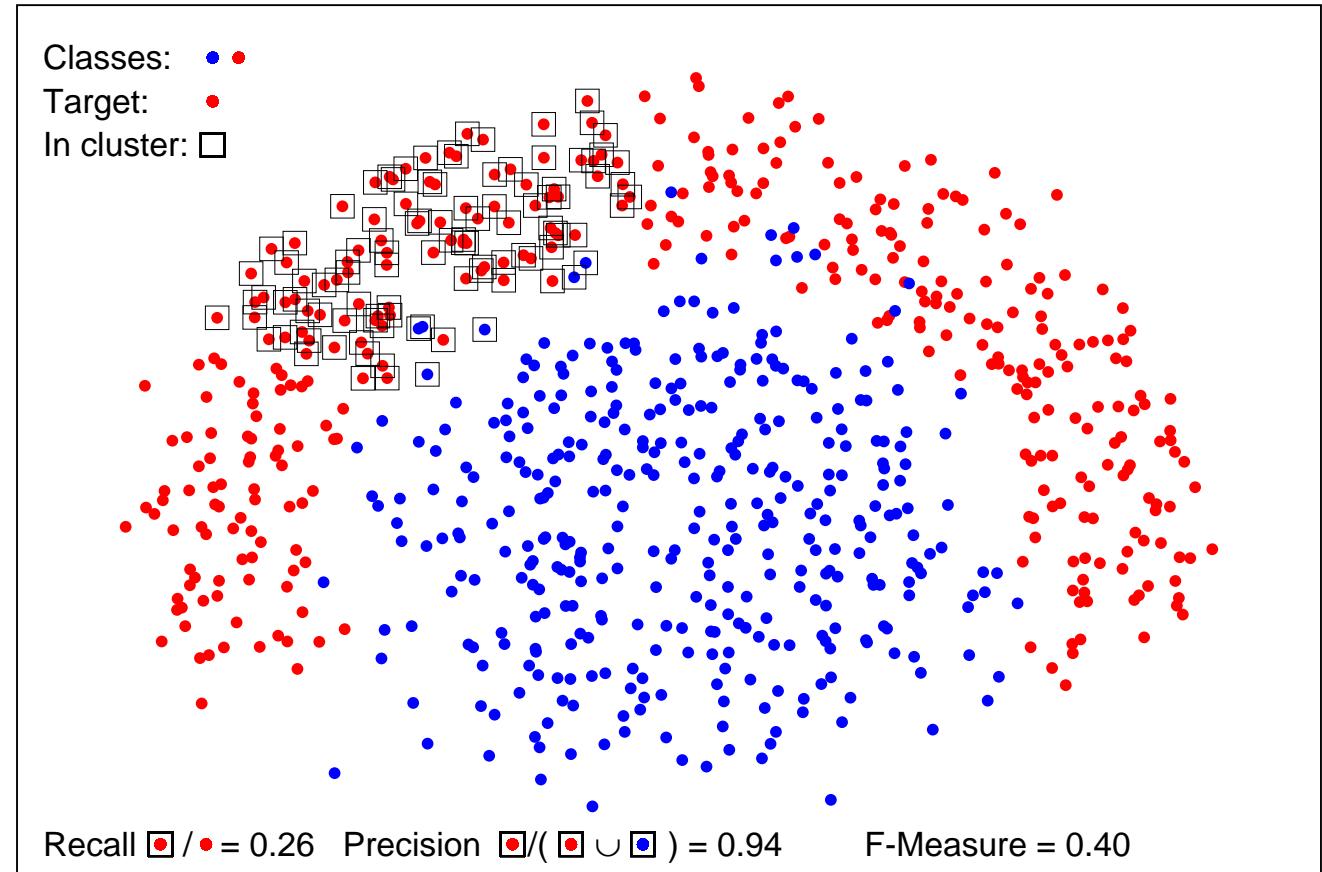


How reliable is our claim of same authorship?

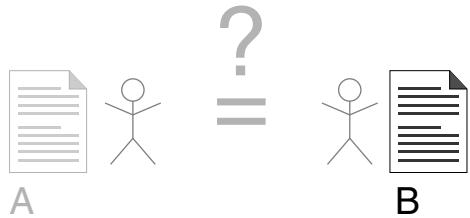
~ **Precision**

Did we get all cases of same authorship?

~ **Recall**



# Measuring Model Effectiveness

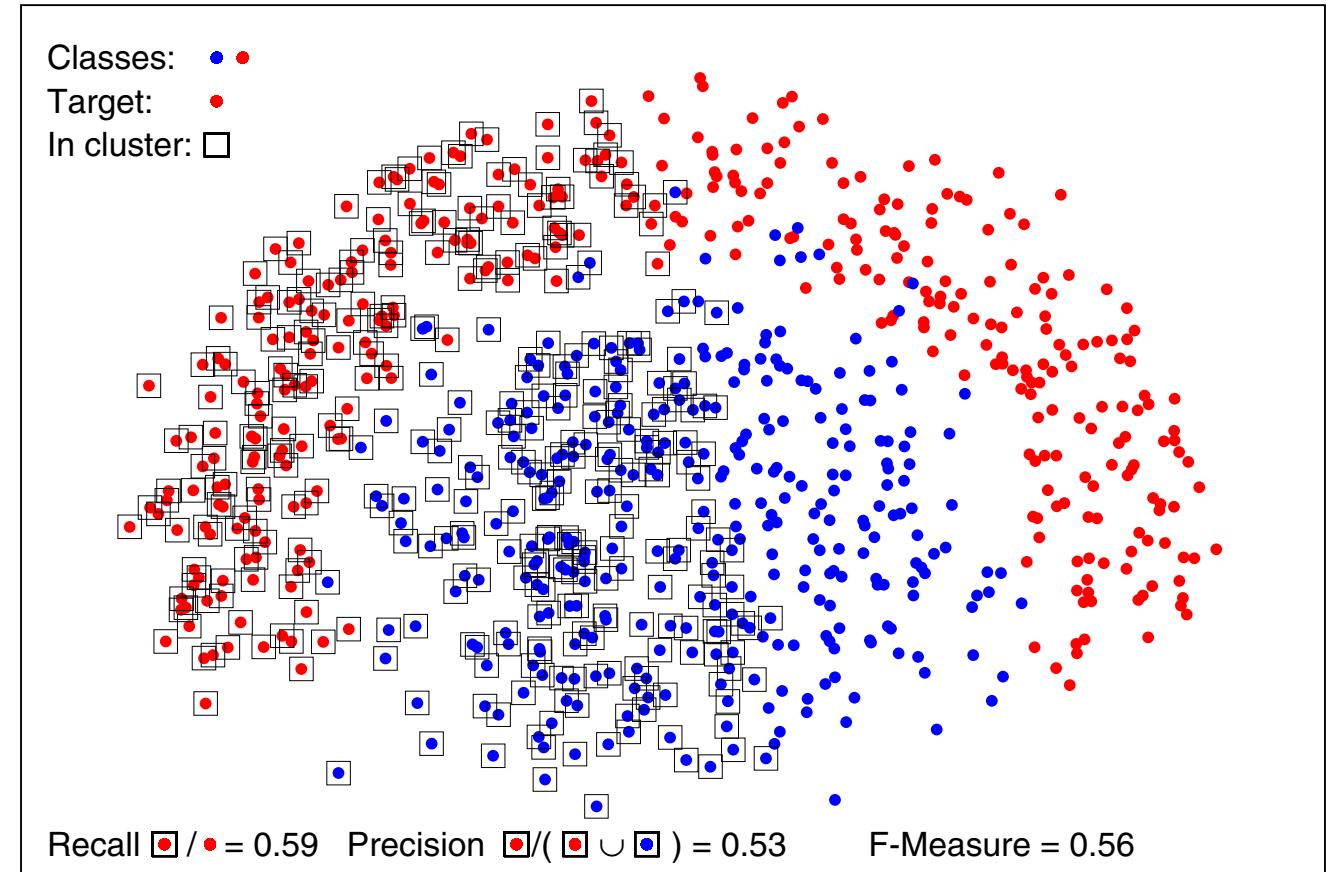


How reliable is our claim of same authorship?

~ **Precision**

Did we get all cases of same authorship?

~ **Recall**

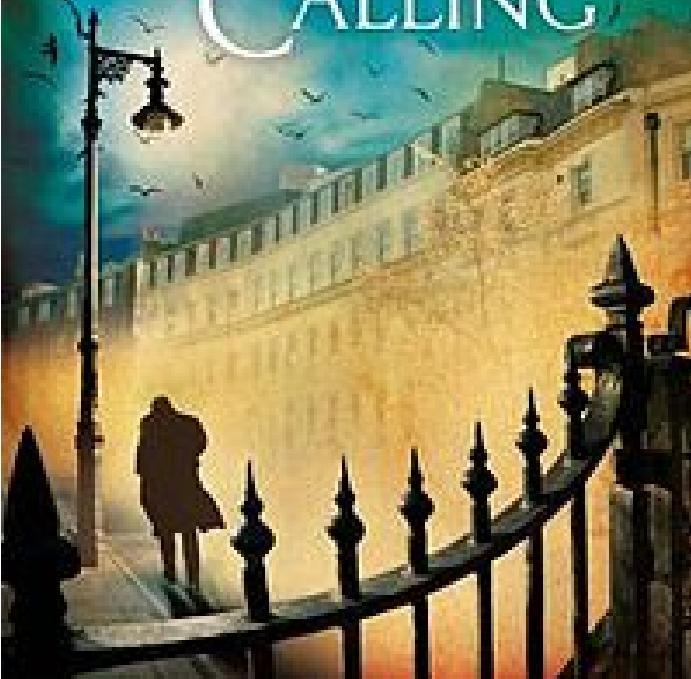




Authorship Identification

ROBERT GALBRAITH

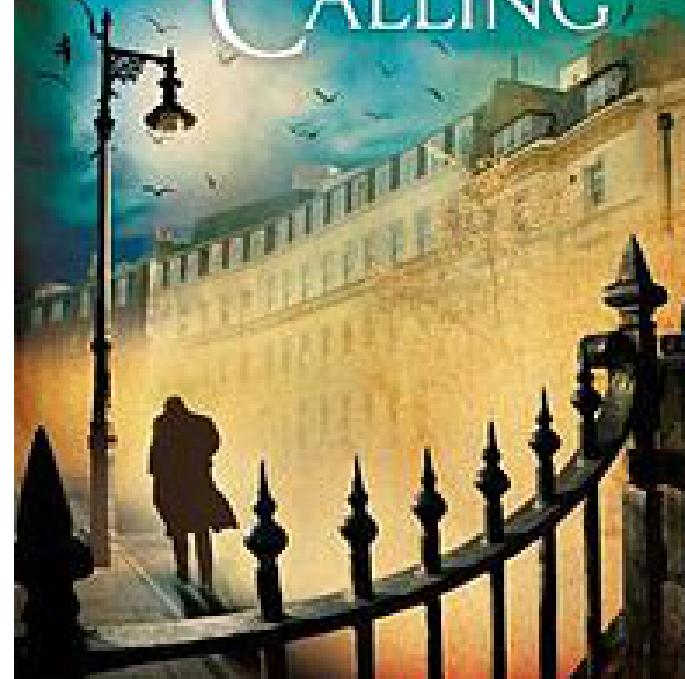
# THE CUCKOO'S CALLING



? =

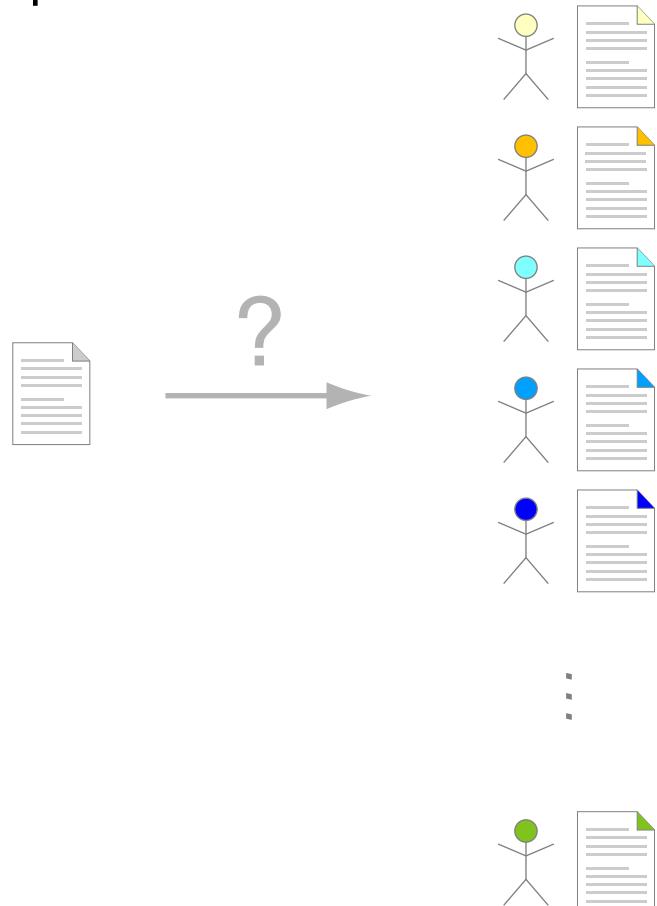
JOANNE K. ROWLING

# THE CUCKOO'S CALLING



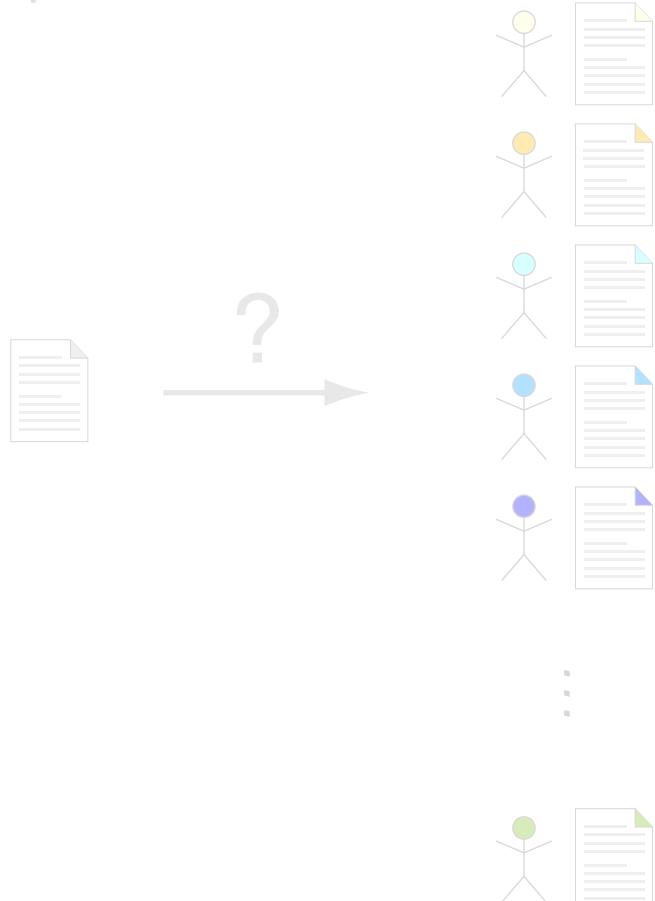
2013, Patrick Juola

# Authorship Attribution



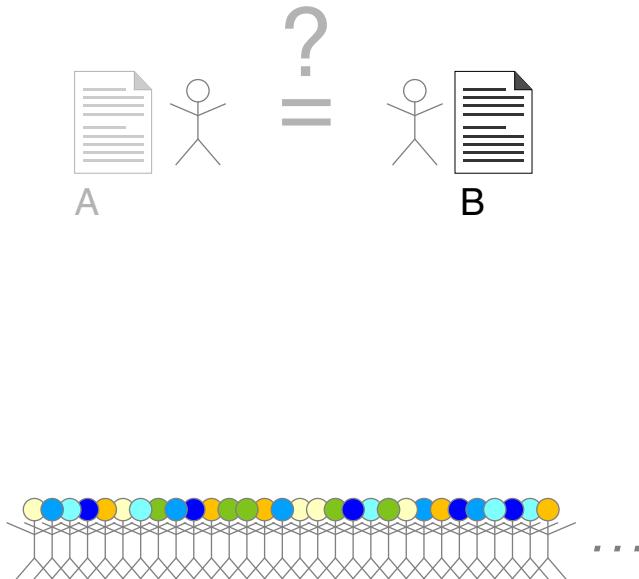
To which author does a text belong?

# Authorship Attribution



To which author does a text belong?

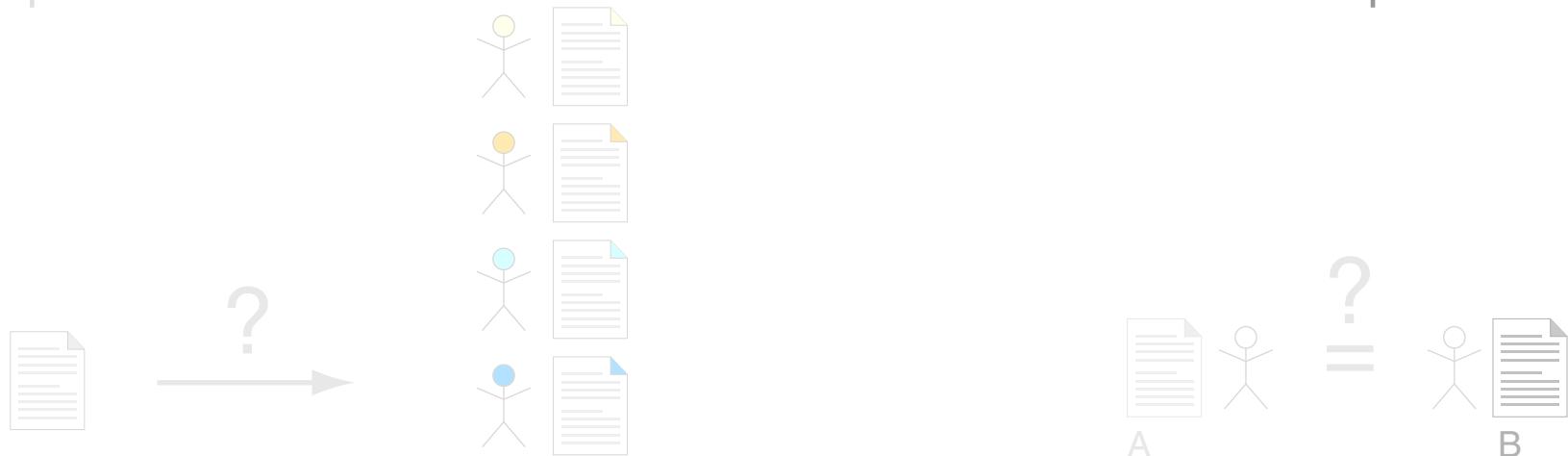
# Authorship Verification



Originate two texts from the same author?

# Authorship Attribution

# Authorship Verification



All authorship identification problems can be  
reduced to the verification problem.

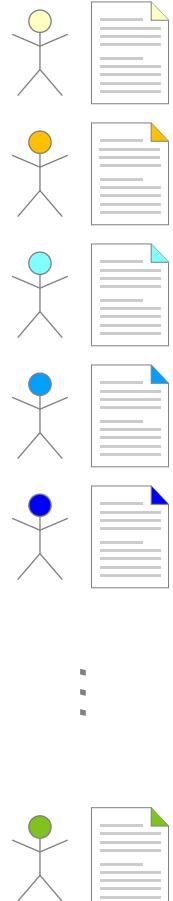


Authorship Verification is the ultimate problem.

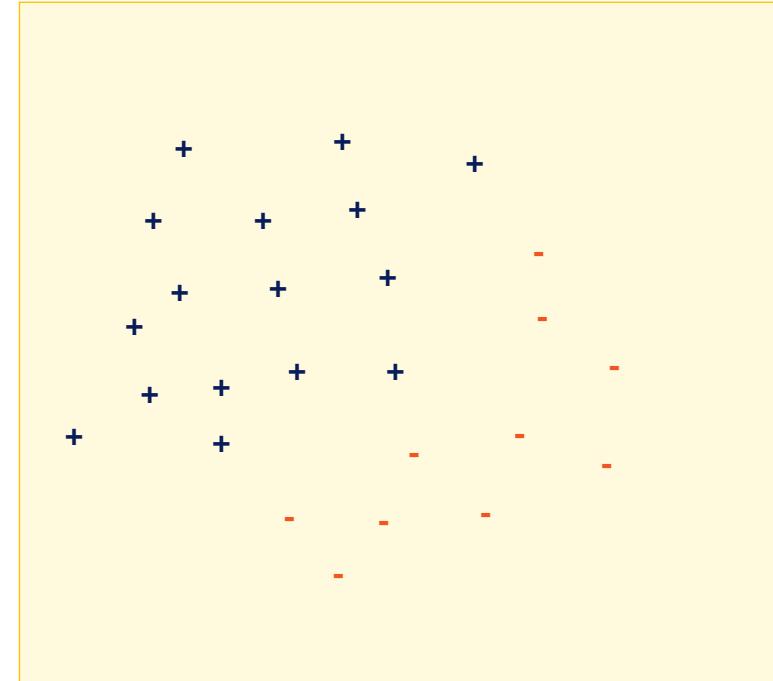
To which author does a text belong?

Originate two texts from the same author?

# Authorship Attribution

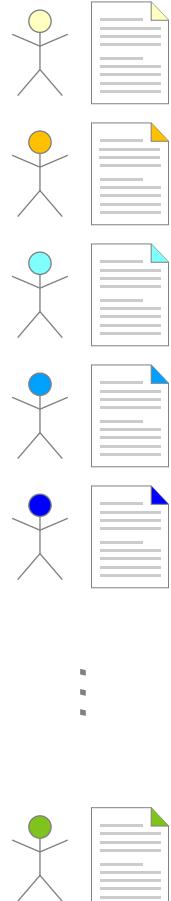


To which author does a text belong?

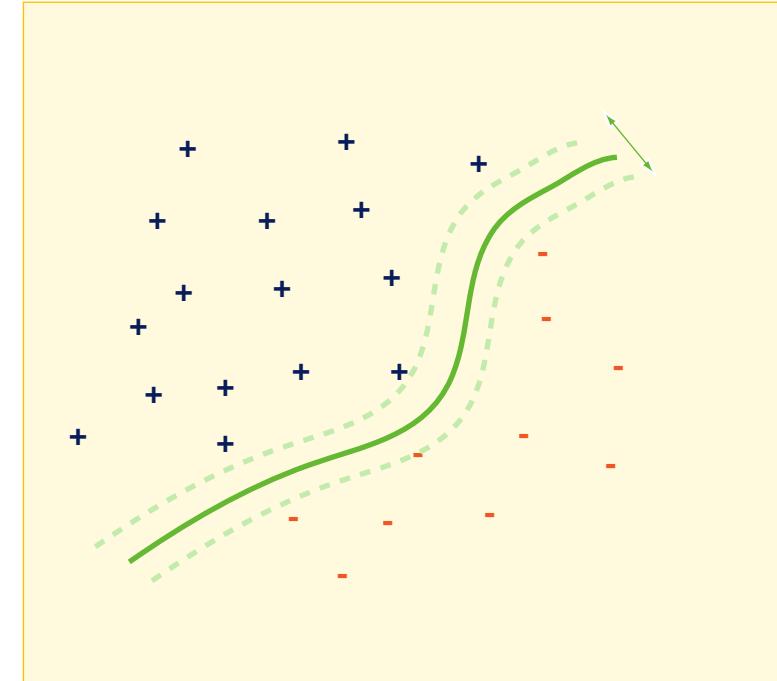


Discrimination-based classification.

# Authorship Attribution

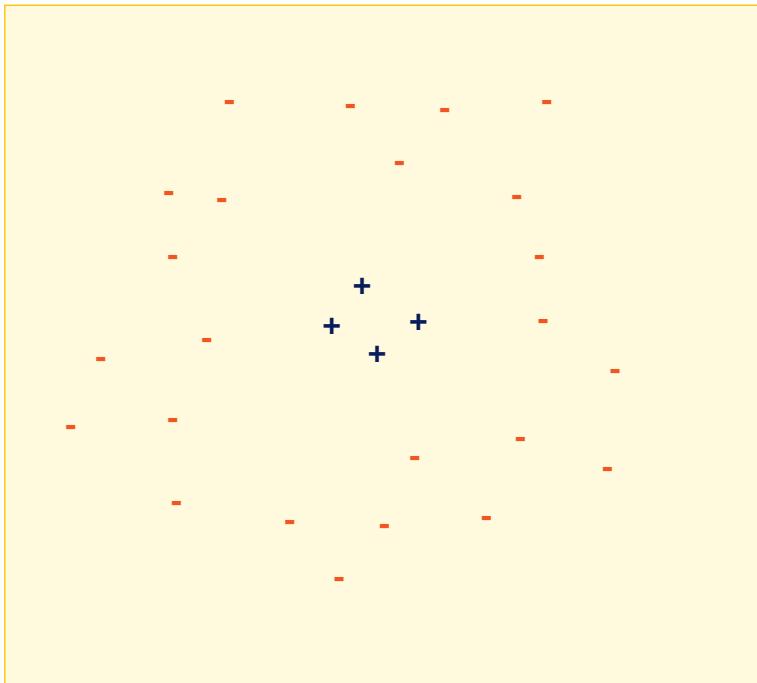


To which author does a text belong?

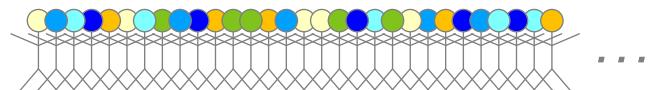
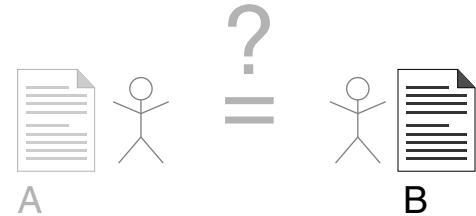


Discrimination-based classification.

# Authorship Verification

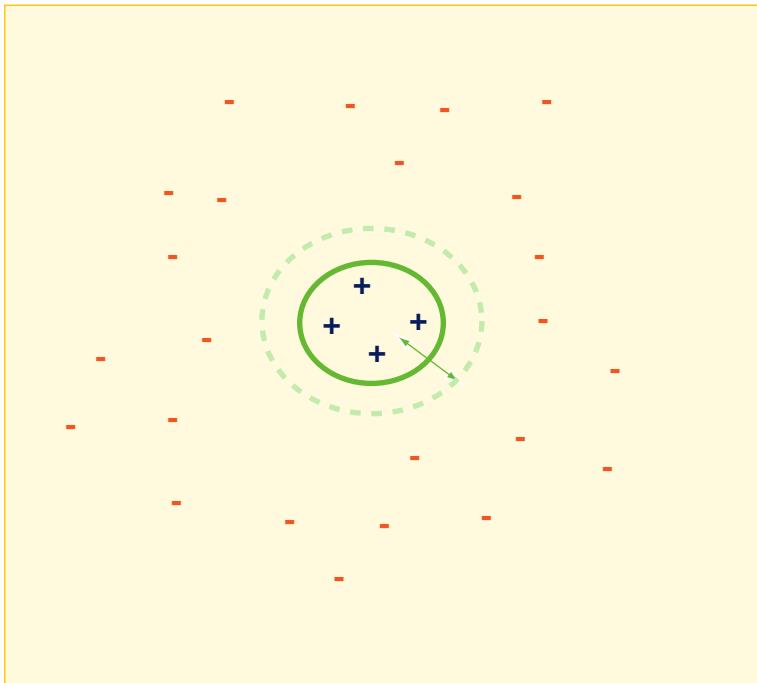


One-class classification.

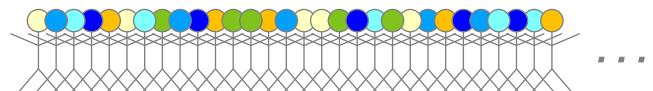
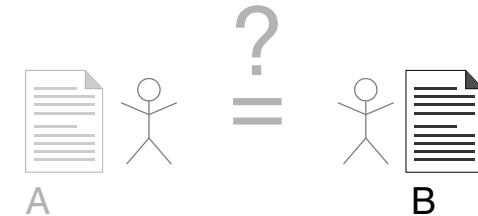


Originate two texts from the same author?

# Authorship Verification

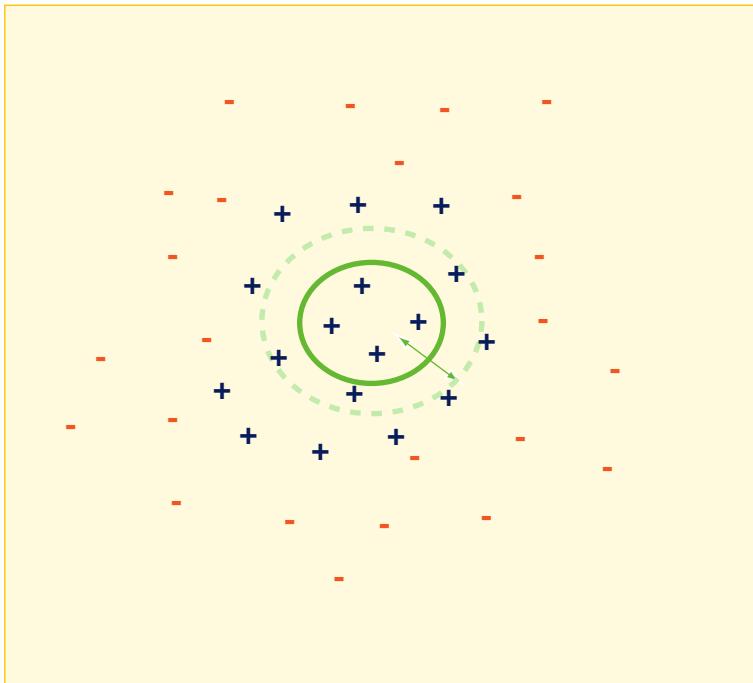


One-class classification.

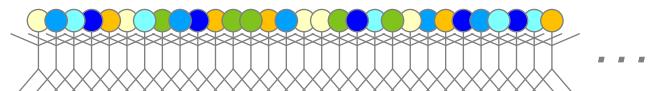
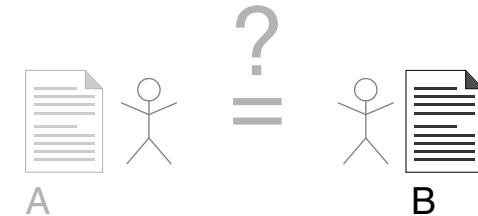


Originate two texts from the same author?

# Authorship Verification



One-class classification.



Originate two texts from the same author?



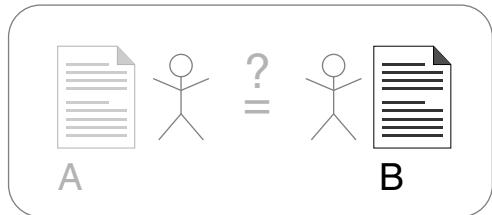
Moshe Koppel

Jonathan Schler

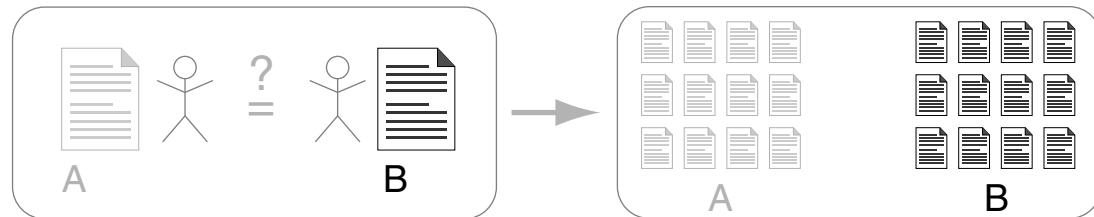
# The “Unmasking” Method for Authorship Verification

# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]

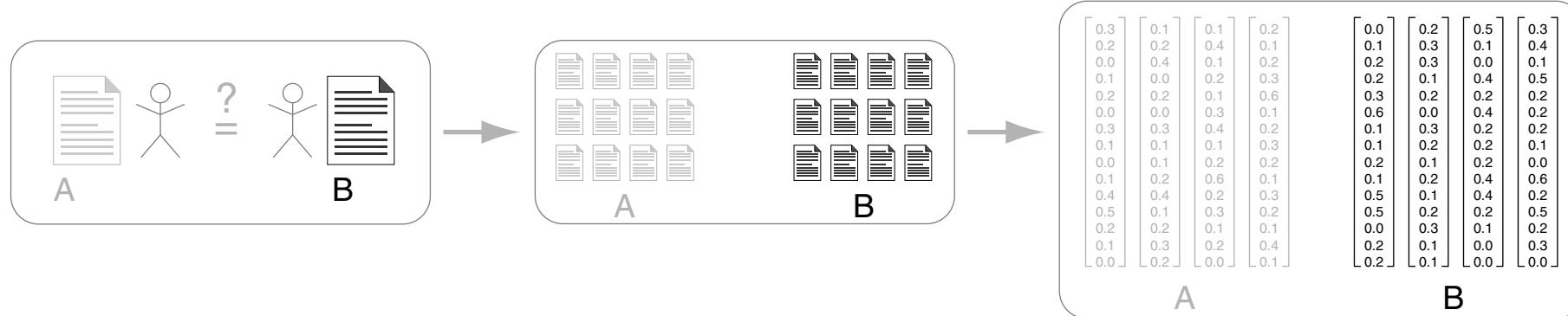


# Authorship Verification “Unmasking” [Koppel/Schler 2004] [BOW model]



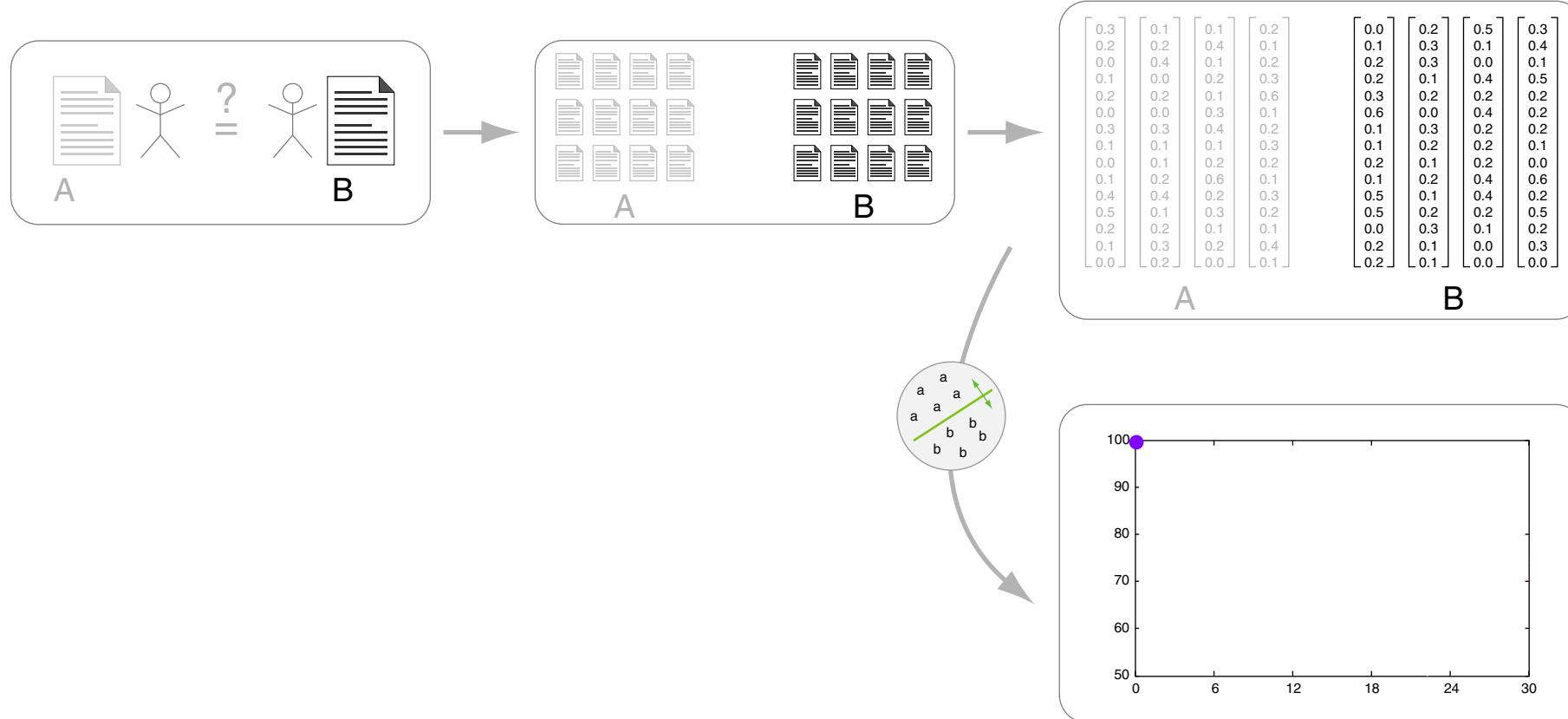
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



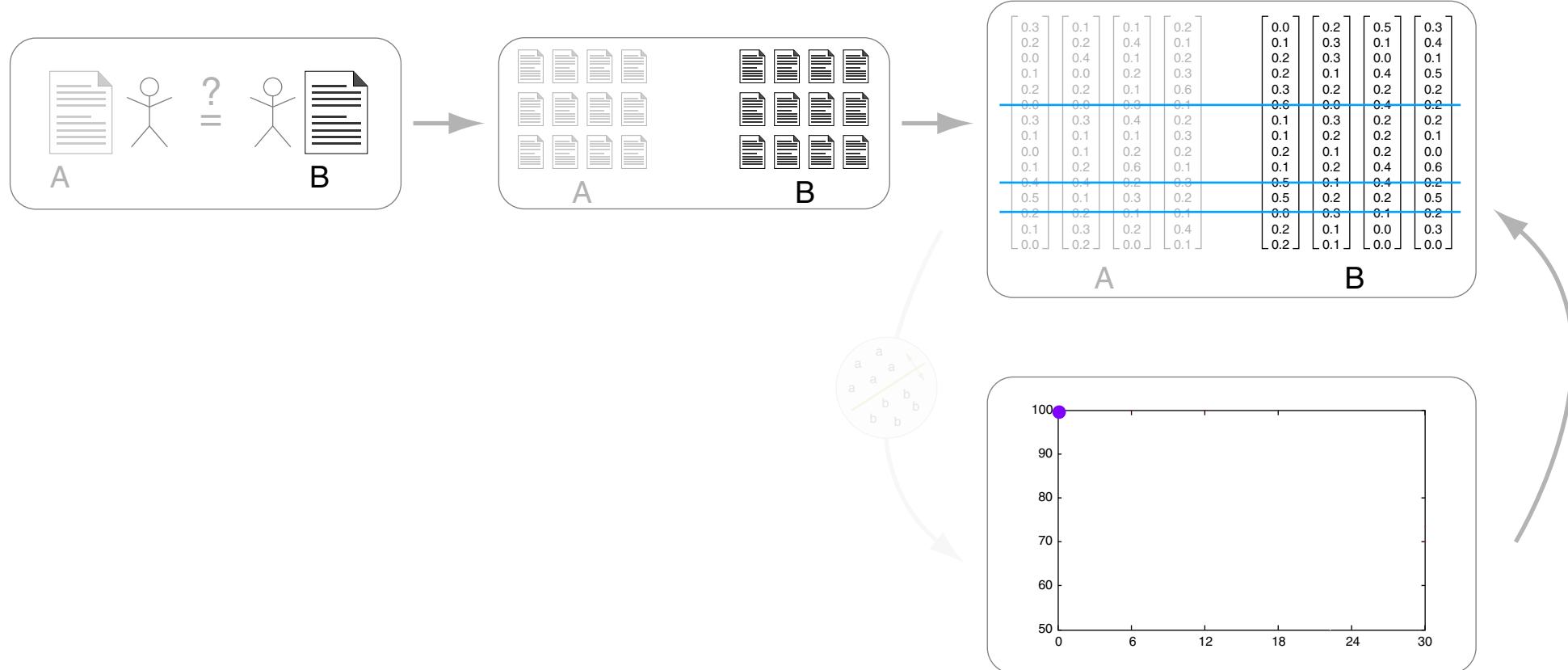
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



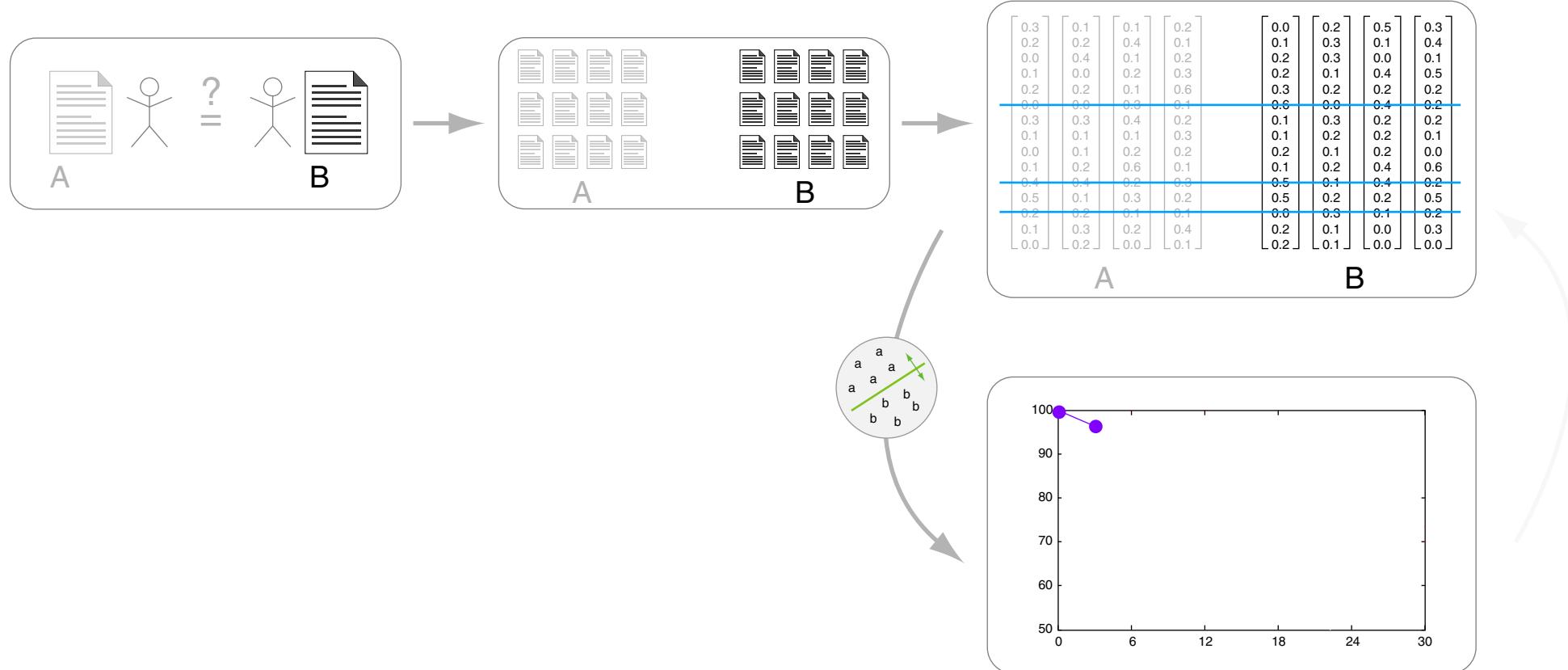
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



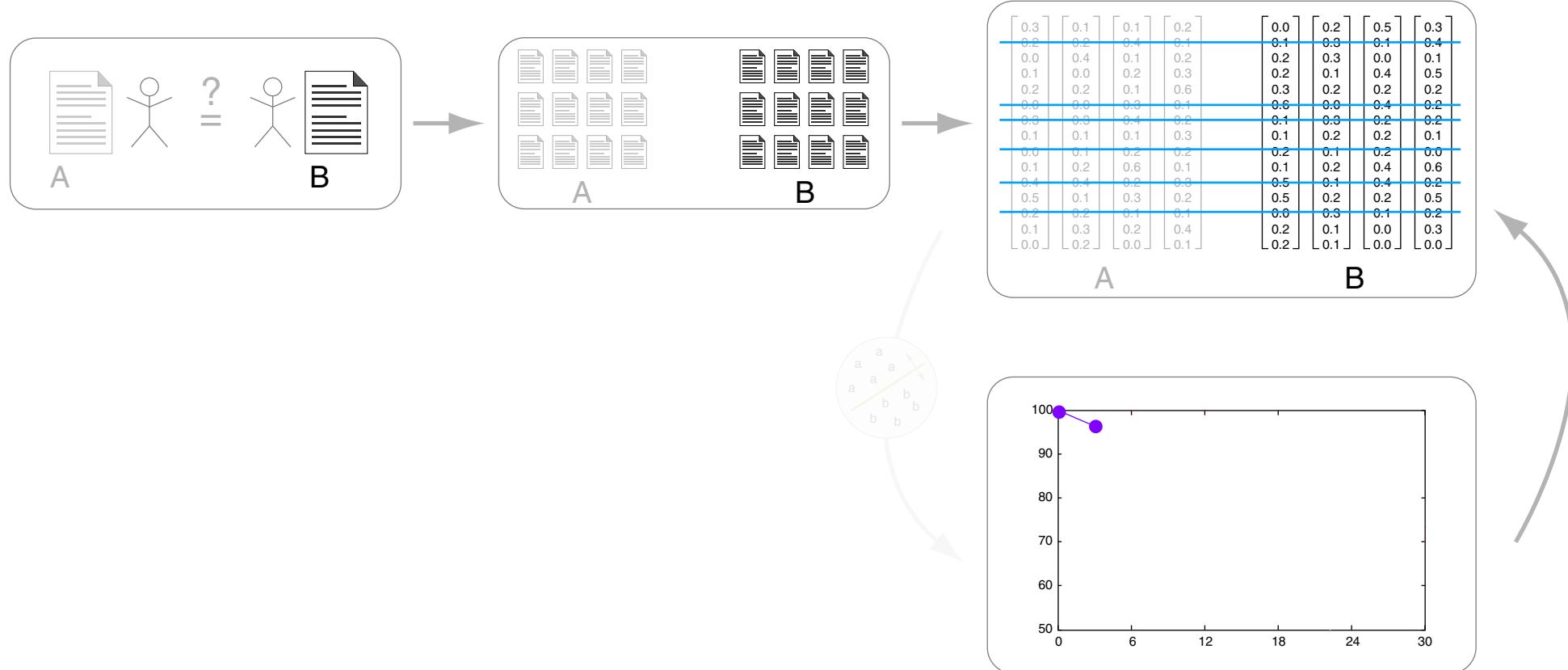
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



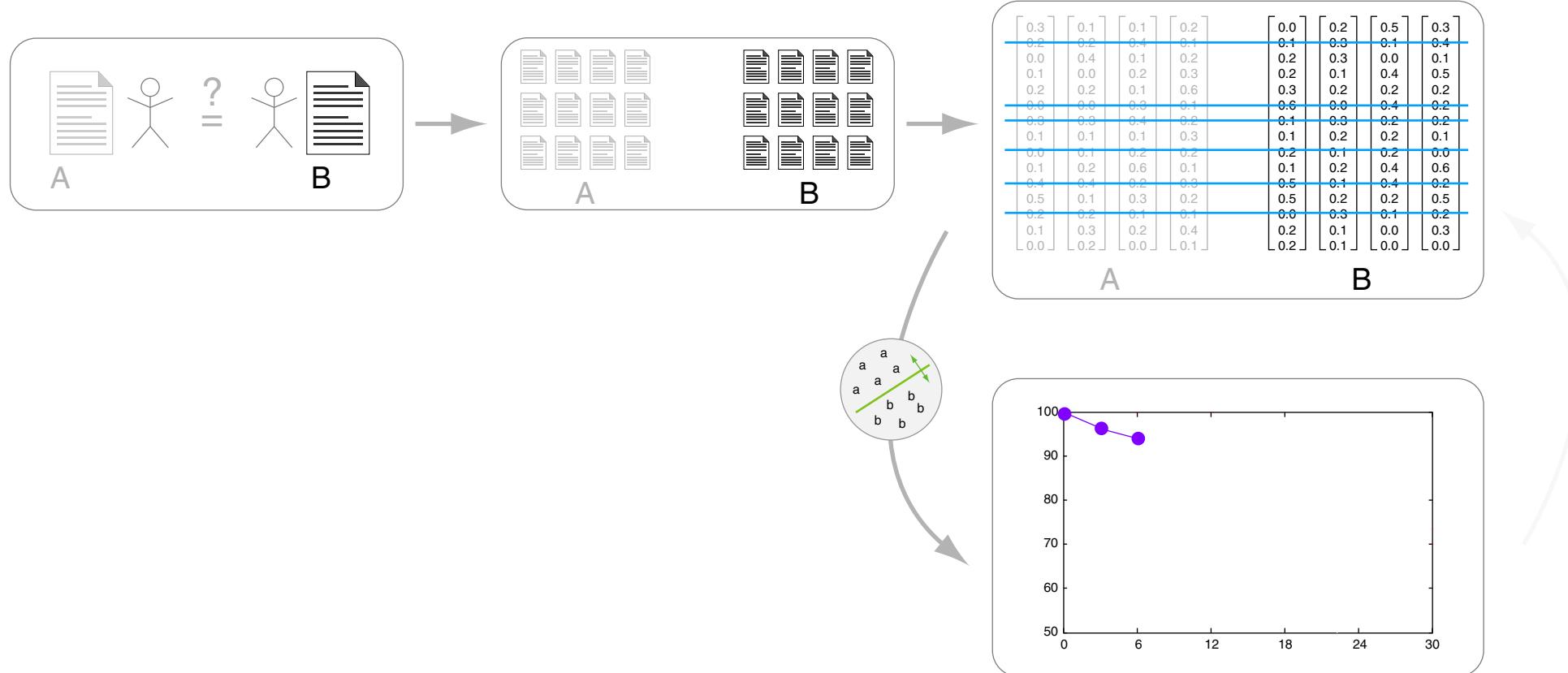
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



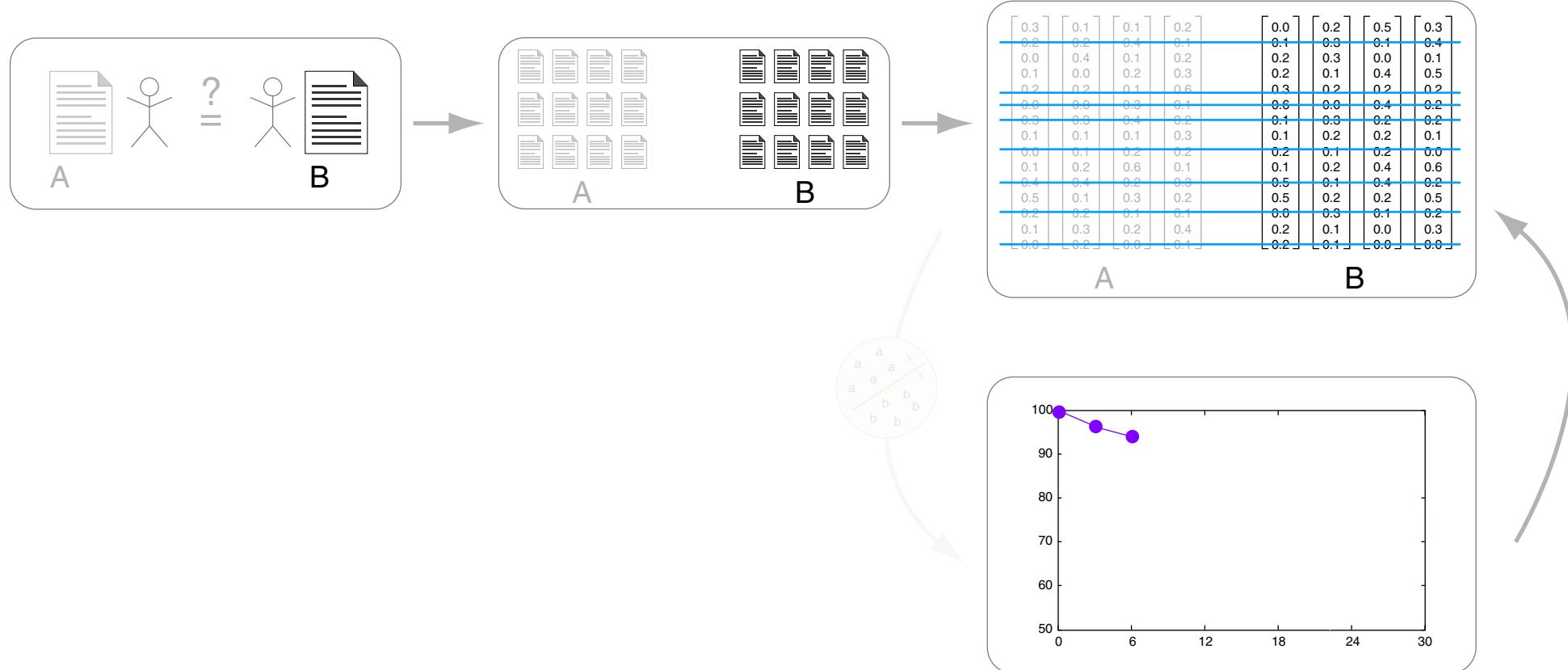
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



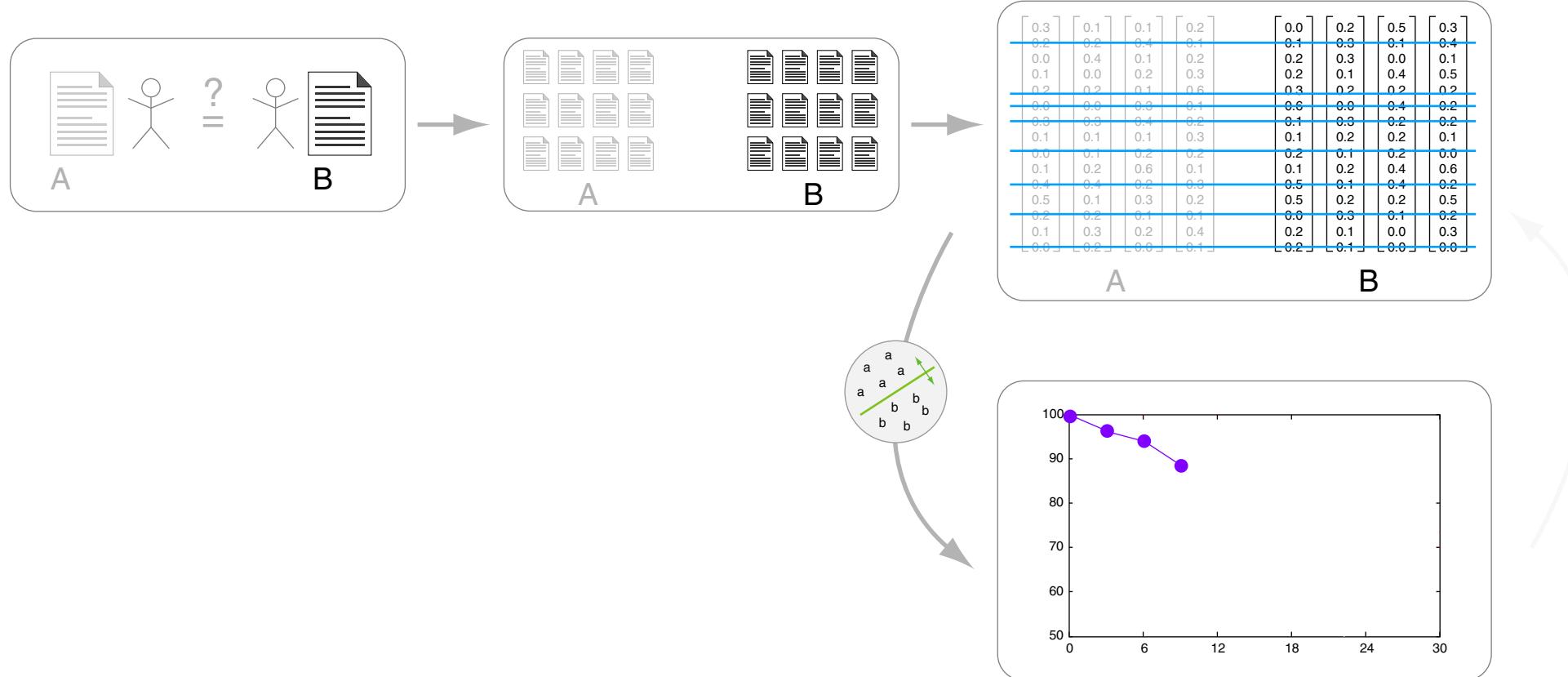
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



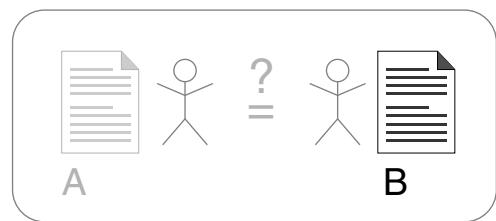
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



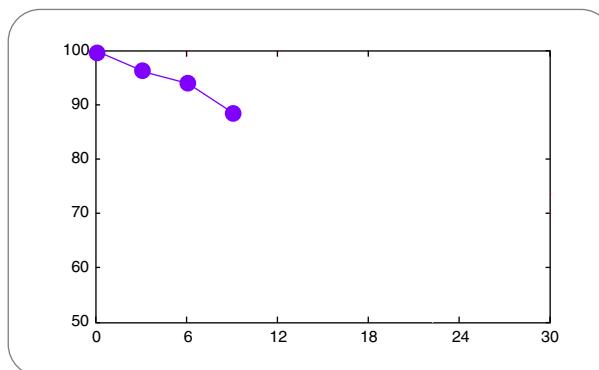
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



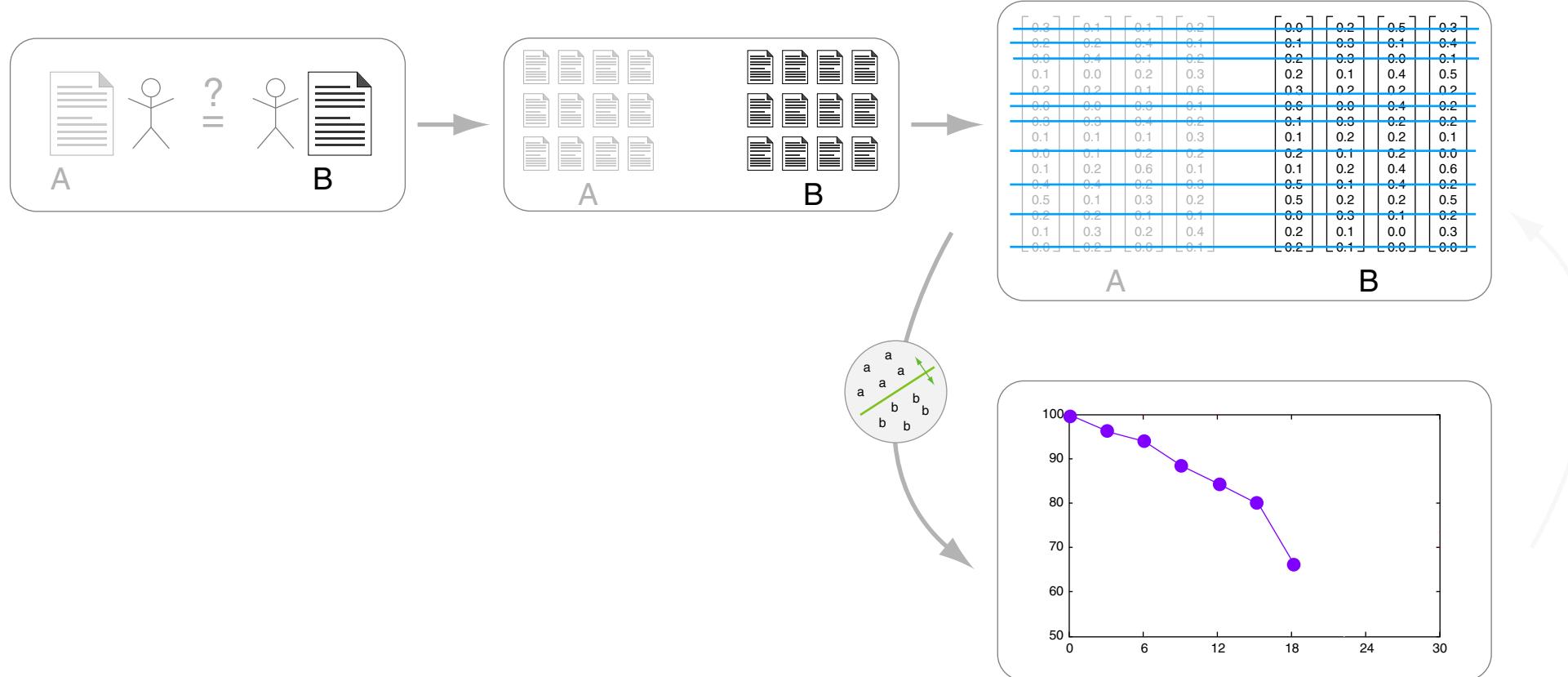
A rounded rectangle containing two tables labeled 'A' and 'B'. Each table has 15 rows and 8 columns. Blue horizontal lines highlight specific rows in both tables. The first row of table A is highlighted, and the last row of table B is highlighted.

[0.3]	[0.1]	[0.1]	[0.2]	[0.0]	[0.2]	[0.5]	[0.3]
0.2	0.2	0.4	0.1	0.1	0.3	0.1	0.4
0.0	0.4	0.1	0.2	0.2	0.3	0.0	0.1
0.1	0.0	0.2	0.3	0.2	0.1	0.4	0.5
0.2	0.2	0.1	0.6	0.3	0.2	0.2	0.2
0.0	0.0	0.3	0.1	0.6	0.0	0.4	0.2
0.3	0.0	0.4	0.2	0.1	0.3	0.2	0.2
0.1	0.1	0.1	0.3	0.1	0.2	0.2	0.1
0.0	0.1	0.2	0.2	0.2	0.1	0.2	0.0
0.1	0.2	0.6	0.1	0.1	0.2	0.4	0.6
0.4	0.4	0.2	0.0	0.5	0.1	0.4	0.2
0.5	0.1	0.3	0.2	0.5	0.2	0.2	0.5
0.2	0.2	0.1	0.1	0.0	0.3	0.1	0.2
0.1	0.3	0.2	0.4	0.2	0.1	0.0	0.3
0.0	0.2	0.0	0.1	0.2	0.1	0.0	0.0



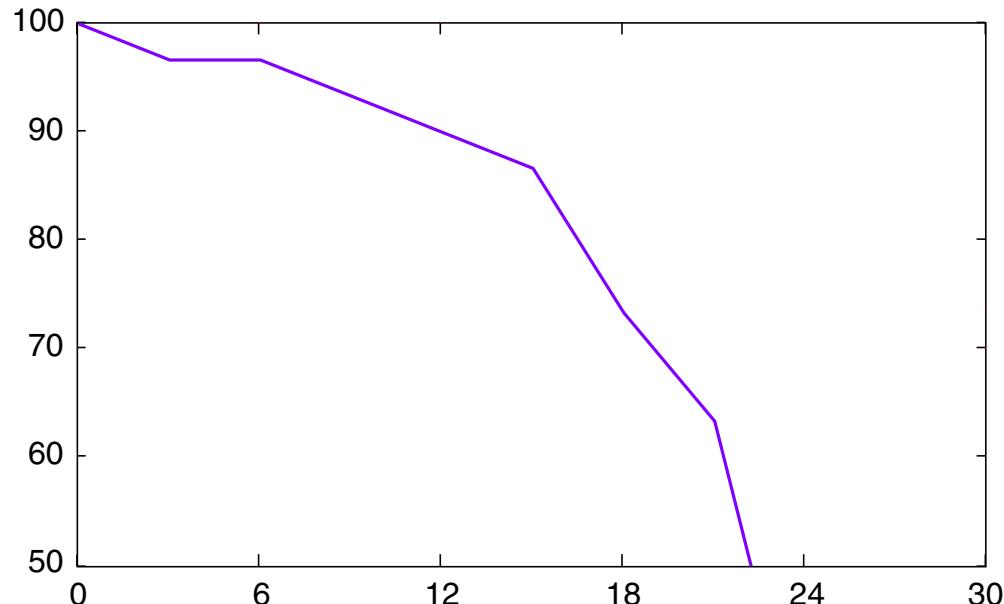
# Authorship Verification

“Unmasking” [Koppel/Schler 2004] [BOW model]



# Authorship Verification “Unmasking” at Work

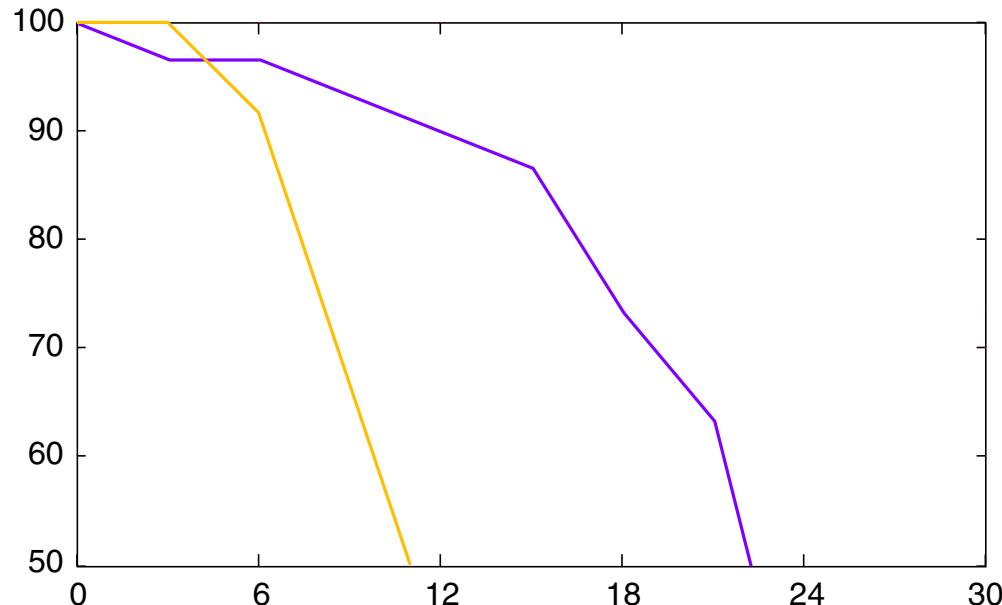
Learning characteristic for a pair of texts . . .



. . . from different authors ( $A \neq B$ ).

# Authorship Verification “Unmasking” at Work

Learning characteristic for two pairs of texts . . .

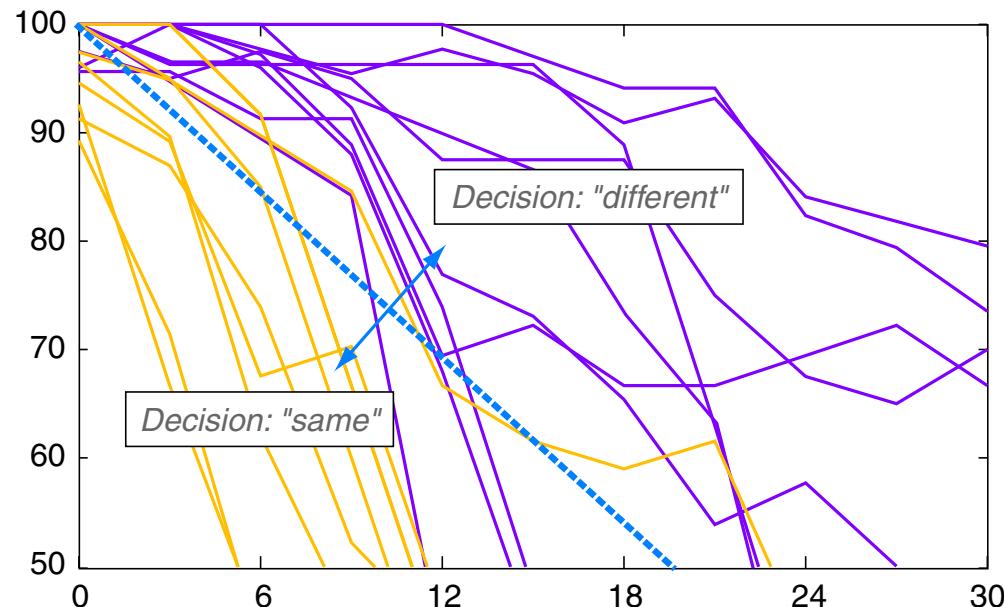


. . . from different authors ( $A \neq B$ ).

. . . from the same author ( $A = B$ ).

# Authorship Verification “Unmasking” at Work

Learning characteristic for many pairs of texts . . .



. . . from different authors ( $A \neq B$ ).

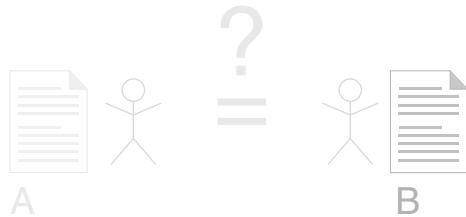
. . . from the same author ( $A = B$ ).

The typical learning characteristic can be learned.



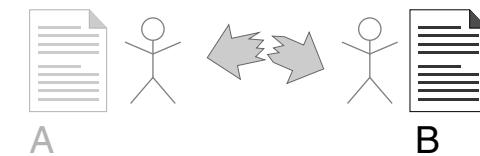
## Authorship Obfuscation

# Authorship Verification



Originate two texts from the same author?

# Authorship Obfuscation



Obfuscate same authorship.

# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

- Requires:
1. a model to measure style distance
  2. a confidence function to assess “same authorship”
  3. a means to manipulate style distance



Janek Bevendorff



Martin Potthast



Matthias Hagen



Benno Stein

# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”
3. a means to manipulate style distance

**Author A**

Trigram Freq.

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping to  
hunger and pain, so he might save us, his lost  
ones, from shame; now if we love him, he bids  
us to feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...

**Author B**

Trigram Freq.

come and see zip, the foremost of freaks! come  
and see palestine's sinister sheiks! eager  
equestriennes, each unexcelled, most mammoth  
menagerie ever beheld, the giant, the fat girl, the  
lion-faced man, aerial artists from far-off japan,  
audacious acrobats shot from a gun, don't ...

# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”
3. a means to manipulate style distance

Author A	Trigram	Freq.	Author B	Trigram	Freq.
beautiful christmas you know jesus our saviour was born here below, patiently stooping to hunger and pain, so he might save us, his lost ones, from _shame; now if we love him, he bids us to feed all his poor brothers and sisters who need. blessed old nick! i was sure if ...	and	4	come and see zip, the foremost of freaks! come	and	2
	to_	3	and see palestine's sinister _sheiks! eager	to_	1
	the	1	equestriennes, each unexcelled, most mammoth	the	4
	our	5	menagerie ever beheld, the giant, the fat girl, the	our	1
	_sh	1	lion-faced man, aerial artists from far-off japan,	_sh	2
	...		audacious acrobats shot from a gun, don't ...	...	

# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”
3. a means to manipulate style distance

Author A	Trigram	Freq.	Author B	Trigram	Freq.
beautiful christmas you know jesus our saviour was born here below, patiently stooping to hunger and pain, so he might save us, his lost ones, from shame; now if we love him, he bids us to feed all his poor brothers and sisters who need. blessed old nick! i was sure if ...	and	4	come and see zip, the foremost of freaks! come	and	2
	to_	3	and see palestine's sinister sheiks! eager	to_	1
	the	1	equestriennes, each unexcelled, most mammoth	the	4
	our	5	menagerie ever beheld, the giant, the fat girl, the	our	1
	_sh	1	lion-faced man, aerial artists from far-off japan,	_sh	2
	...		audacious acrobats shot from a gun, don't ...	...	

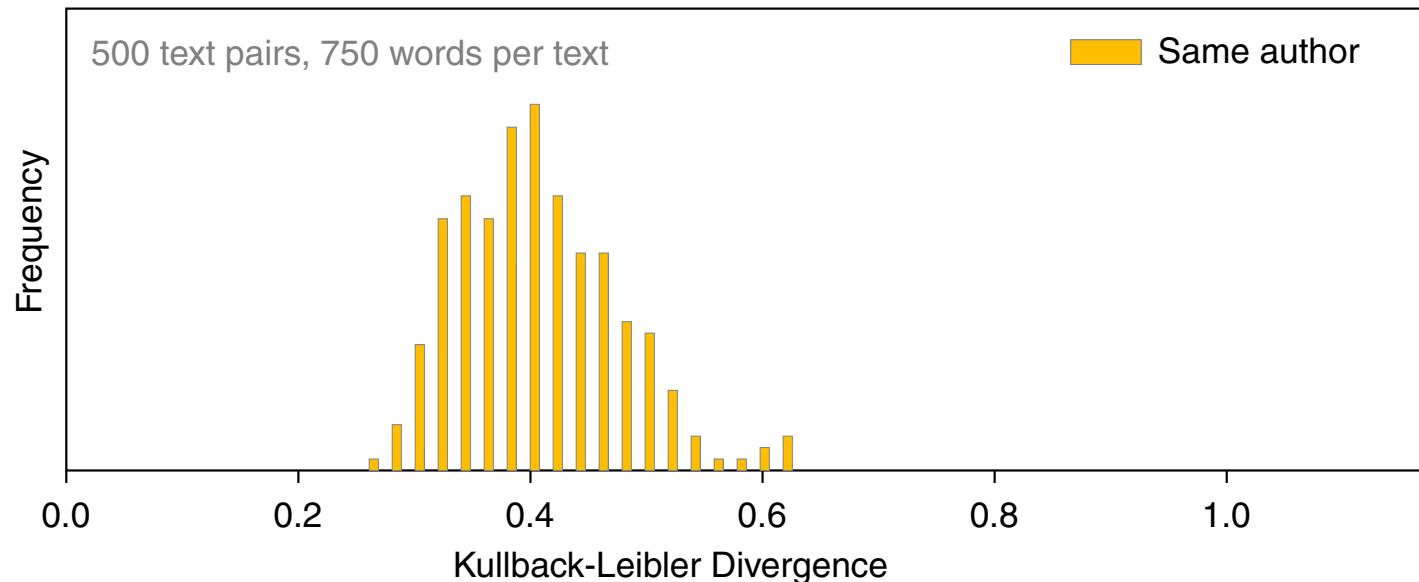
Kullback-Leibler Divergence:  $\text{KLD}(P \mid Q) = \sum_{i \in \text{trigrams}} P[i] \log \frac{P[i]}{Q[i]}$

# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”
3. a means to manipulate style distance

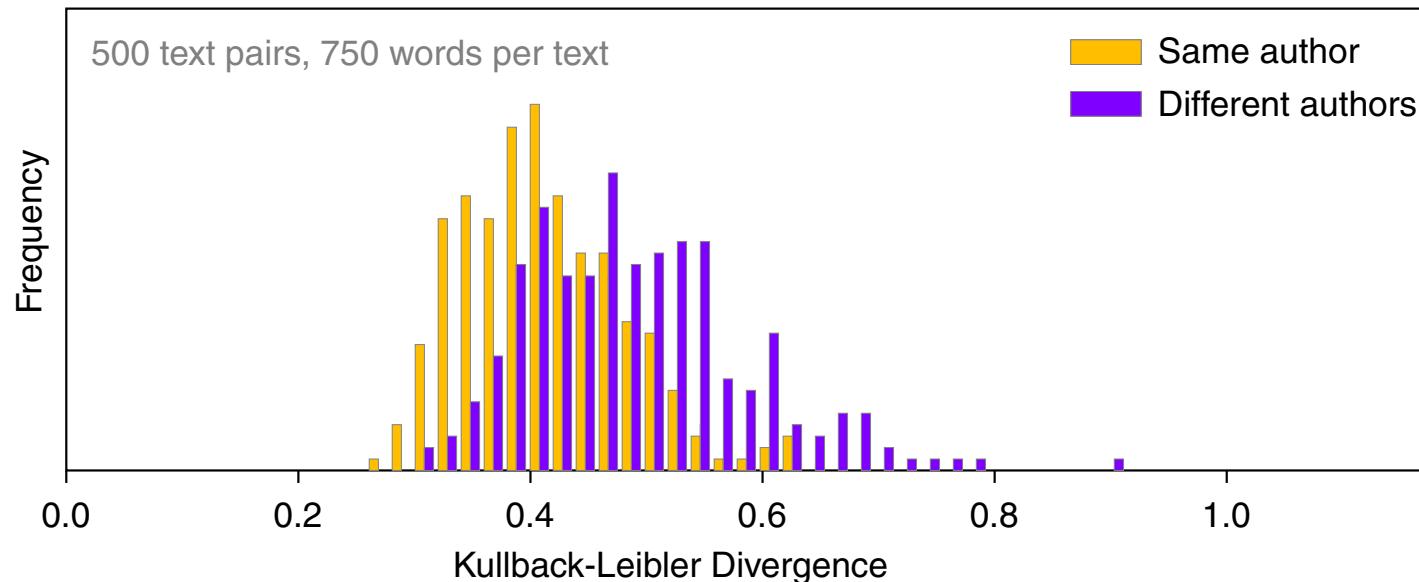


# Authorship Obfuscation

[Bevendorff/Potthast/Hagen/Stein 2019] [char-trigram model]

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”
3. a means to manipulate style distance

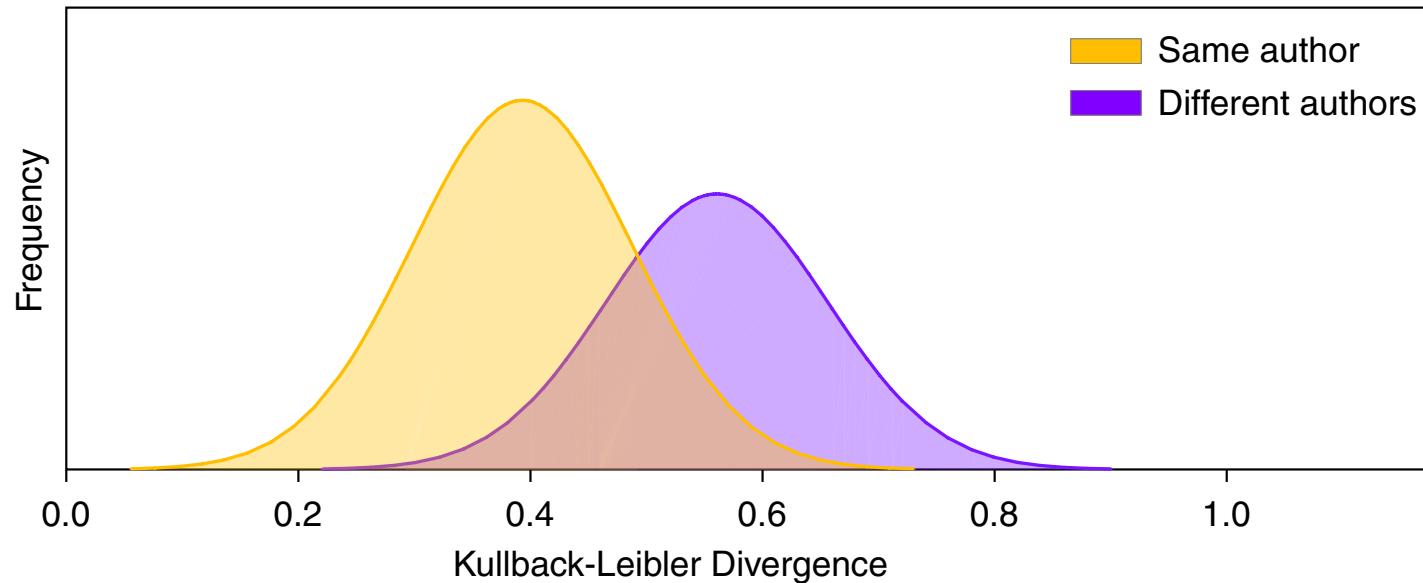


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

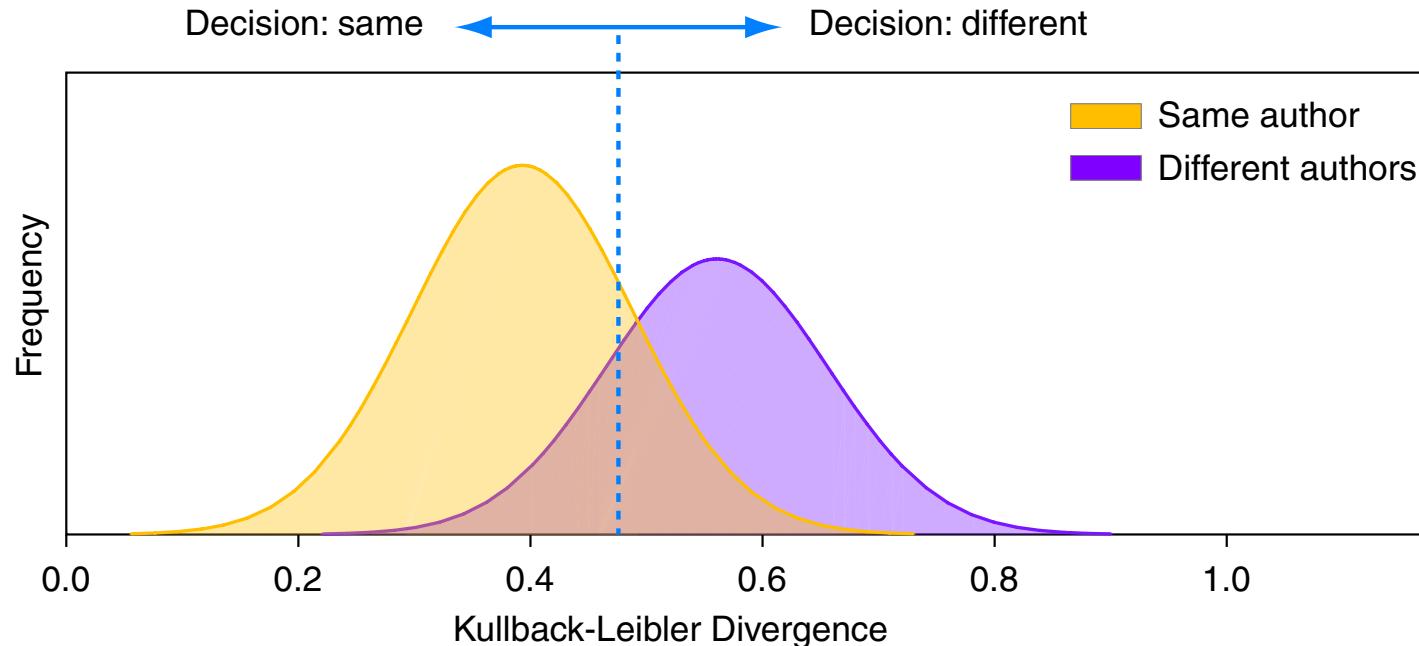


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

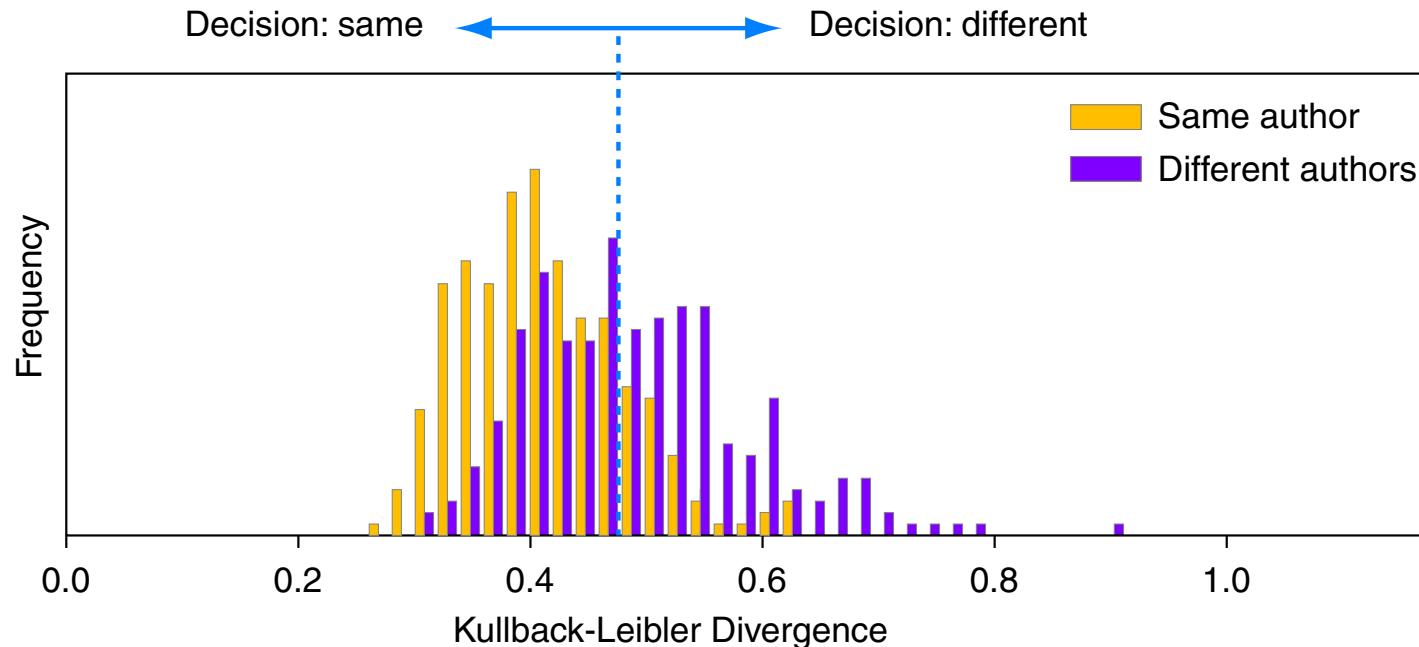


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

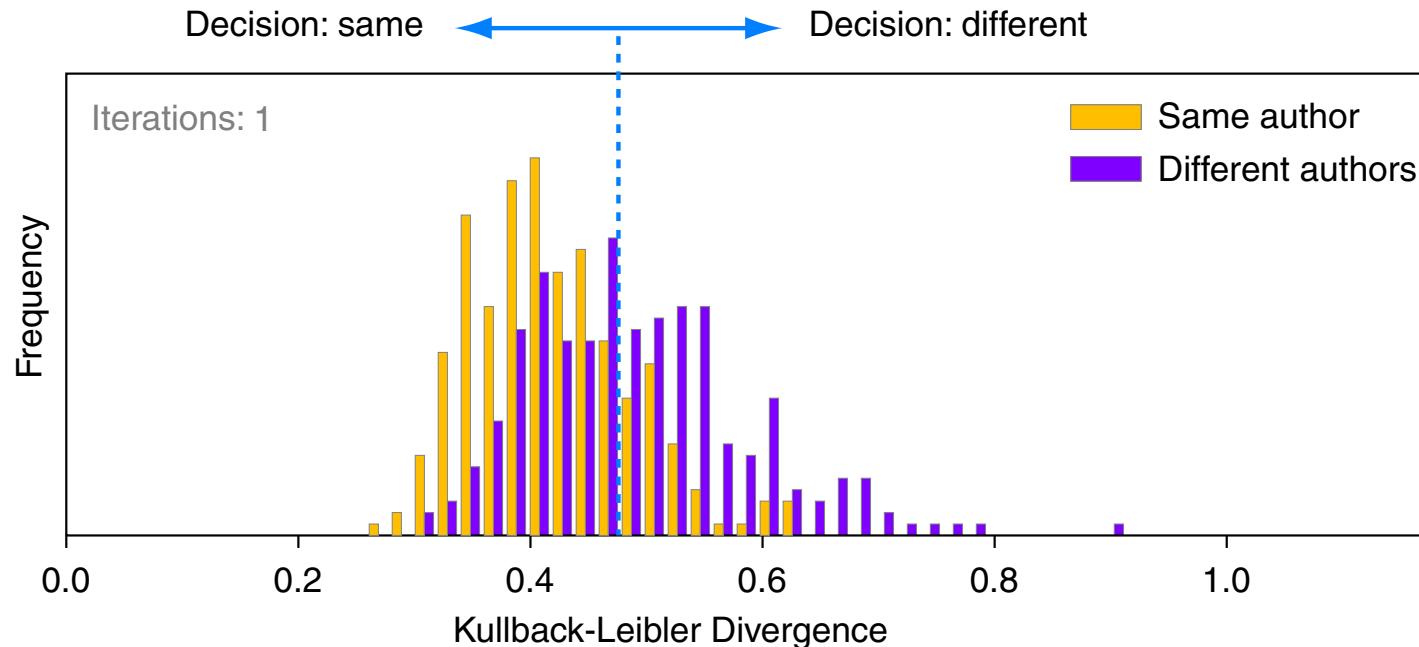


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

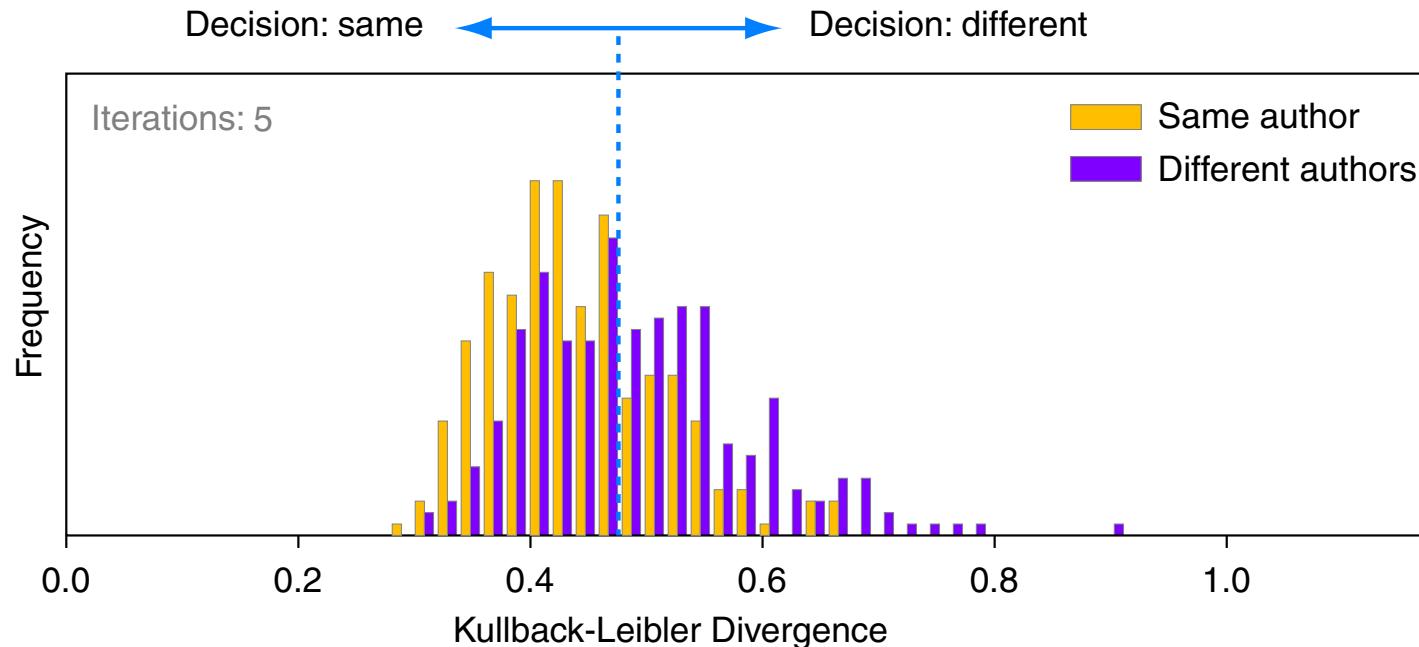


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

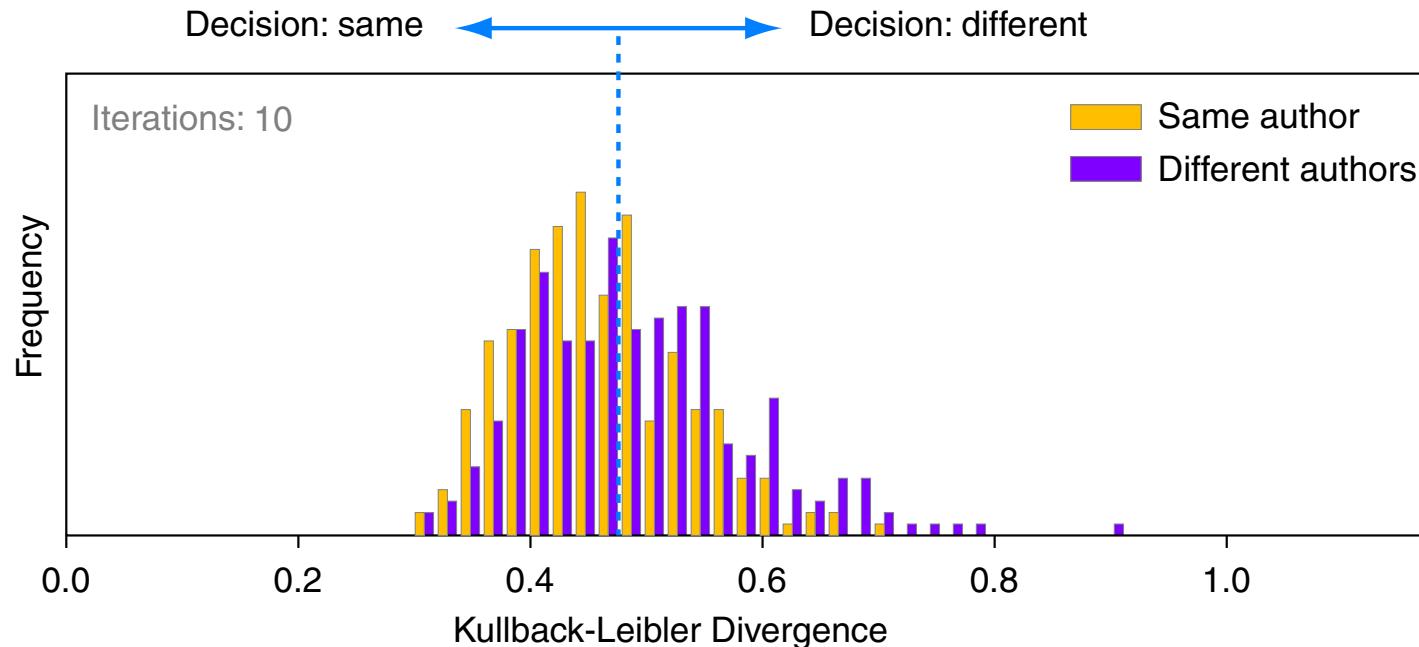


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

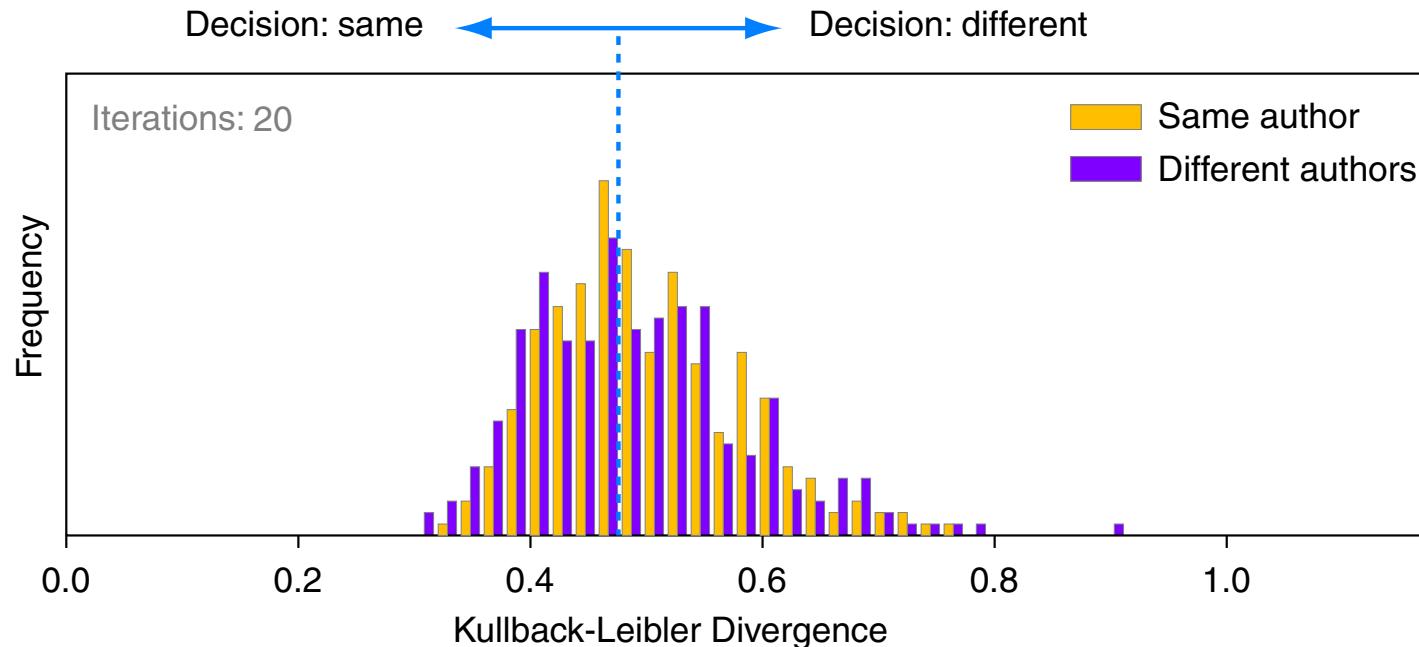


# Authorship Obfuscation

Requires: 1. a model to measure style distance

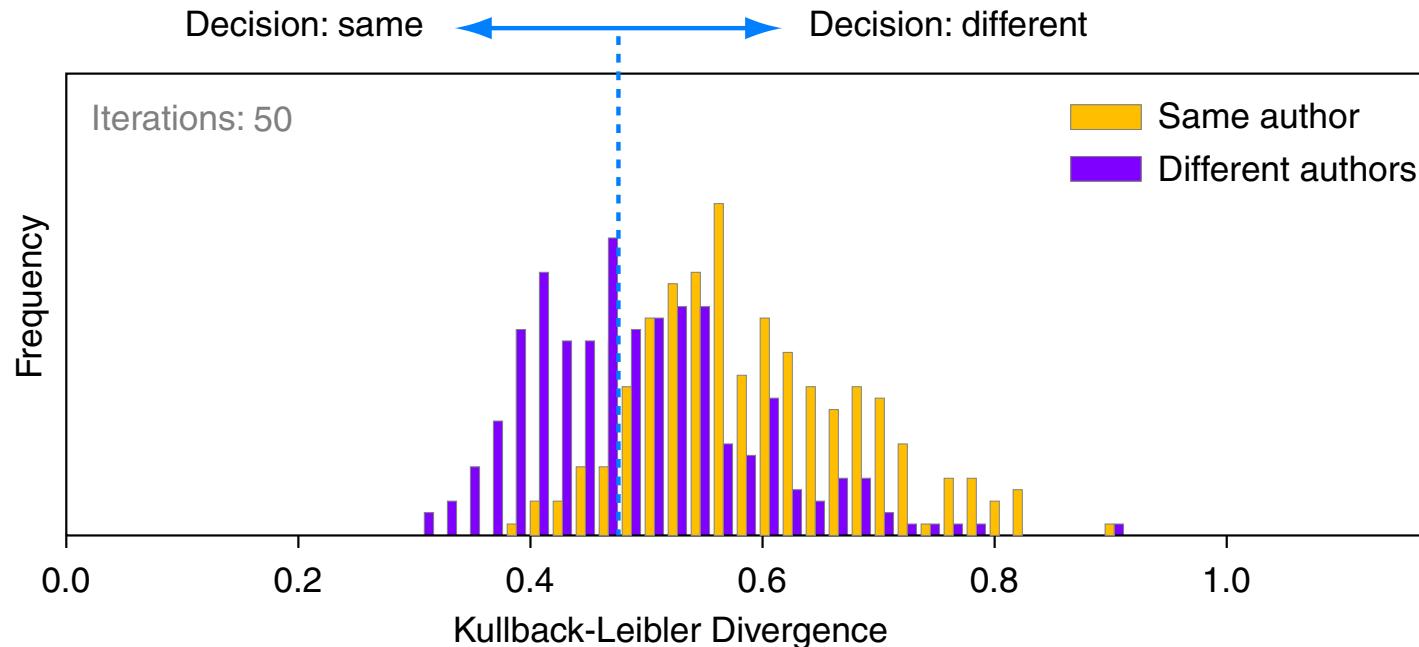
2. a confidence function to assess “same authorship”

3. a means to manipulate style distance



# Authorship Obfuscation

- Requires:
1. a model to measure style distance
  2. a confidence function to assess “same authorship”
  3. a means to manipulate style distance

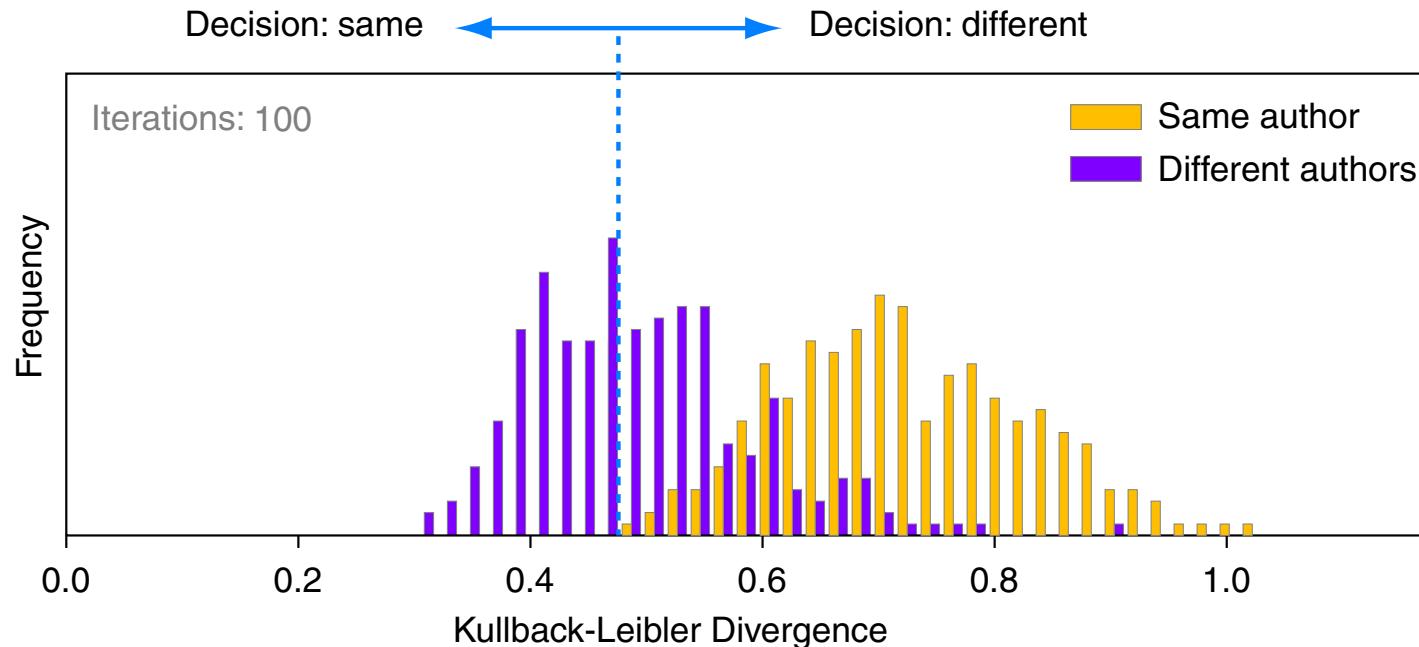


# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance



# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization:

$$\text{KLD}(P \mid Q) = \sum_{i \in \text{trigrams}} P[i] \log \frac{P[i]}{Q[i]} \rightarrow \max$$

~ Compute a ranking among trigrams regarding their KLD impact:

$$\underset{i \in \text{trigrams}}{\operatorname{argmax}} \frac{\partial}{\partial Q[i]} (\text{KLD}(P \mid Q))$$

Derivation omitted

# Authorship Obfuscation

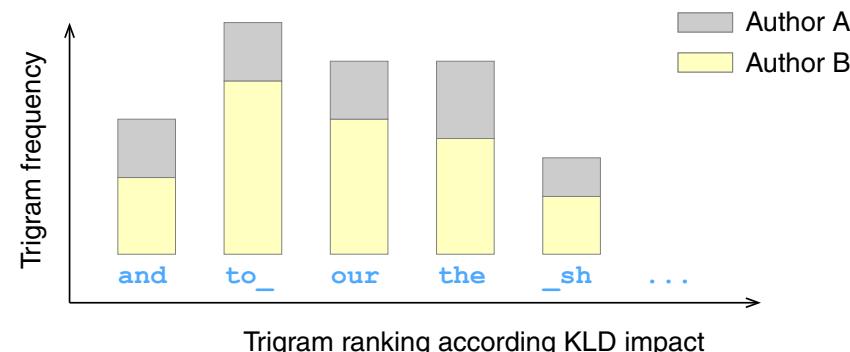
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

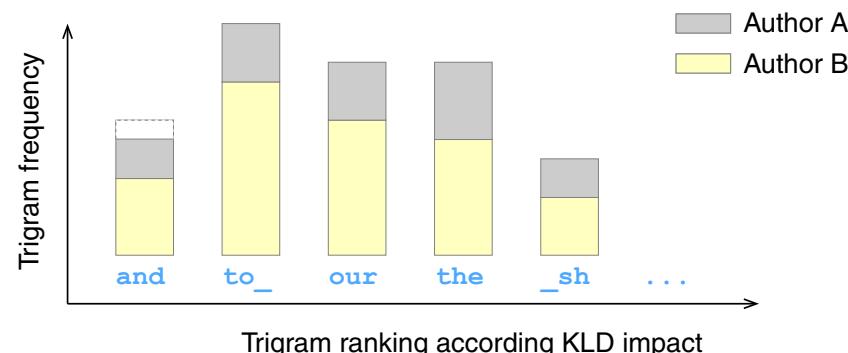
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

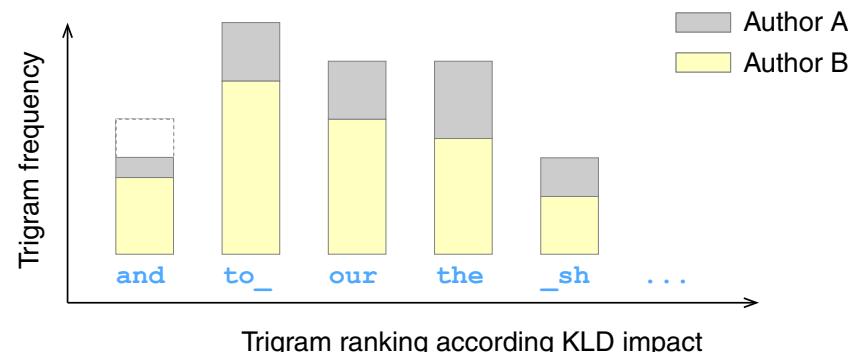
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

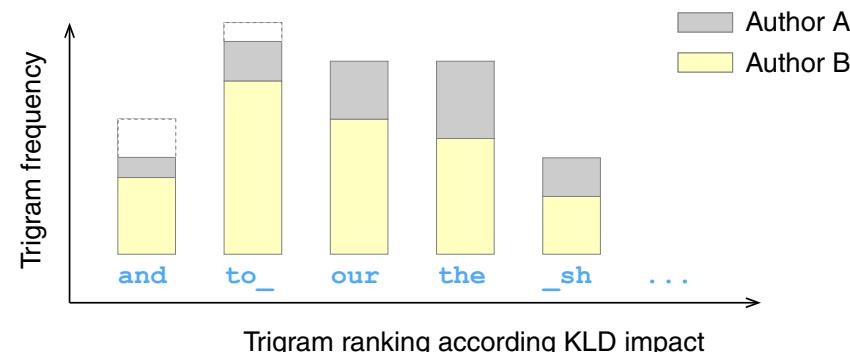
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

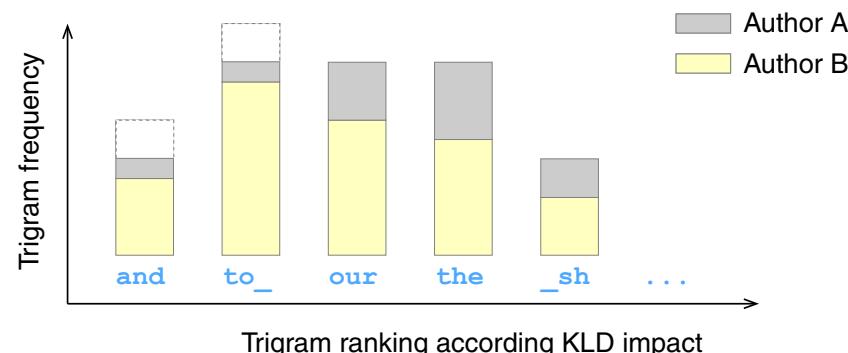
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

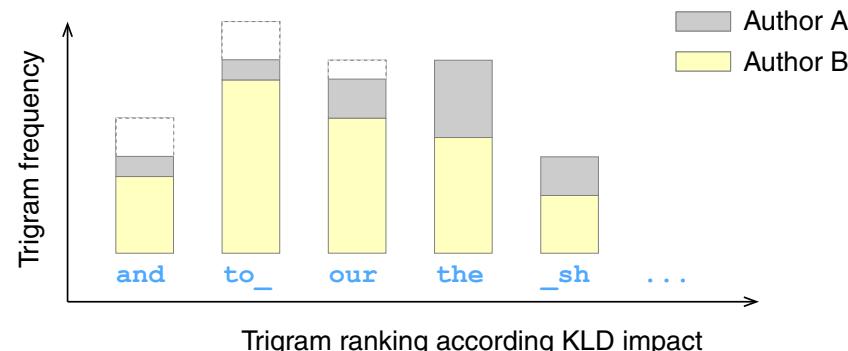
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



# Authorship Obfuscation

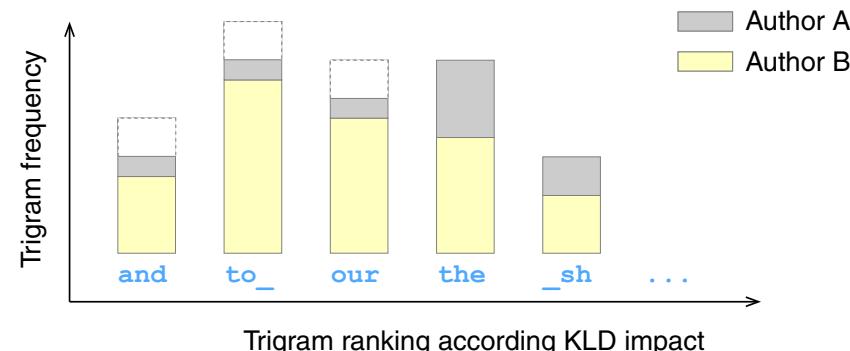
Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Style distance manipulation  $\equiv$  KLD maximization  $\rightarrow$  Remove trigrams:

beautiful christmas you know jesus our saviour  
was born here below, patiently stooping  
to\_hunger and pain, so he might save us, his lost  
ones, from\_shame; now if we love him, he bids  
us to\_feed all his poor brothers and sisters who  
need. blessed old nick! i was sure if ...



But: How to render style distance manipulation unsuspicious?

# Authorship Obfuscation

Requires: 1. a model to measure style distance

2. a confidence function to assess “same authorship”

3. a means to manipulate style distance

Examples of how to eliminate the trigram **our**:

Operator	removal	character flip	synonym	number	paraphrase
	savi <b>our</b> ↓ savi	<b>sour</b> ly ↓ suorly	<b>cour</b> se ↓ lesson	<b>our</b> home ↓ my home	for 24 <b>hours</b> ↓ round-the-clock
Text quality loss	8	4	4	2	2
Obfuscation impact			(to be computed, depends on the entire text)		

But: How to render style distance manipulation unsuspicious?

# Authorship Obfuscation

## Example 1

Original:

It was the only chance **we had** to win.' Duke swallowed the idea slowly. He couldn't picture a **planet** giving up its last protection for a desperate effort to end the war on purely offensive drive. Three billion people watching the home fleet take off, **knowing** the skies were **open** for all the **hell** that a savage enemy could send! On Earth, the World Senate hadn't permitted the building of one **battleship**, for fear of reprisal. [...]

Obfuscated @  $\epsilon_{0.7}$ :

It was the only chance **w ehad** to win.' Duke swallowed the idea slowly. He couldn't picture a **satellite** giving up its last protection for **phi** desperate effort to end the war on purely offensive drive. Three billion people watching the home fleet take off, **deciding** the skies were **resort** for all the **mischief** that a savage enemy could send! On Earth, the World Senate hadn't permitted the building of one **frigate**, for fear of reprisal. [...]

*Victory* by Lester del Rey

# Authorship Obfuscation

Example 1 [summary]

Original:

It was the only chance **we had** to win.' Duke swallowed the idea slowly. He couldn't picture a **planet** giving up its last protection for a desperate effort to end the war on purely offensive drive. Three billion people watching the home fleet take off, **knowing** the skies were **open** for all the **hell** that a savage enemy could send! On Earth, the World Senate hadn't permitted the building of one **battleship**, for fear of reprisal. [...]

Obfuscated @  $\varepsilon_{0.7}$ :

It was the only chance **w ehad** to win.' Duke swallowed the idea slowly. He couldn't picture a **satellite** giving up its last protection for **phi** desperate effort to end the war on purely offensive drive. Three billion people watching the home fleet take off, **deciding** the skies were **resort** for all the **mischief** that a savage enemy could send! On Earth, the World Senate hadn't permitted the building of one **frigate**, for fear of reprisal. [...]

} like a typo

} unsuspicious

} suspicious

} unsuspicious

*Victory* by Lester del Rey

Optimum operator sequence determined via heuristic search.

# Authorship Obfuscation

## Example 2 [summary]

Original:

With a furtive glance around him, he clapped the other half of the clay sphere over the filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across **green** hills to the main hospital. The attendants made a quick **count** and then unlocked the door. The group shuffled out into the warm, **afternoon** sunlight and door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart **book** of patient progress.

Moving slowly down the line of **benches**, she made **short**, precise notes on the day's work accomplished **by each** patient. [...]

Obfuscated @  $\epsilon_{0.7}$ :

With a furtive glance around him, he clapped the other half of the clay sphere over filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across **site** hills to the main hospital. The attendants made a quick **investigation** and then unlocked the door. The group shuffled out into the warm, **daylight** sunlight and the door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart **forward** of patient progress.

Moving slowly down the line of **bens**, she made **parcel**, precise notes on the day's work accomplished **b y aehc** patient. [...]

} unsuspicious

} suspicious

# Authorship Obfuscation Case Study on Effectiveness

Text pairs

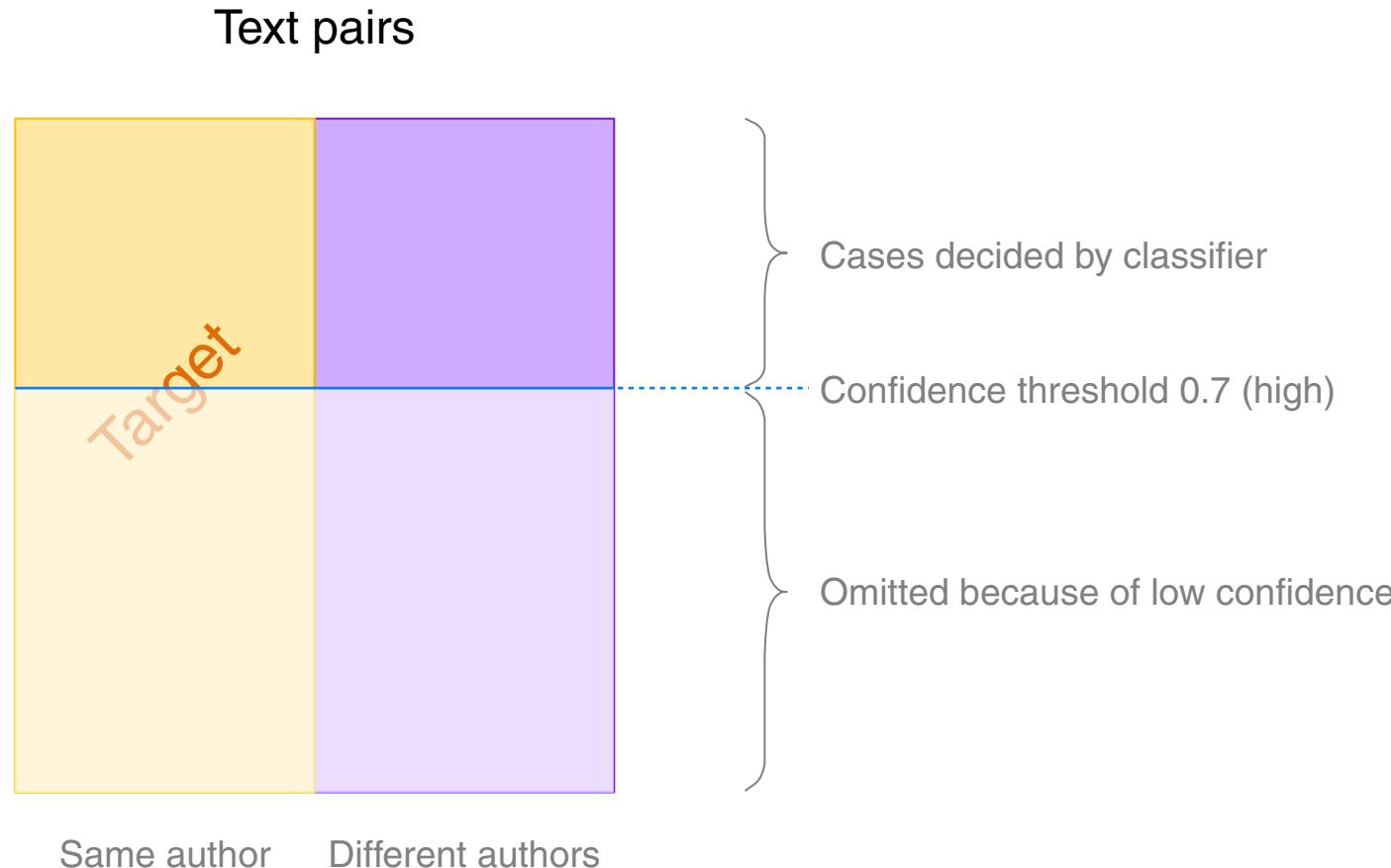


Same author

Different authors

# Authorship Obfuscation

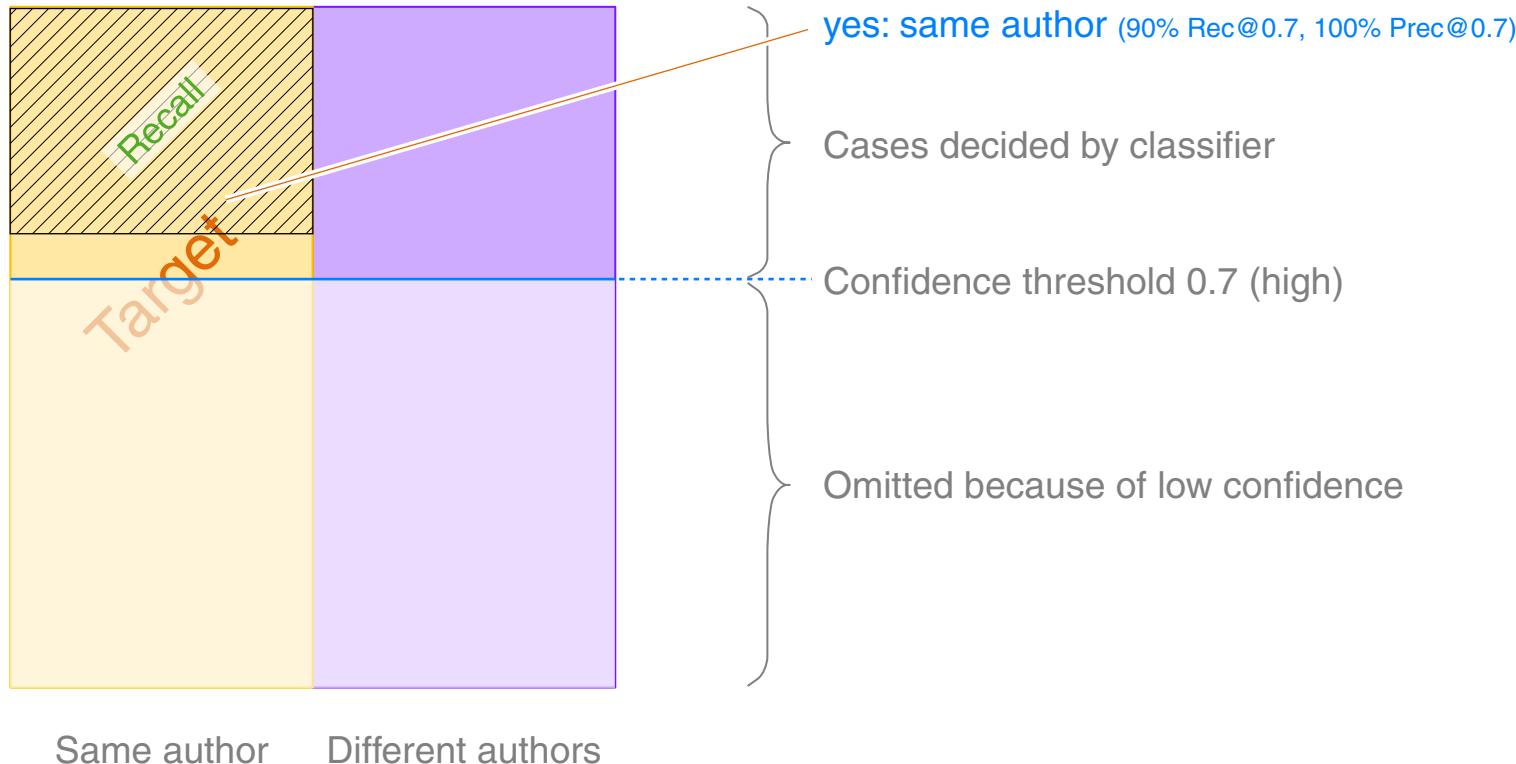
## Case Study on Effectiveness



# Authorship Obfuscation

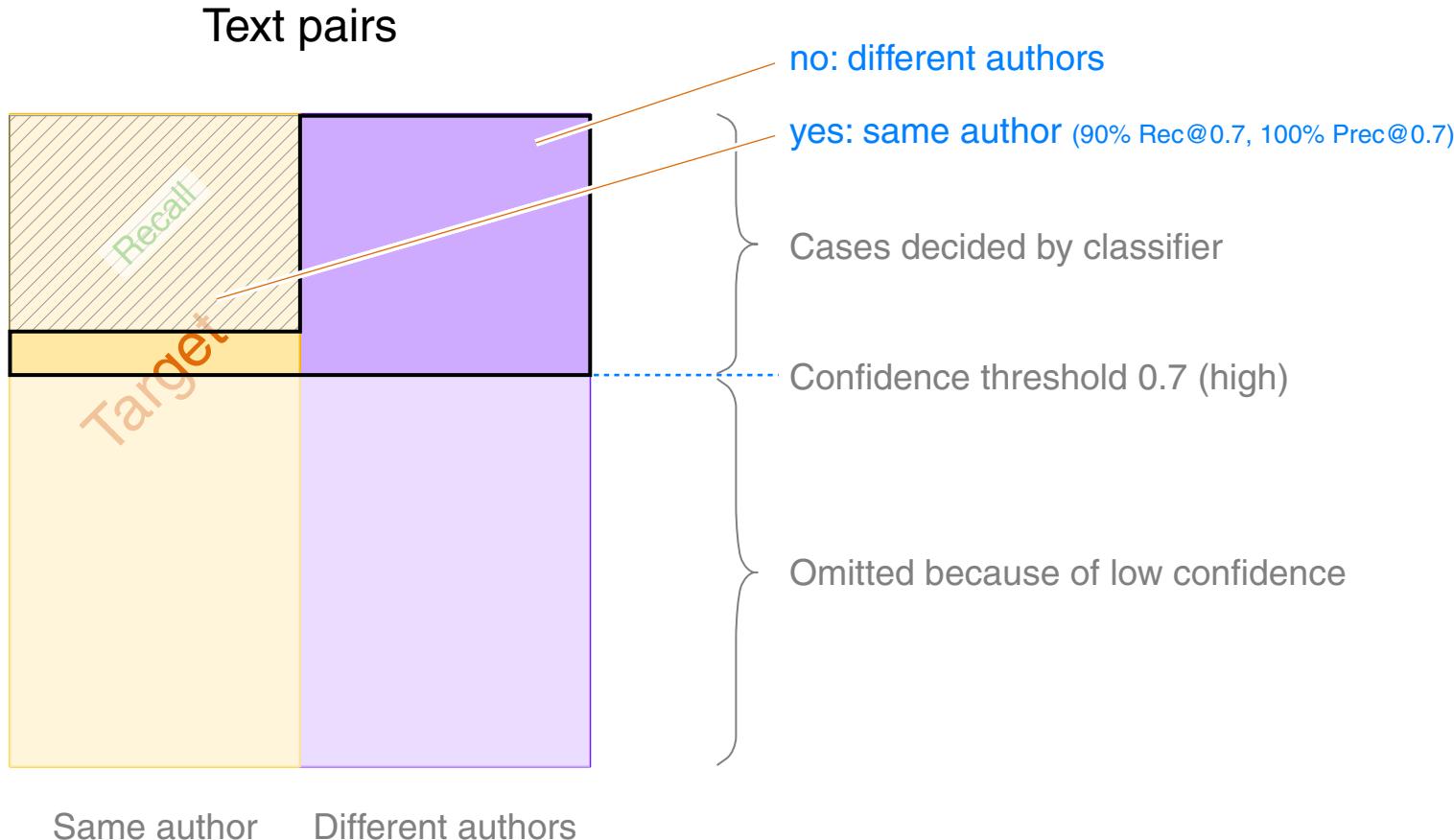
## Case Study on Effectiveness

### Text pairs



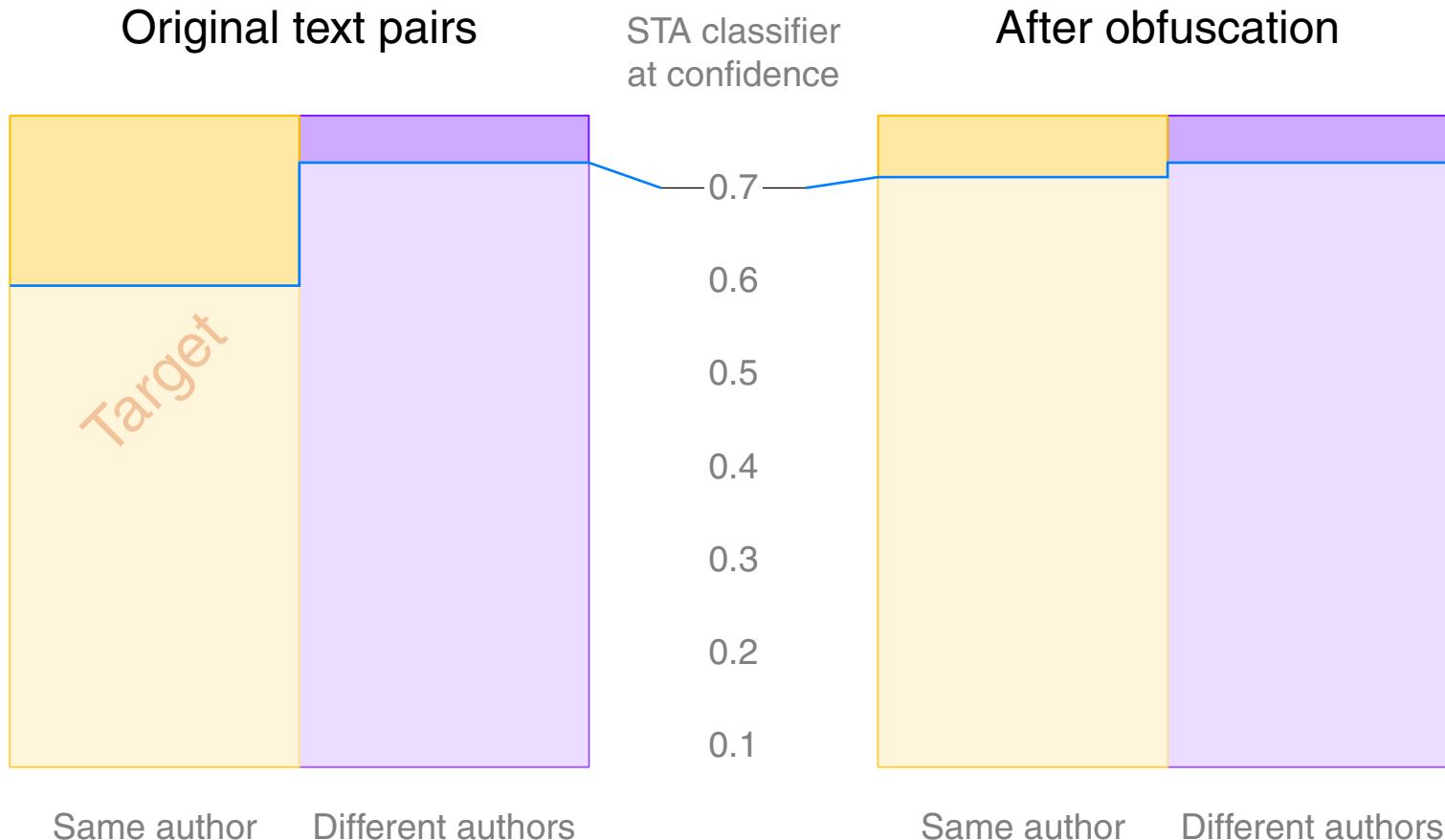
# Authorship Obfuscation

## Case Study on Effectiveness



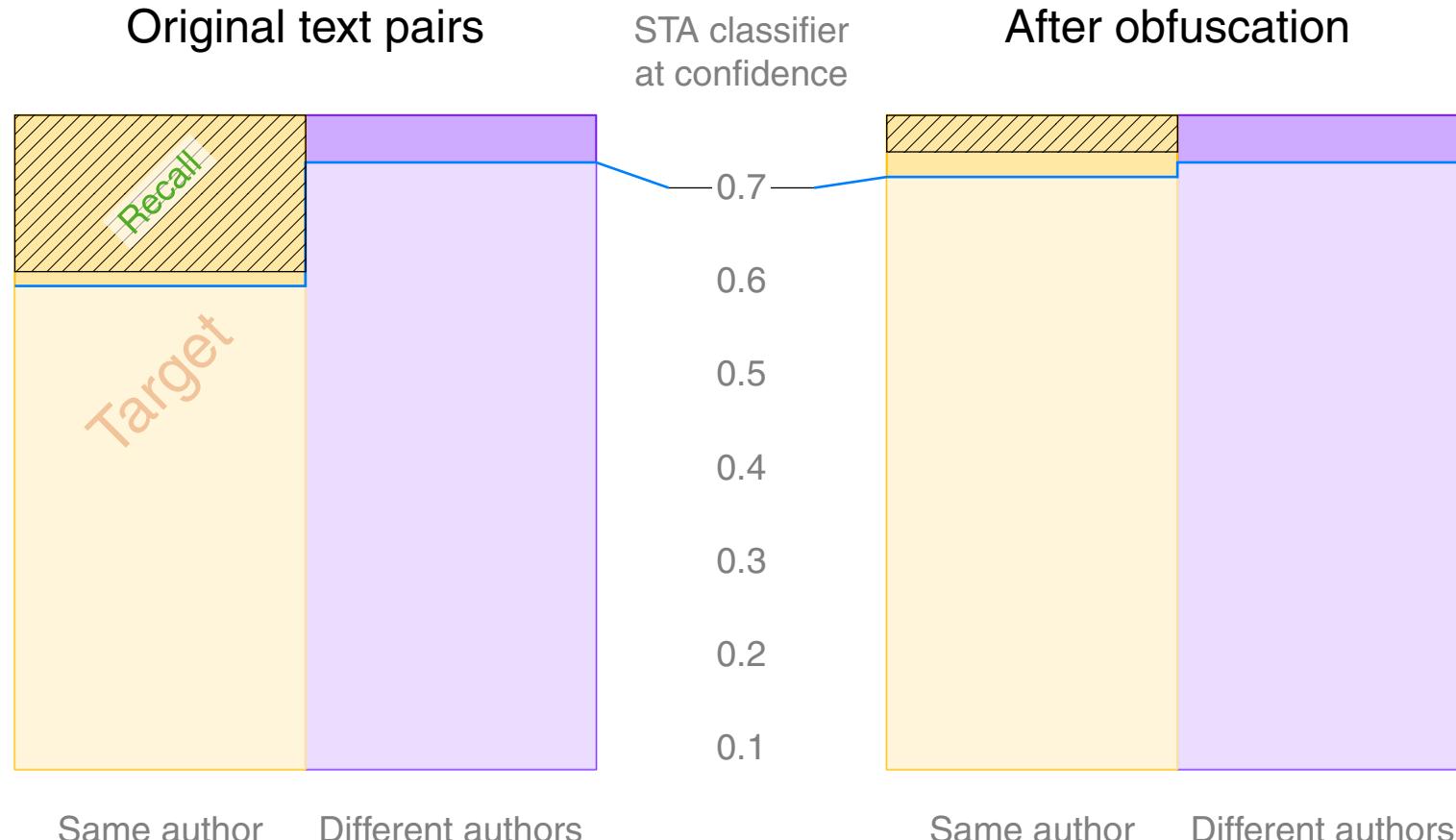
# Authorship Obfuscation

## Case Study on Effectiveness

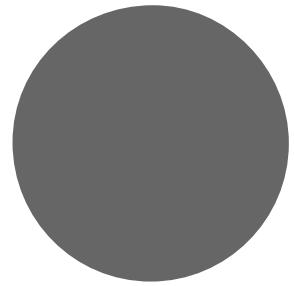


# Authorship Obfuscation

## Case Study on Effectiveness



Reduction in recall by more than factor 4 for state-of-the-art classifiers.



Summary

# PAN

[pan.webis.de](http://pan.webis.de)

# PAN Competitions

# PAN Competitions

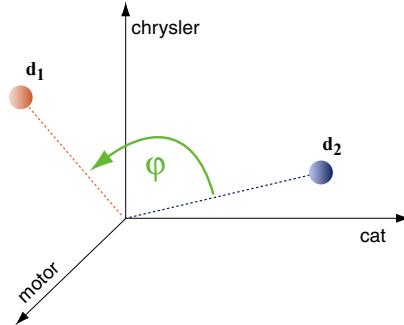
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	PAN
Intrinsic plagiarism detection											
External plagiarism detection											
		Cross-language text reuse detection									
Text Reuse and Plagiarism Detection			Source retrieval								
			Text alignment								
		Authorship attribution						Cross-domain authorship attribution			
Author Identification				Authorship verification							
								Author clustering			
									Author diarization / Style change detection		
			Age prediction								
Author Profiling			Gender prediction				Personality prediction				
									Language variety analysis		
								Author masking			
Author Obfuscation											
		Wikipedia vandalism detection									
Credibility Analysis				Wikipedia quality flaw prediction							
				Sexual predator identification							

Originality

Authorship

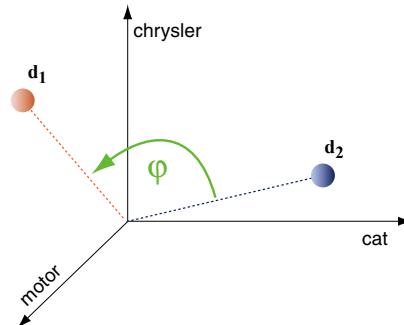
Trust

# Take Away Messages



Text models: bag of words and char-trigrams

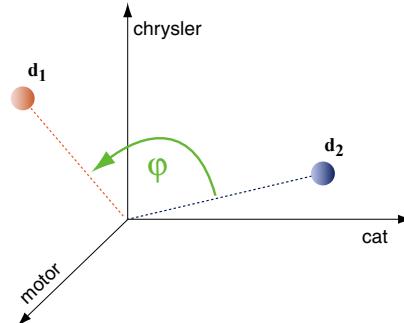
# Take Away Messages



Text models: bag of words and char-trigrams



# Take Away Messages

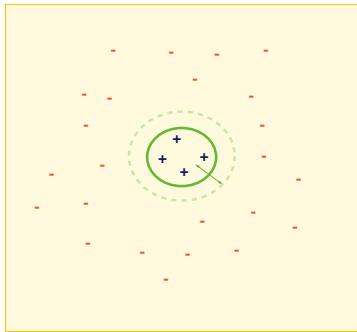


Text models: bag of words and char-trigrams



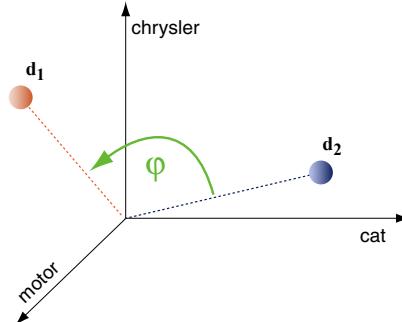
Authorship attribution

Authorship verification



Verification means one-class classification

# Take Away Messages

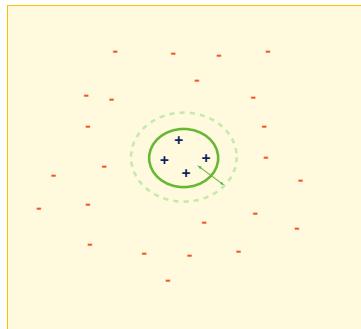


Text models: bag of words and char-trigrams

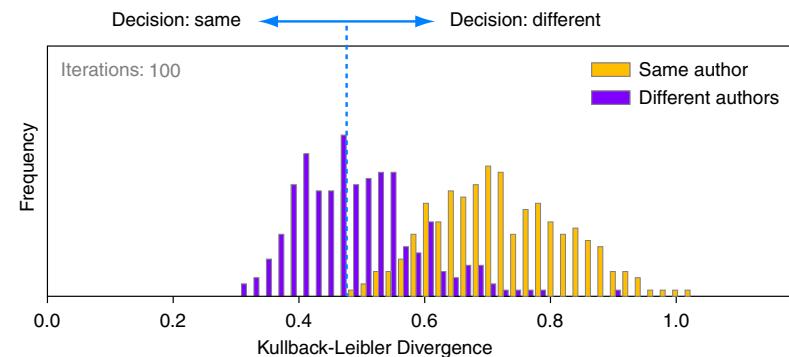


Authorship attribution

Authorship verification



Verification means one-class classification



Obfuscation by KLD manipulation

Thank you!