Authorship Obfuscation
Using Heuristic Search

Master's Thesis Defence by Janek Bevendorff on 20 June 2018

Supervisors: Prof. Dr. Benno Stein, PD Dr. Andreas Jakoby

Contributions

- Unmasking for short texts
- Obfuscation against unmasking
- Obfuscation against compression models
- Authorship verification quality measure proposal
- Obfuscation safety analysis and definitions
- Side effect analysis
- JS_{Δ} as authorship metric
- Adaptive obfuscation
- Design of an admissible obfuscation heuristic
- Analysis of consistency and monotonicity properties
- Design and implementation of an efficient obfuscation framework
- Development of obfuscation operators
- Inspection of search space challenges and solutions

Contributions

- Unmasking for short texts
- Obfuscation against unmasking
- Obfuscation against compression models
- Authorship verification quality measure proposal
- Obfuscation safety analysis and definitions
- Side effect analysis
- JS_{Δ} as authorship metric
- Adaptive obfuscation
- Design of an admissible obfuscation heuristic
- Analysis of consistency and monotonicity properties
- Design and implementation of an efficient obfuscation framework
- Development of obfuscation operators
- Inspection of search space challenges and solutions


SHERLOCK HOLMES
MẬT MÃ MAYA
THIÊN THẦN VÀ ÁC QUỶ
CUỐN THEO CHIỀU GIÓ
ĐỒI GIÓ HÚ
THẦN THOẠI HY LẠP
MONG TO CRIXTO
NELSON DEMILLE
Chiến tranh và hòa bình
NGUYÊN HUY
Việt Nam Phong Tục
Mark Twain
TAY SUNG
CON

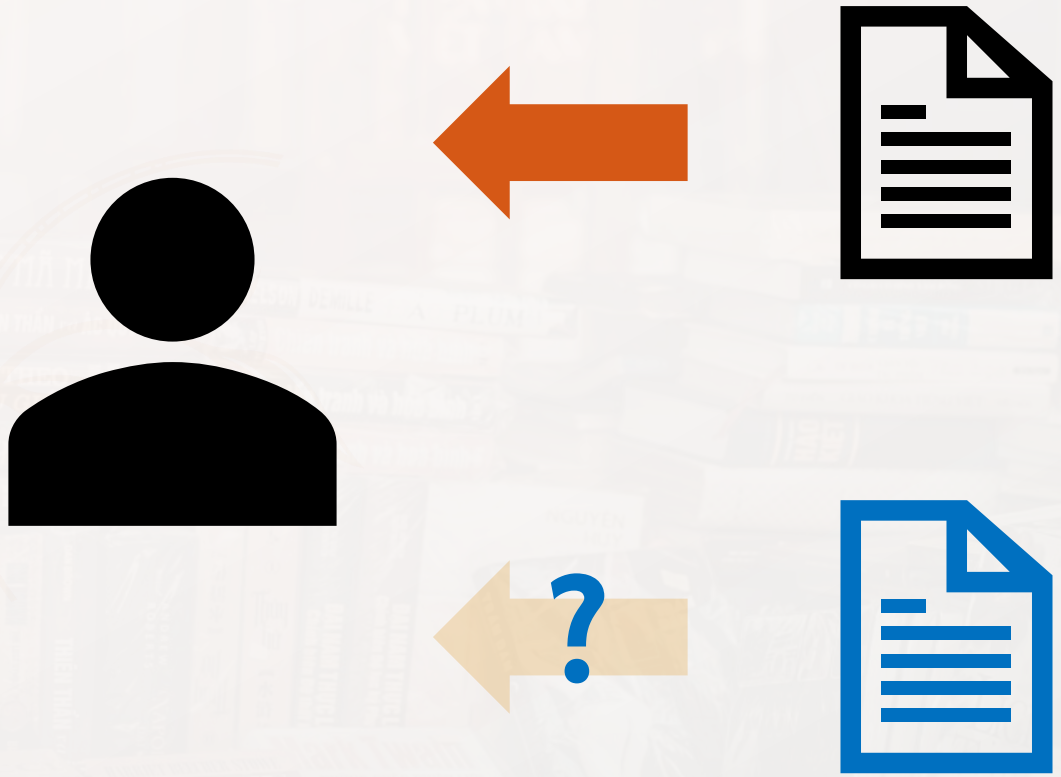
Authorship

Authorship Verification

Authorship Verification

Authorship Verification

Authorship Verification



Unmasking

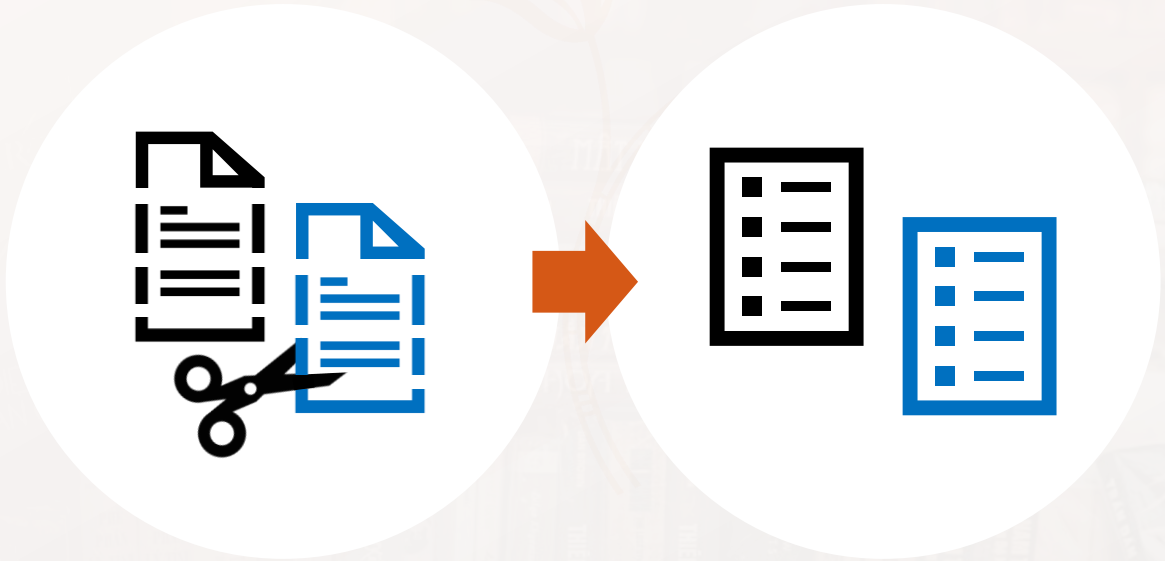
Koppel and Schler, Authorship verification as a one-class problem, 2004

Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004

Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004

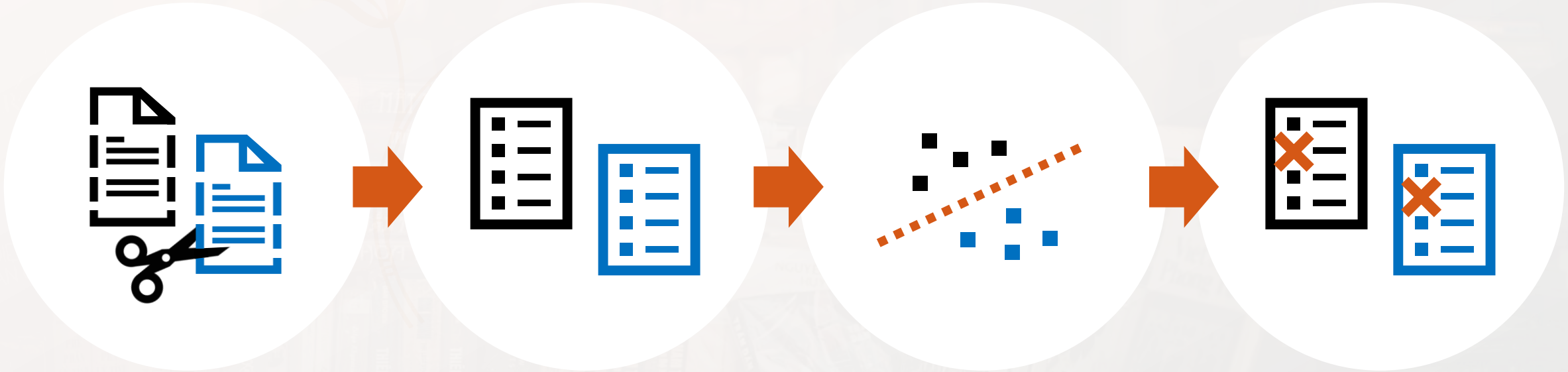


Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004

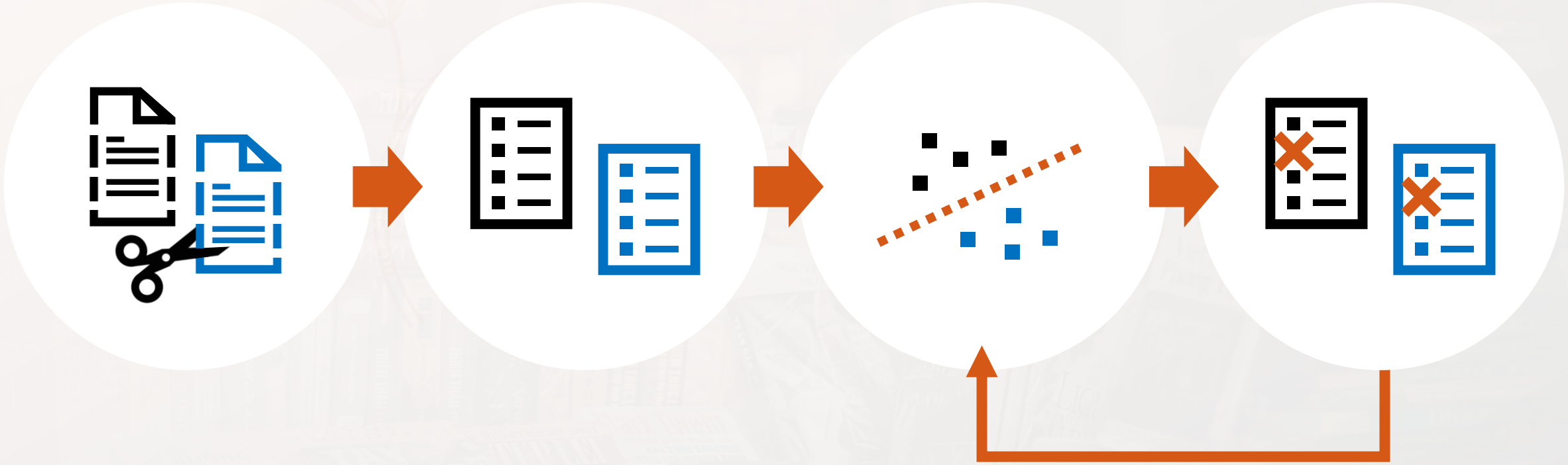
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004



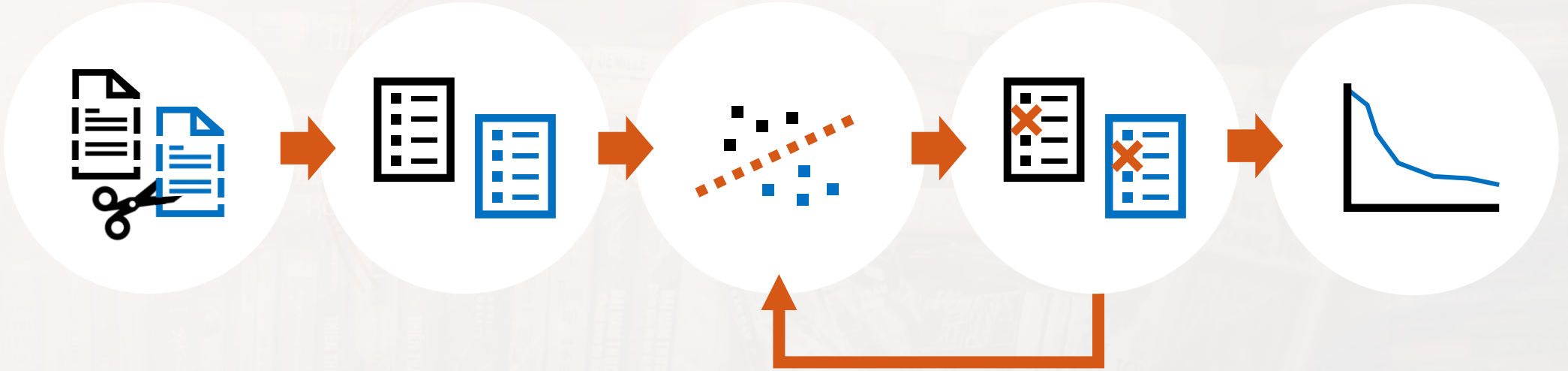
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004

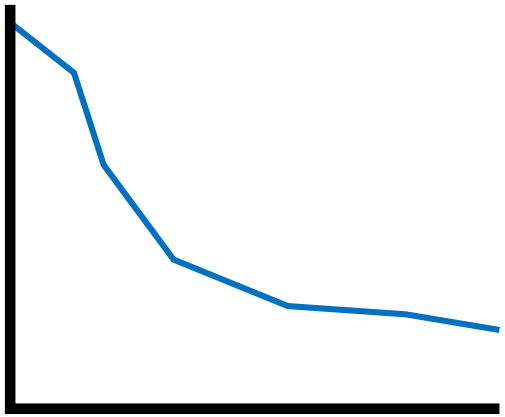


Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004



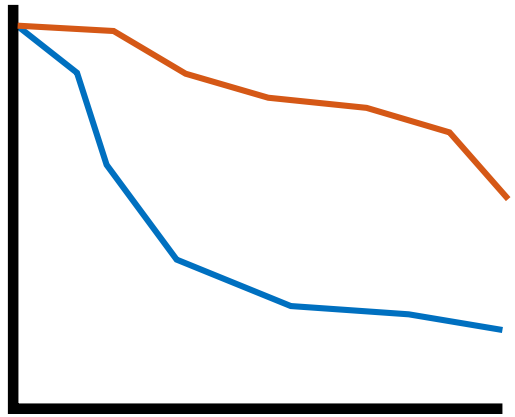
Unmasking



Same author

Unmasking

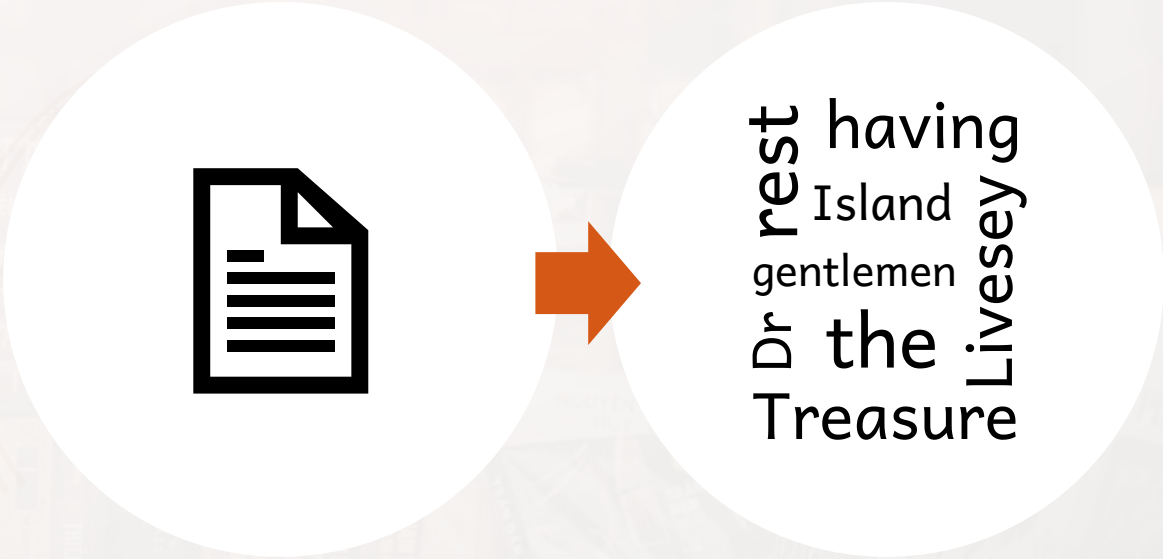
Different authors



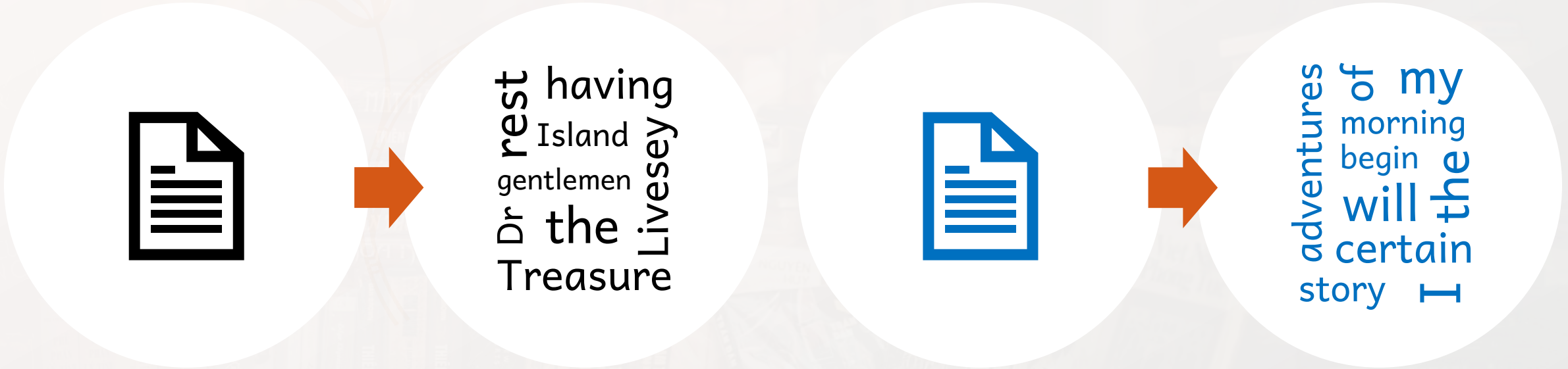
Same author

Chunk Expansion

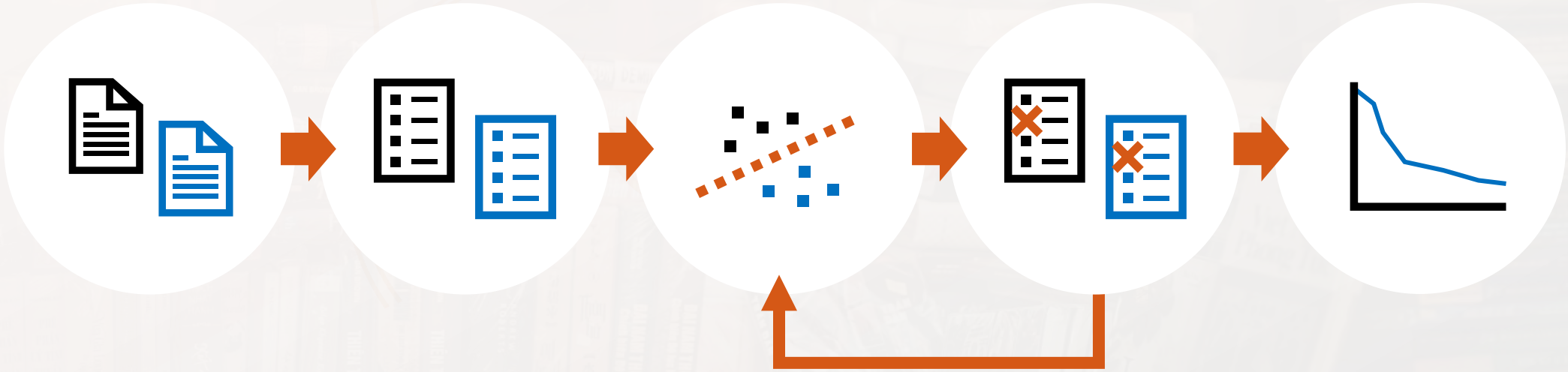
Chunk Expansion



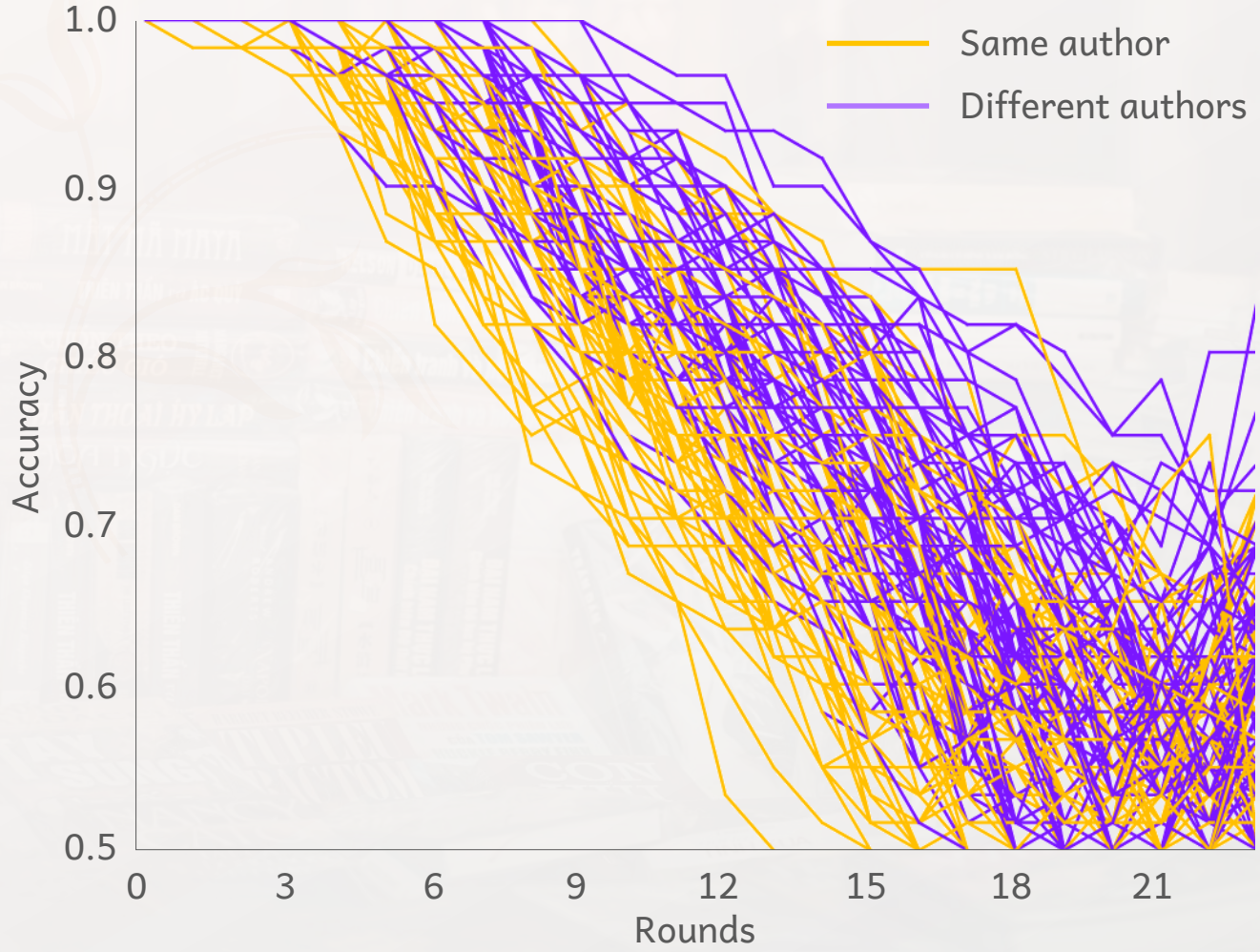
Chunk Expansion



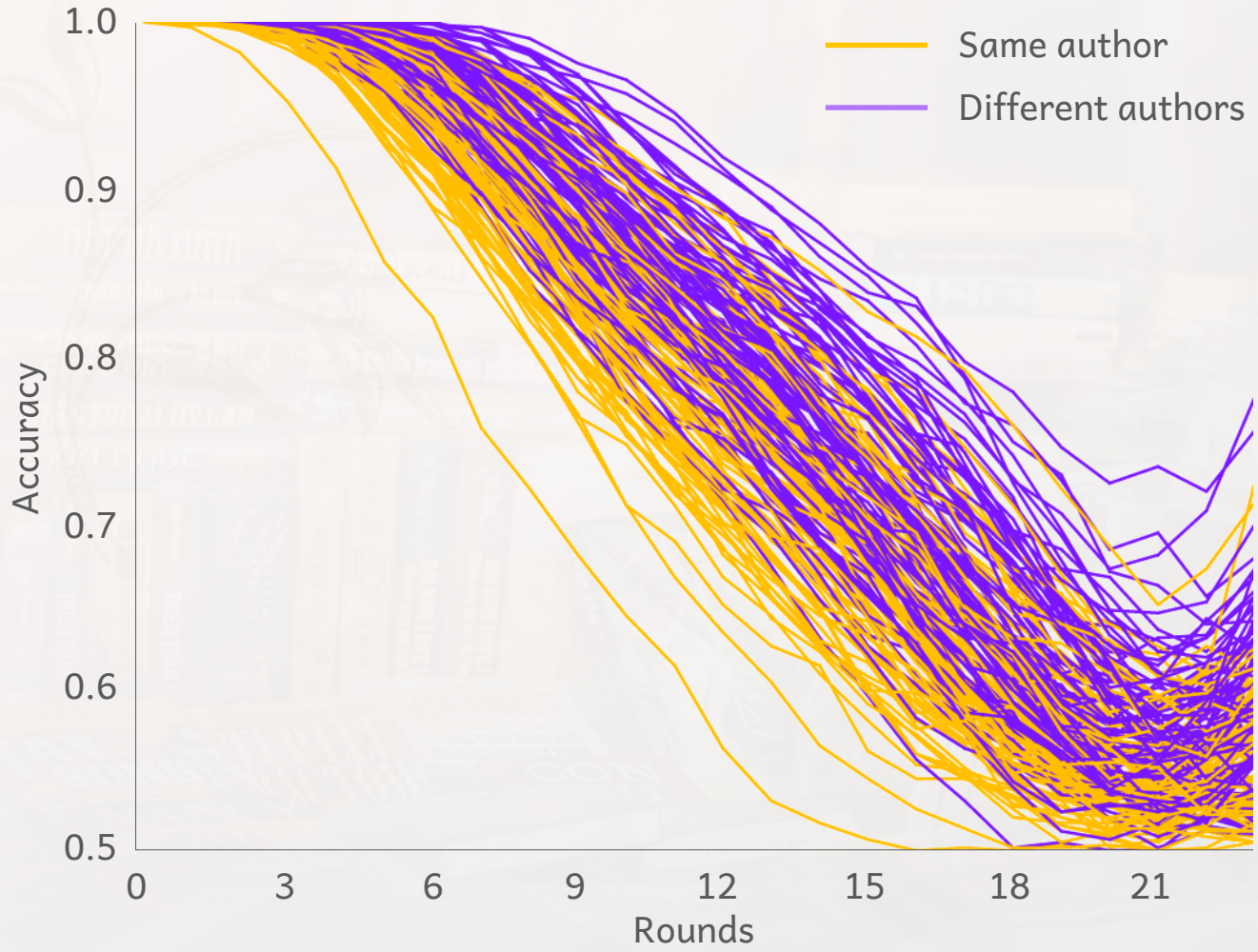
Chunk Expansion



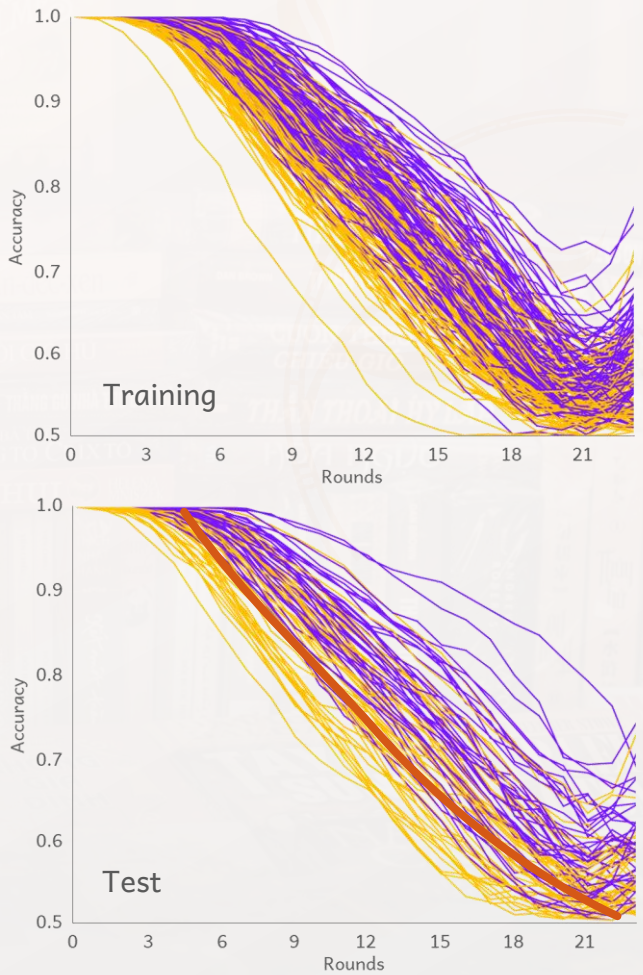
Unmasking with Chunk Expansion



Unmasking with Chunk Expansion



Unmasking with Chunk Expansion

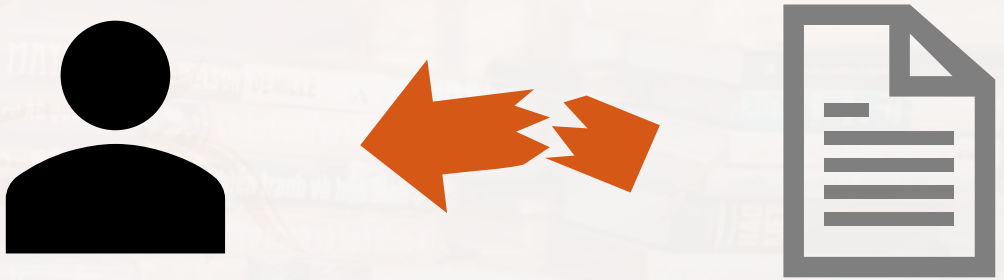


Confidence Level	Threshold	Precision	% Classified
Very High	0.9	1.00	6.2
	0.8	1.00	12.5
	0.7	1.00	13.8
High	0.6	1.00	18.8
	0.5	1.00	30.0
Moderate	0.4	0.93	43.8
	0.3	0.83	55.0
	0.2	0.68	70.0
Low	0.1	0.82	87.5
	0.0	0.76	100.0

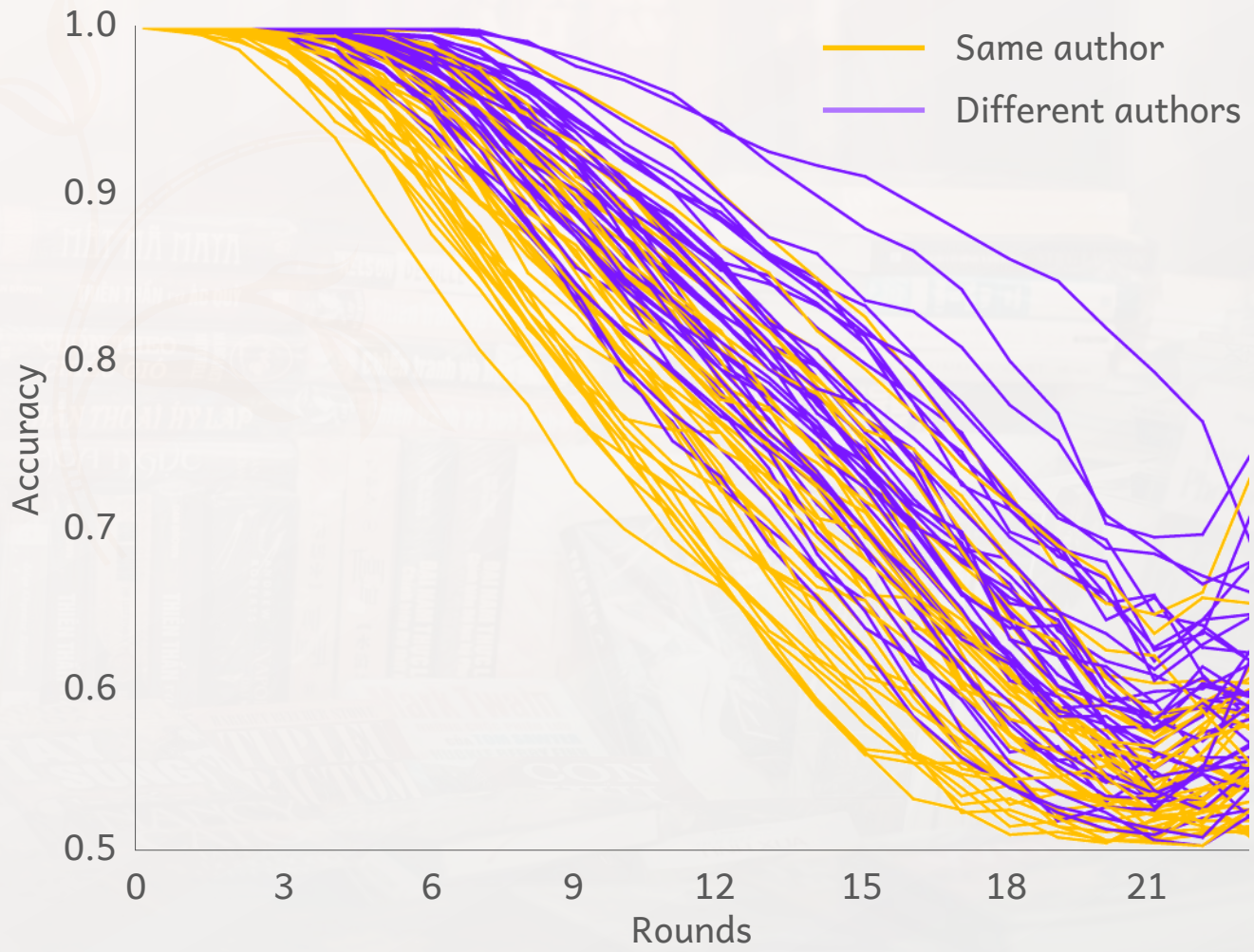
Obfuscation

Authorship Obfuscation

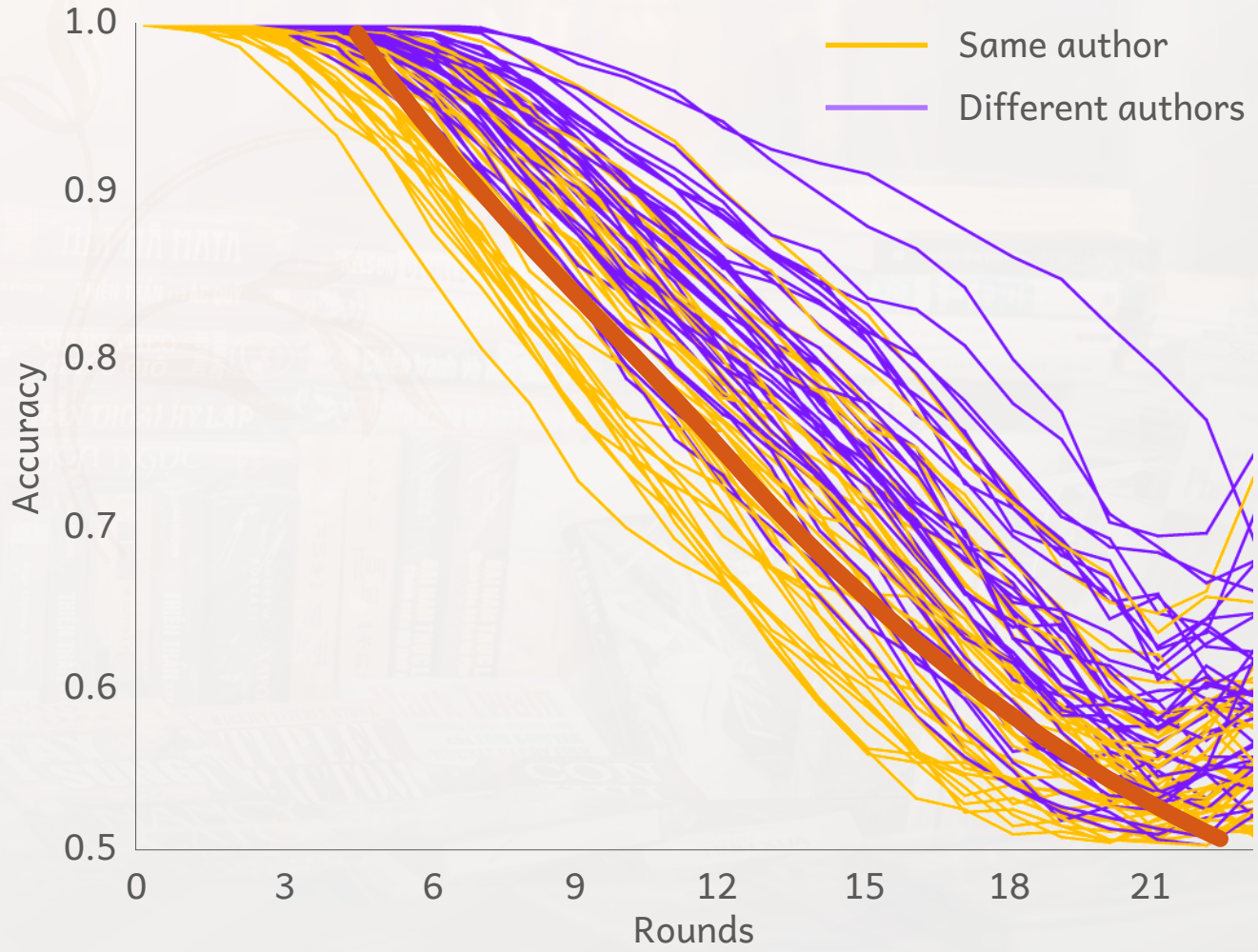
Authorship Obfuscation



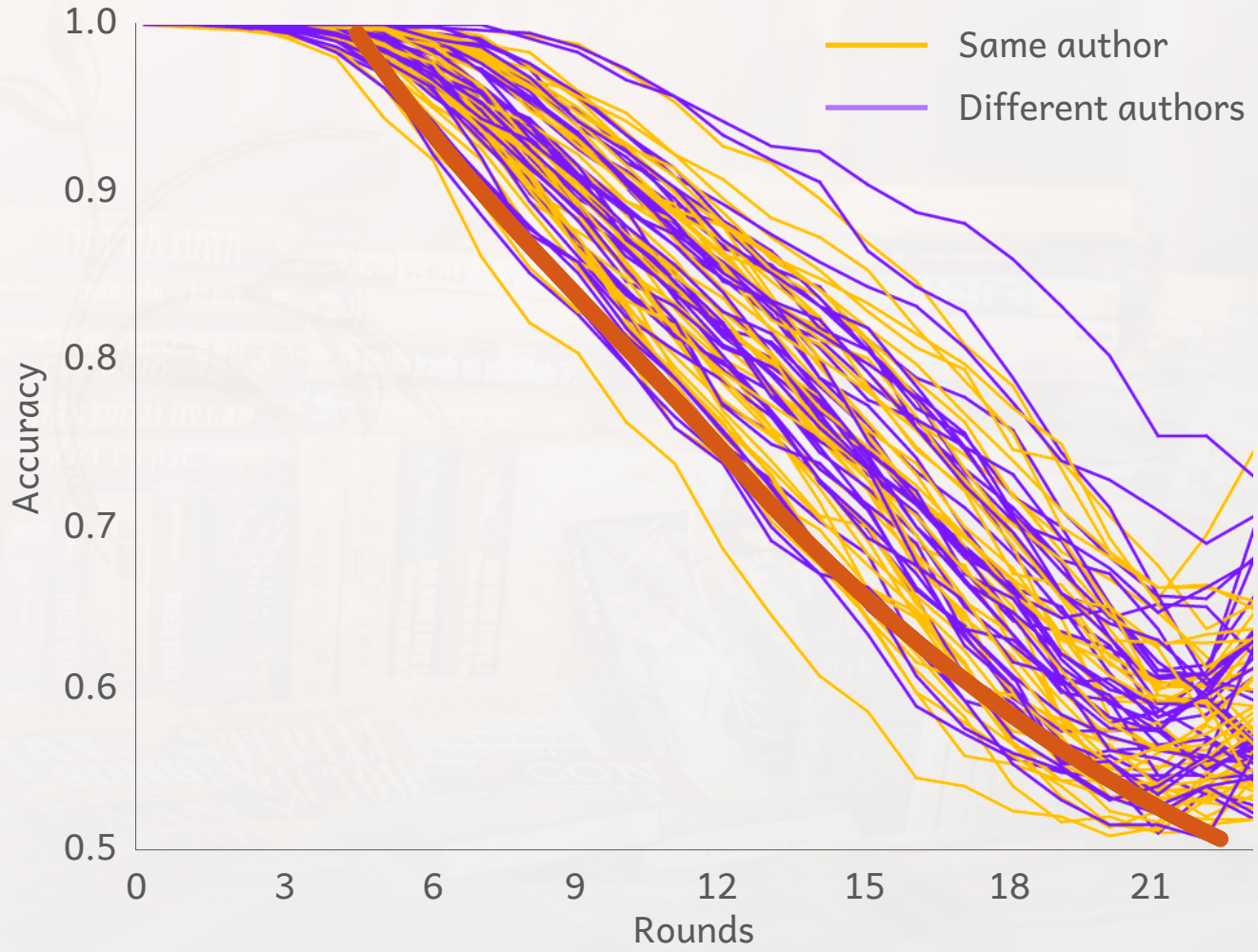
Authorship Obfuscation



Authorship Obfuscation



Authorship Obfuscation



Kullback-Leibler Divergence

$$\mathrm{KLD}(P\|Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]}$$

Jensen-Shannon Divergence

$$\mathrm{KLD}(P\|Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]}$$

$$\mathrm{JSD}(P\|Q) = \frac{\mathrm{KLD}(P\|M) + \mathrm{KLD}(Q\|M)}{2}$$

$$M = \frac{P + Q}{2}$$

Jensen-Shannon Divergence

$$\mathrm{KLD}(P\|Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]}$$

$$\mathrm{JSD}(P\|Q) = \frac{\mathrm{KLD}(P\|M) + \mathrm{KLD}(Q\|M)}{2} \qquad \rightarrow \text{maximize}$$

$$M = \frac{P + Q}{2}$$

Basic Obfuscation

$$\frac{\partial}{\partial Q[i]}\left(P[i]\log_2\frac{P[i]}{Q[i]}\right) = -\frac{P[i]}{Q[i]\ln 2}$$

Basic Obfuscation

$$\frac{\partial}{\partial Q[i]}\left(P[i]\log_2\frac{P[i]}{Q[i]}\right) = -\frac{P[i]}{Q[i]\ln 2}$$

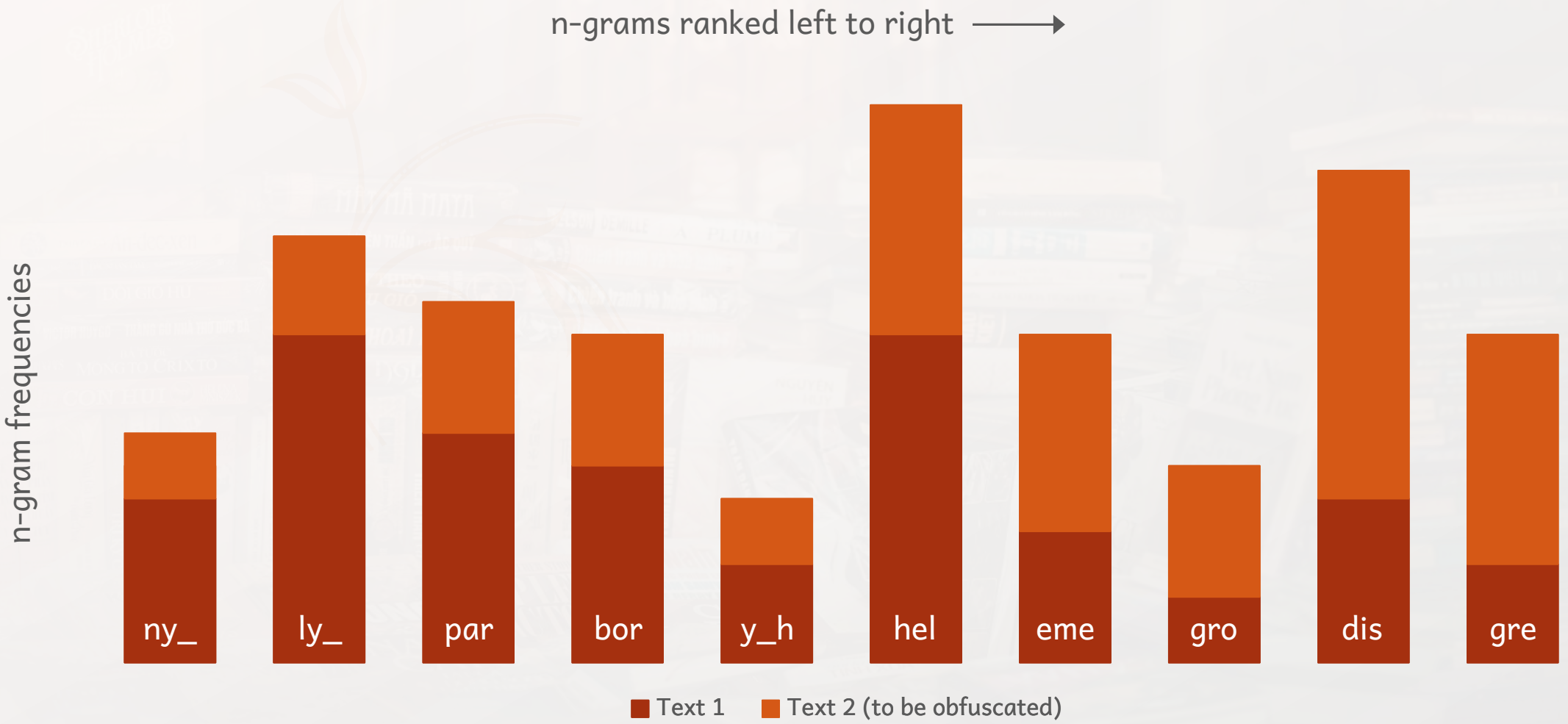
$$\mathrm{R_{KL}}\,(i) = \frac{P[i]}{Q[i]}$$

Basic Obfuscation

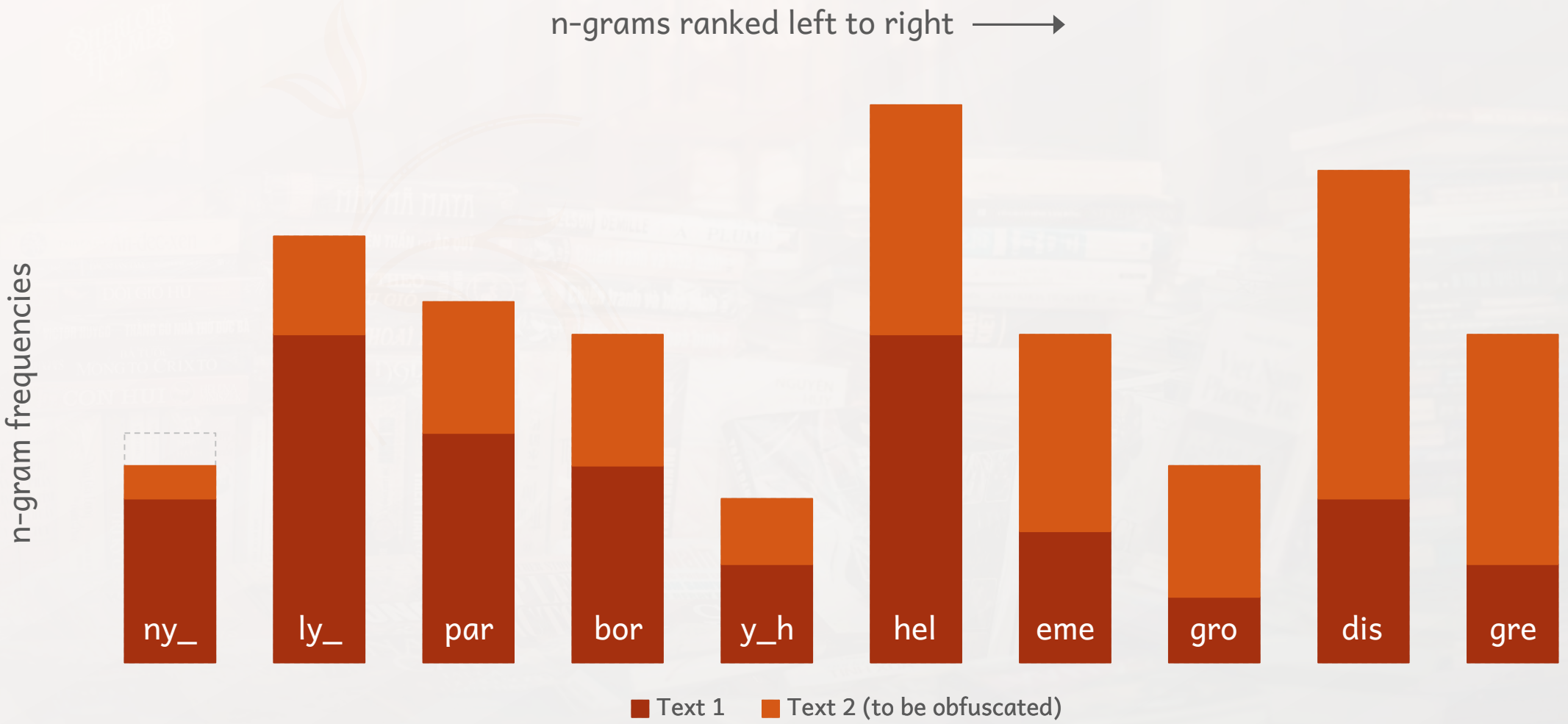
$$\frac{\partial}{\partial Q[i]}\left(P[i]\log_2\frac{P[i]}{Q[i]}\right) = -\frac{P[i]}{Q[i]\ln 2}$$

$$\mathrm{R_{KL}}\,(i) = \frac{P[i]}{Q[i]} \qquad \rightarrow \text{maximize}$$

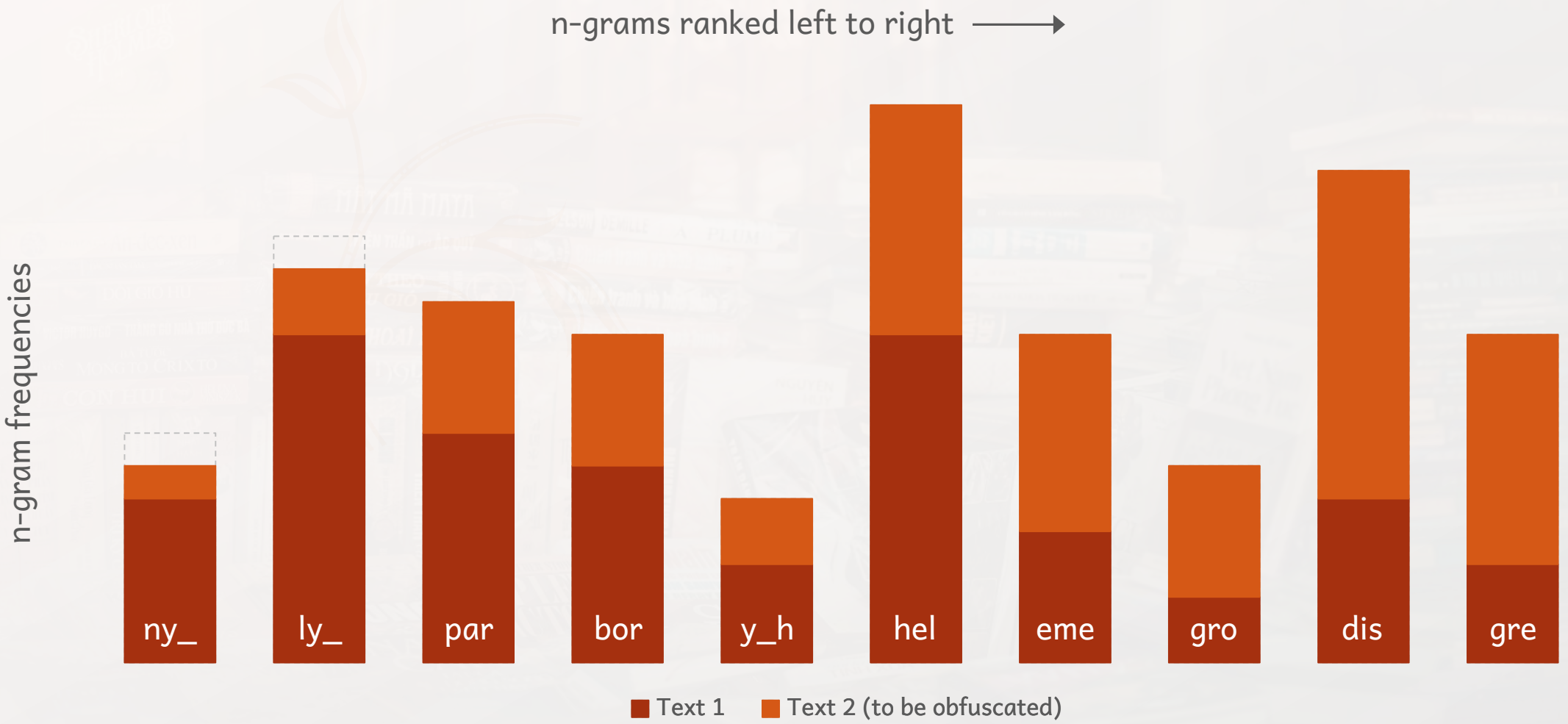
Basic Obfuscation



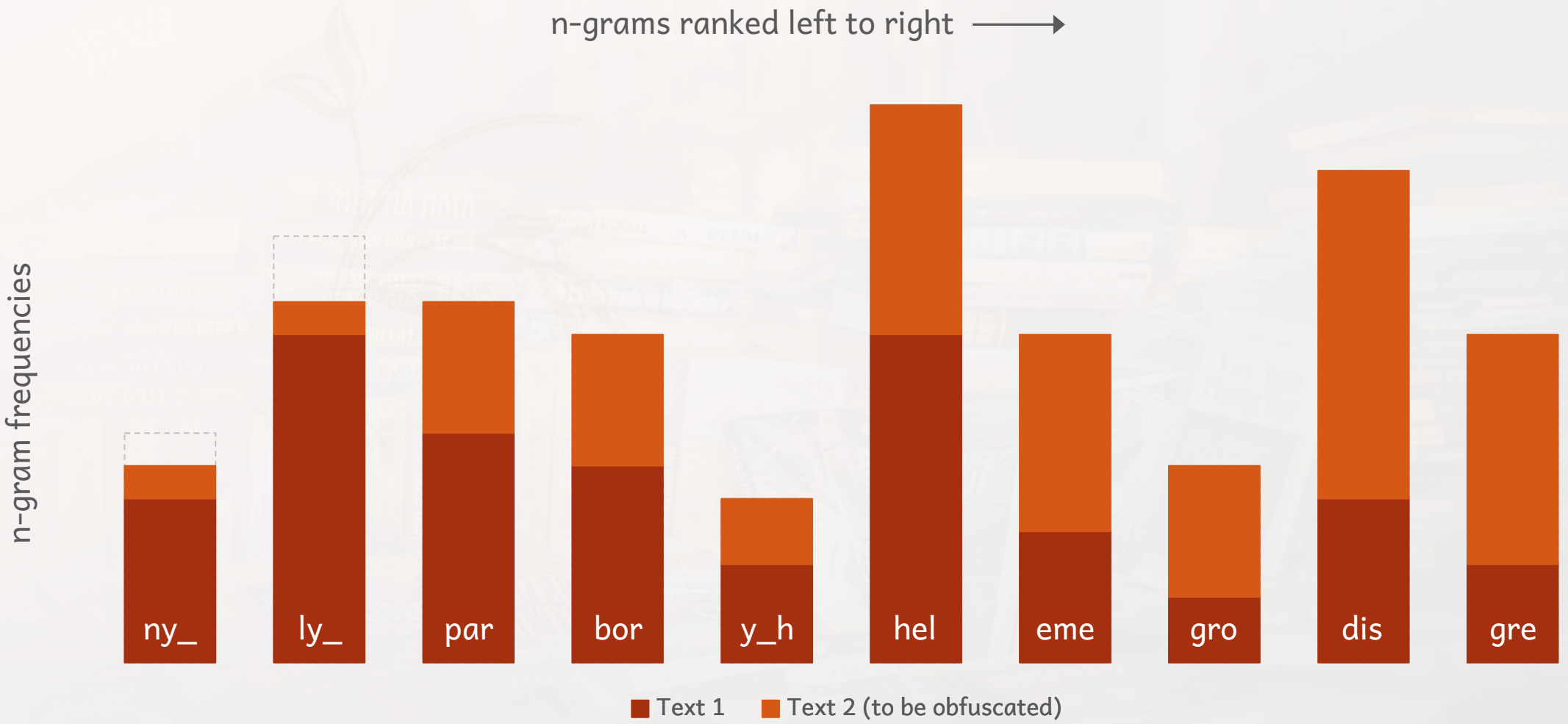
Basic Obfuscation



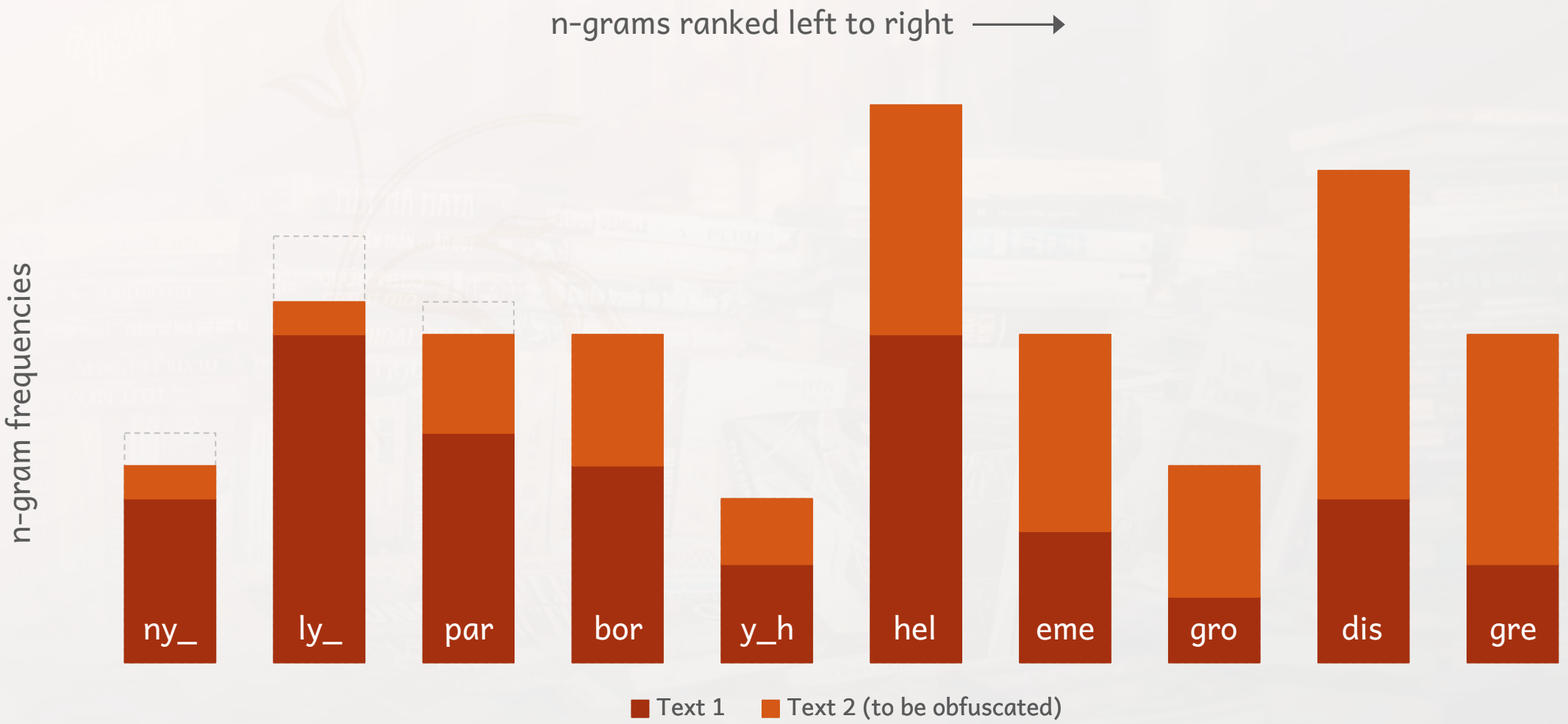
Basic Obfuscation



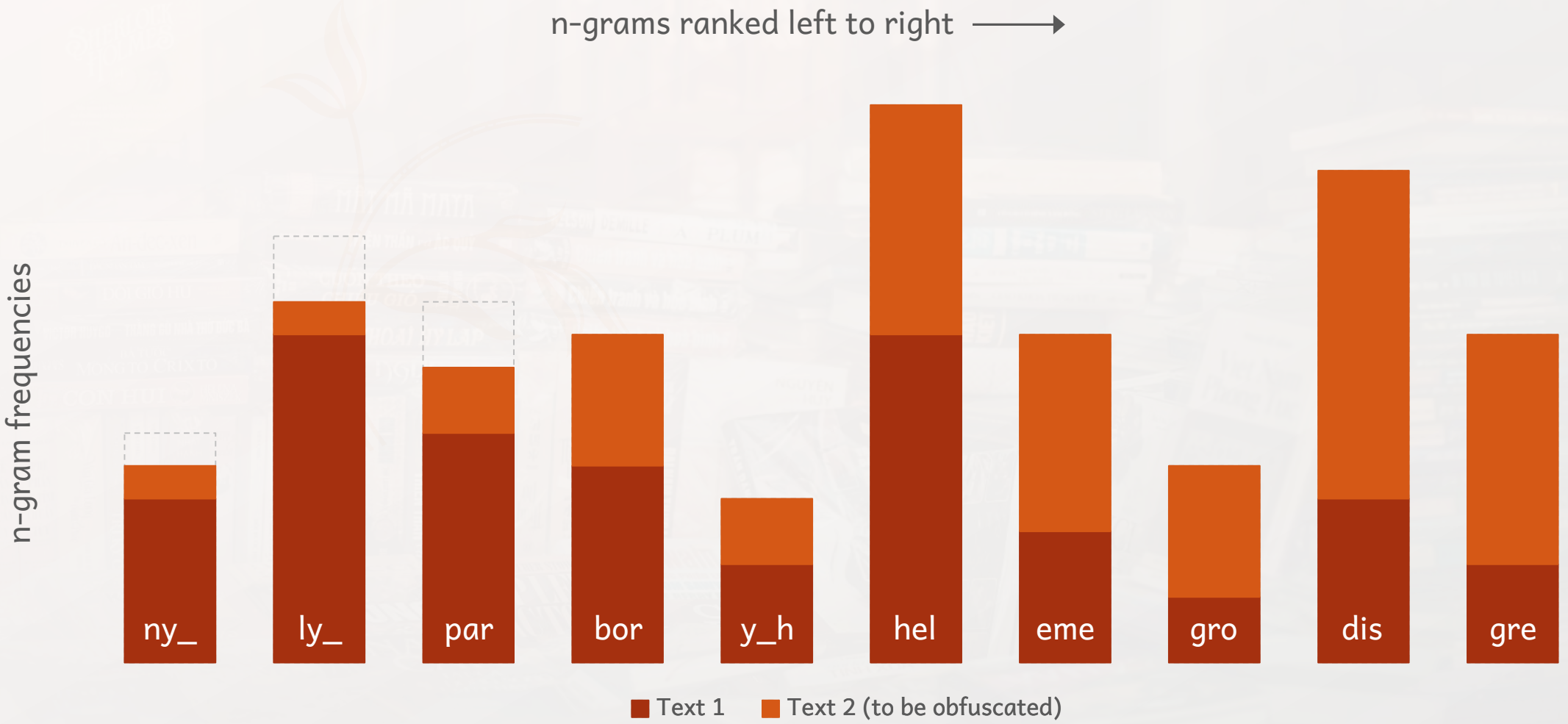
Basic Obfuscation



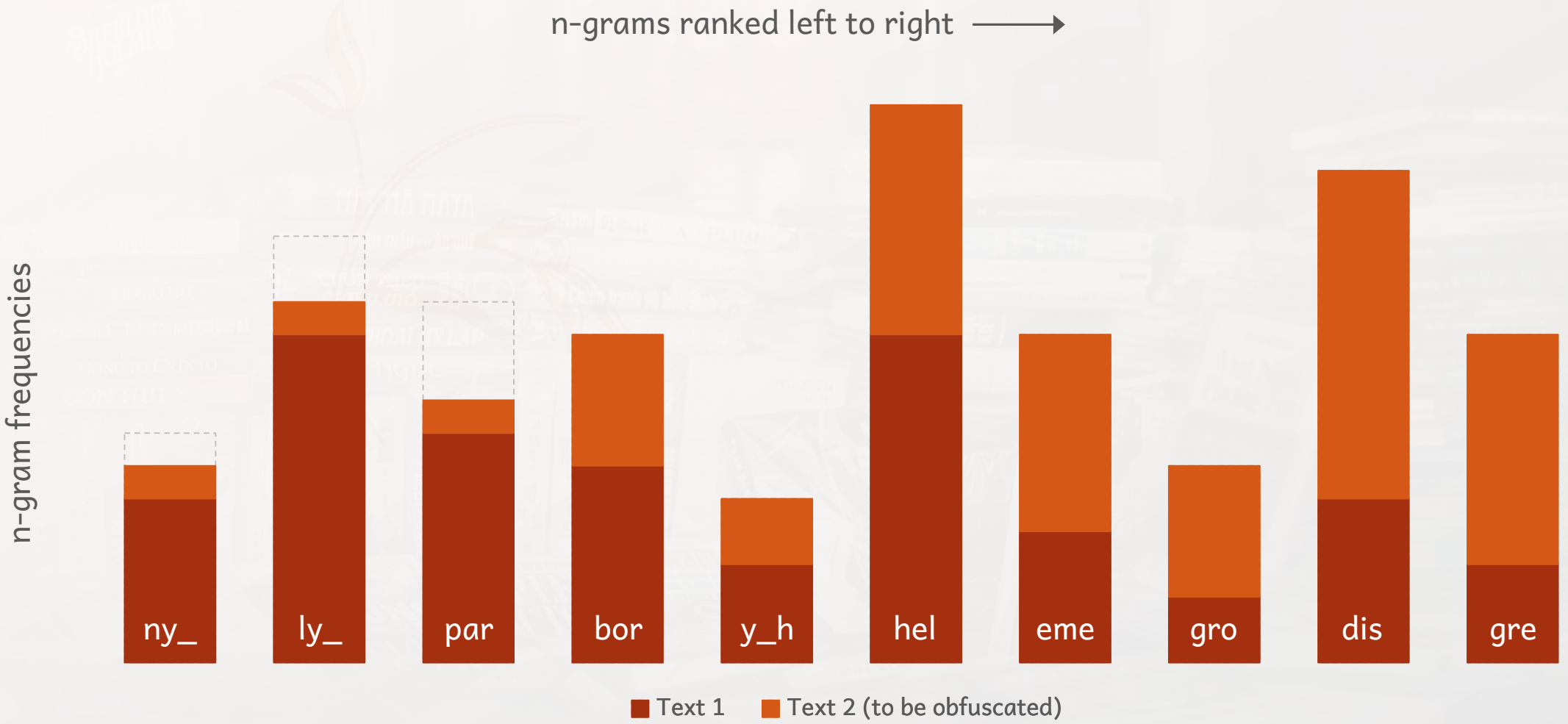
Basic Obfuscation



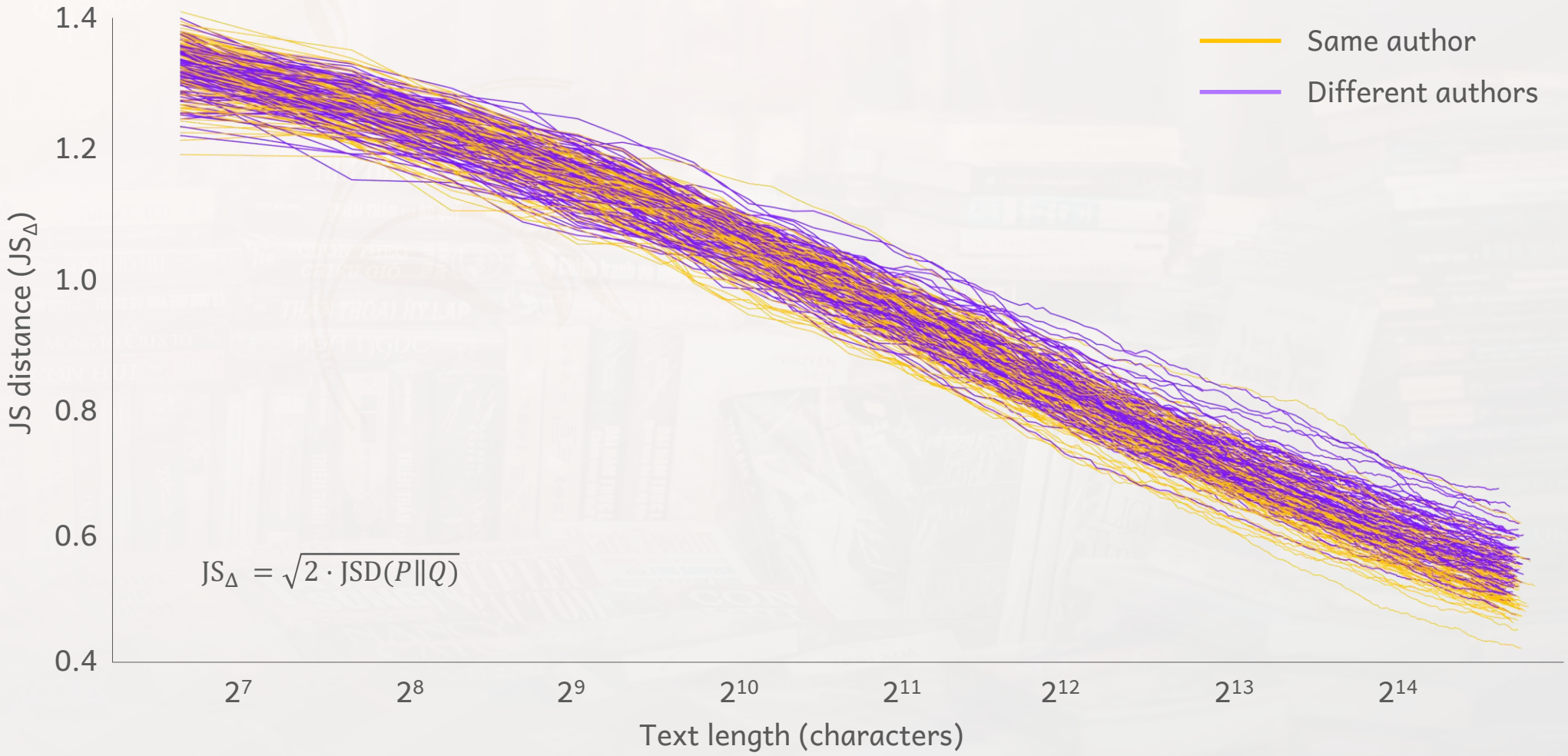
Basic Obfuscation



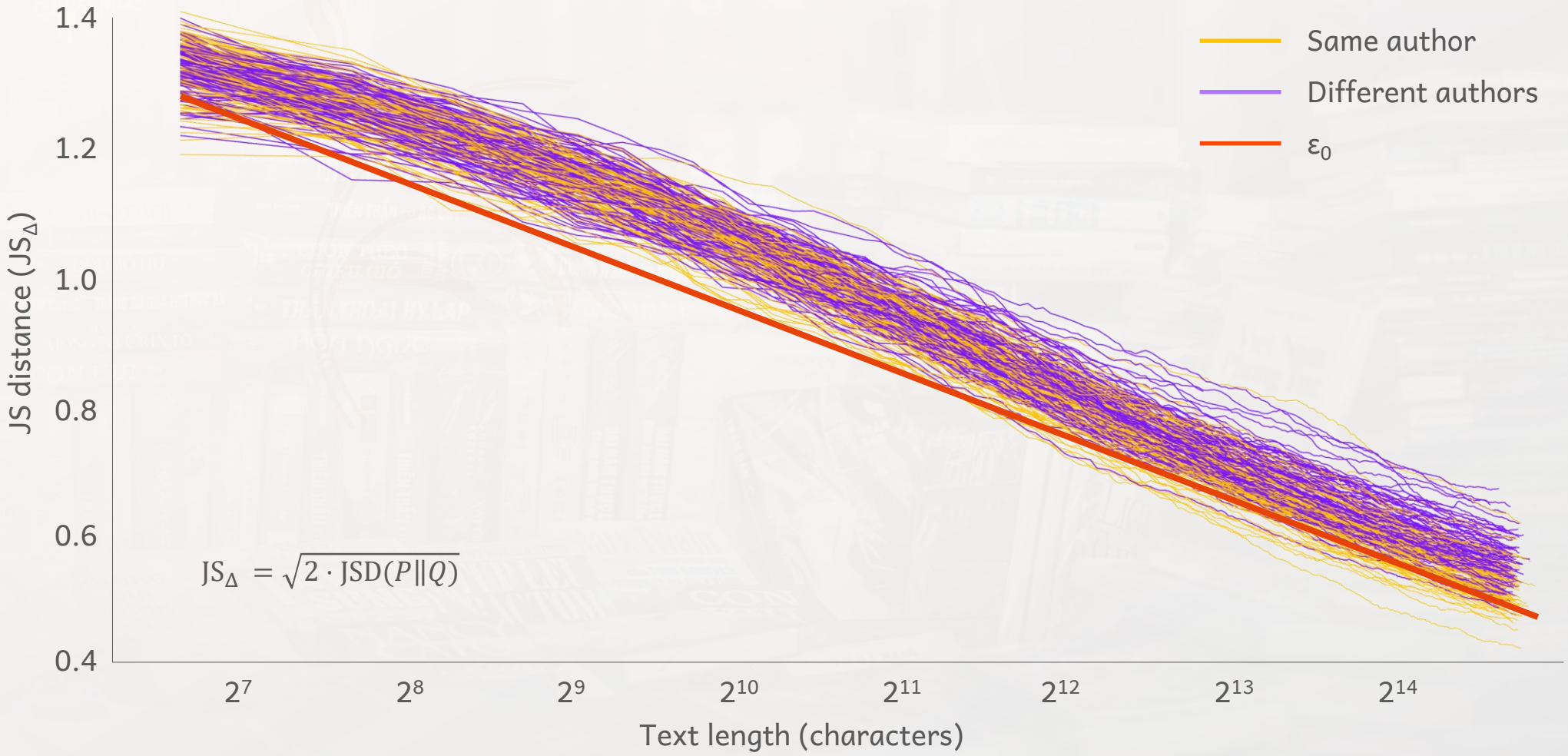
Basic Obfuscation



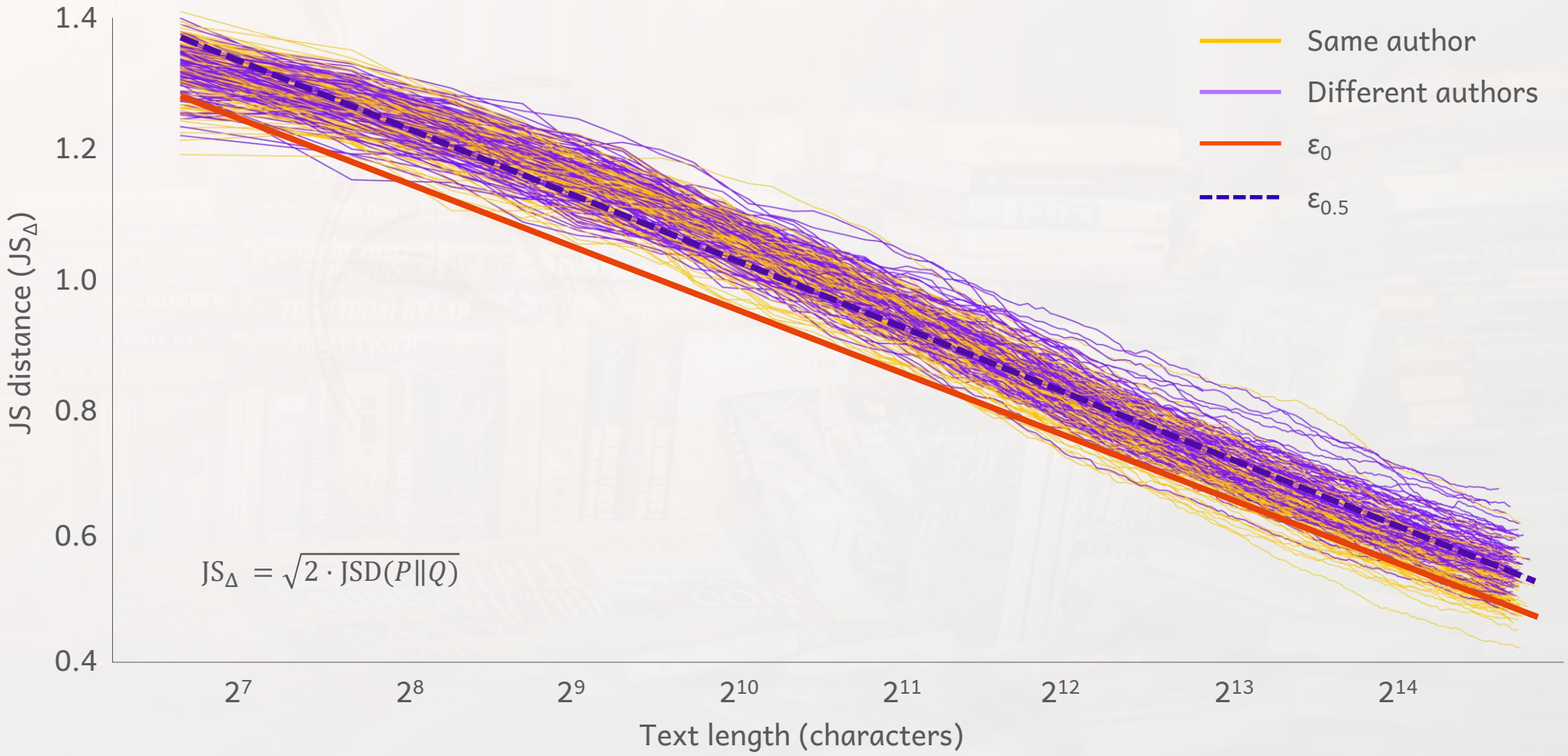
Adaptive Obfuscation



Adaptive Obfuscation



Adaptive Obfuscation



Obfuscation Results

Confidence Level	Threshold	Precision	% Classified
Very High	0.9	1.00	6.2
	0.8	1.00	12.5
	0.7	1.00	13.8
High	0.6	1.00	18.8
	0.5	1.00	30.0
Moderate	0.4	0.93	43.8
	0.3	0.83	55.0
	0.2	0.68	70.0
Low	0.1	0.82	87.5
	0.0	0.76	100.0

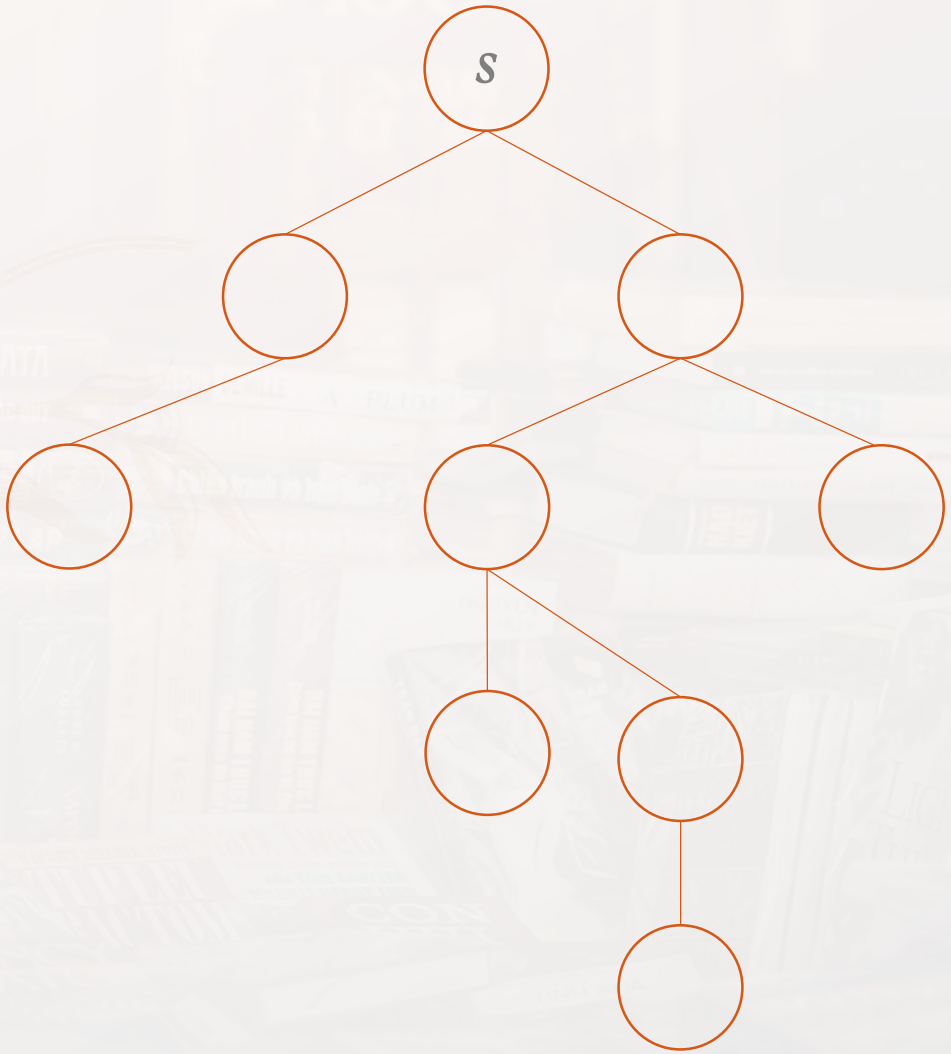
Obfuscation Results

Confidence Level	Threshold	Precision	% Classified
Very High	0.9	1.00	6.2
	0.8	1.00	12.5
	0.7	1.00	13.8
High	0.6	1.00	18.8
	0.5	1.00	30.0
Moderate	0.4	0.93	43.8
	0.3	0.83	55.0
	0.2	0.68	70.0
Low	0.1	0.82	87.5
	0.0	0.76	100.0

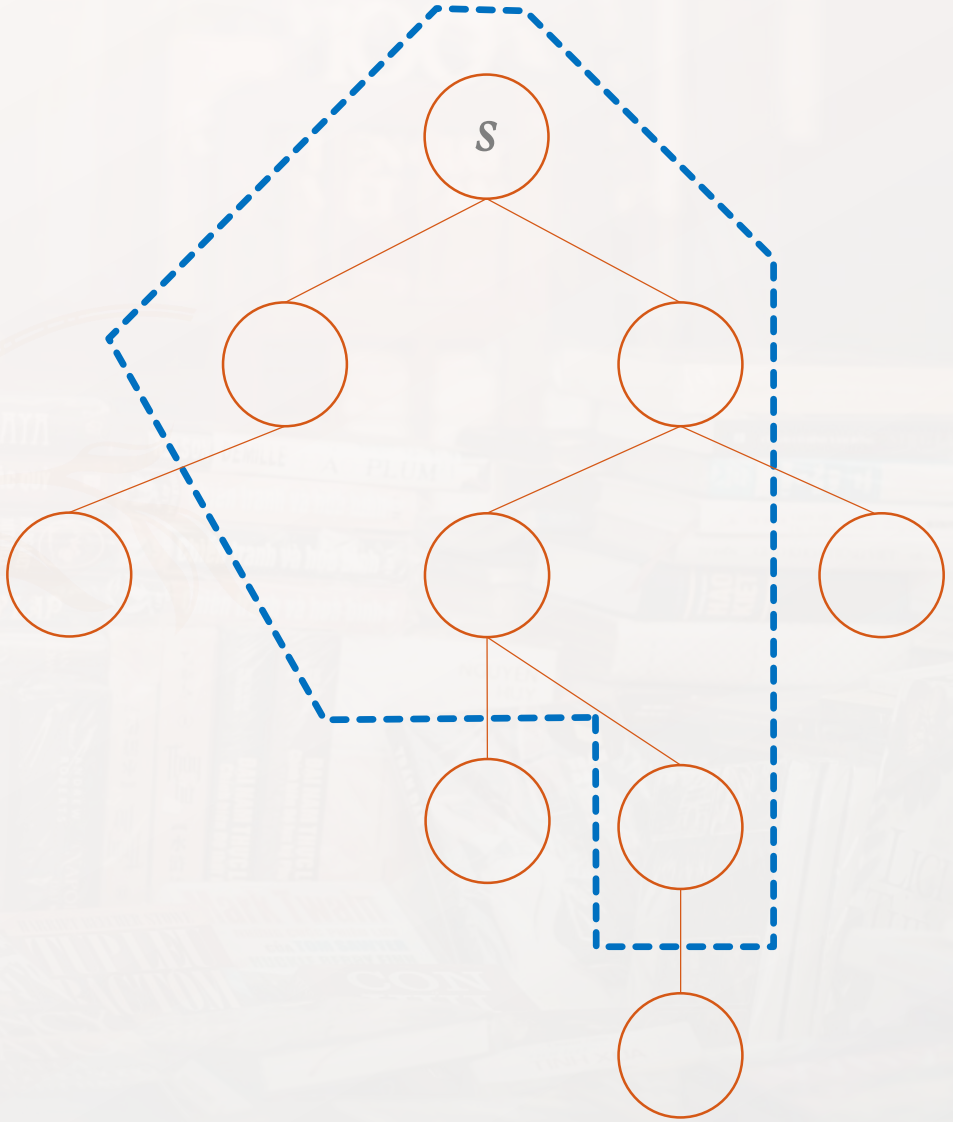
Confidence Level	Threshold	Precision	% Classified
Very High	0.9	0.00	2.5
	0.8	0.00	5.0
	0.7	0.00	8.7
High	0.6	0.00	17.5
	0.5	0.00	27.5
Moderate	0.4	0.00	42.5
	0.3	0.67	66.7
	0.2	0.50	70.0
Low	0.1	0.42	85.0
	0.0	0.53	100.0

Heuristic Search

Best-first Search

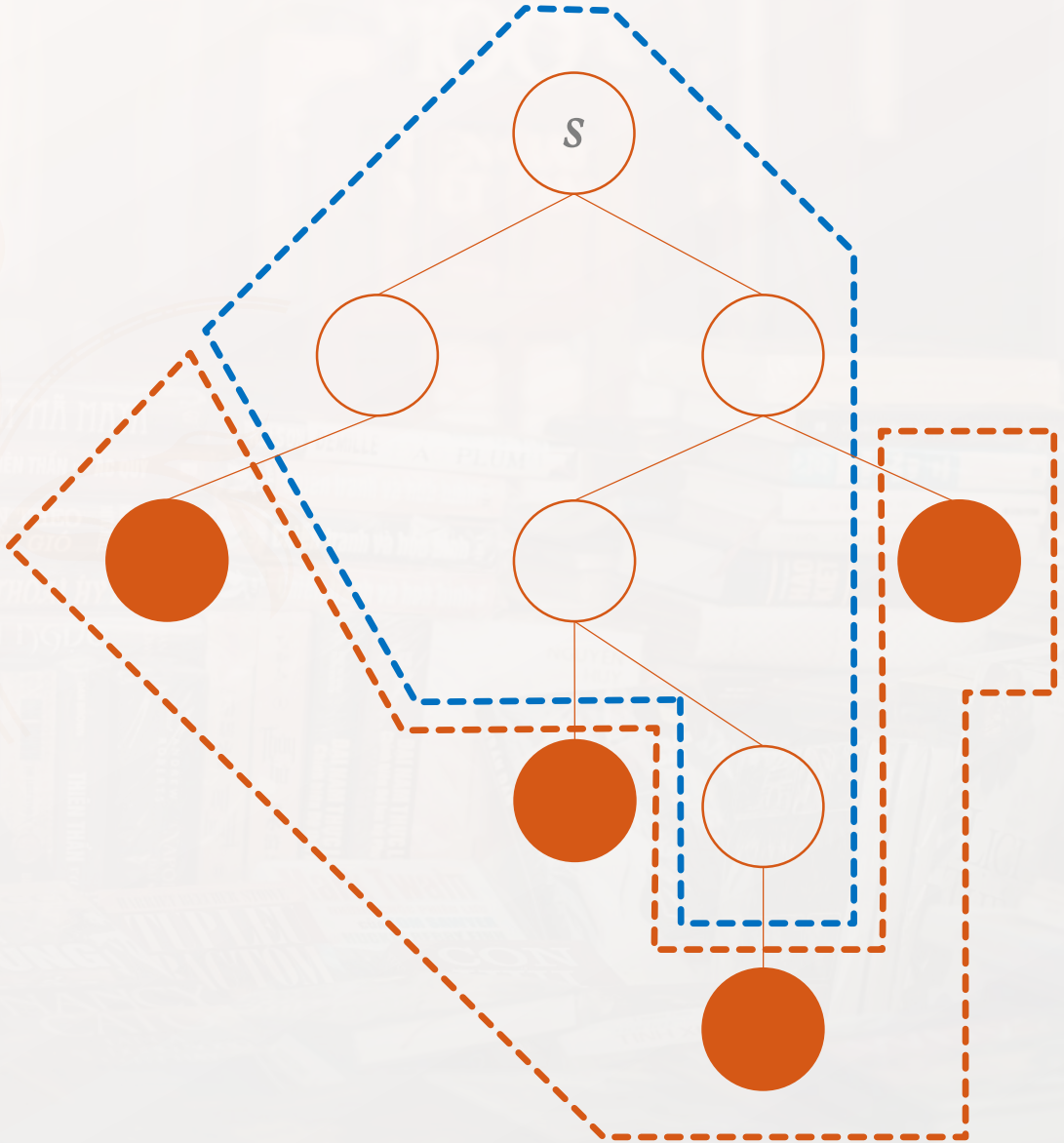


Best-first Search



CLOSED

Best-first Search

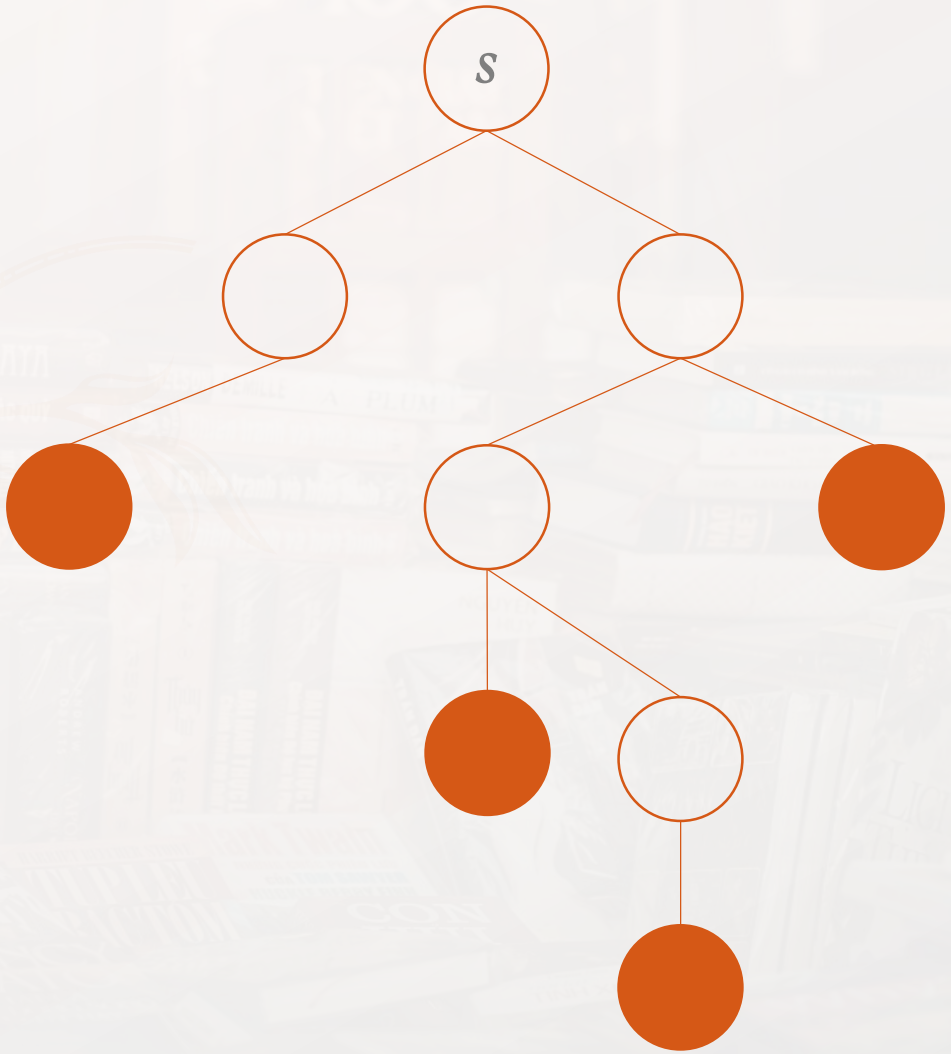


CLOSED

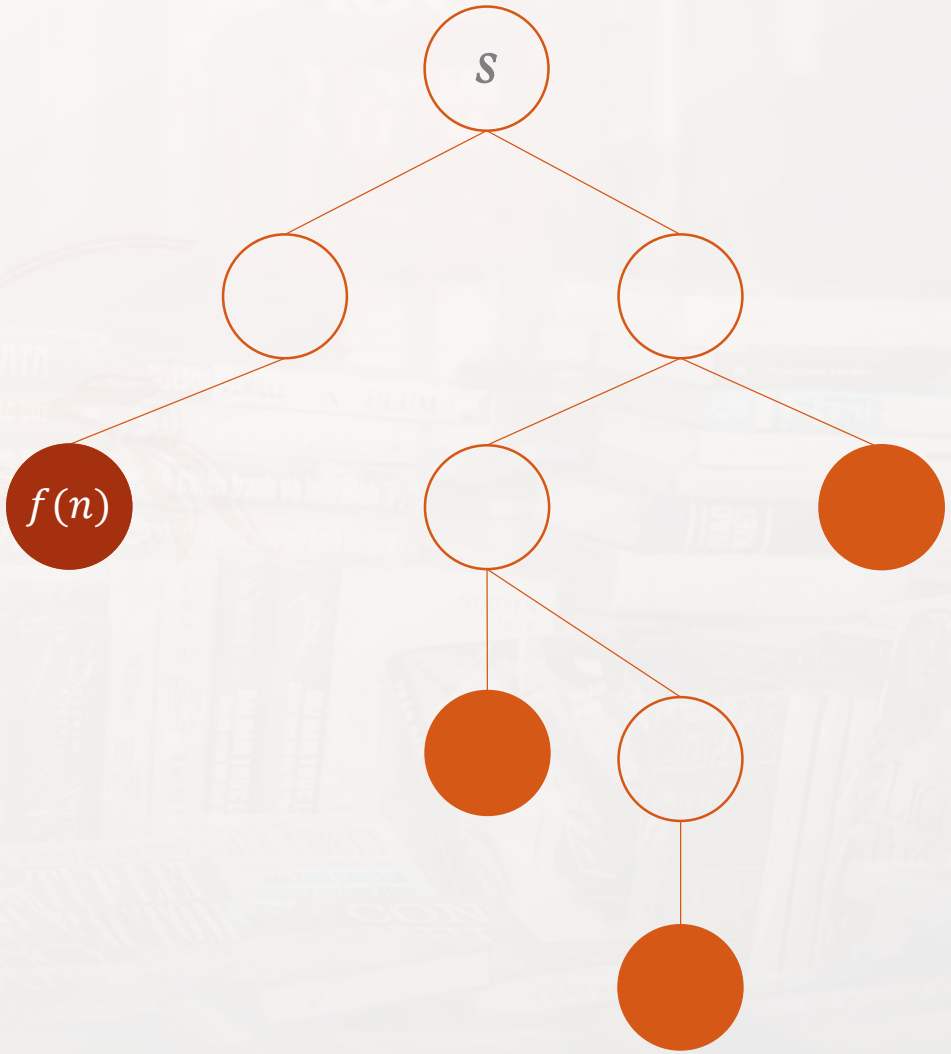
OPEN

Best-first Search

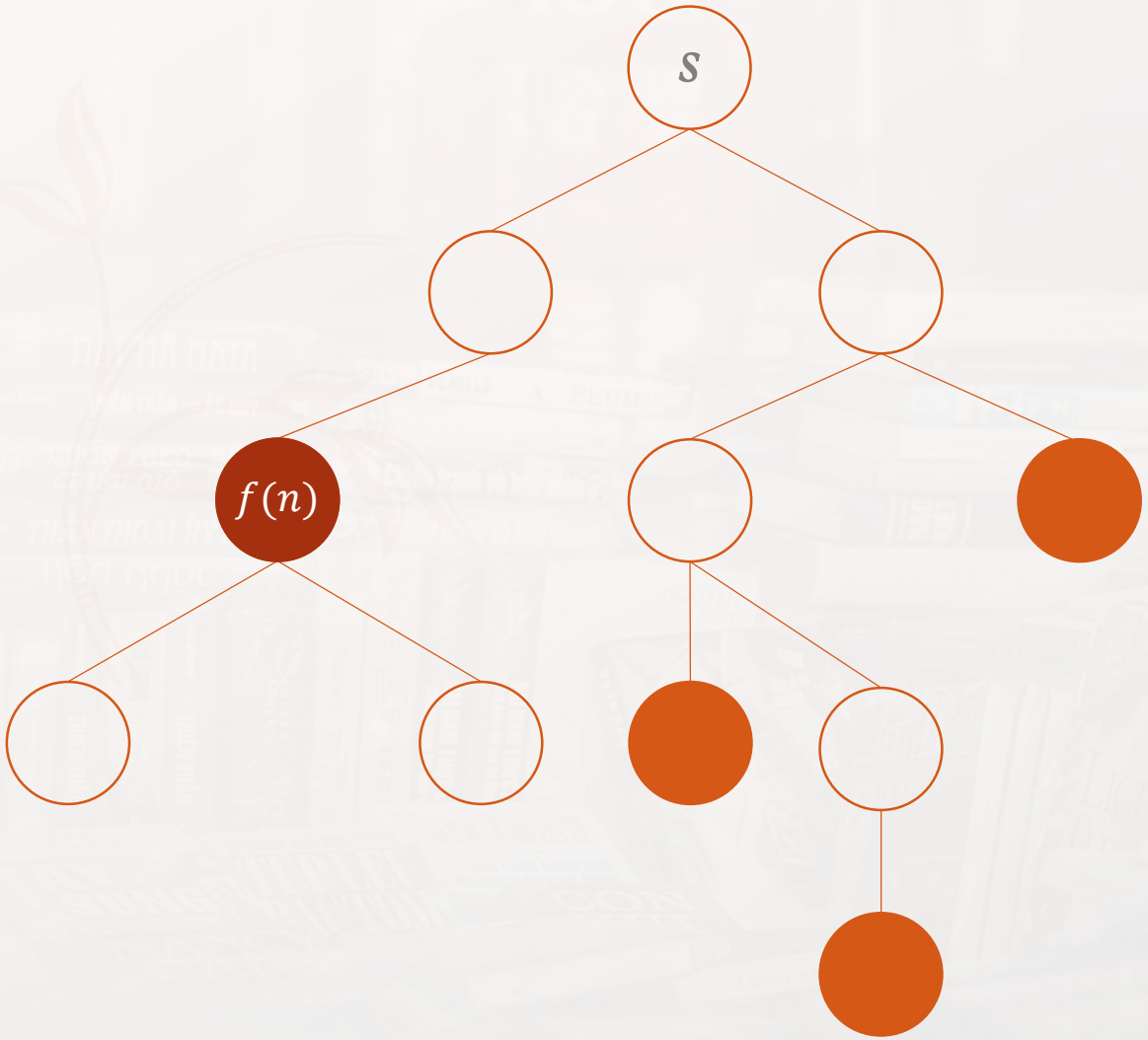
$f(n)$



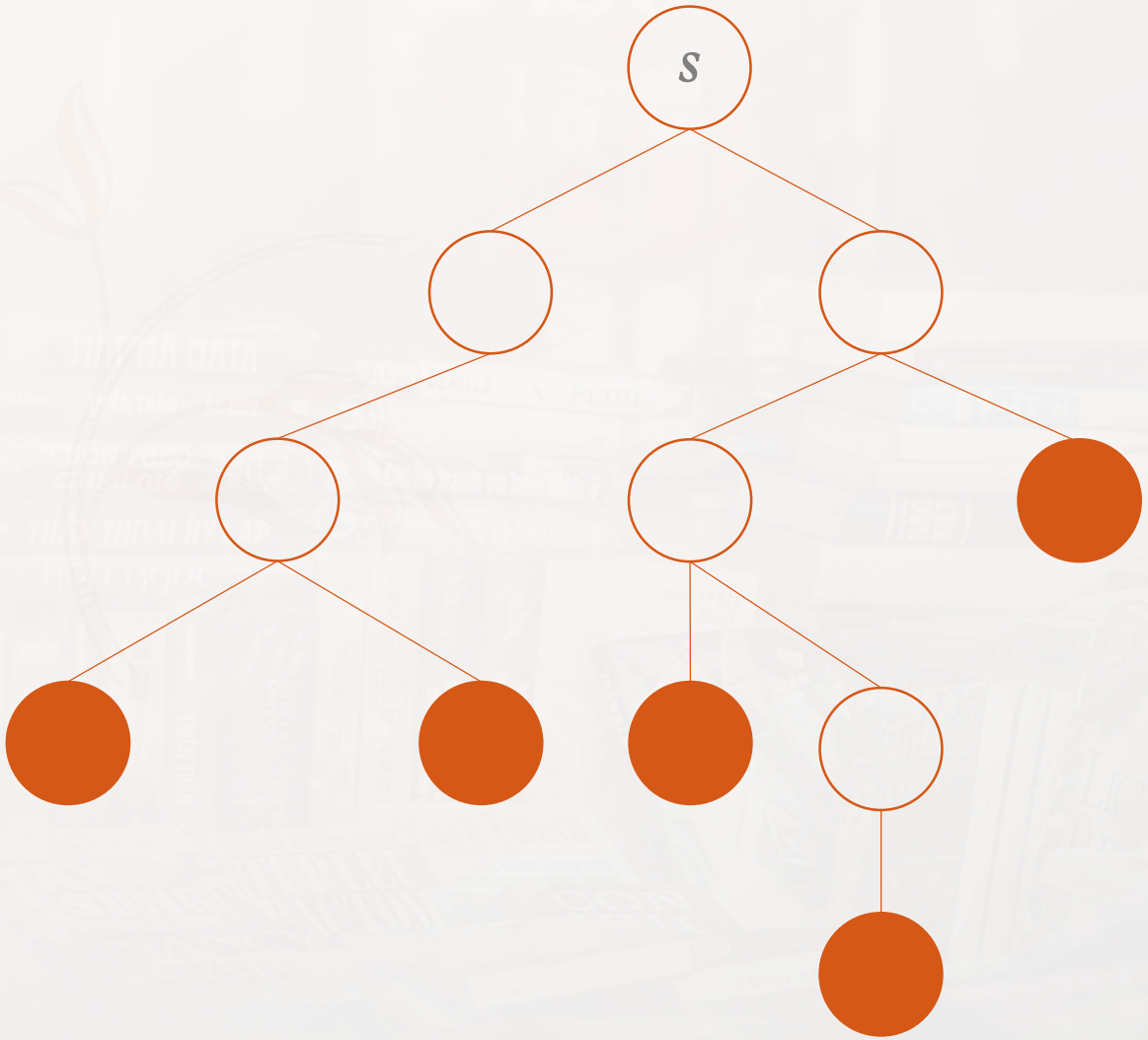
Best-first Search



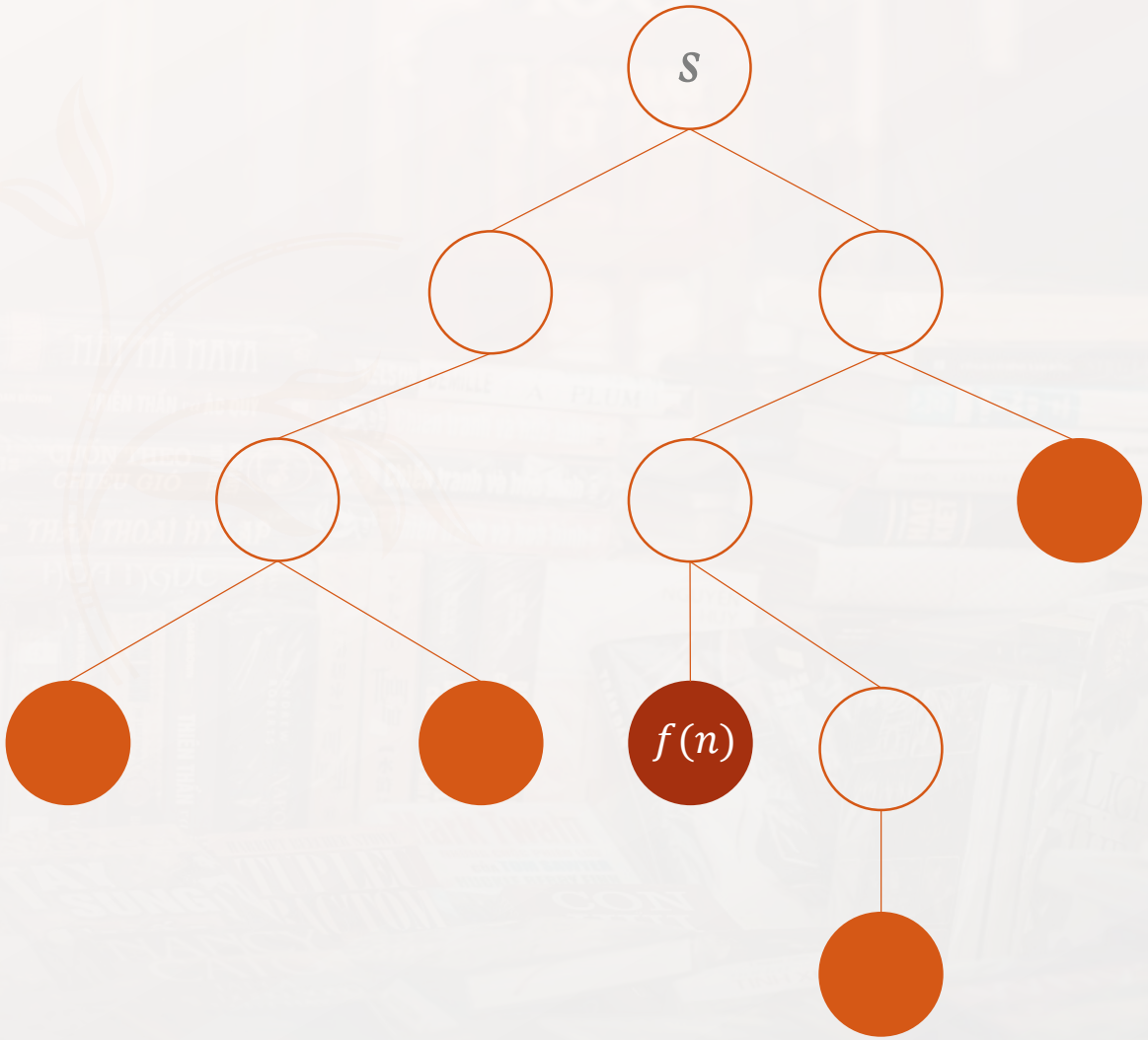
Best-first Search



Best-first Search



Best-first Search



The A Algorithm*

$$f(n) = g(n) + h(n)$$

The A Algorithm*

$$f(n) = g(n) + h(n)$$

$$h(n) \leq h^*(n)$$

The Heuristic

$$h_{prior}(n) = \varepsilon - \mathrm{JS}_{\Delta_n}$$

The Heuristic

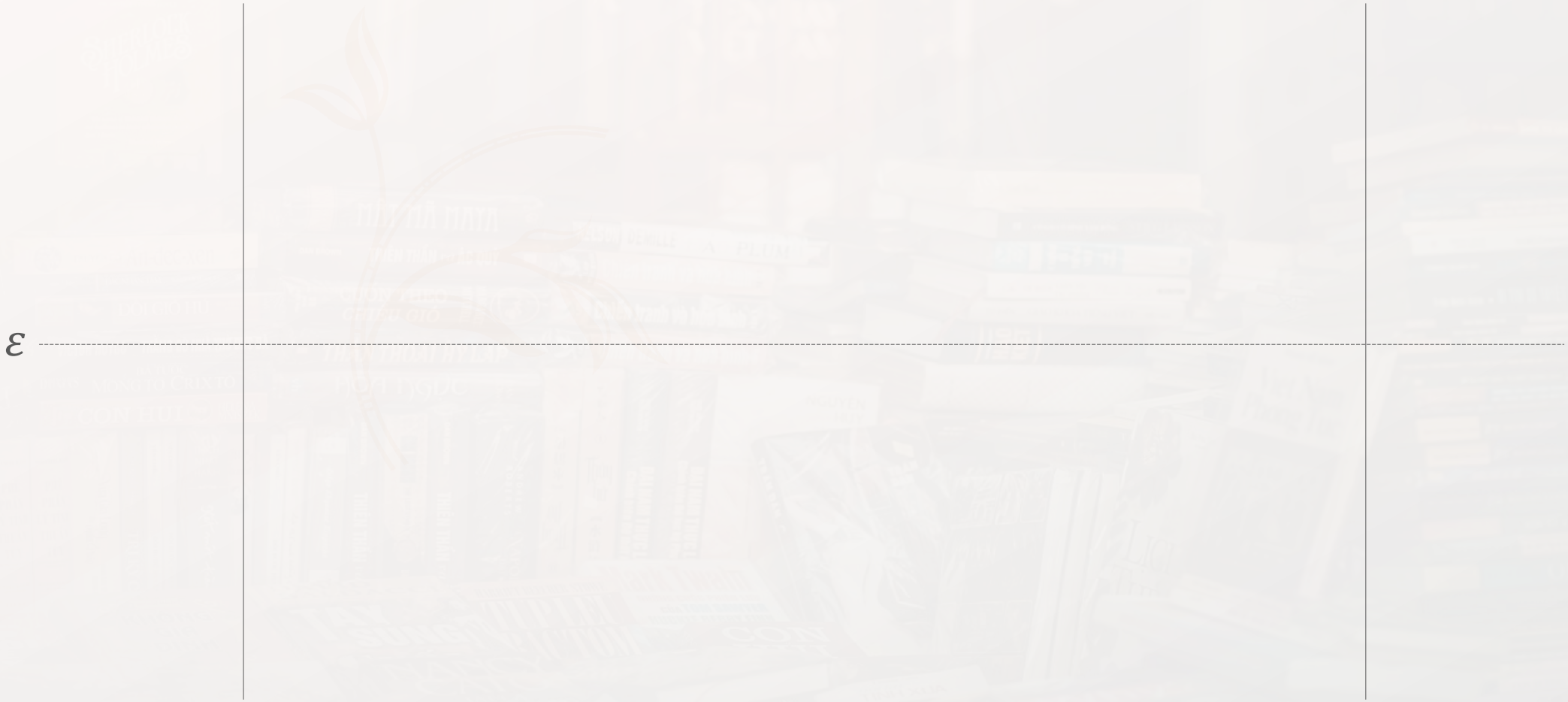
$$h_{prior}(n) = \varepsilon - \mathrm{JS}_{\Delta_n}$$

$$g_{norm}(n) = \frac{g(n)}{\mathrm{JS}_{\Delta_n} - \mathrm{JS}_{\Delta_0}}$$

The Heuristic

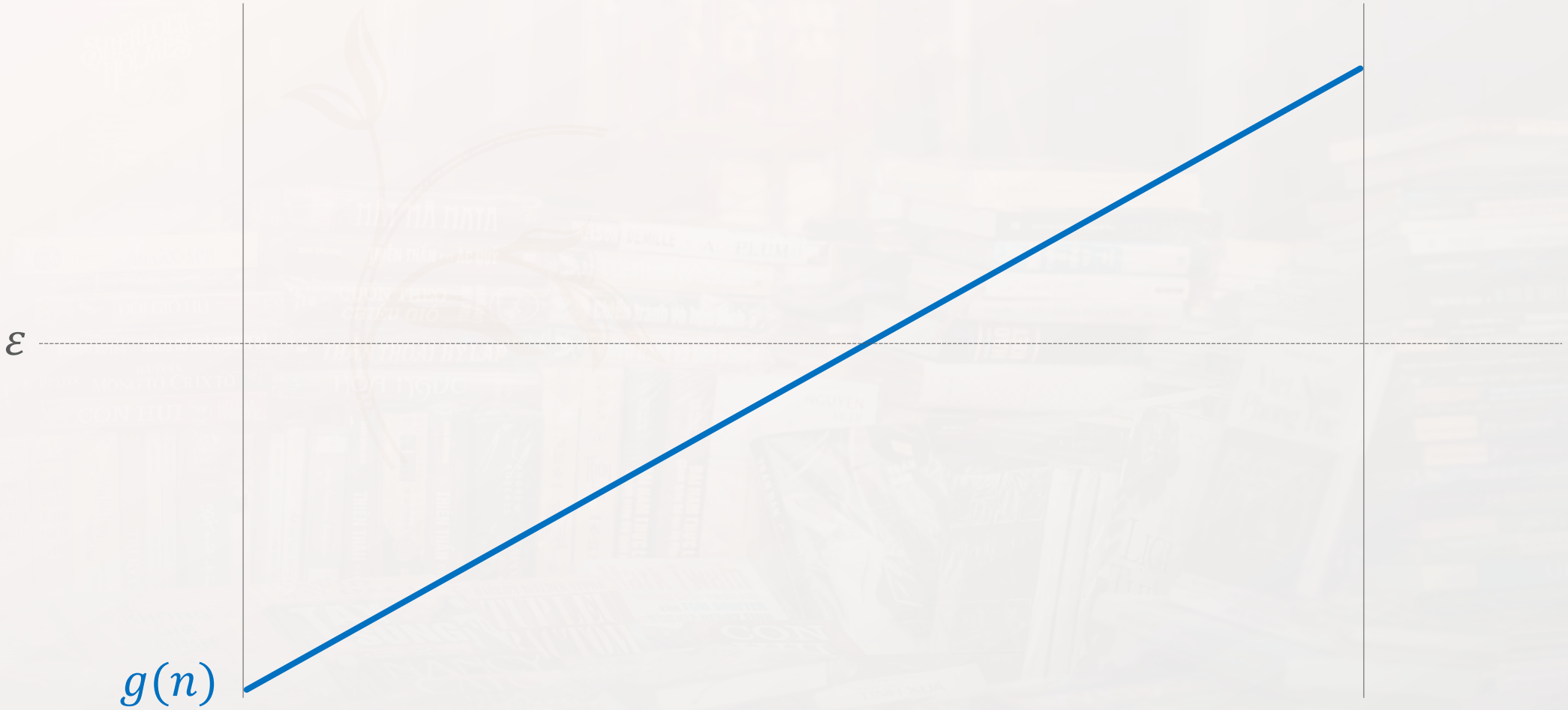
$$h(n) = h_{prior}(n) \cdot g_{norm}(n)$$

The Heuristic



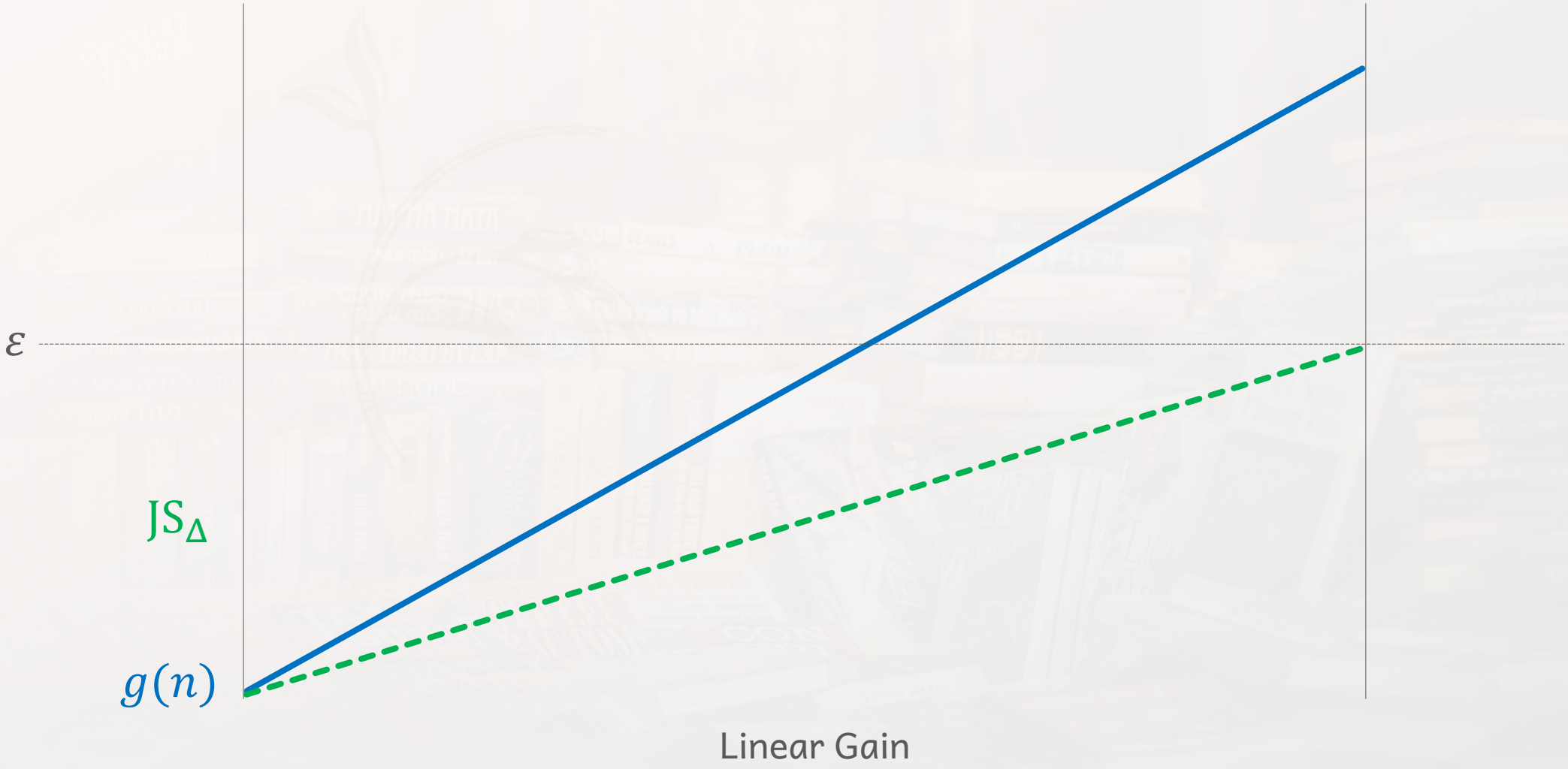
Linear Gain

The Heuristic

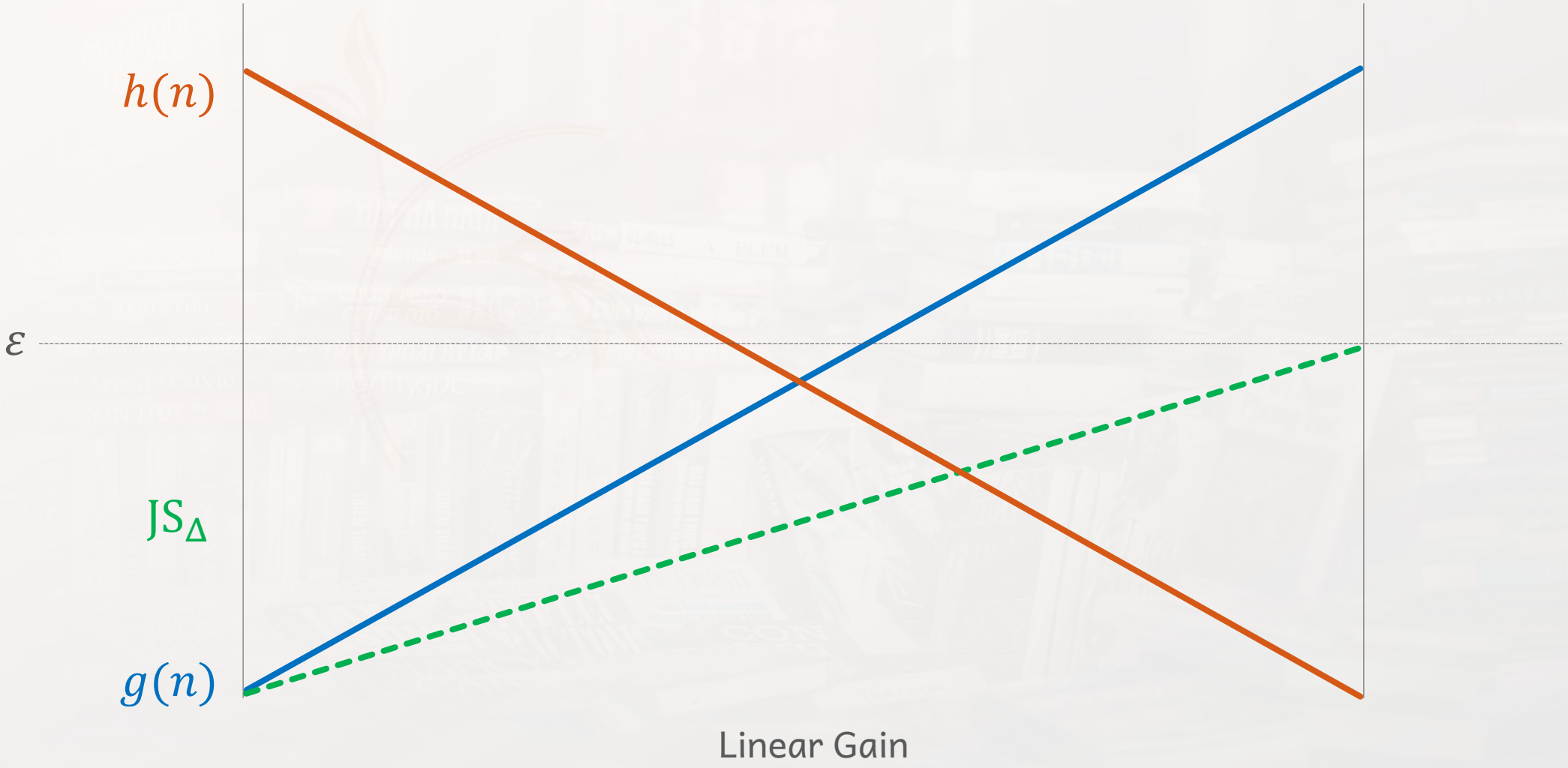


Linear Gain

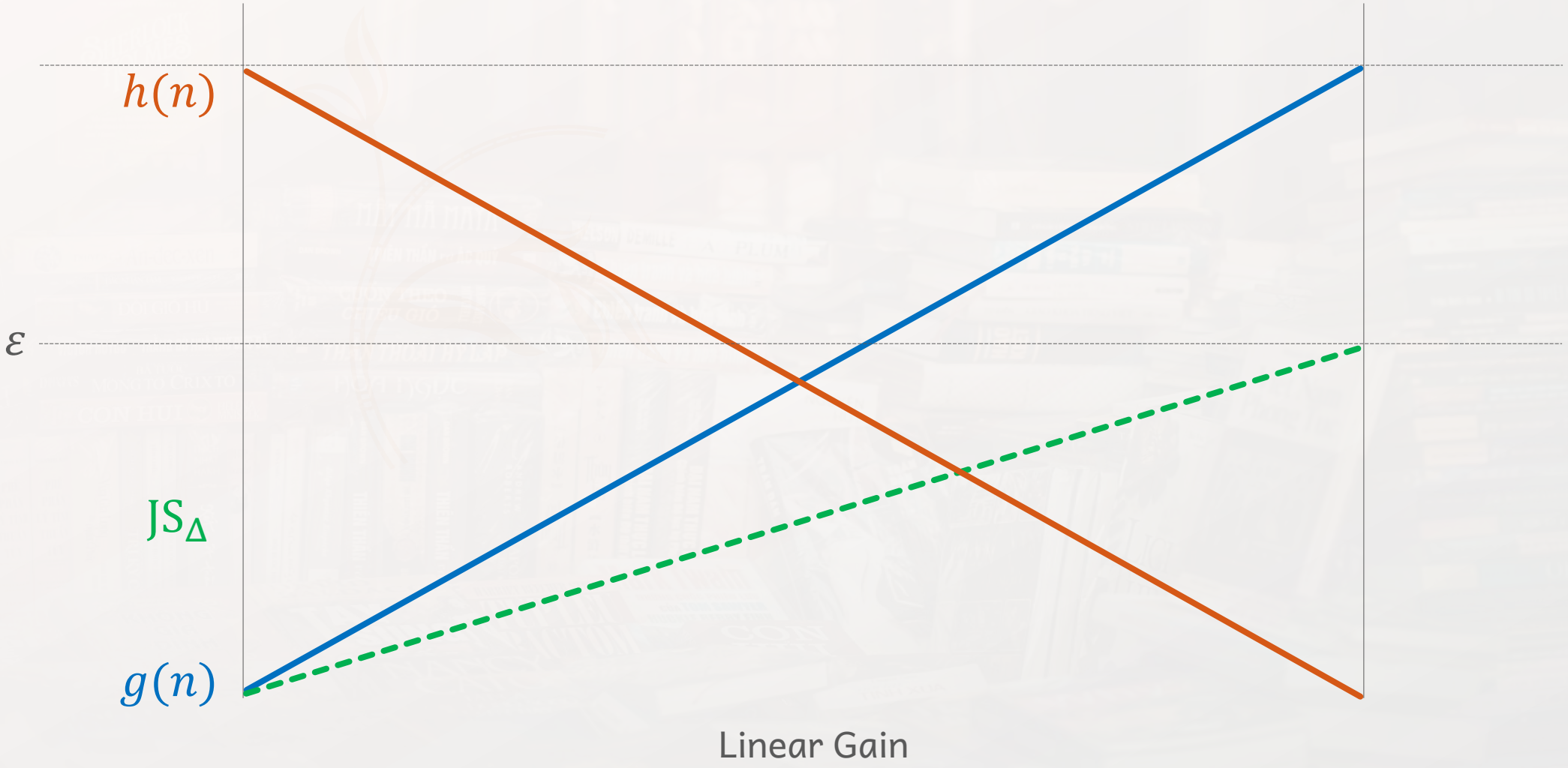
The Heuristic



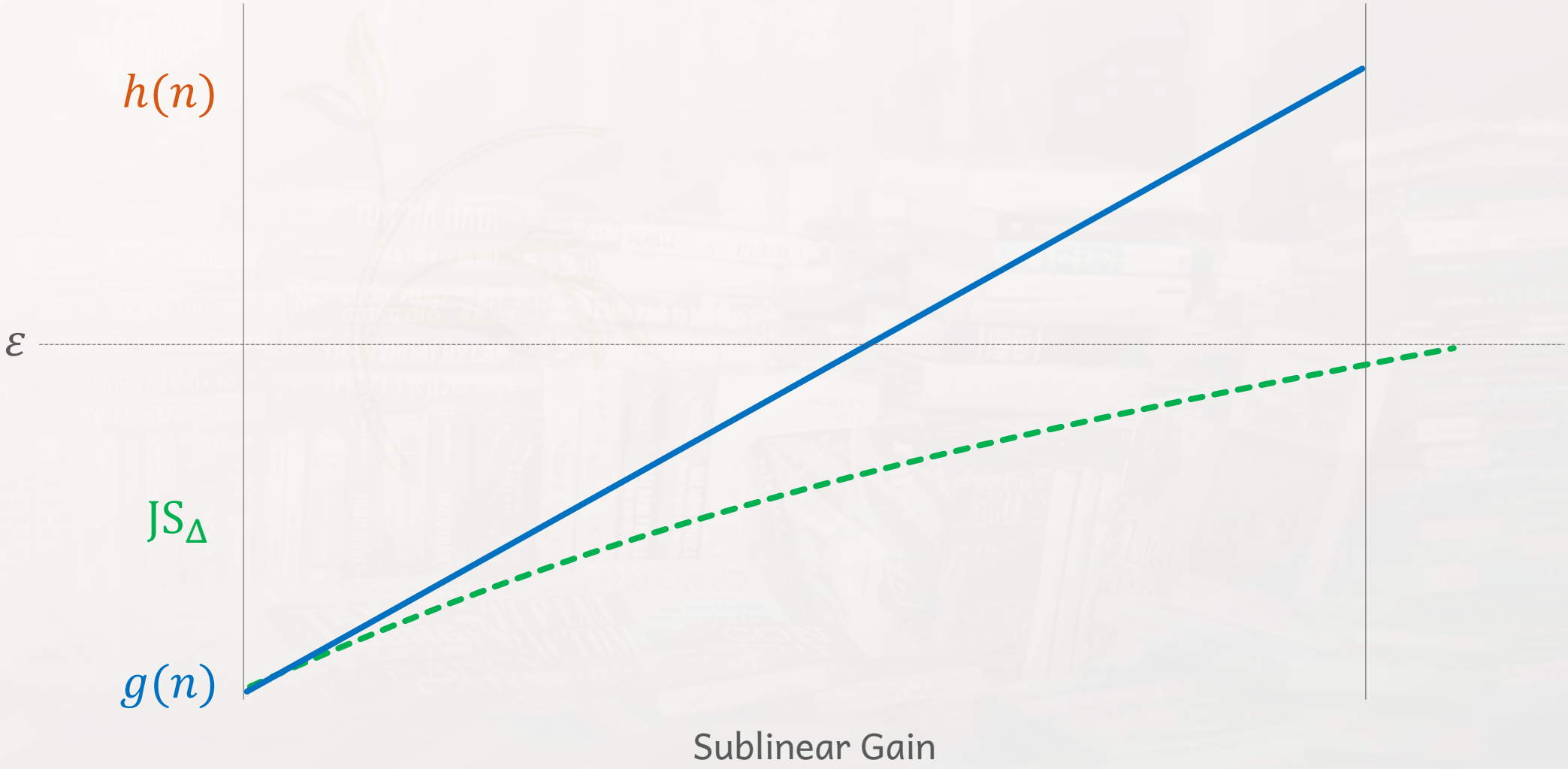
The Heuristic



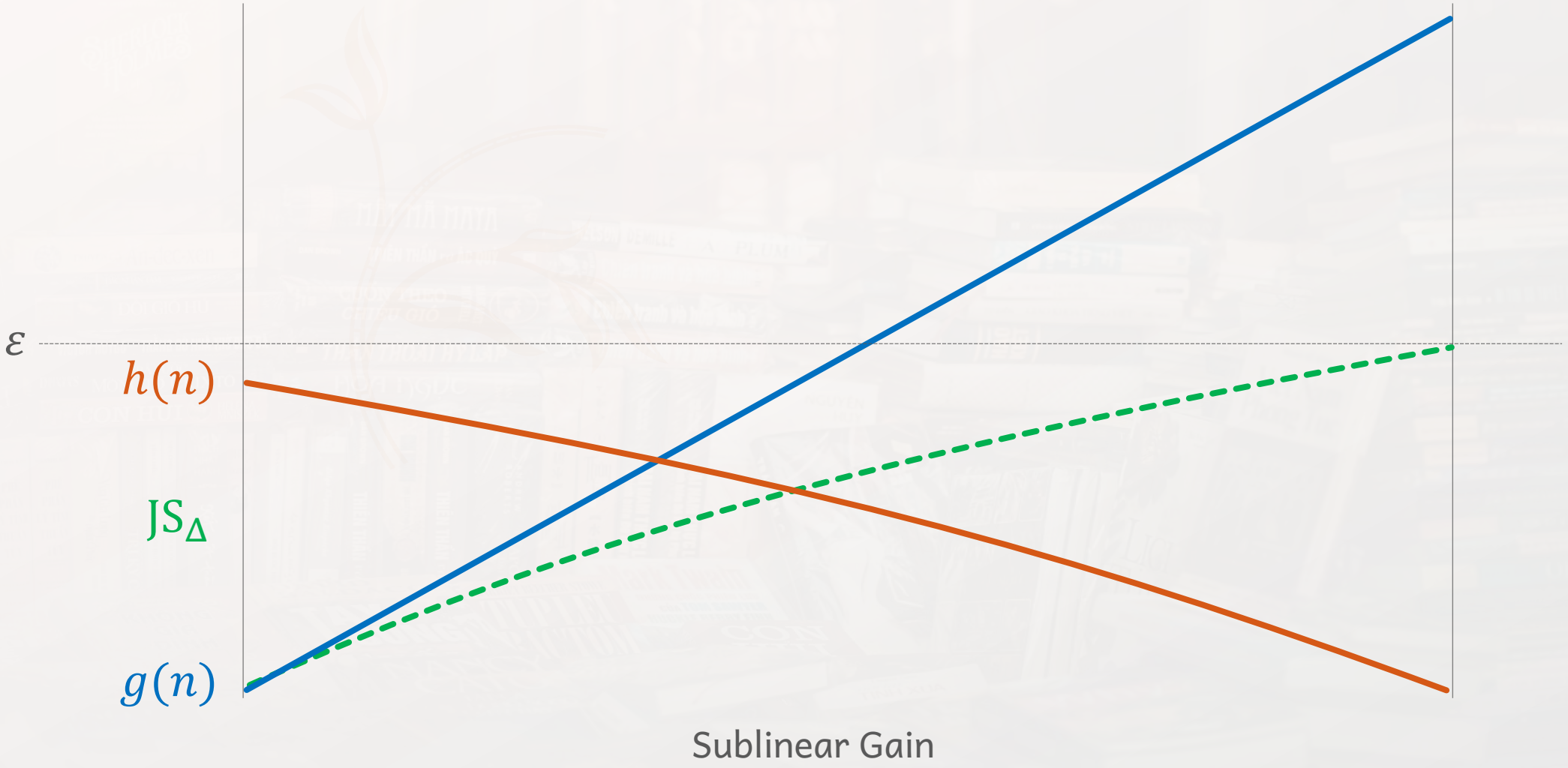
The Heuristic



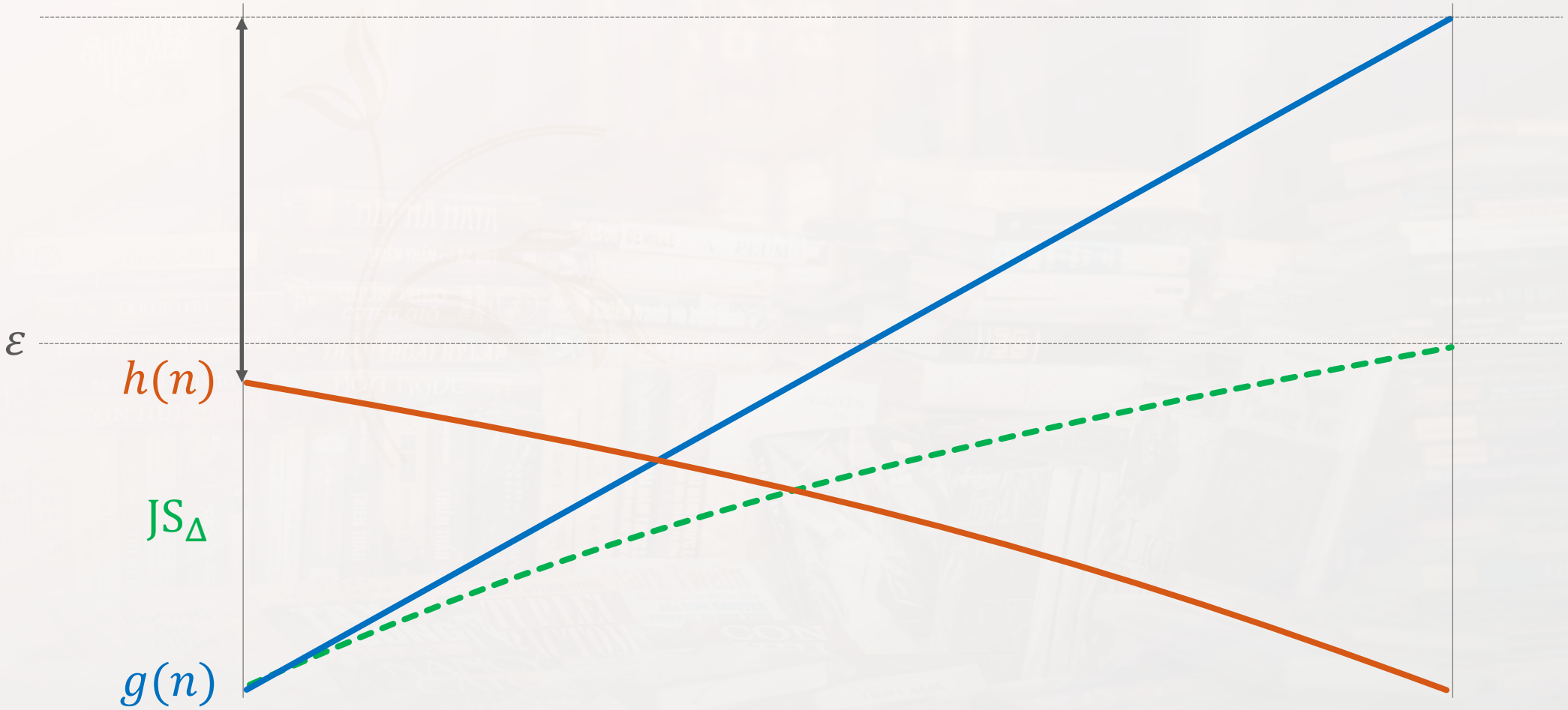
The Heuristic



The Heuristic

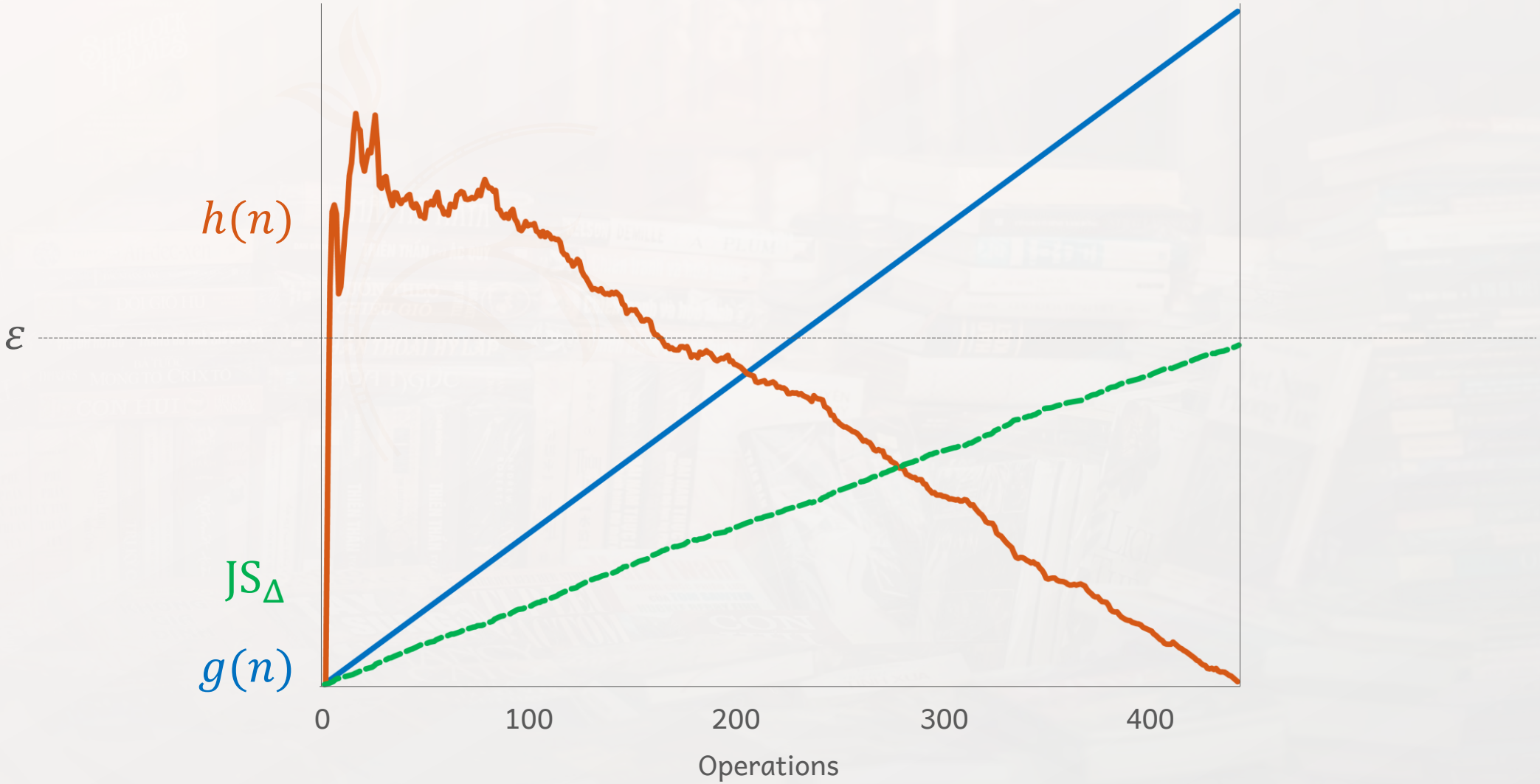


The Heuristic

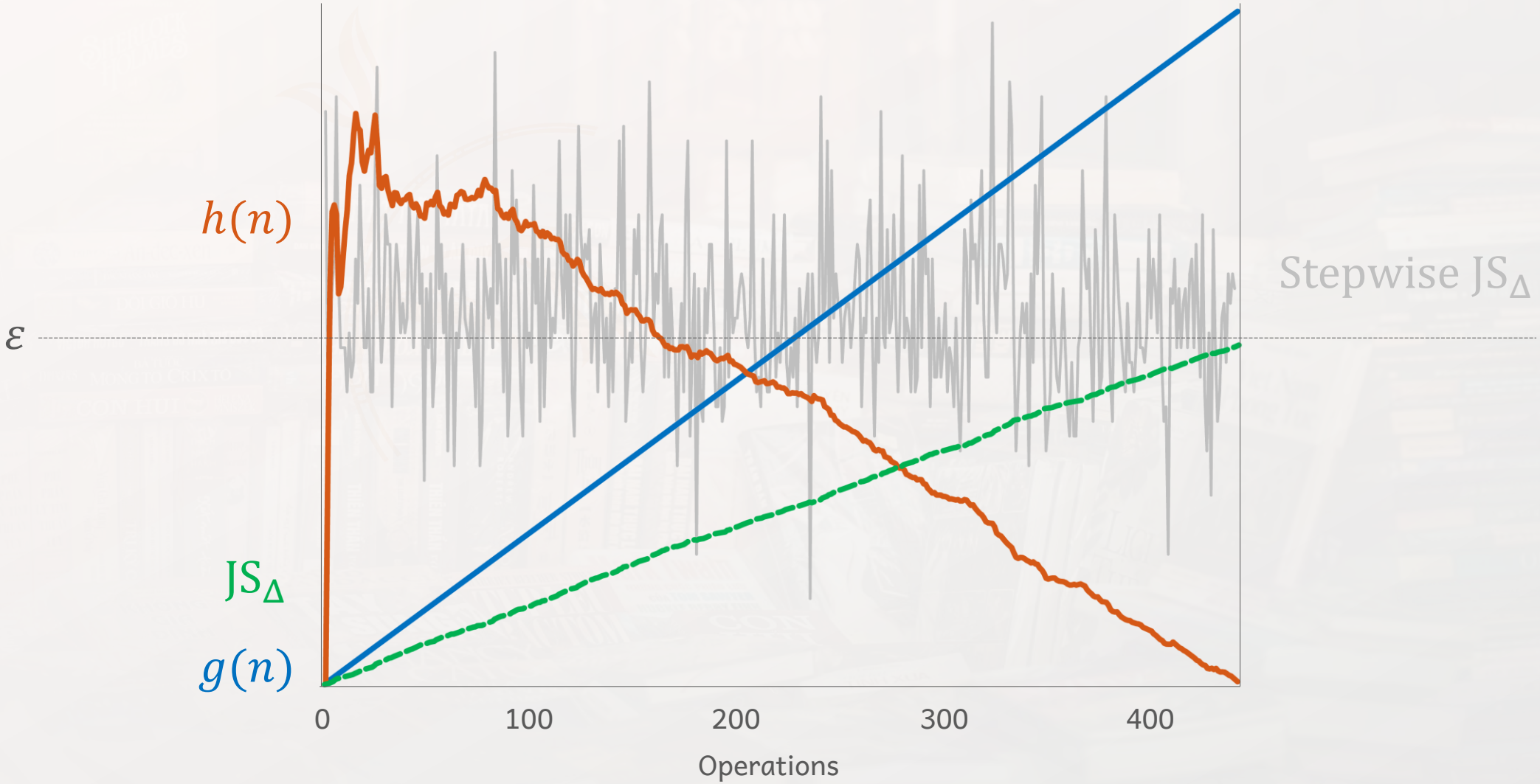


Sublinear Gain

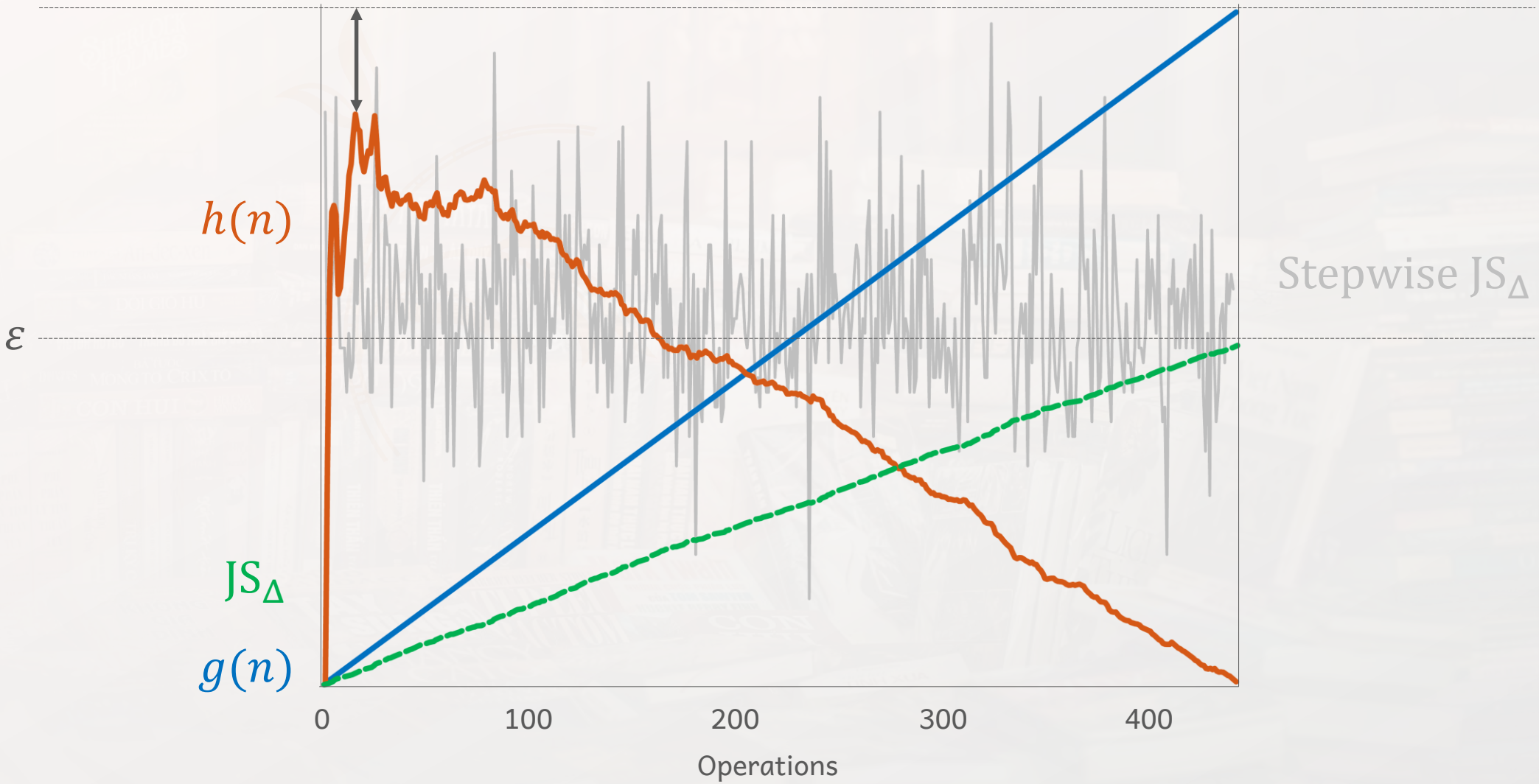
The Heuristic – Actual Example



The Heuristic – Actual Example



The Heuristic – Actual Example



Operators

n-gram removal

Operators

n-gram removal

abfg

Operators

n-gram removal

abfg

character flip

wizard

Operators

n-gram removal

abfg

character flip

wiazrd

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End.

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End!

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End!

house

synonym

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End!

synonym

home

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End!

home

synonym

author

Netspeak

Operators

n-gram removal

abfg

character flip

wiazrd

character map

The End!

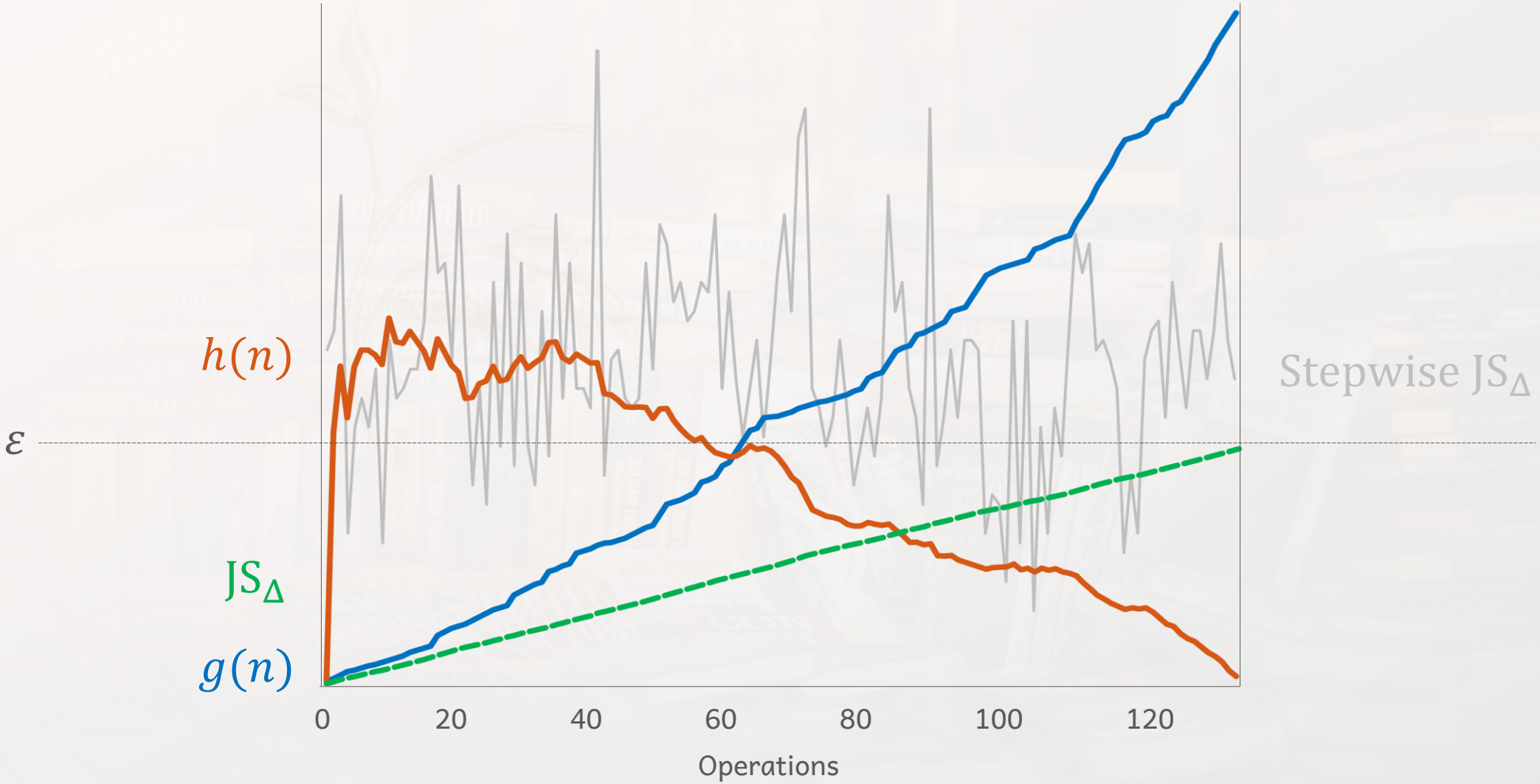
home

synonym

author of

Netspeak

Obfuscation Using Heuristic Search



Obfuscation Using Heuristic Search

> ‘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 the 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 the 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. […]’

A Filbert Is a Nut by Rick Raphael

Obfuscation Using Heuristic Search

> ‘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 the 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 accomplishedb y aehc patient. […]’

A Filbert Is a Nut by Rick Raphael

Obfuscation Using Heuristic Search

> ‘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 the 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 accomplishedb y aehc patient. […]’

A Filbert Is a Nut by Rick Raphael

Obfuscation Using Heuristic Search

> ‘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 the 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 accomplishedb y aehc patient. […]’

A Filbert Is a Nut by Rick Raphael

Summary

- Unmasking can be attacked by KLD obfuscation,

Summary

- Unmasking can be attacked by KLD obfuscation,
- JS_{Δ} is an effective authorship metric,

Summary

- Unmasking can be attacked by KLD obfuscation,
- JS_{Δ} is an effective authorship metric,
  - important building block: length-invariant thresholds.

Summary

- Unmasking can be attacked by KLD obfuscation,
- JS_{Δ} is an effective authorship metric,
  - important building block: length-invariant thresholds.
- Design of an admissible heuristic search function:

Summary

- Unmasking can be attacked by KLD obfuscation,
- JS_{Δ} is an effective authorship metric,
  - important building block: length-invariant thresholds.
- Design of an admissible heuristic search function:
  - significant reduction of text modifications at the same effect,

Summary

- Unmasking can be attacked by KLD obfuscation,
- JS_{Δ} is an effective authorship metric,
  - important building block: length-invariant thresholds.
- Design of an admissible heuristic search function:
  - significant reduction of text modifications at the same effect,
  - better text quality.

Thank you

for your attention

Image Credits: Min An, Pexels.com