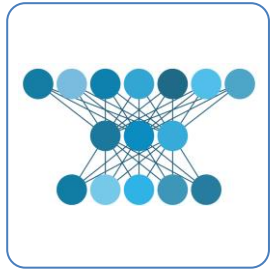
Phil Weber

Justifying AI in court:
Human or machine analysis of evidence?



Aston Centre for Artificial Intelligence Research and Applications (**ACAIRA**)

Aston Institute for Forensic Linguistics (**AIFL**)

- → **Forensic Data Science Laboratory**
- → Forensic Speech Science Laboratory

http://forensic-data-science.net/

Q1: Who do you want in your corner?

AI?

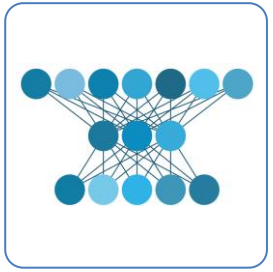
or the

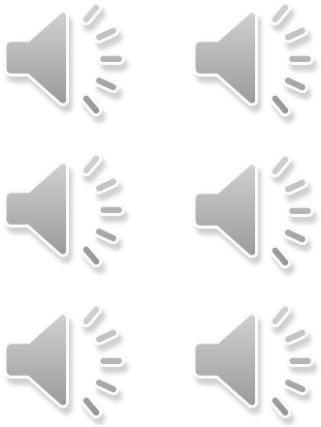
Human expert?

John Morgan, Public domain, via Wikimedia Commons

Q2: Same speaker or different speaker?

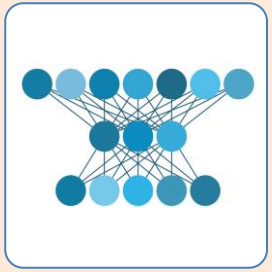
How good are you
at assessing speakers from audio?





1. Forensic data science 101

→ let's use data!



2. Artificial intelligence

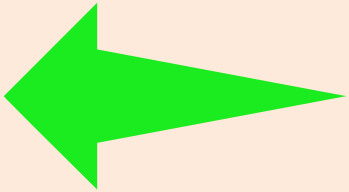
→ Aston's E^3FS^3 forensic voice comparison system

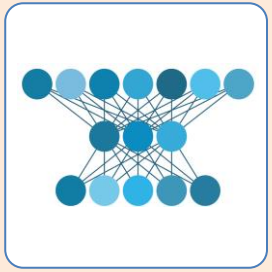
3. Human intelligence

→ but surely humans can do that?

1. Forensic data science 101

→ let's use data!





2. Artificial intelligence

→ Aston's E^3FS^3 forensic voice comparison system

3. Human intelligence

→ but surely humans can do that?

What is forensic data science?

Question for the court:

Is the defendant guilty?
Is the defendant the speaker on both recordings?

Questioned-speaker recordings

e.g. crime scene

specific “population”
specific recording conditions

Known-speaker recordings

i.e. the suspect

different conditions?

What is forensic data science?

Moving away from reliance on experts

to methods based on

- **relevant** data
- quantitative measurements
- statistical models / machine learning

transparent & reproducible

resistant to “cognitive bias”

logically correct framework for interpreting evidence:
the **likelihood ratio**

empirically validated under casework conditions

What is a likelihood ratio?

"How much more likely is it that we would observe the features of the questioned-speaker recordings Q and known-speaker recordings K"

if they were made by the **same** speaker (**randomly selected from the relevant population**)

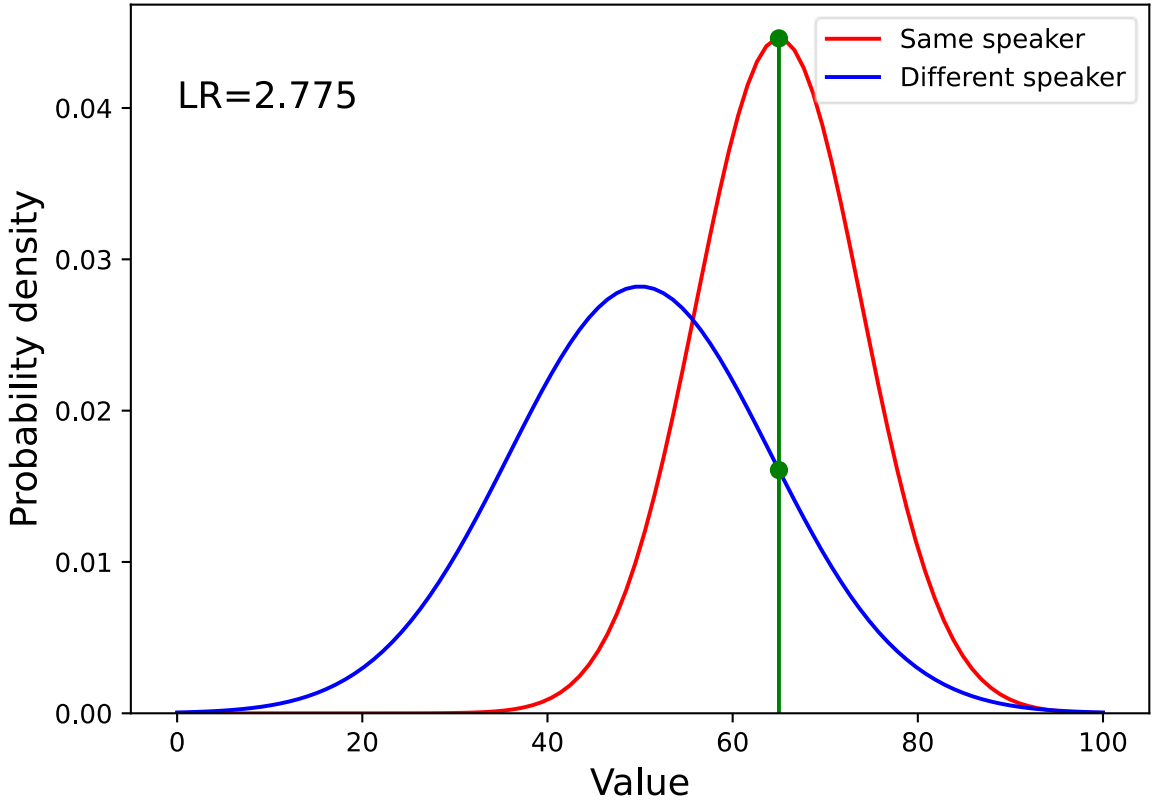
versus

if they were made by **different** speakers (**randomly selected from the relevant population**)?

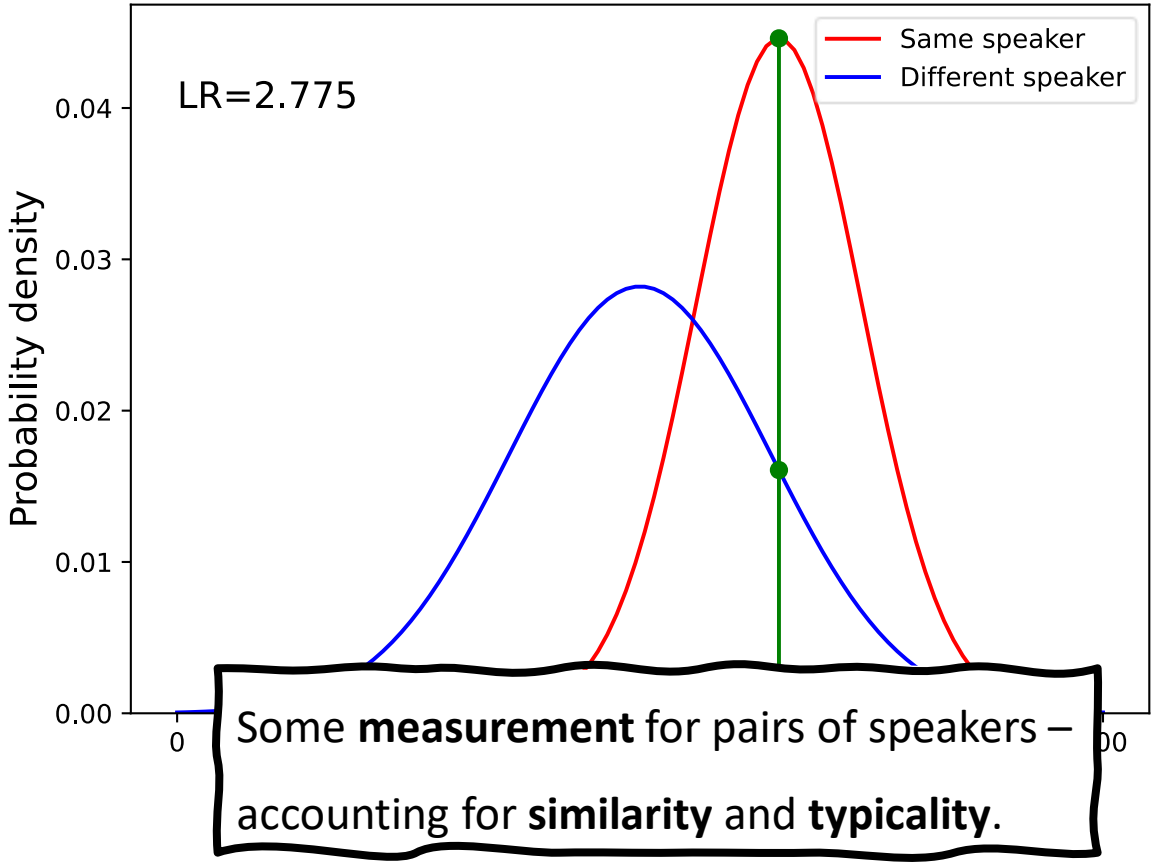
$$\frac{\Pr(hypothesis_1 \mid evidence)}{\Pr(hypothesis_2 \mid evidence)} = \frac{\Pr(hypothesis_1)}{\Pr(hypothesis_2)} \times \frac{\Pr(evidence \mid hypothesis_1)}{\Pr(evidence \mid hypothesis_2)}$$

Posterior odds = **Prior odds** × **Likelihood ratio**

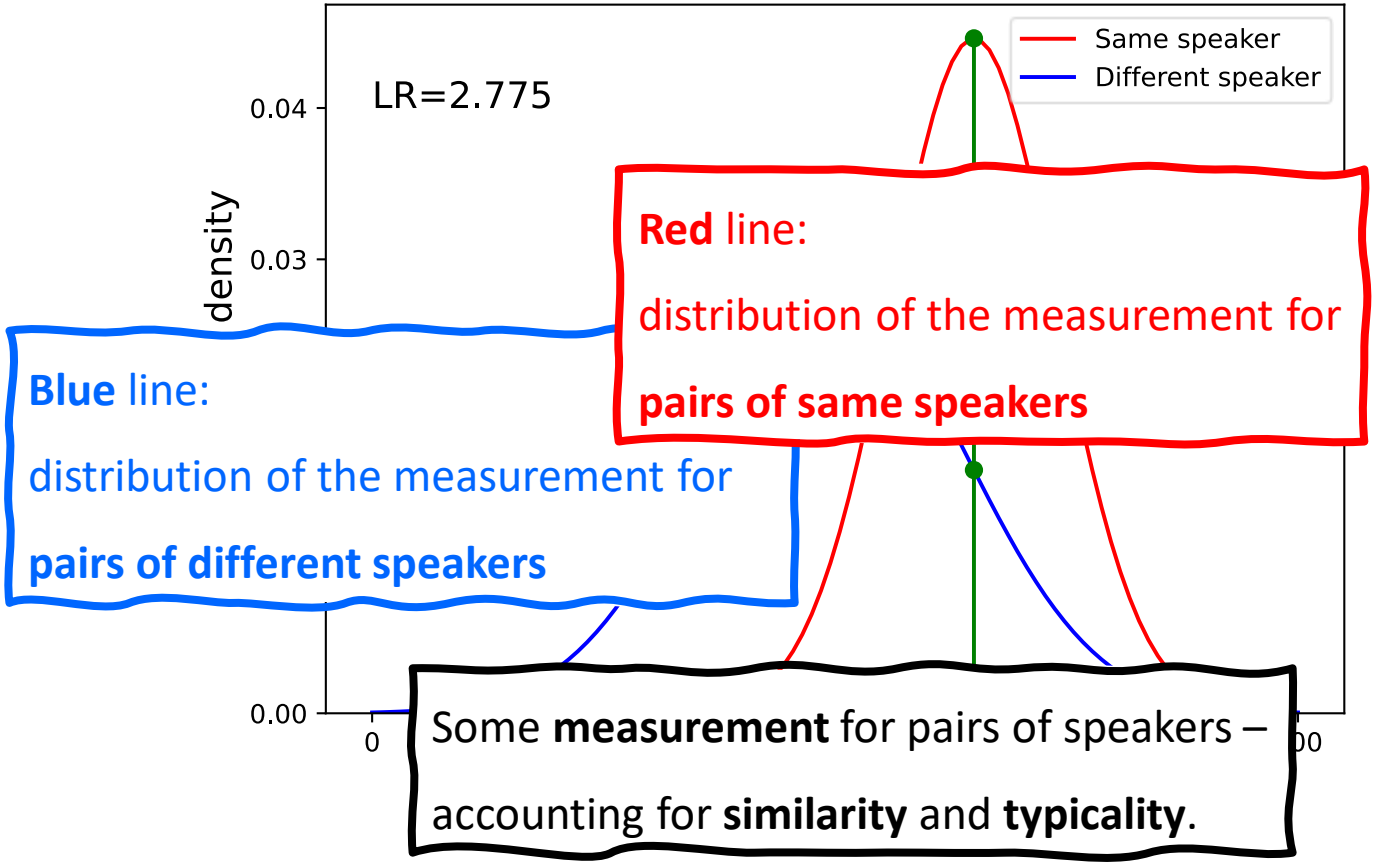
What is a likelihood ratio?



What is a likelihood ratio?



What is a likelihood ratio?

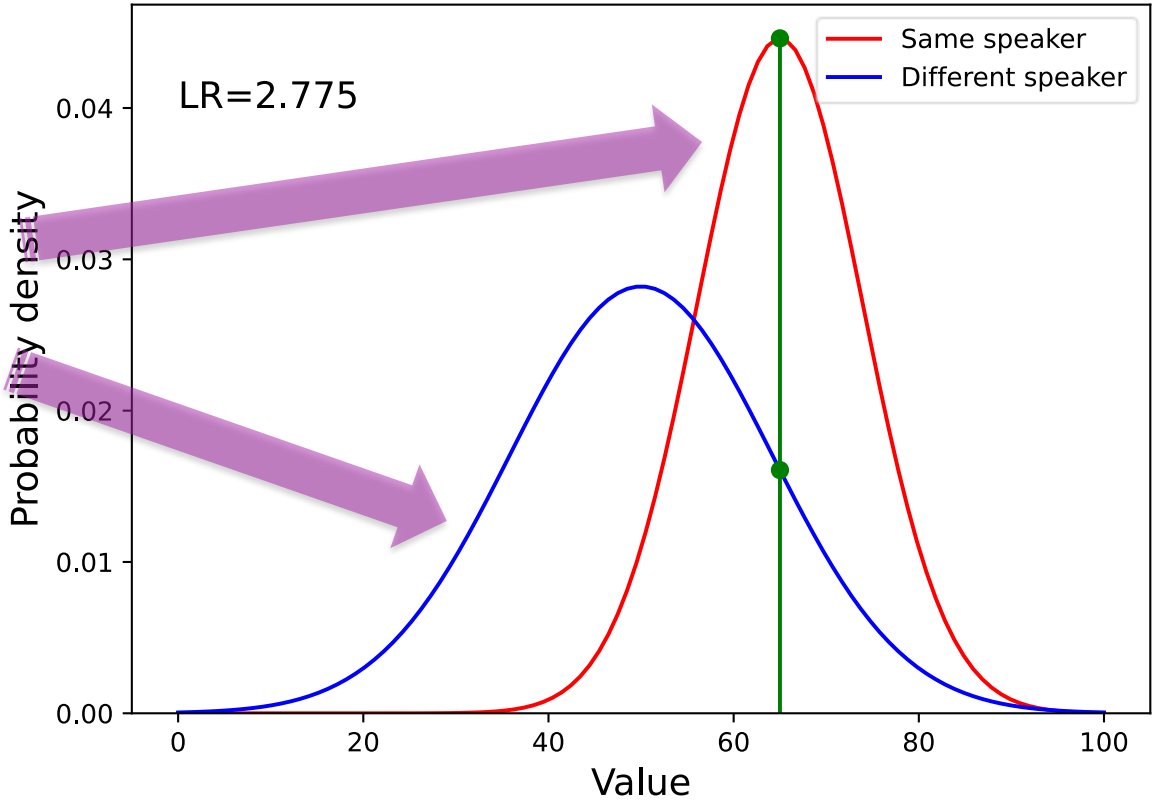


What is a likelihood ratio?

But these must be **specific**:

- Speakers from the relevant **population**
- Recordings in the **conditions** of the **case**

Note the difference in variability



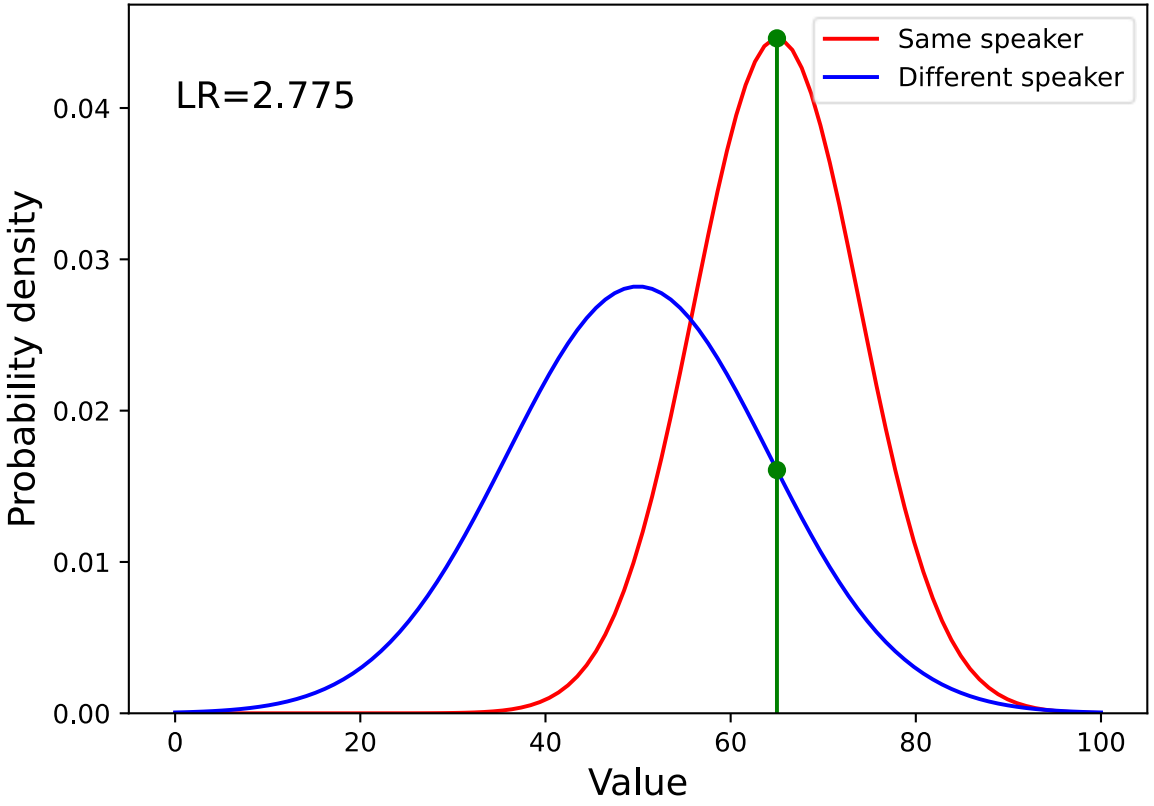
What is a likelihood ratio?

Bigger than one : $LR \in (1, \infty]$

Weight of evidence points

towards same-speaker

Log LR > 0.



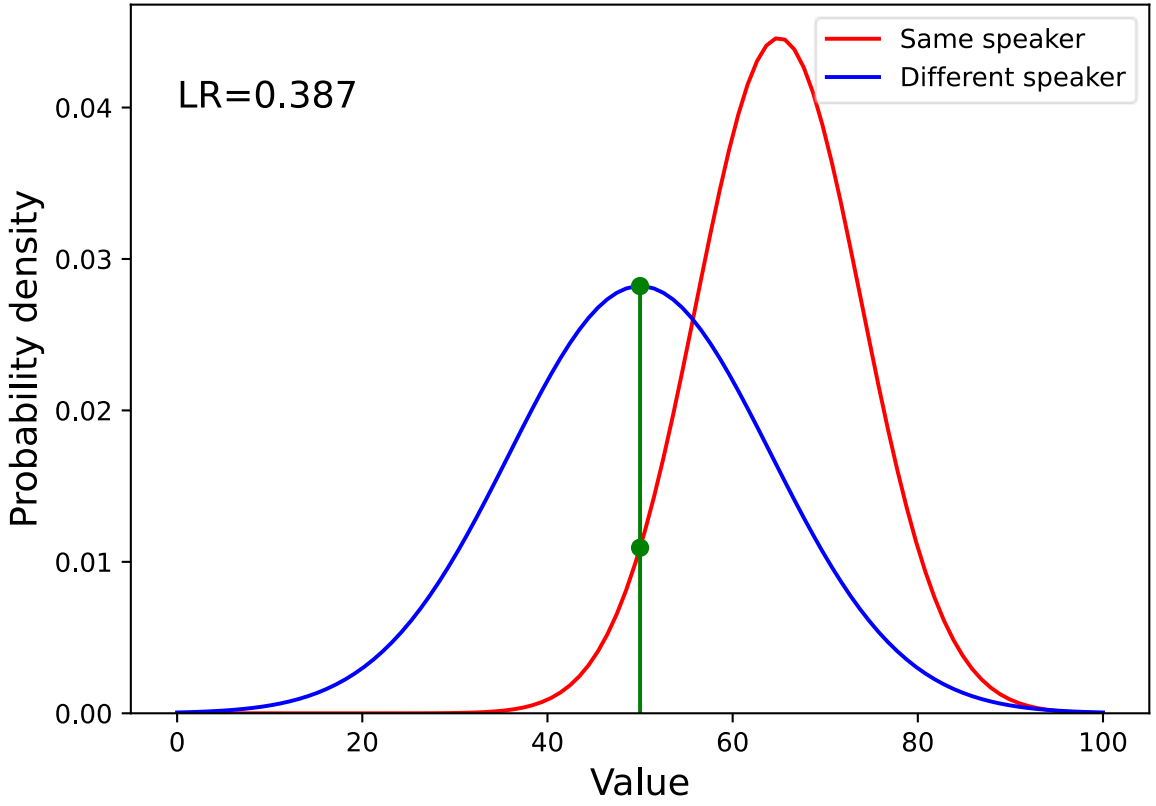
What is a likelihood ratio?

Less than one : $LR \in [0, 1)$

Weight of evidence points

towards different-speaker

Log LR < 0.

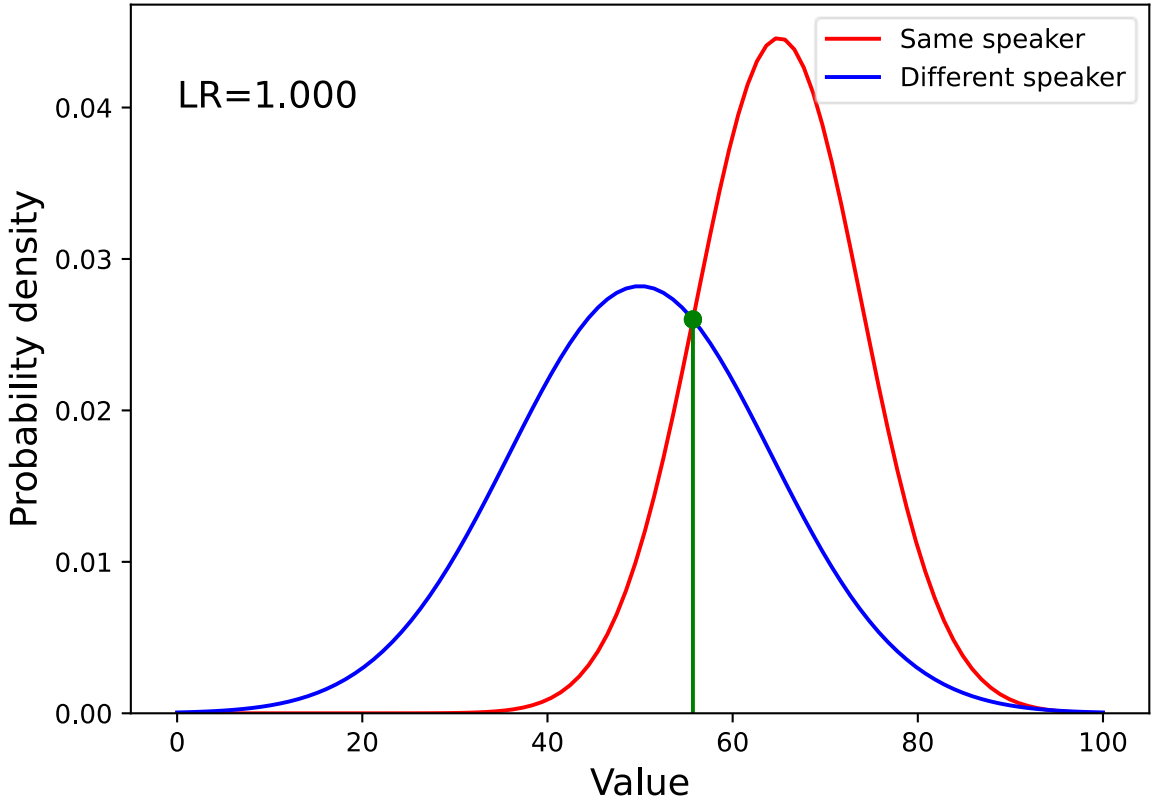


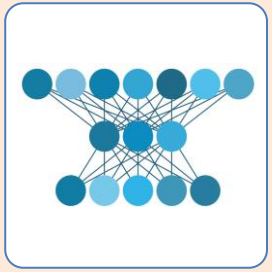
What is a likelihood ratio?

Less than one : $LR = 1$

No conclusive evidence

Log LR = 0.





1. Forensic data science 101

→ let’s use data!

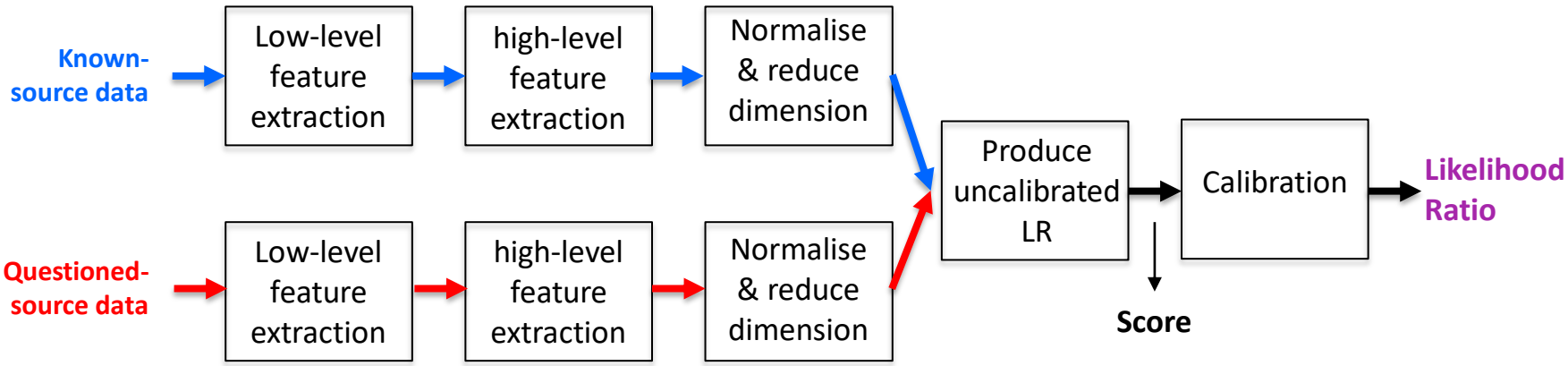
2. Artificial intelligence

→ Aston’s E^{3}FS^3 forensic voice comparison system

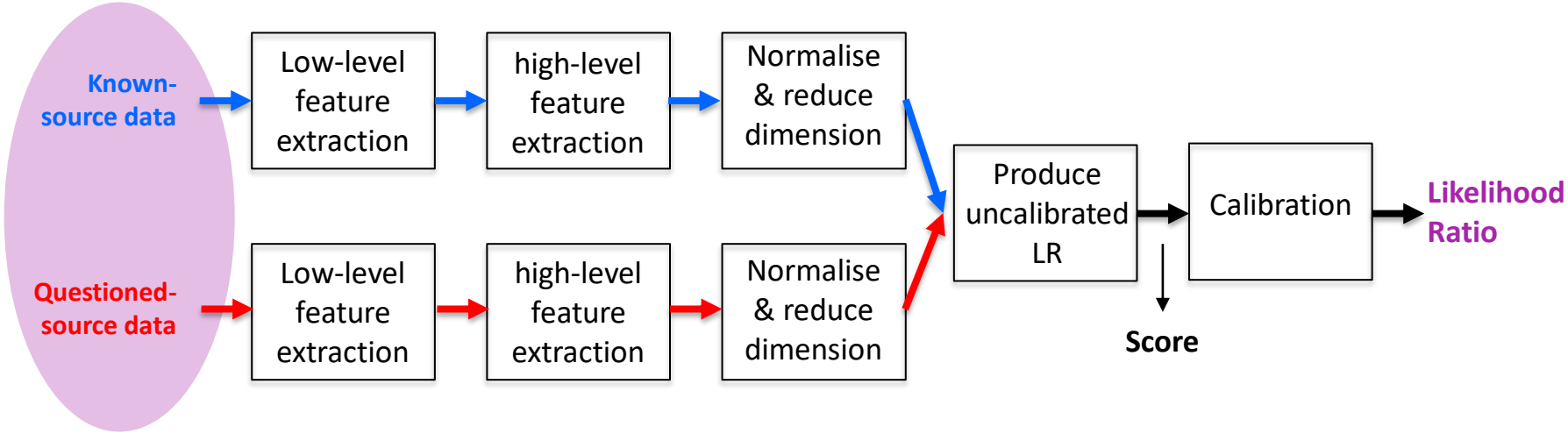
3. Human intelligence

→ but surely humans can do that?

AI for forensic voice comparison

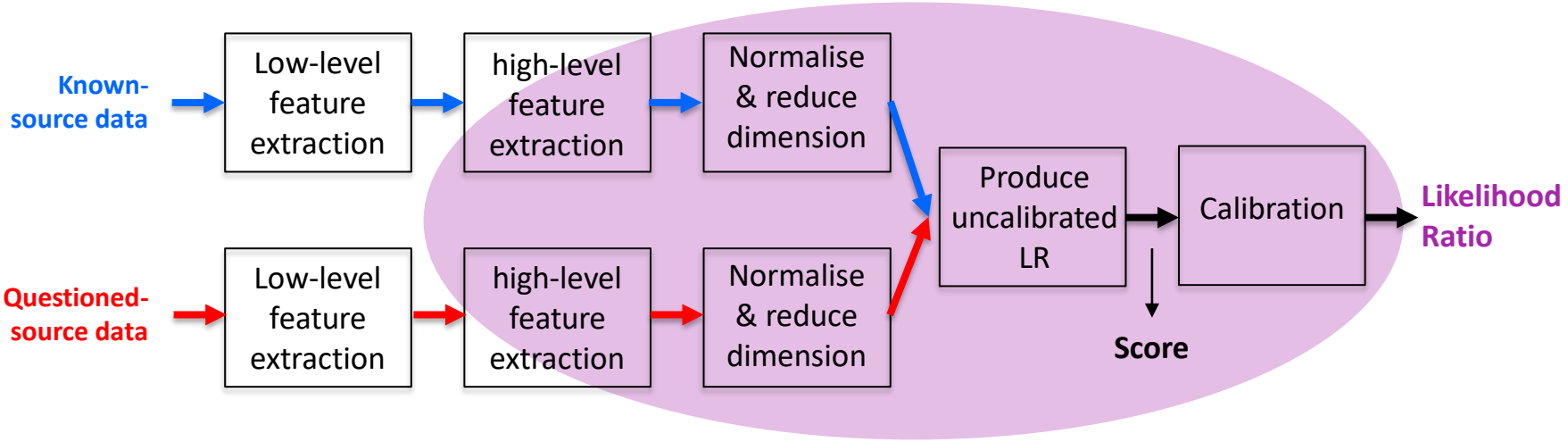


AI for forensic voice comparison



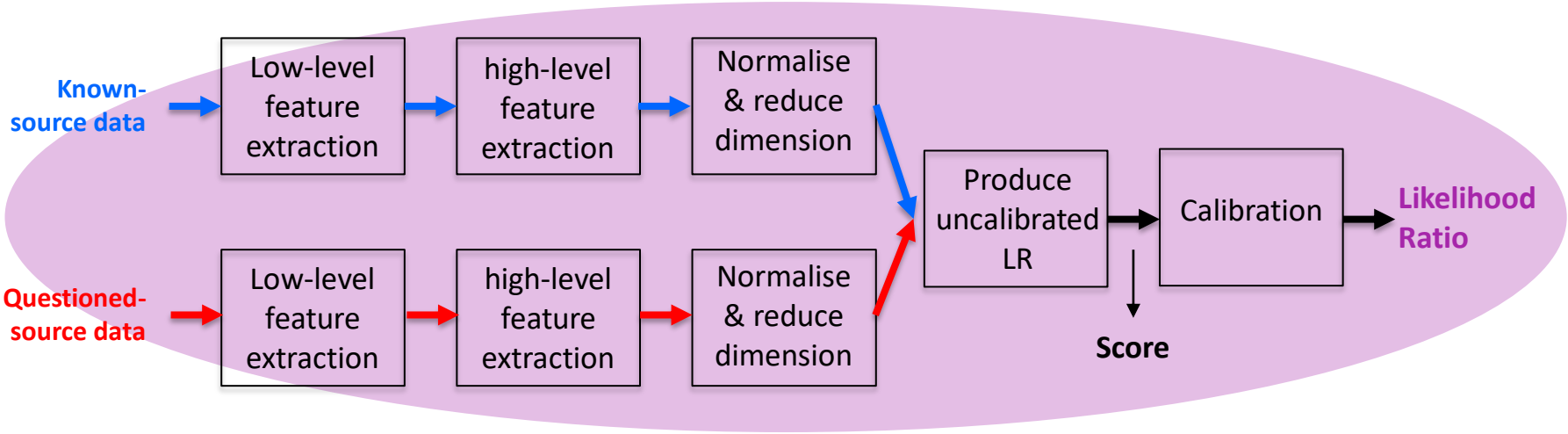
Understand
the case data

AI for forensic voice comparison



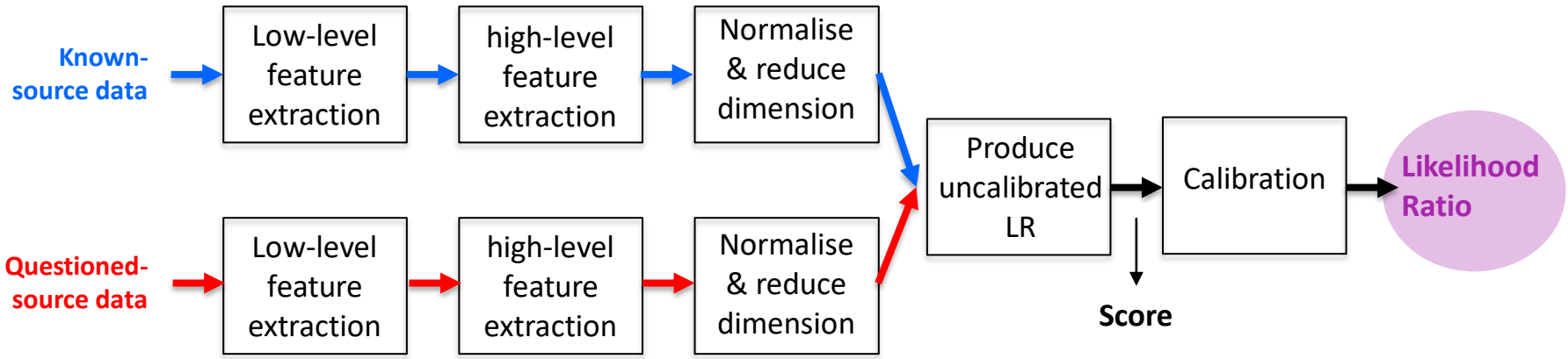
Collect data and train models

AI for forensic voice comparison



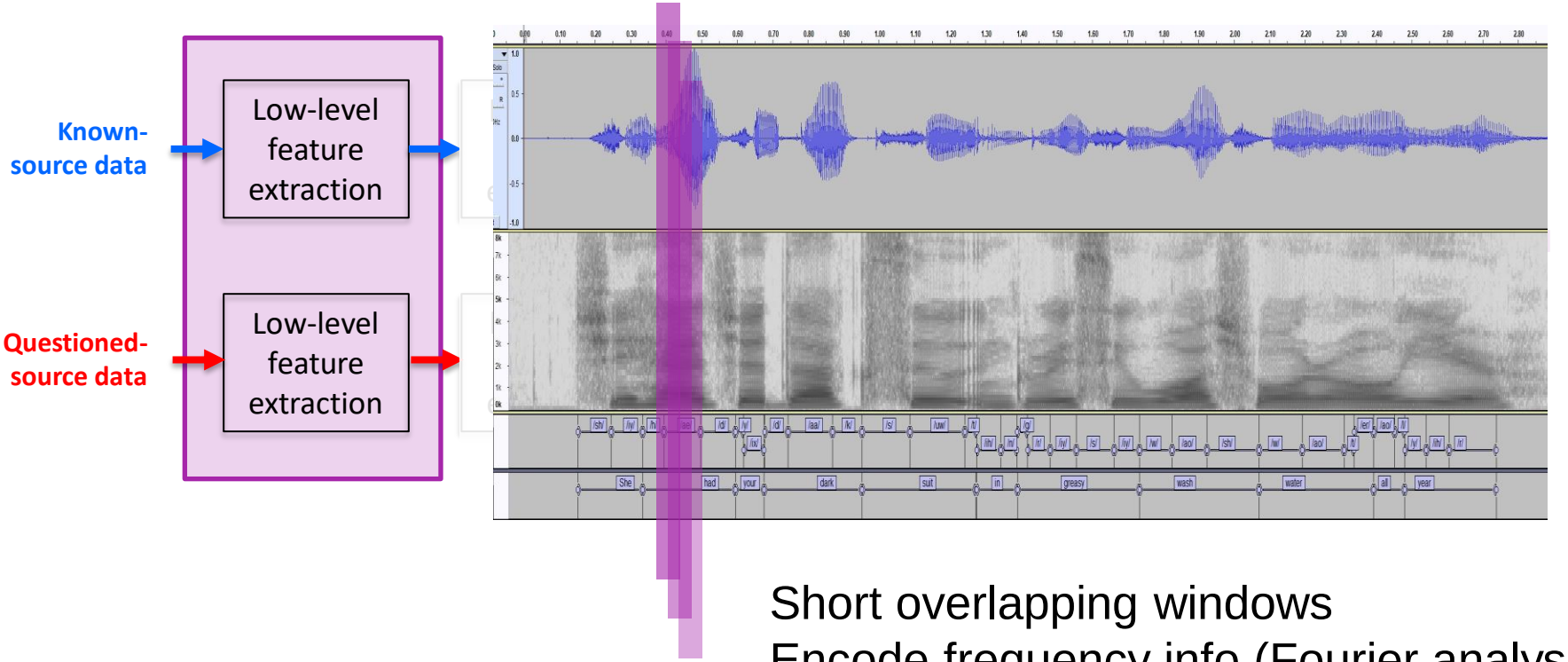
Collect (more) data and
validate the system

AI for forensic voice comparison



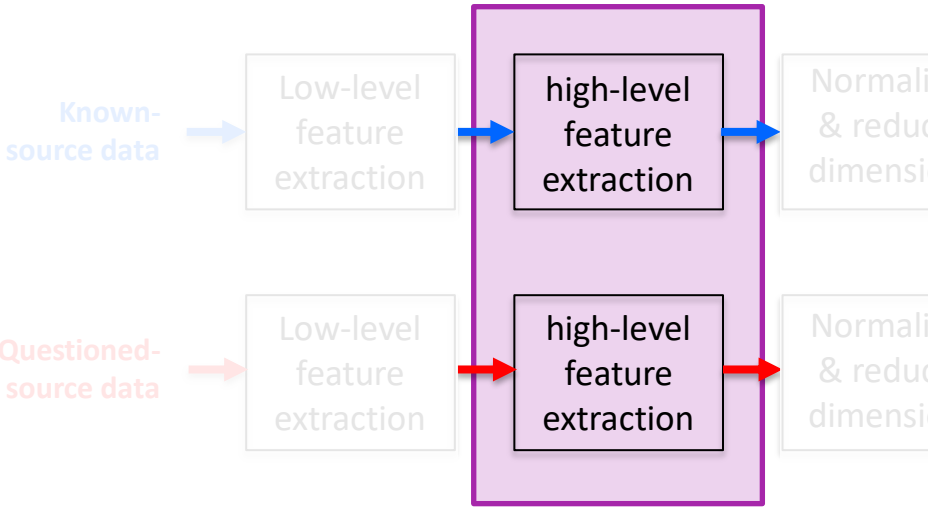
Report the likelihood ratio for the case

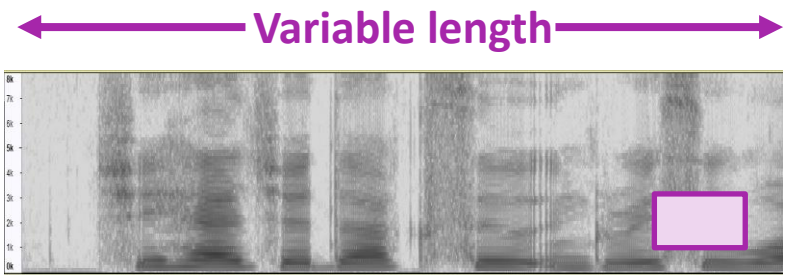
AI for forensic voice comparison

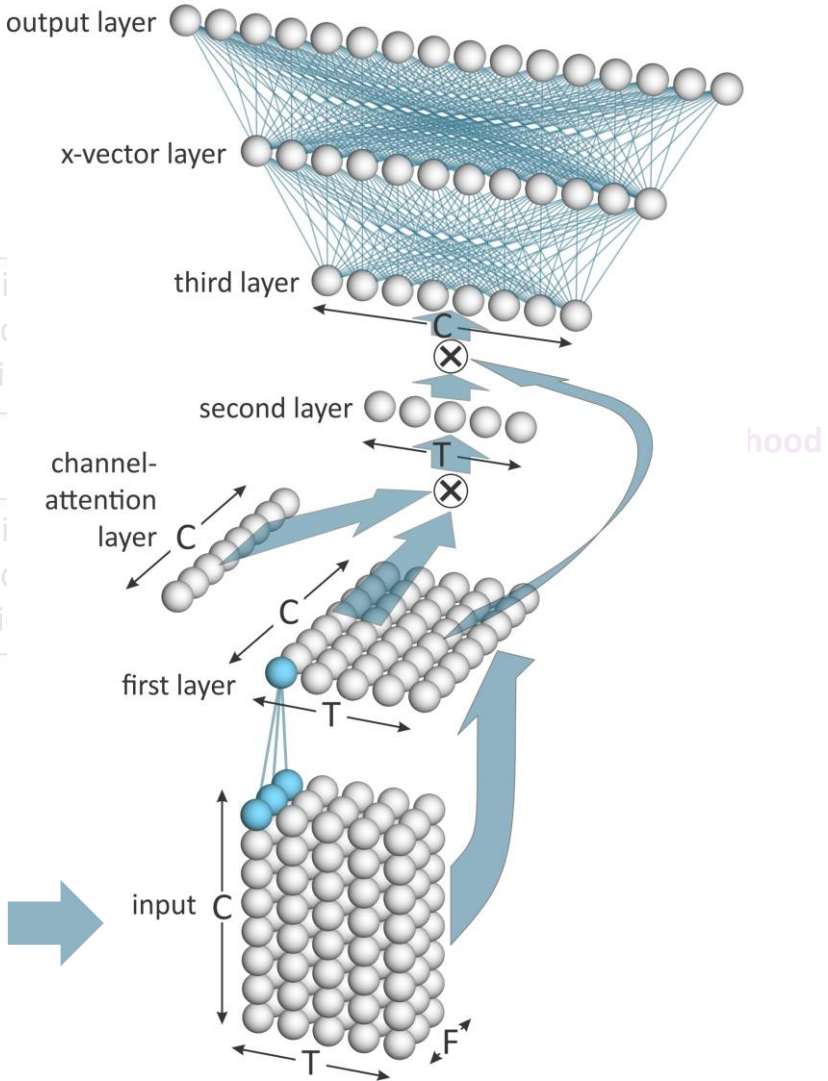


Short overlapping windows
Encode frequency info (Fourier analysis)

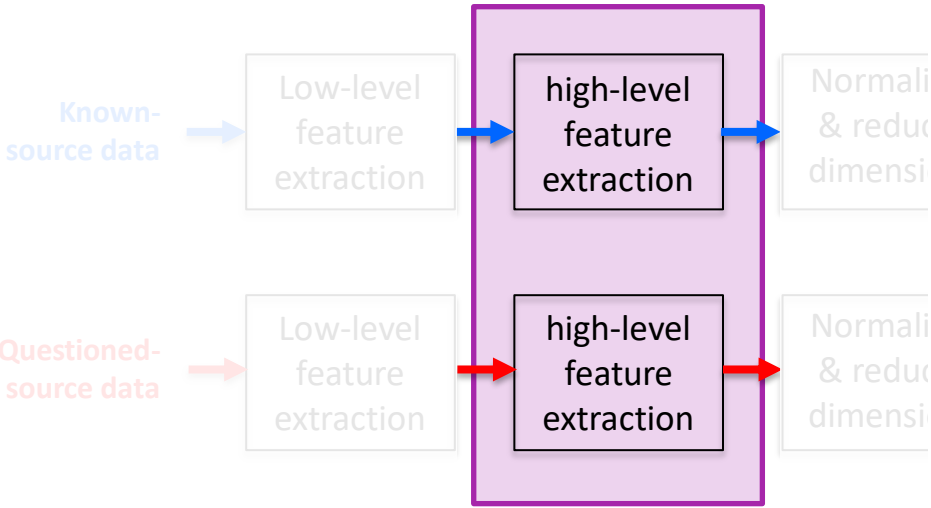
AI for forensic …

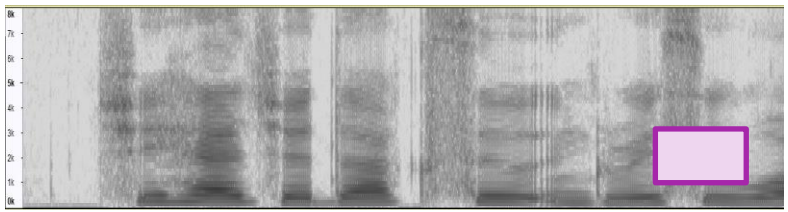


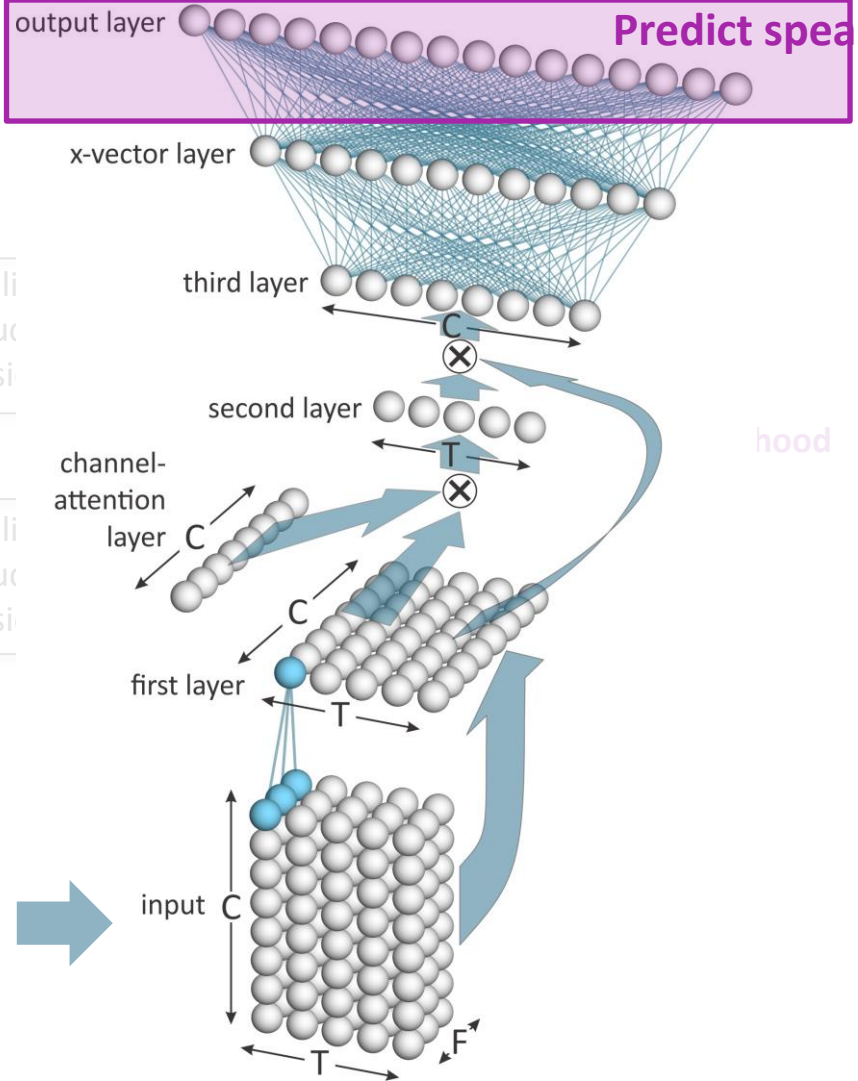




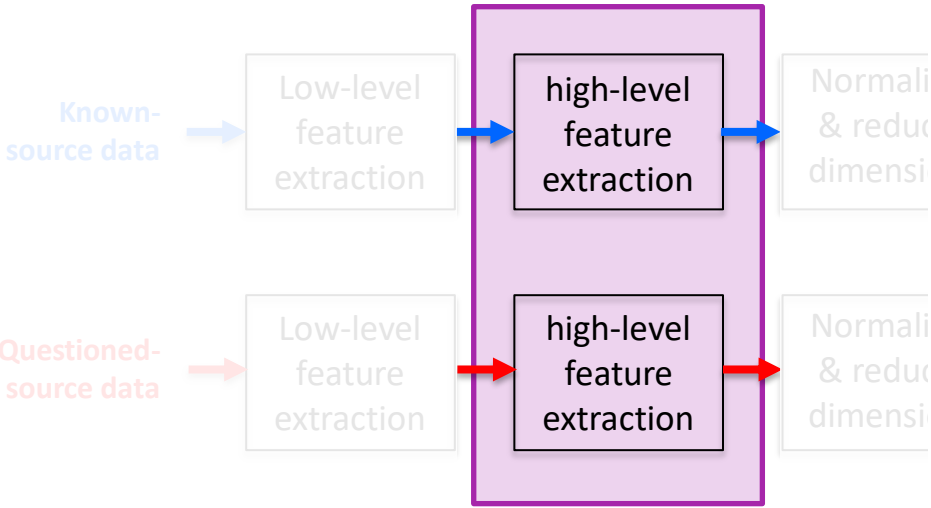
AI for forensic …

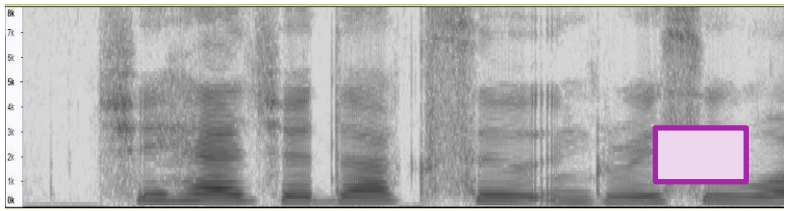


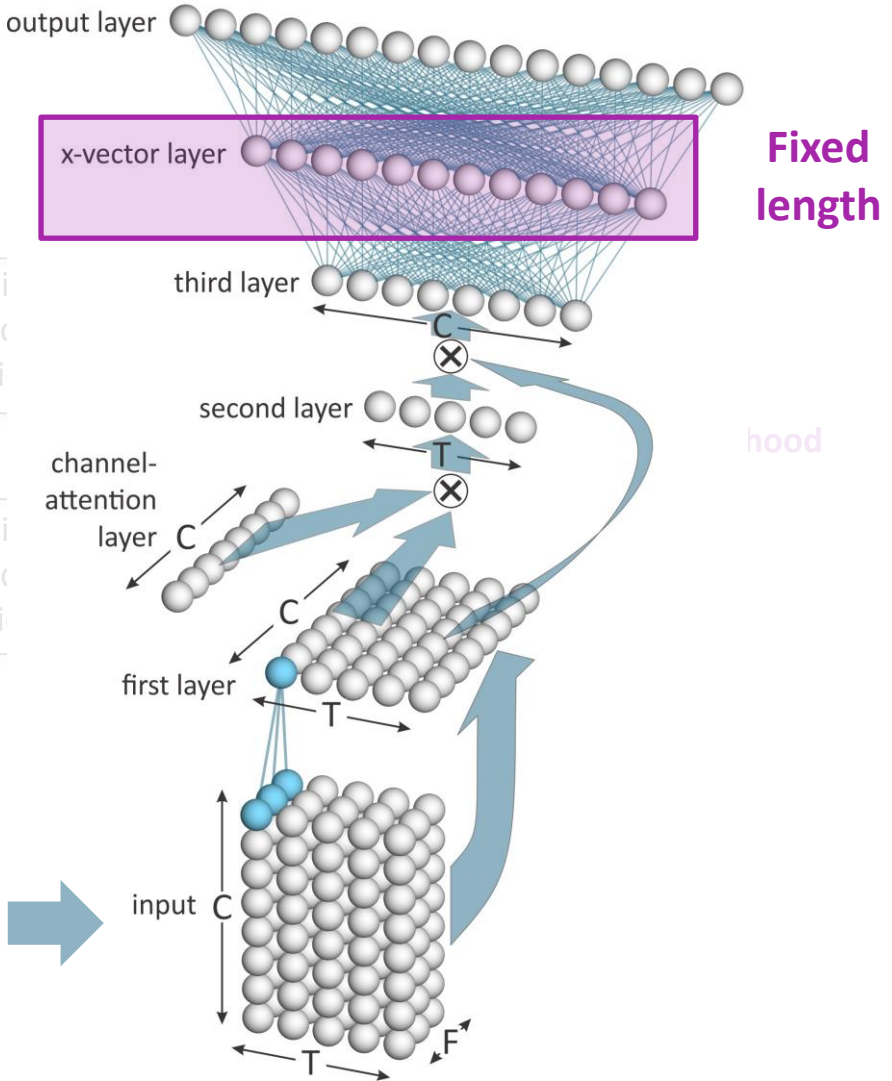




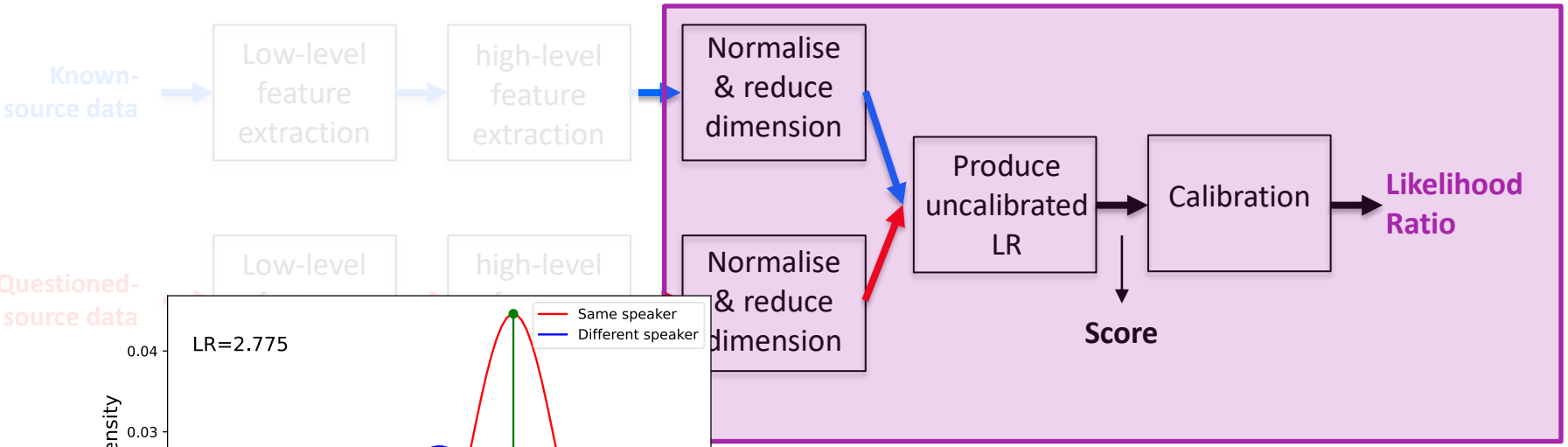
AI for forensic …

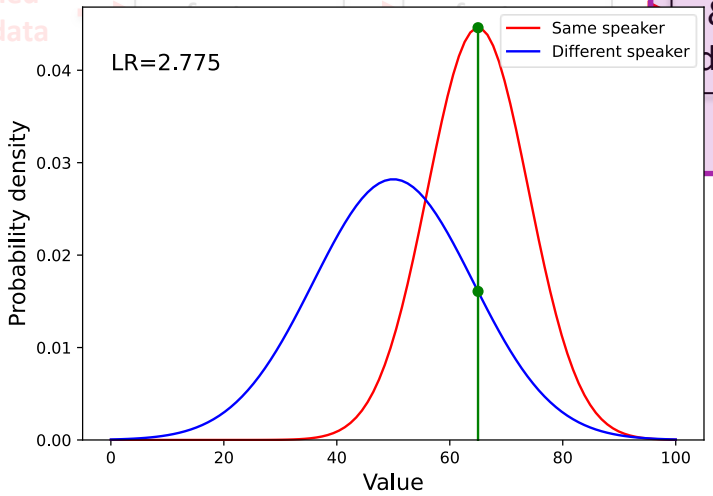






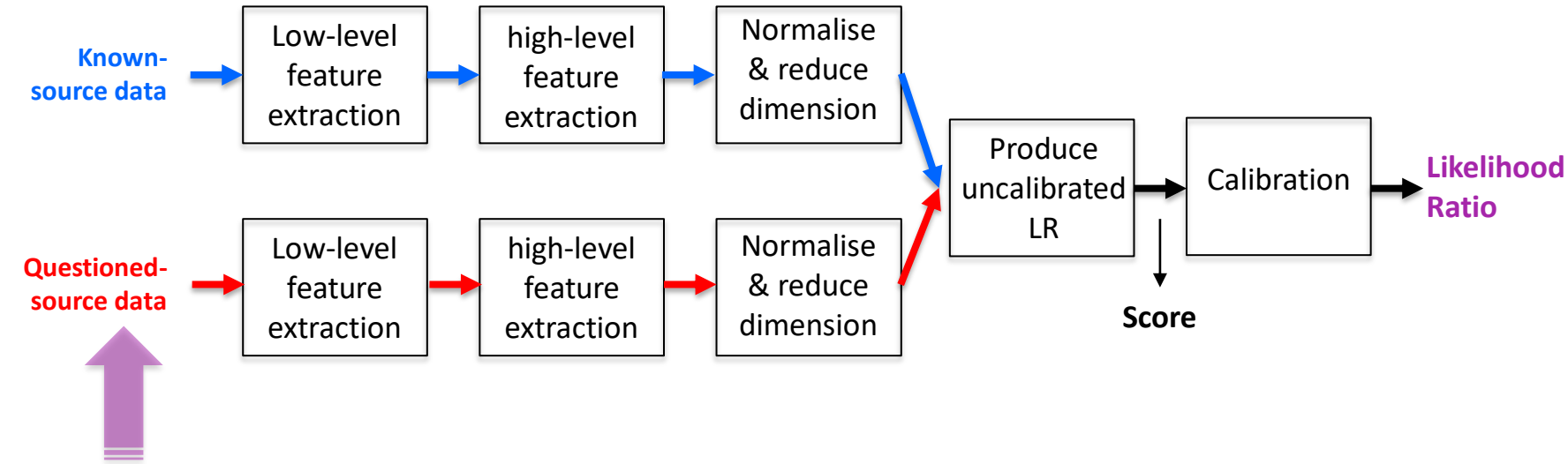
AI for forensic voice comparison





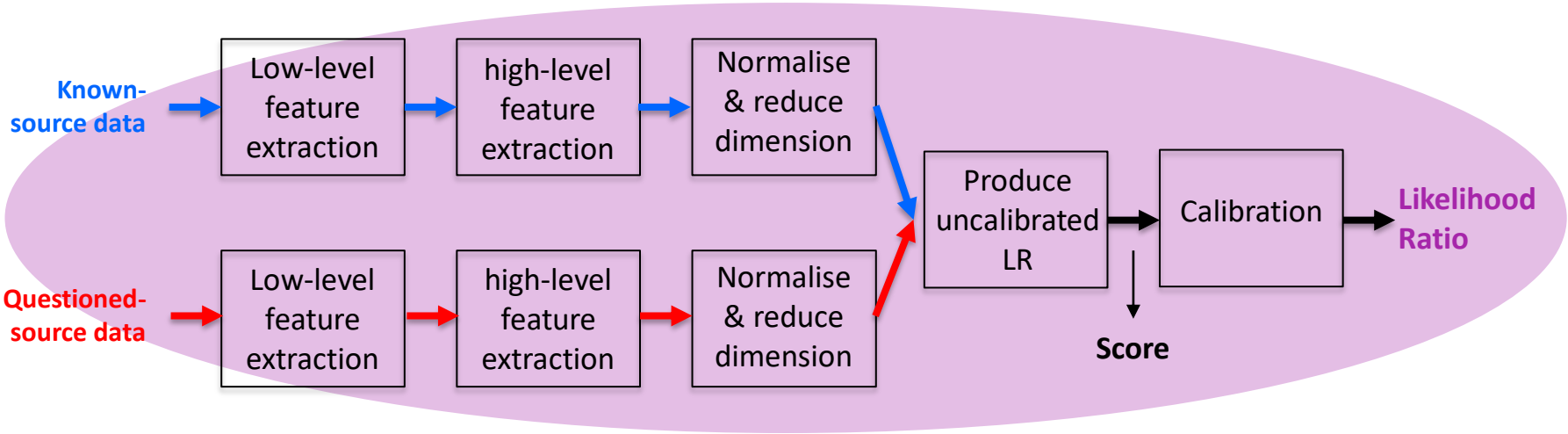
Bias in AI

Bias in AI


Known-source data
Low-level feature extraction
high-level feature extraction
Normalise & reduce dimension
Questioned-source data
Low-level feature extraction
high-level feature extraction
Normalise & reduce dimension
Produce uncalibrated LR
Calibration
Likelihood Ratio
Score
Human decisions are made here
Automated
Repeatable
Testable

Explainable AI

Validation



Empirical validation under casework conditions

Validation – data

Validated using **new data**

recording pairs : same-speaker and different-speaker

case conditions : questioned- and known-speaker.

Sufficient data.

Sufficiently **representative** data.

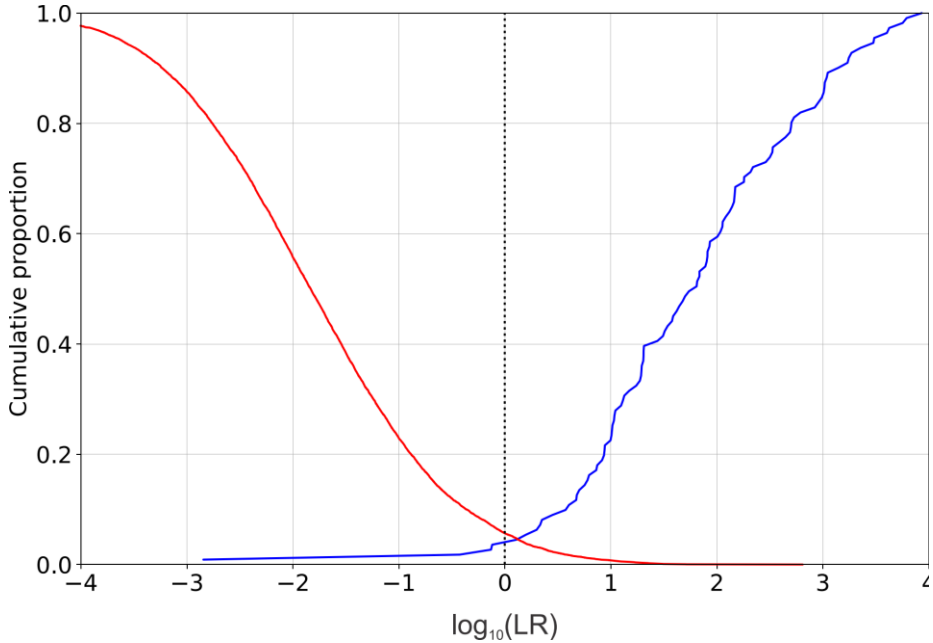
Validation – measurement

Tippett plot

Log LRs

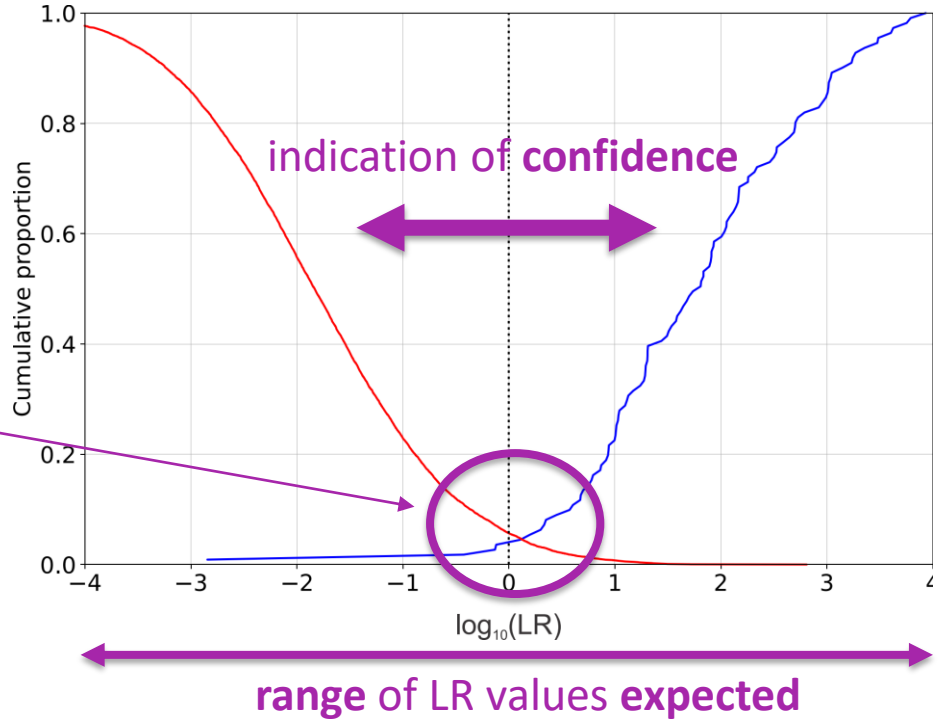
Red: from pairs of recordings from different speakers

Blue: from pairs of recordings from same speakers



Validation – measurement

indication of **calibration**



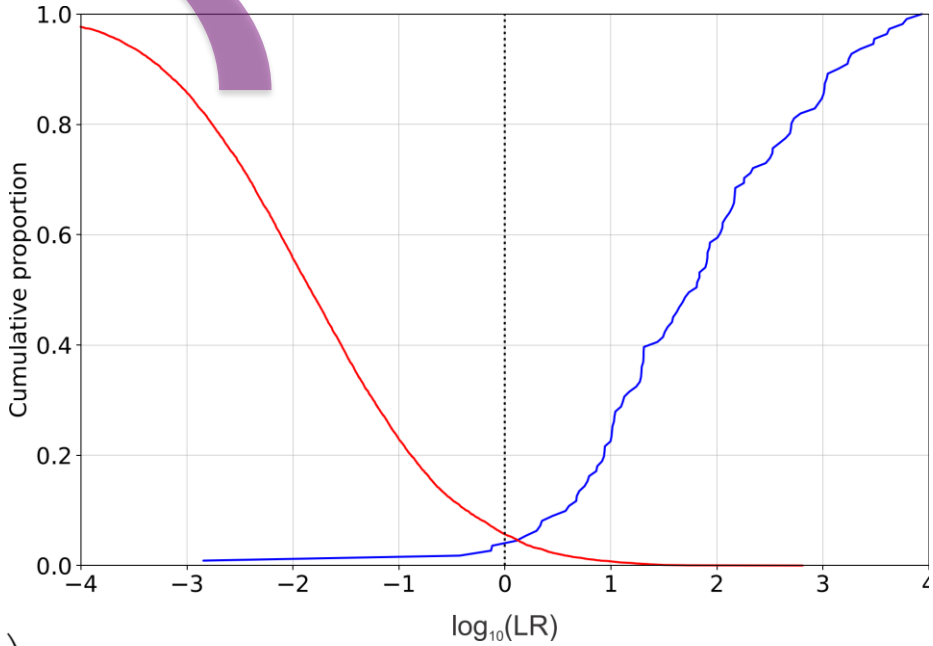
Validation – measurement

Cost of Log Likelihood Ratio C_{llr}

$C_{llr} \rightarrow 0$: more informative

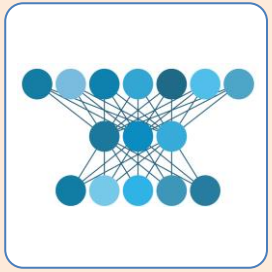
$C_{llr} \rightarrow 1$: less informative

$C_{llr} > 1$: error



For **this system**
for **this case**

$$C_{\mathrm{llr}} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{SO}}\log_2\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\log_2\left(1+LR_{do_j}\right)\right)$$



1. Forensic data science 101

→ let's use data!

2. Artificial intelligence

→ Aston's E^{3}FS^3 forensic voice comparison system

3. Human intelligence

→ but surely humans can do that?

Speaker identification by humans

Human listeners *vs* the Aston E^3FS^3 system.

How did you do?

ID	Same speaker?	Human	E^3FS^3 System
25	Yes	33/53 ✗	✓
4	Yes	33/53 ✗	✗
40	No	25/53 ✗	✓

Speaker identification by humans

A challenge to human-supervised automatic forensic voice comparison:

“I could do just as well”

Speaker identification by humans

A challenge to human-supervised automatic forensic voice comparison:

“I could do just as well”

Expert testimony is only admissible in common-law jurisdictions

if it will potentially assist the “trier of fact” to make a decision.

Speaker identification by humans

Basu N., Bali A.S., **Weber P.**, Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2022). Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. Forensic Science International, 111499.

Basu N., **Weber P.**, Bali A.S., Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2023). Speaker identification in courtroom contexts – Part II: Investigation of bias in individual listeners' responses. Forensic Science International, 349, 111768. https://doi.org/10.1016/j.forsciint.2023.111768

Bali A.S., Basu N., **Weber P.**, Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2023). Speaker identification in courtroom contexts – Part III: Groups of collaborating listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. Manuscript submitted for publication.

Speaker identification by humans

Research questions:

1. Is **speaker identification** by a **lay listener (judge) listening alone** more or less accurate than the output of a **forensic-voice-comparison system** that is based on state-of-the-art automatic-speaker-recognition technology?

Speaker identification by humans

Research questions:

1. Is **speaker identification** by a **lay listener (judge) listening alone** more or less accurate than the output of a **forensic-voice-comparison system** that is based on state-of-the-art automatic-speaker-recognition technology?

Speaker identification by lay listeners:

a listener unfamiliar with the speaker(s) listens to:

- a voice they hear on two different occasions, or
- two voice recordings, or
- one voice recording and a live speaker

and **attempts to determine whether they are the same speaker or different speakers**.

Speaker identification by humans

Research questions:

1. Is **speaker identification** by a **lay listener (judge) listening alone** more or less accurate than the output of a **forensic-voice-comparison system** that is based on state-of-the-art automatic-speaker-recognition technology?

(i.e. is the human better than the AI?)

Speaker identification by humans

Research questions:

1. Is **speaker identification** by a **lay listener (judge) listening alone** more or less accurate than the output of a **forensic-voice-comparison system** that is based on state-of-the-art automatic-speaker-recognition technology?

(i.e. is the human better than the AI?)

What about when the **accent** is unfamiliar?

What about when the **language** is unfamiliar?

Speaker identification by humans

Research questions:

1. Is speaker identification by a lay listener (judge) listening alone more or less accurate than the output of a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology?
2. What happens if the lay listener **listens to the recordings** and **considers the output of the system**?

Speaker identification by humans

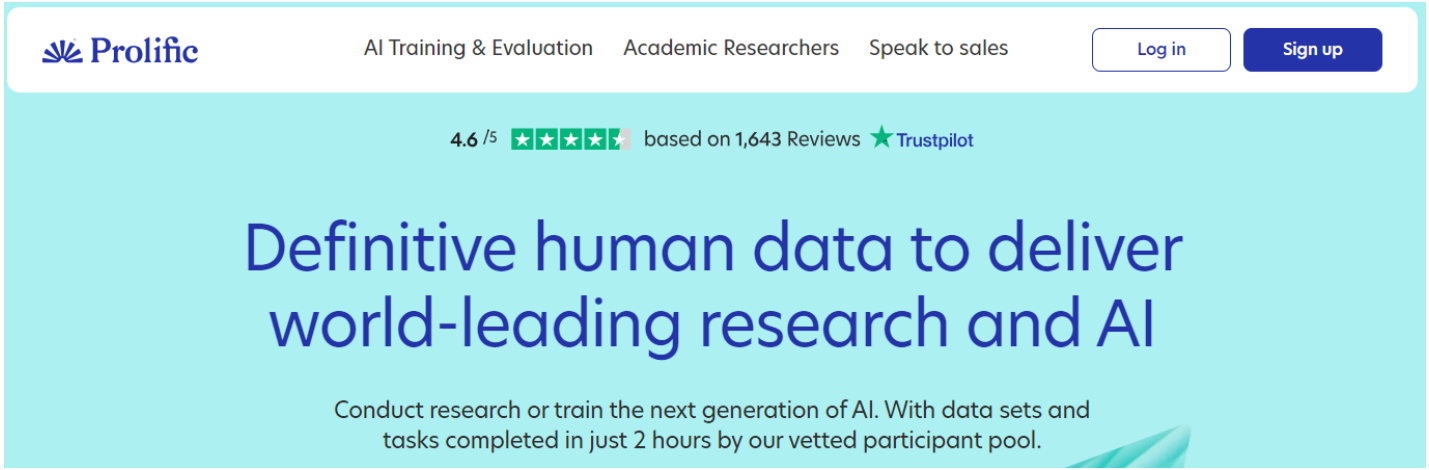
Research questions:

1. Is speaker identification by a lay listener (judge) listening alone
   more or less accurate than the output of a forensic-voice-comparison system that is
   based on state-of-the-art automatic-speaker-recognition technology?
2. What happens if the lay listener listens to the recordings and
   considers the output of the system?
3. What about when a **jury** listens and **collaboratively** makes a judgement?

Speaker identification by humans

Recruitment

Participants (“listeners”) recruited through Prolific.com.



Speaker identification by humans

53 x **Australian-English** listeners
61 x **North-American-English** listeners
55 x **Spanish-language** listeners

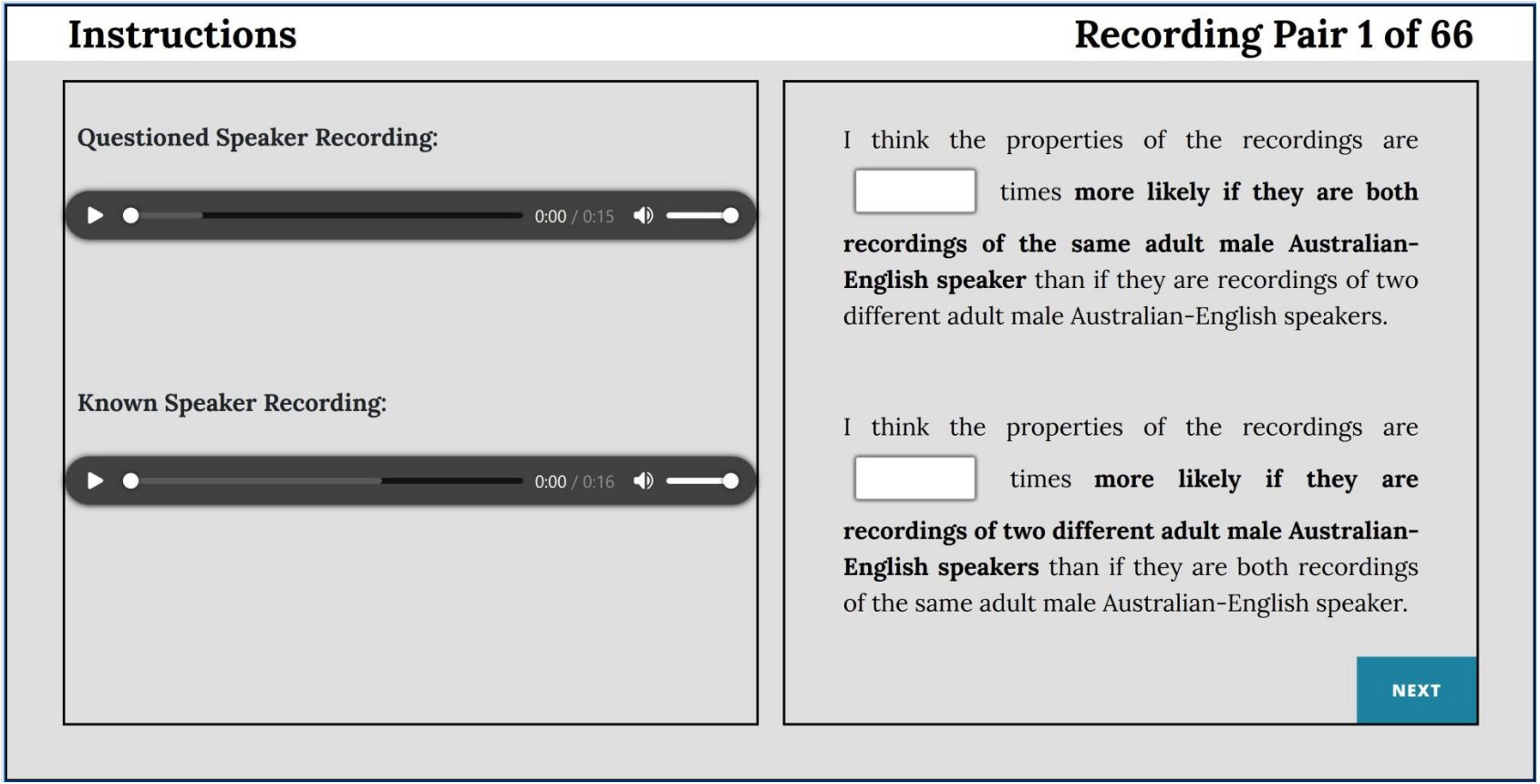
Pairs of 15s recordings: 31 x **same-speaker**, 30 x **different-speaker**

Adult male speaker of Australian English.

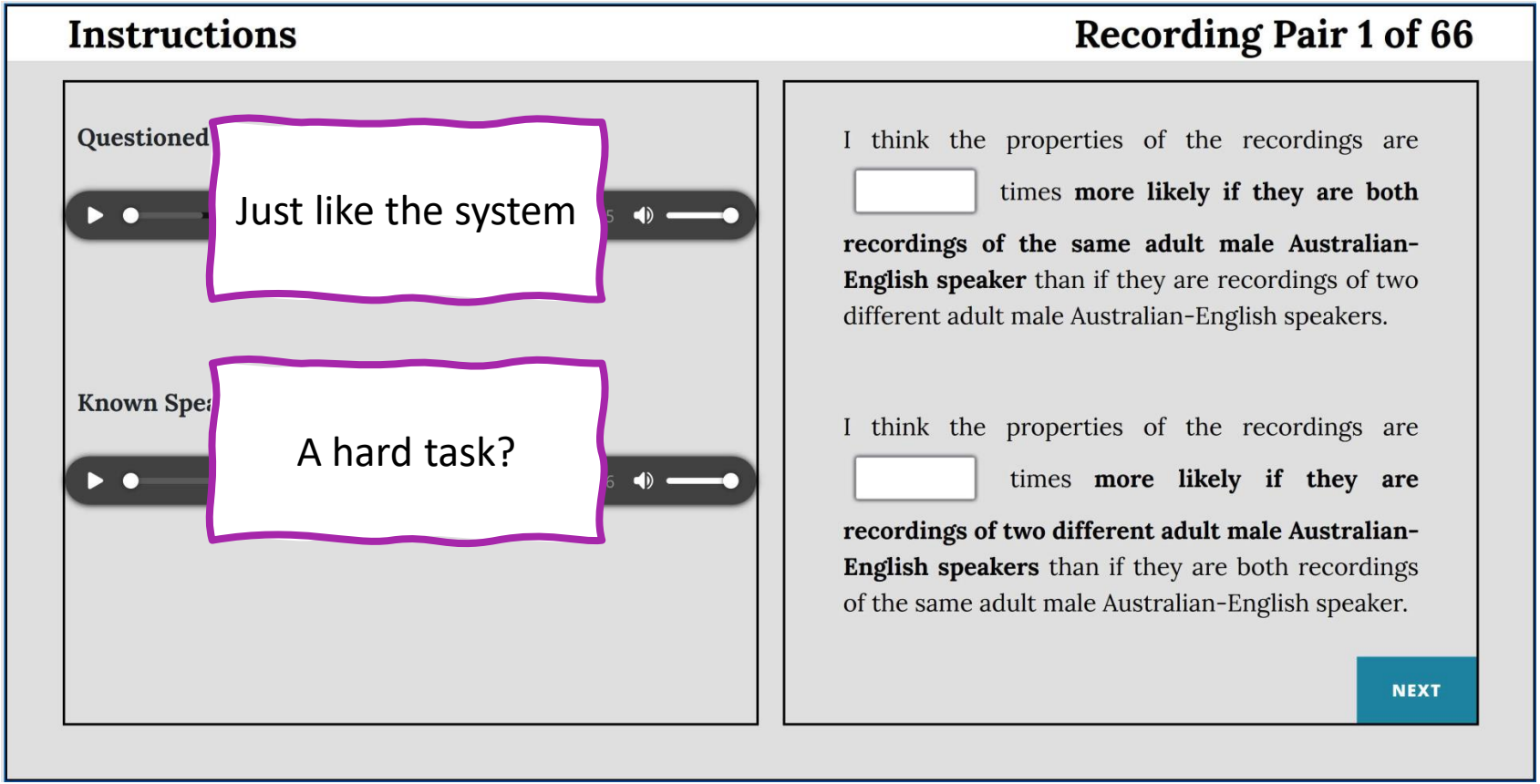
Conditions of a real case:

Questioned-speaker condition:	landline telephone call, babble noise, saved using lossy compression.
Known-speaker condition:	interview recorded in a reverberant room, ventilation system noise.

Speaker identification by humans



Speaker identification by humans

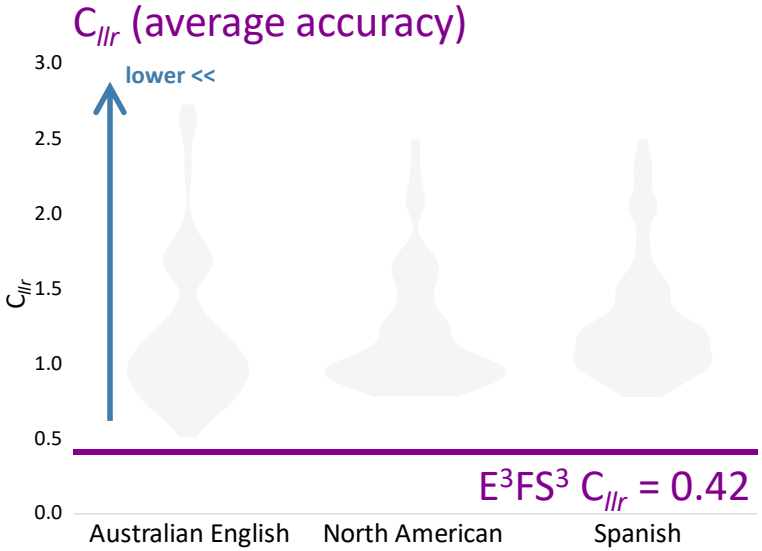


Speaker identification by humans

C_{llr} system: 0.42 (lower than our baseline 0.18 due to short sections of speech)

C_{llr} **best listener**: 0.51

C_{llr} many listeners: > 1 (worse than uninformative)

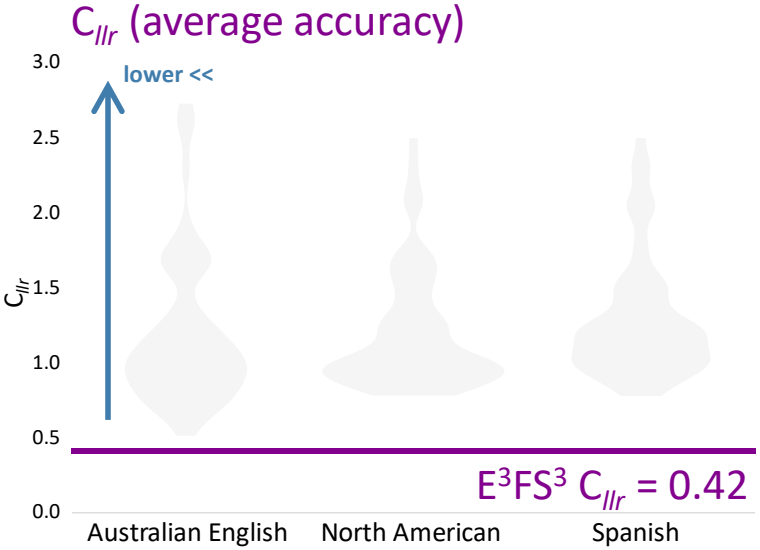


Speaker identification by humans

C_{llr} system: 0.42 (lower than our baseline 0.18 due to short sections of speech)

C_{llr} **best listener**: 0.51

C_{llr} many listeners: > 1 (worse than uninformative)



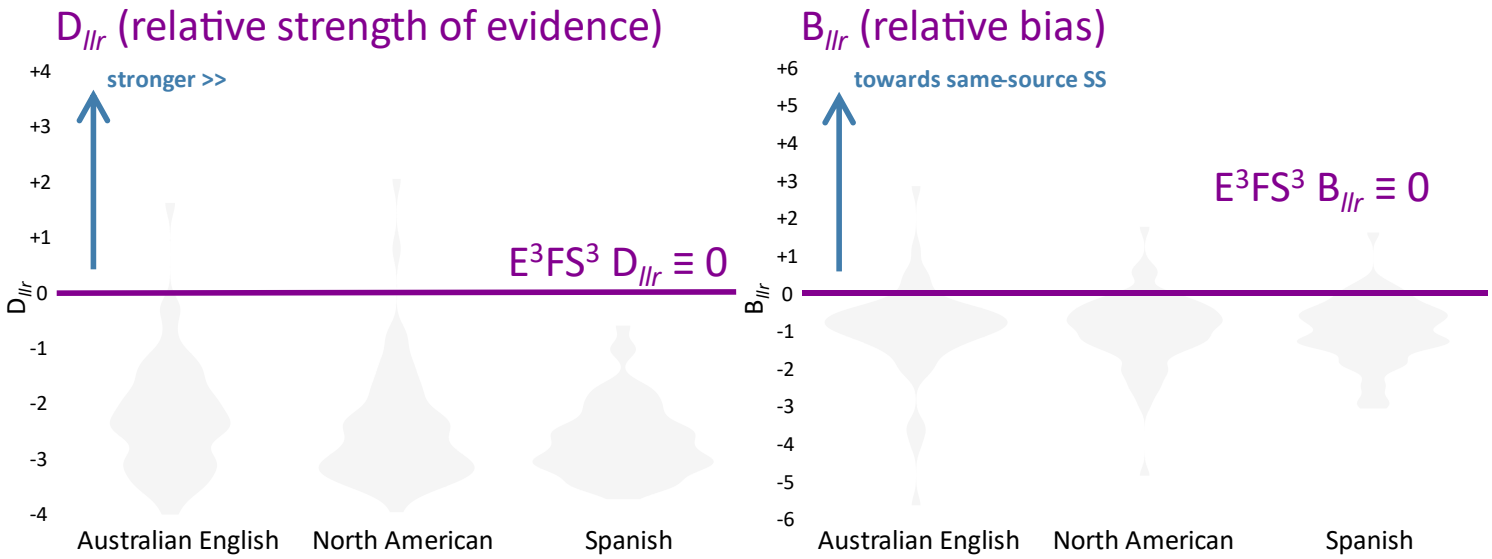
Accent makes it **harder** (North American listeners)

Language makes it **harder** (Spanish listeners)

Speaker identification by humans

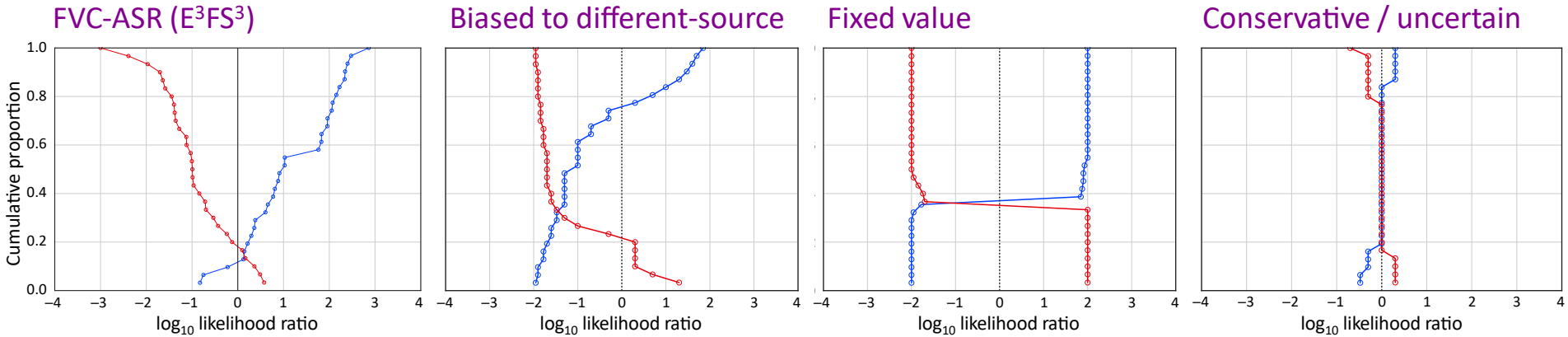
Listeners generally **less discriminative** than the system ($\mathrm{D}_{llr} < 0$)

Listeners generally **biased towards different-speaker hypothesis** ($\mathrm{B}_{llr} < 0$)



Speaker identification by humans

Common human “strategies” – is this informative?



Speaker identification by humans

“But it would help me to know how the system performed”

2. What happens if the lay listener **listens to the recordings** and **considers the output of the system**?

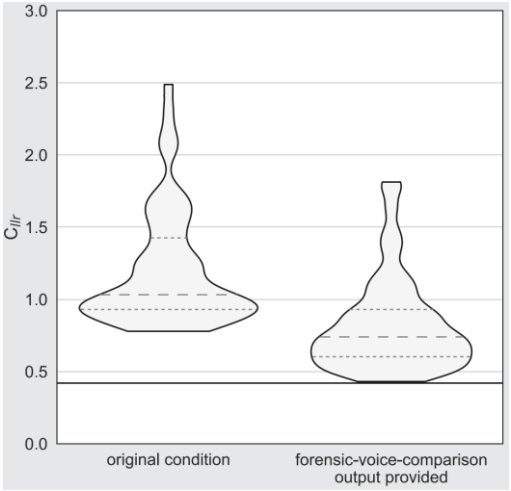
Speaker identification by humans

C_{llr} system: 0.42 (unchanged)

C_{llr} **best listener**: down to 0.43 (was 0.51)

This replicates patterns seen in other domains:

algorithm > algorithm + subjective judgement > subjective judgement



Speaker identification by humans

Challenge: ***"Perhaps the bias towards the different-speaker hypothesis is due to the poor recording conditions and mismatch between the conditions of the questioned- and known-speaker recordings?"***

Additional experiments:

- provide information about the recording conditions in advance
- provide high-quality recordings

Speaker identification by humans

C_{llr} system: 0.40 (slightly reduced set of stimuli)

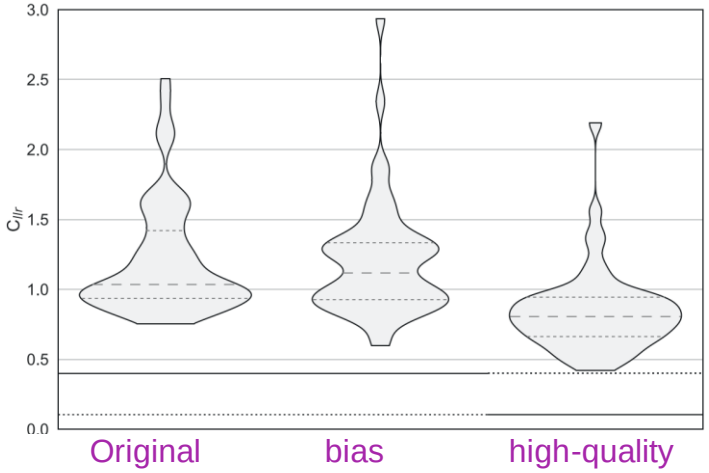
biasing (providing info about conditions) had **no obvious effect**

C_{llr} system – high quality: 0.10

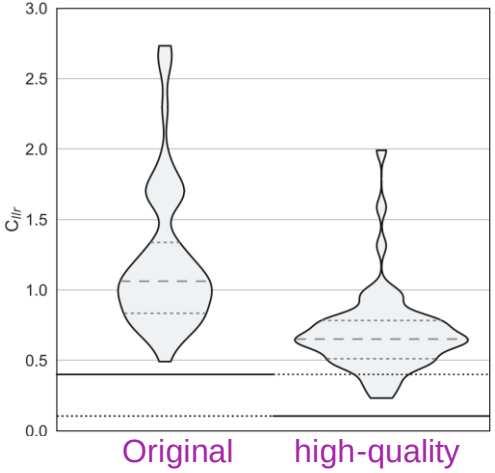
C_{llr} **best American-English listener**: 0.43

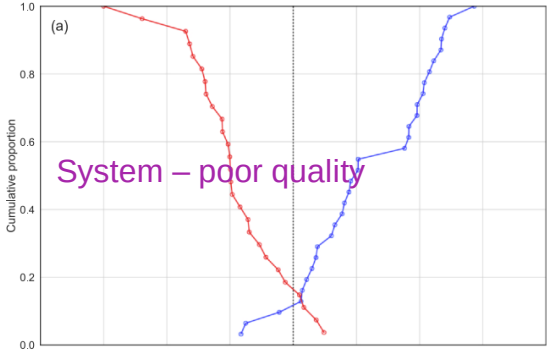
C_{llr} **best Australian-English listener**: 0.23

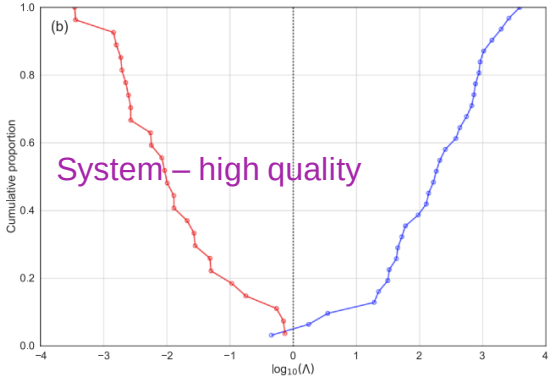
American-English listeners



Australian-English







Speaker identification by humans

“What about group consensus (jury decision)?”

3. What about when a **jury** listens and **collaboratively** makes a judgement?

12 same-speaker pairs, 12 different-speaker pairs

23 groups of size 5 – 12 listeners

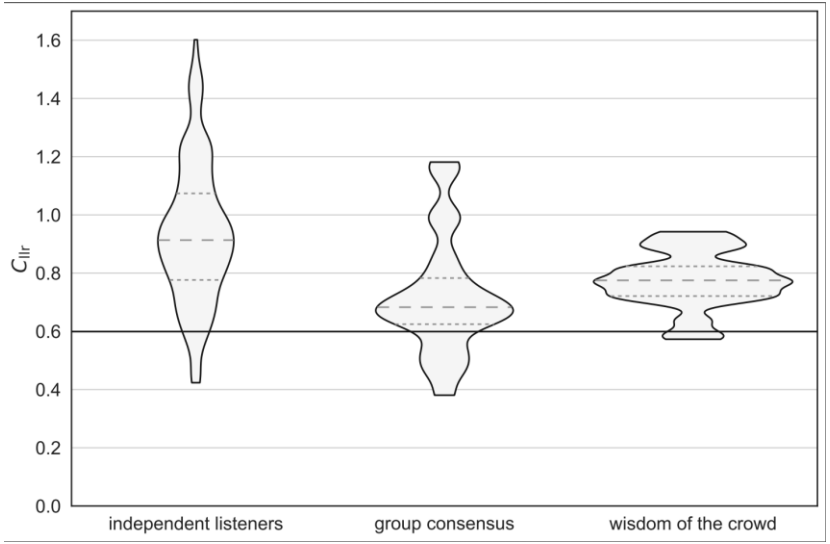
(difficulty recruiting groups of 12 )

Speaker identification by humans

C_{llr} system: 0.60 (very reduced set of stimuli – and biased sample)

No clear relationship between results and group size

Collaboration: 78% groups C_{llr} > system



"Wisdom of the crowd" (average individual responses) on average **worse** than group consensus

(unexpected – small sample size)

Speaker identification by humans

Conclusions

Forensic voice comparison based on state-of-the-art automatic speaker recognition is

more accurate than speaker identification by **individual** lay listeners

Note: even when they may also consider the system output
Note: listeners overestimate their own ability.

more accurate than speaker identification by **collaborating groups** of listeners

Lay listeners perform worse when speech is in an **unfamiliar accent** and even worse when an **unfamiliar language**

Speaker identification by humans

Conclusions

Forensic voice comparison based on state-of-the-art automatic speaker recognition is

more accurate than speaker identification by **individual** lay listeners

Note: even when they may also consider the system output
Note: listeners overestimate their own ability.

more accurate than speaker identification by **collaborating groups** of listeners

Lay listeners perform worse when speech is in an **unfamiliar accent** and even worse when an **unfamiliar language**

➔ **Judges and juries should not attempt to perform their own speaker ID**

➔ They should **rely exclusively on expert testimony** based on a calibrated and validated forensic-voice-comparison system.

So what?

- What **lessons** to draw for AI more generally?
- And the **interaction** between AI and “human” intelligence?
- Where and when should humans be **involved** in AI systems?
- Where and when should humans **overrule** AI systems?
- What is the role of **XAI** and/or black-box **validation**?
- How do we know if we have the **right data**?
- How do we know if we have **enough data**?

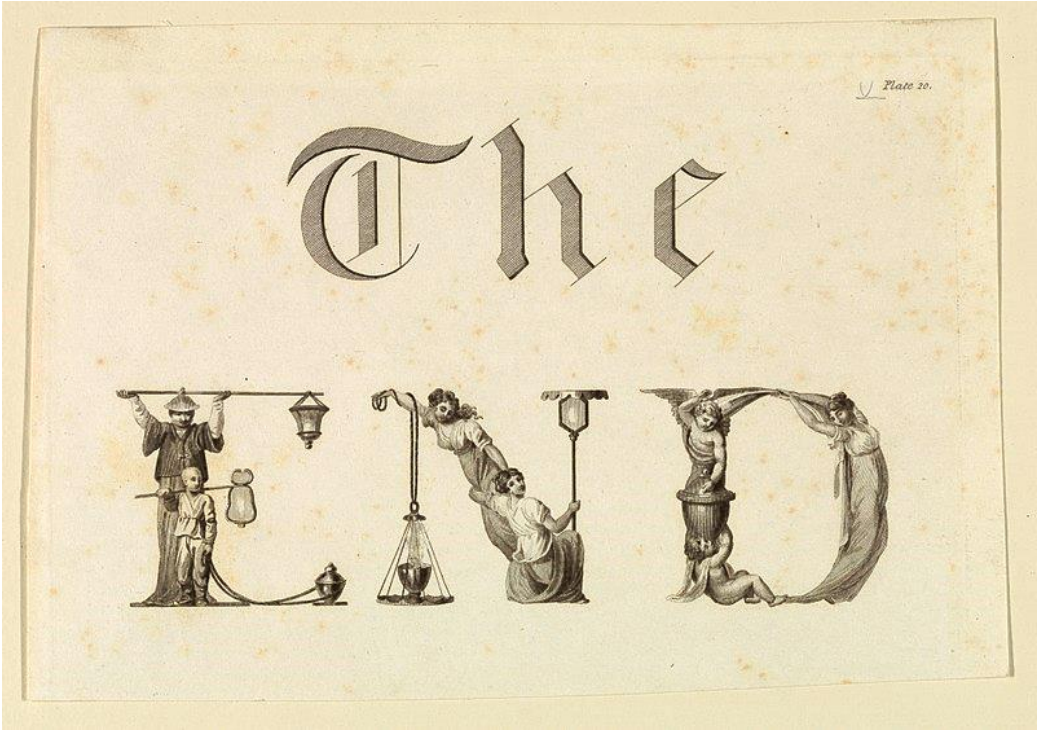
Phil Weber

Thank you!

Contact:

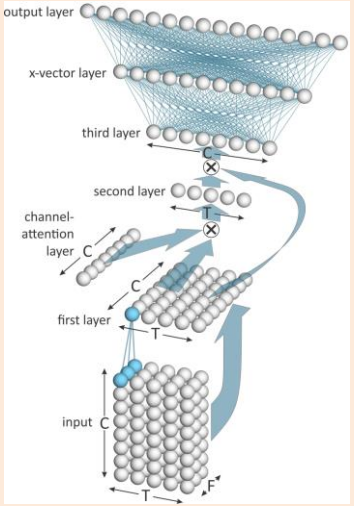
p.weber1@aston.ac.uk

http://forensic-data-science.net/



Where next – Research & development

1. **E³FS³ Forensic voice comparison system**
    - Beta testing → end-partner testing.
        - Packaging – help!
            - Training – Continuous Professional Development.
                - Application to casework
            - State of the art advances
        - New methods for calibration
    - LLM-based speech-processing systems (e.g. HuBERT)



2. **Applying proven methods to other biometrics / data sources**

Research – Use of data – empirical

More generally:

Barrier to justice:

necessity & cost of per-case validation and data collection.

can we build a library of (empirical) “pre-emptive validation reports”?

conditions and populations of interest to our partners.

question: how do we know if a new case is “close enough”?

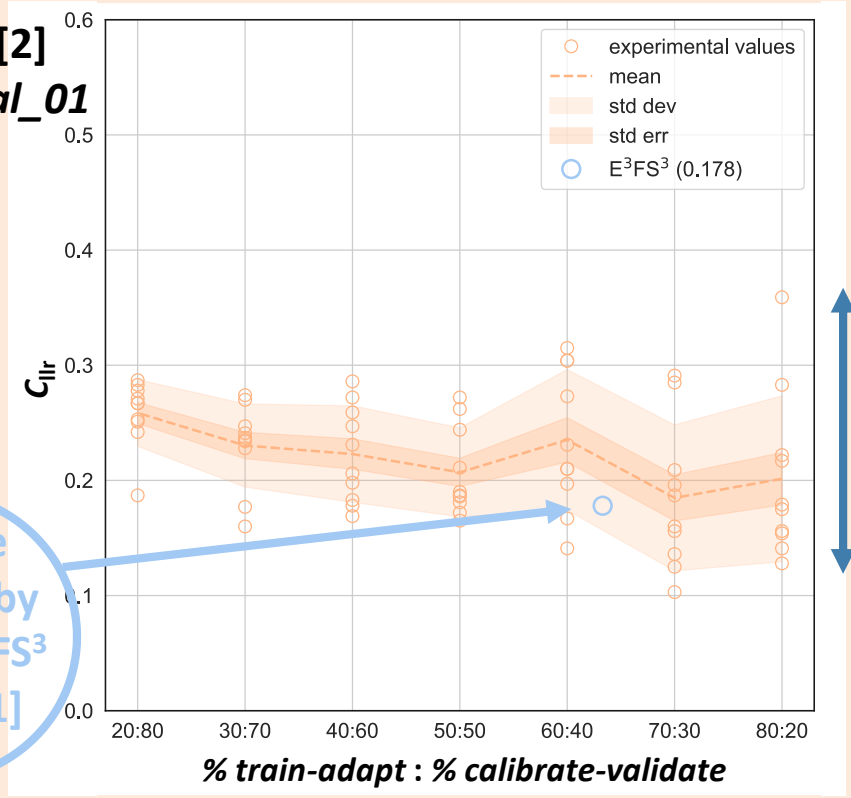
Research – Use of data – theoretical

Still more generally:

- What are the effects of sampling variability at many points?
- What are the impacts of decisions made in training and data selection?
- Black box validation ➔ not “explainable AI”
- but (how) **do we know we can trust** the output?

Research – Use of data – empirical

Benchmark [2]
forensic_eval_01
Varying and re-sampling speaker split for training & validation



C_{llr} value reported by Aston's E³FS³ system [1]

sampling variability

"decision" variability

Research – Use of data – theoretical

Can we

- In a **principled manner** understand how to use **data** and the effects on the **confidence** which can be placed in results?
- Provide **rigorous methods** to determine how much data is needed, what characteristics it requires, and how close a new case is to an existing case?
- Develop **measures and metrics** for the above? (How “close” are two audio data sets?)
- Learn from and influence other areas of AI / machine learning (theory)?