Voice-print transformation for migration between automatic speaker identification systems

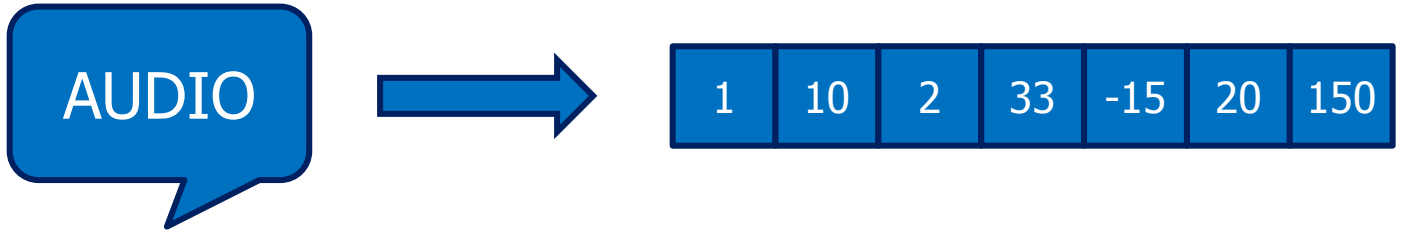
Ondřej Glembek, Pavel Matějka, Olda Plchot,
Jan Pešán, Lukáš Burget, Jan Černocký, Vláďa Malenovský,
and Petr Schwarz

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

i-vector transformation for migration between automatic speaker identification systems

Ondřej Glembek, Pavel Matějka, Olda Plchot,
Jan Pešán, Lukáš Burget, Jan Černocký, Vláďa Malenovský,
and Petr Schwarz

BRNO UNIVERSITY OF TECHNOLOGY
FIT
FACULTY OF INFORMATION TECHNOLOGY

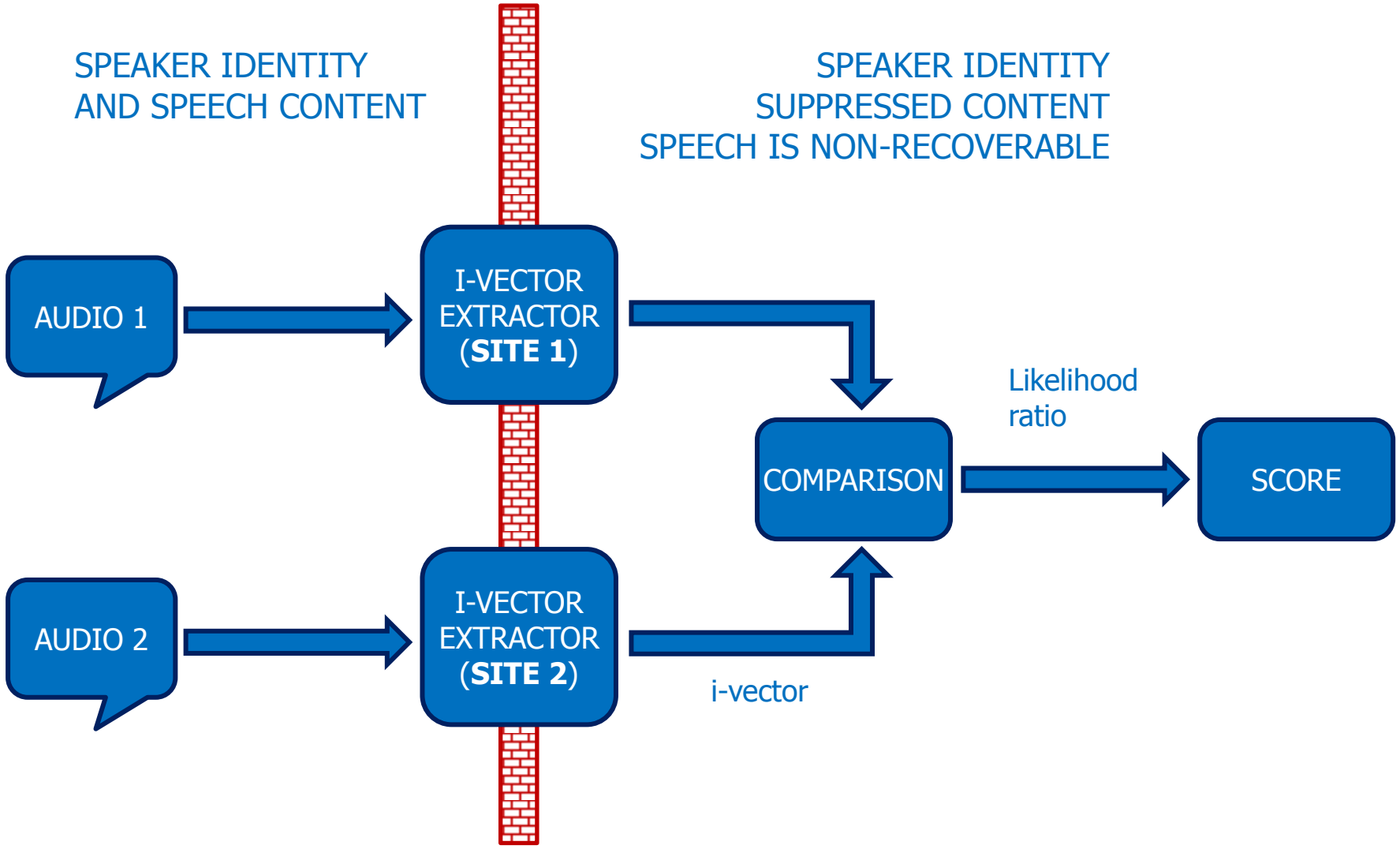


- Information-rich
- Low-dimensional
- Fixed-length
- Vector of real numbers

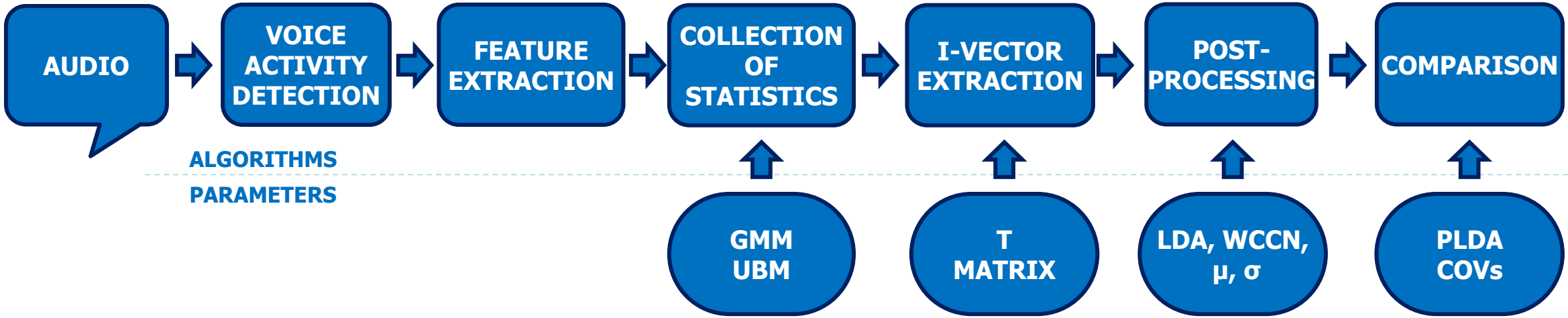
- Based on statistical model
- Easy to compare
- Easy to store
- Not recoverable to speech

Dehak, N., et al., Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification In Proc Interspeech 2009, Brighton, UK, September 2009

System block-diagram



System block-diagram - detail



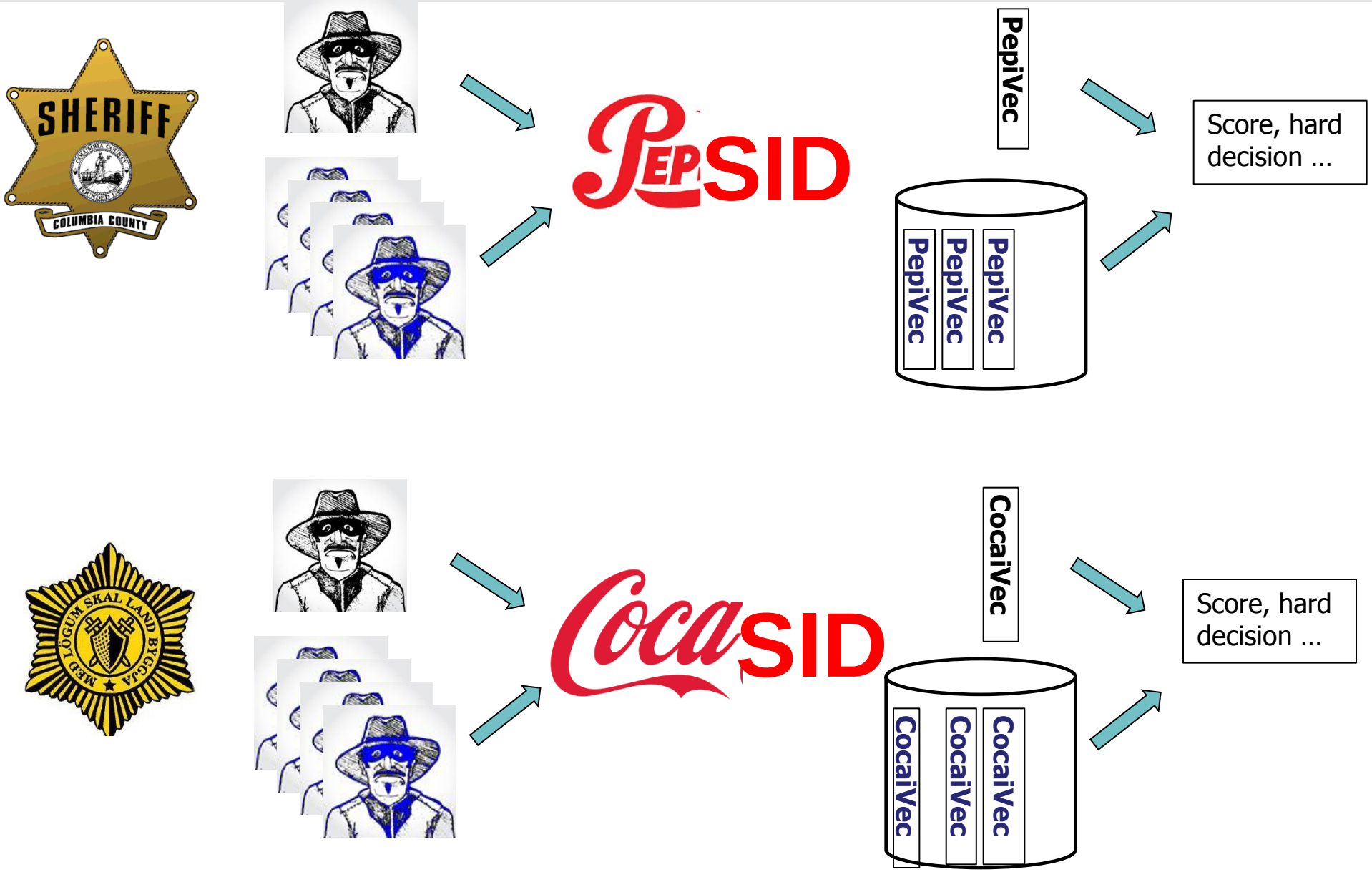
Interpretation of i-vectors

- The interpretation of i-vectors change from system to system
- This depends on many factors
  - Feature extraction
  - The way the GMM Universal background model (UBM) has been trained (initialization, EM algorithm, Gaussian splitting protocol, …)
  - The way the i-vector extractor has been trained (initialization, involves numerical EM algorithm, MD, …)

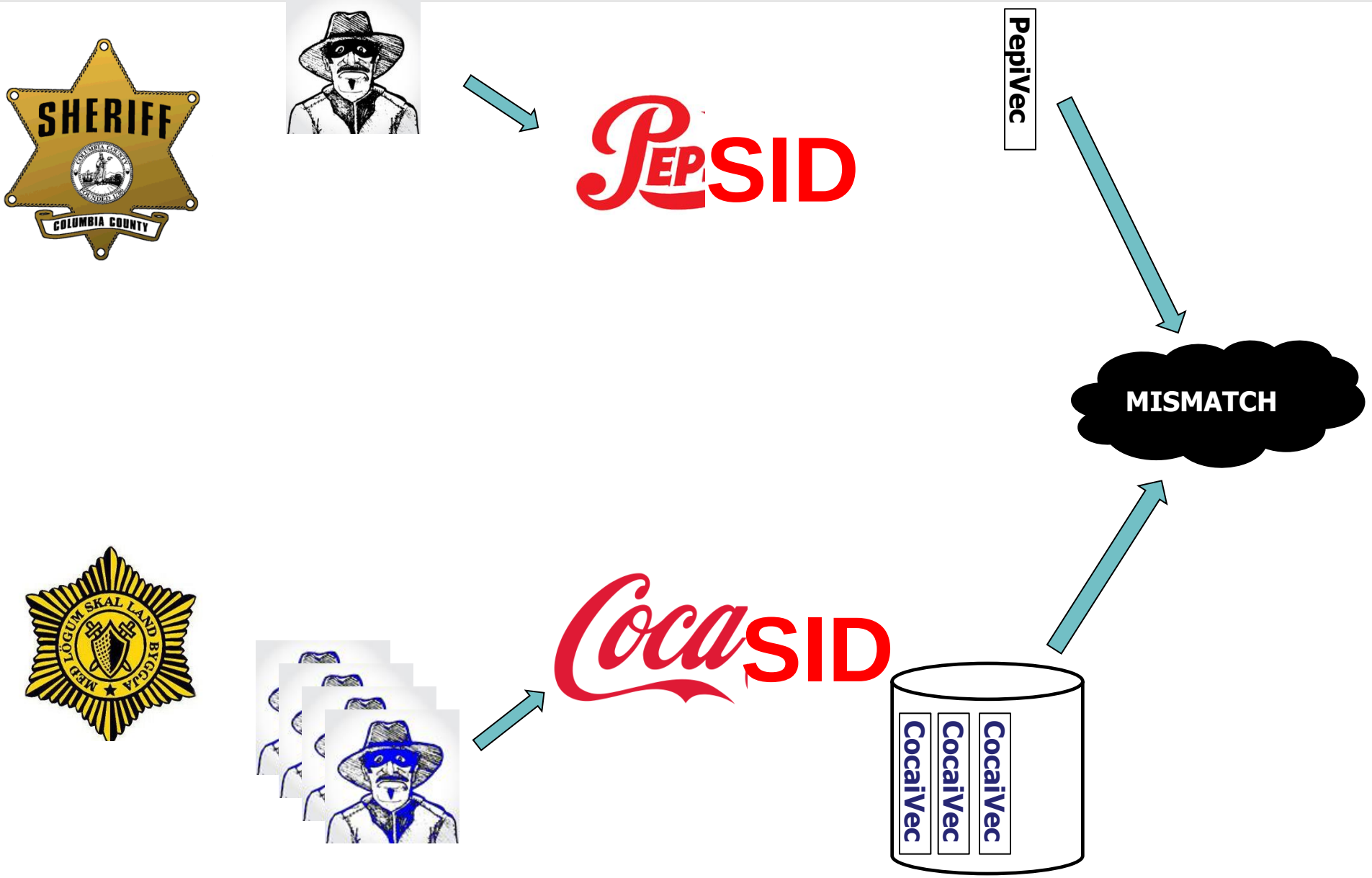
Can we use i-vectors produced by one system for scoring on another system?

- Inter-site data exchange
- i-vector standard
- i-vector extraction upgrade
- …

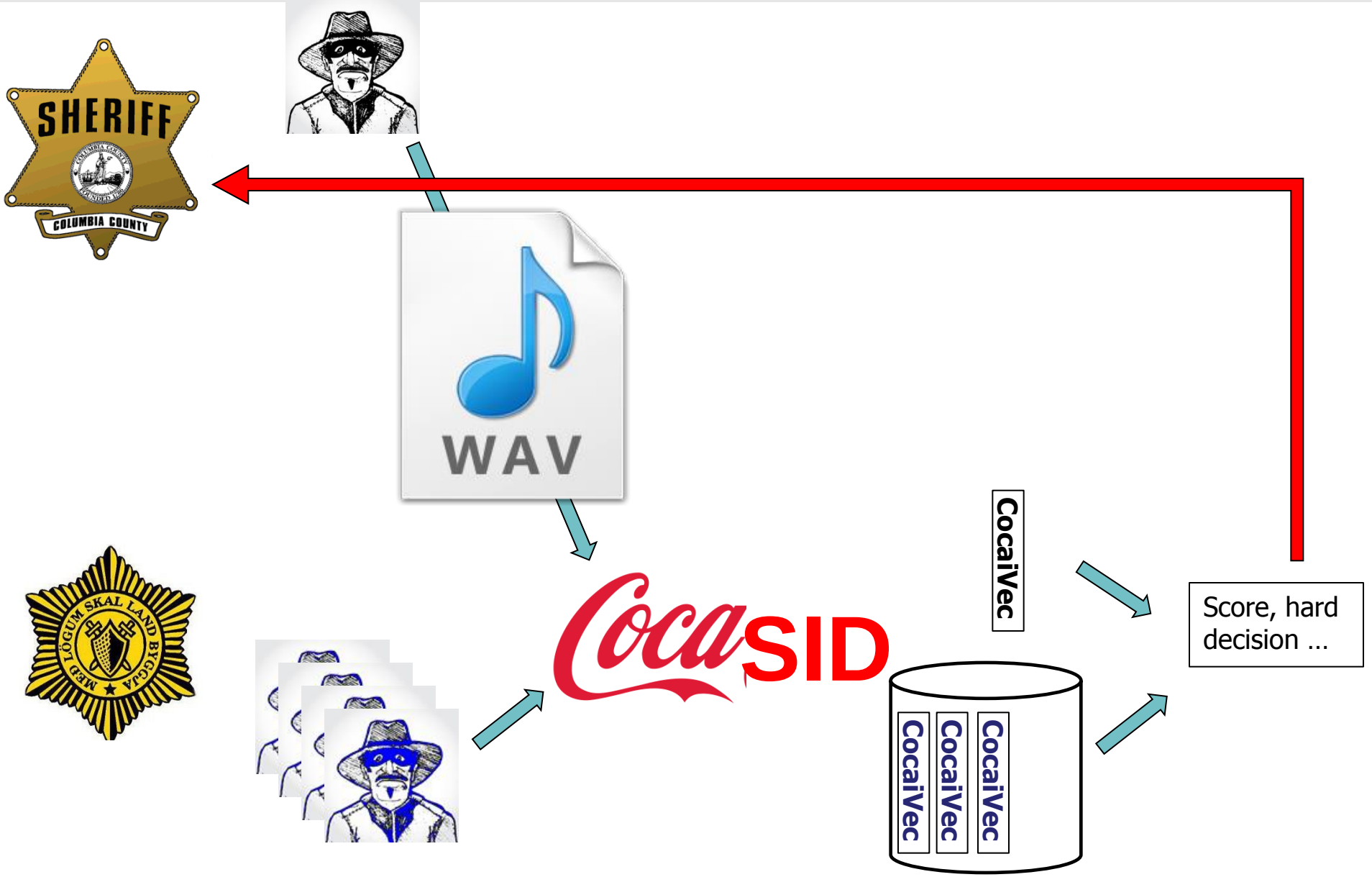
It works ☺


SHERIFF
COLUMBIA COUNTY
PepSID
PepiVec
PepiVec
PepiVec
PepiVec
PepiVec
Score, hard decision ...
MEÐ LÖGUM SKAL LAND BYGGJA
CocaSID
CocaiVec
CocaiVec
CocaiVec
CocaiVec
CocaiVec
Score, hard decision ...

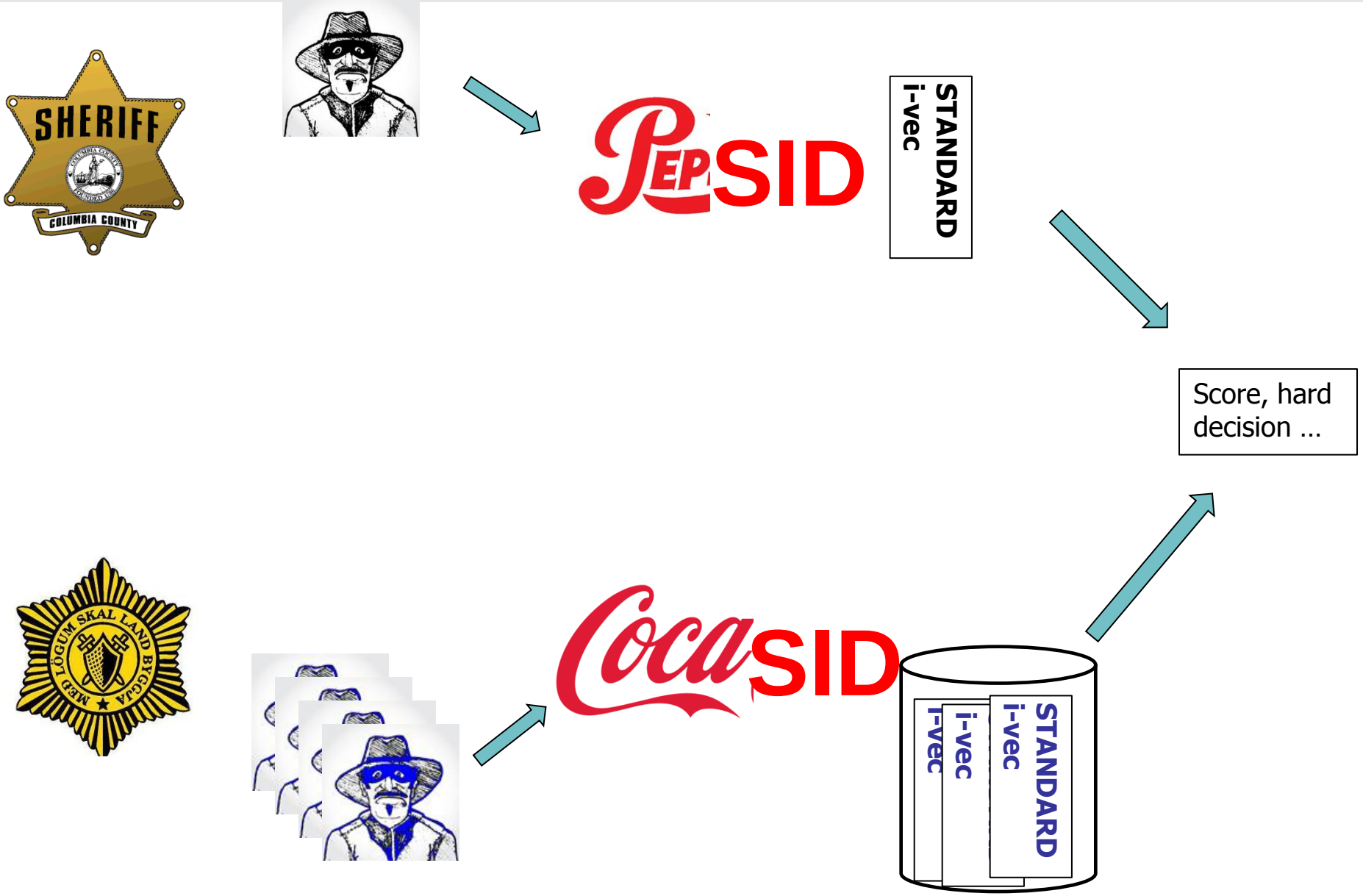
It does not work ☹



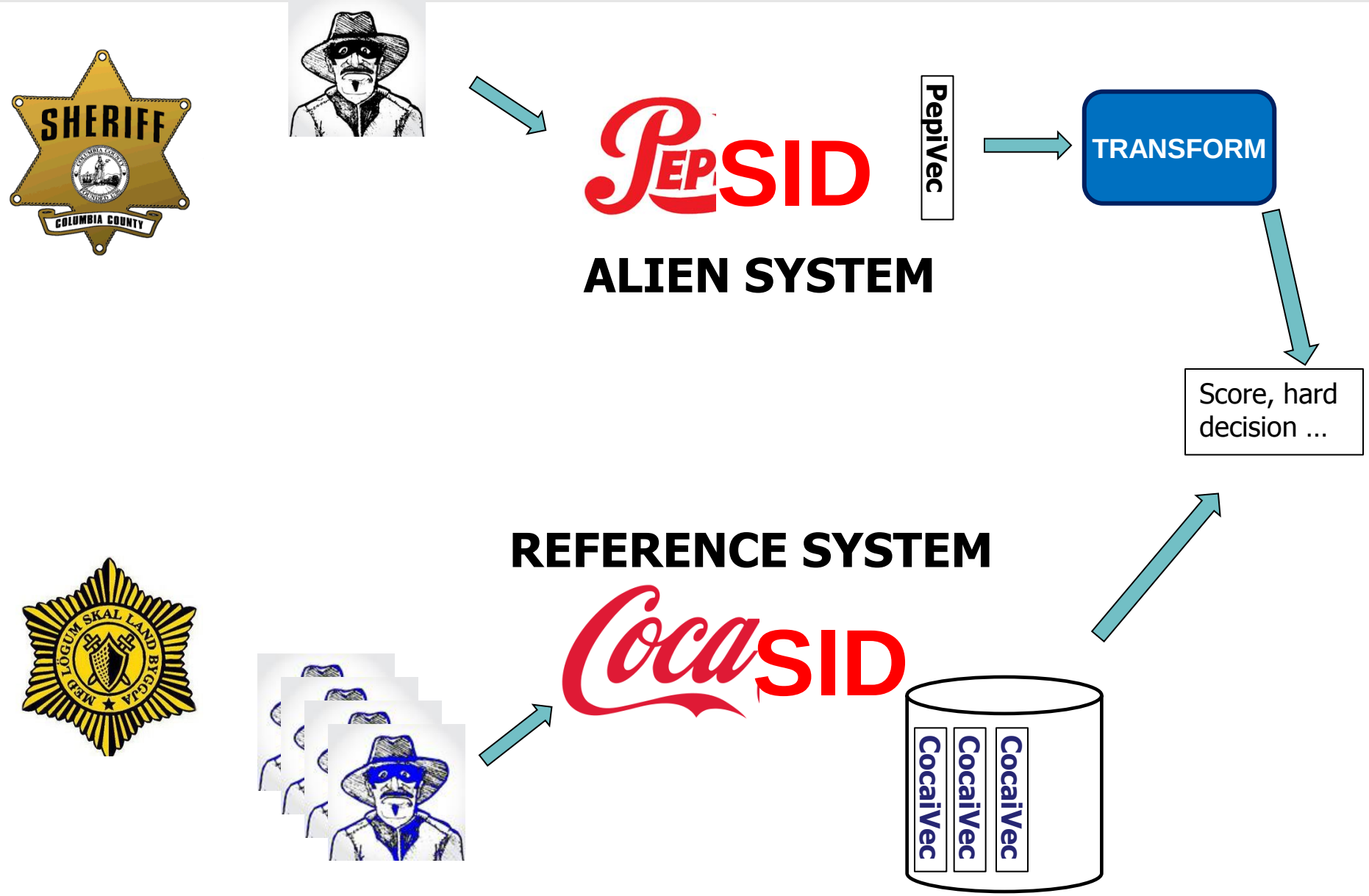
Making it work



Making it work – standardized iVectors



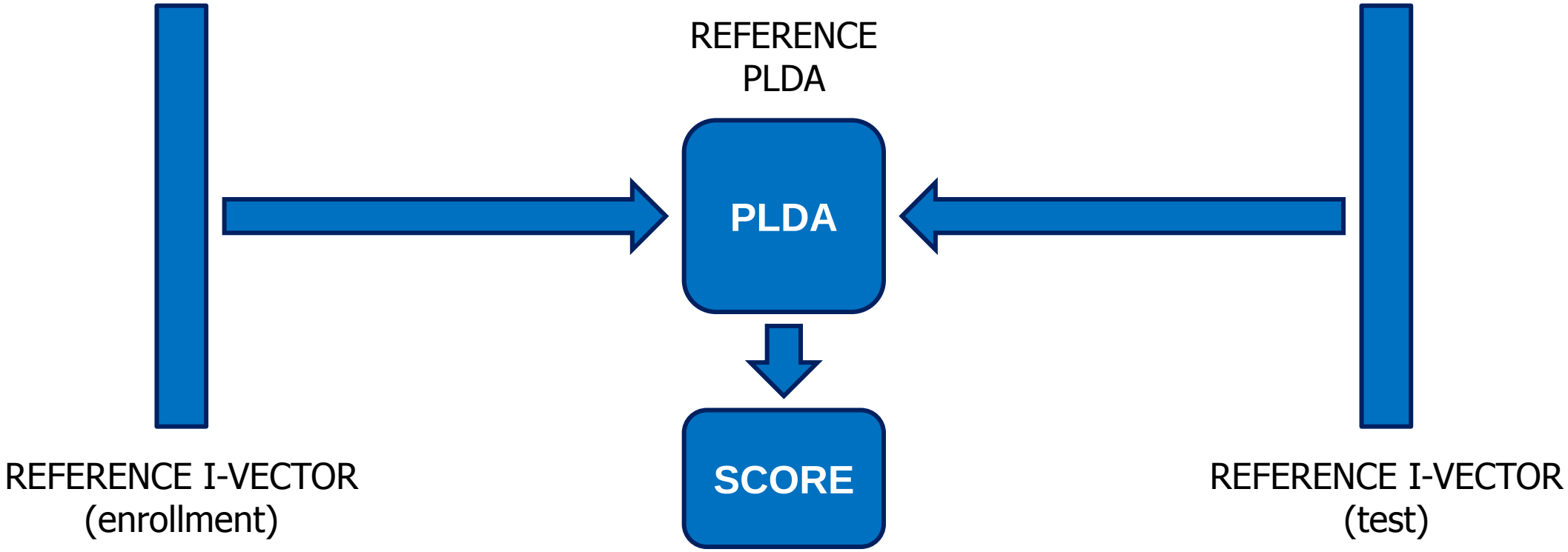
Making it work – standardized iVectors



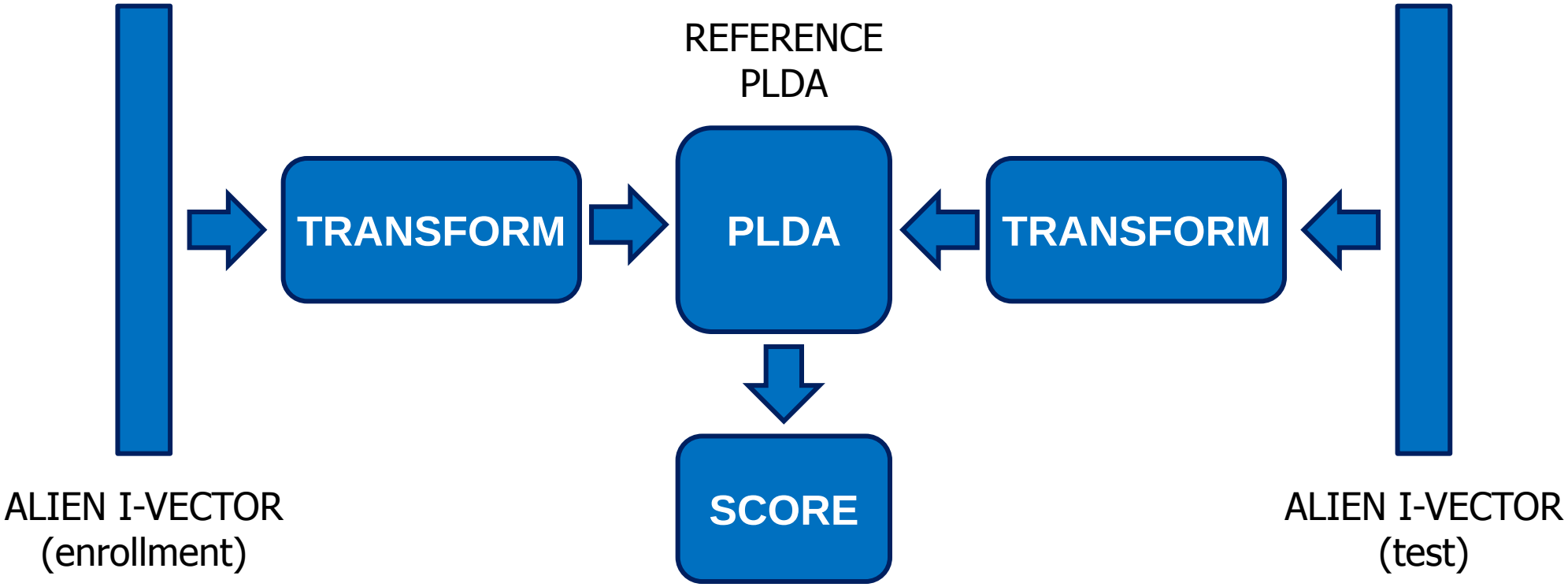
The systems

- **reference**
  - 19 MFCC + C0 + delta + double delta
  - 2048-component GMM
  - 600 dimensional i-vector
  - 9k hours of data (MIX+SW+Fish)
- **Red-ref**
  - 19 MFCC + C0 + delta + double delta
  - 2048-component GMM
  - 600 dimensional i-vector
  - 2k hours of data (MIX only)
- **512/400**
  - 19 MFCC + C0 + delta + double delta
  - 512-component GMM
  - 400 dimensional i-vector
  - 9k hours of data (MIX+SW+Fish)
- **Perseus**
  - 20 Perseus coefs + delta+double delta
  - 2048-component GMM
  - 600 dimensional i-vector
  - 2k hours of data (MIX only)

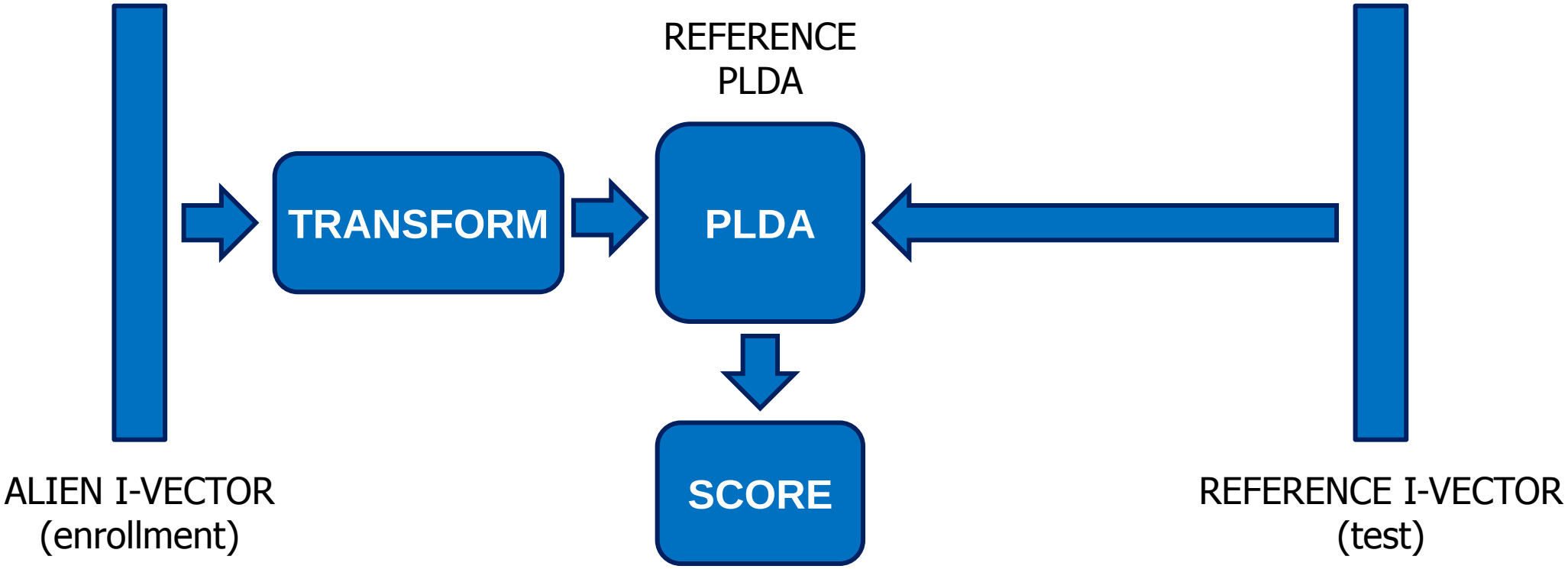
BASELINE TEST



MATCHED TEST



HYBRID TEST


REFERENCE
PLDA
TRANSFORM
PLDA
SCORE
ALIEN I-VECTOR
(enrollment)
REFERENCE I-VECTOR
(test)

The transformation

- We started with one-hidden-layer NN's (simply off-the shelf scripts)
- We experimented with multi-layer NN's
- In the end – no hidden layer = simple linear regression works the best

- Trained using THEANO
- Fixed at MIX+SW+Fish, 9k hours audio data

Performance on various NN topologies

System	DCF^{min}_{new}	DCF^{min}_{old}	eer
reference	0.3834	0.1089	2.13
Perseus on reference	1.0000	0.7834	23.12
Perseus baseline	0.4924	0.1494	2.86
600-600	0.4662	0.1522	2.85
600-600*	0.4490	0.1360	2.64
600-600-600	0.5596	0.1799	3.48
600-600-600*	0.5039	0.1526	2.96
600-1200-600	0.5794	0.1732	3.56
600-1200-600*	0.4834	0.1467	2.93
600-600-600-600	0.5845	0.1898	3.66
600-600-600-600*	0.5045	0.1587	3.09

NIST SRE 2010, condition 5, female

* Hybrid test

Performance of different systems

System		DCF^{min}_{new}	DCF^{min}_{old}	eer
reference	baseline	0.3834	0.1089	2.13
512/400	baseline	0.5711	0.1742	3.78
	400-600	0.5011	0.1548	3.12
	400-600*	0.4555	0.1387	2.76
Red-Ref	baseline	0.4475	0.1283	2.64
	600-600	0.4392	0.1299	2.73
	600-600*	0.4224	0.1213	2.53
Perseus	baseline	0.4924	0.1494	2.86
	600-600	0.4662	0.1522	2.85
	600-600*	0.4490	0.1360	2.64

NIST SRE 2010, condition 5, female

* Hybrid test

Conclusion

- We tried to explore whether i-vectors can be transformed in order to be compatible
- For a selected set of systems, we found that simple **linear regression** can be used to transform the i-vectors
- **Hybrid test** performs generally **better**
- In some cases it can be beneficial to extract i-vectors using one system and score using other

- Future work:
  - more feature types
  - different topologies
  - more experiments

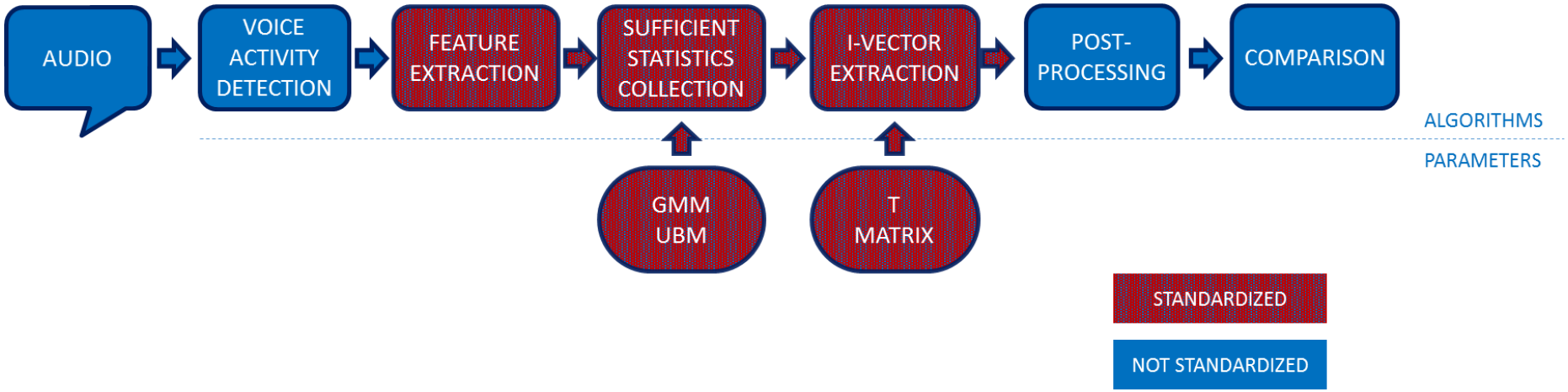
Situation

- In the last 10 years, scientific advances in speaker recognition (JFA, i-vectors, PLDA) allowed for producing precise and robust SRE systems
- Quickly adopted by vendors, producing solutions that **are successful**
- R&D never stopping
  - Everyone continuously improving performance of their system, robustness, calibration, etc
  - New versions of engines released

A vibrant community working in cooperative/competitive mode both for R&D labs and vendors.

What is needed

- Fix the core iVector extraction algorithms
- Fix the necessary parameters
- Do the necessary minimum, let people freedom to use their (own, best) VAD and scoring.



We WANT

- Users
  - Having interoperable systems
  - Being able to exchange speaker information without compromising content
  - within companies/agencies, across companies/agencies and across borders
- Vendors
  - Increasing the whole market (think about introduction of USB!)
- R&D labs
  - sharing i-vectors between labs without lengthy discussions on configuration (not excluded though!)
  - Giving a working recipe to juniors to play with.
  - Obtaining massive data from the users

We DON'T WANT

- stop R&D (both academic and commercial) of speaker recognition technology by saying that this will be the only iVector extraction scheme forever.
  - all of us are trying to push the field further, sometimes as collaborators, sometimes as competitors.
  - We want to define a snap-shot of the best practice up to day on which we could agree.
- Earn money on licenses or patents – the proposed standard is license and patent-free
- Have something too complex and too relying on a proprietary and/or 3rd party technology.
- Present this as an ultimate forensic solution.

What is there

- http://voicebiometry.org/ - technical description, Python code with all necessary parameters
  (feature extraction, UBM, T-matrix)
- Google group http://groups.google.com/d/forum/voice-biometry-standard - please subscribe

THANK YOU