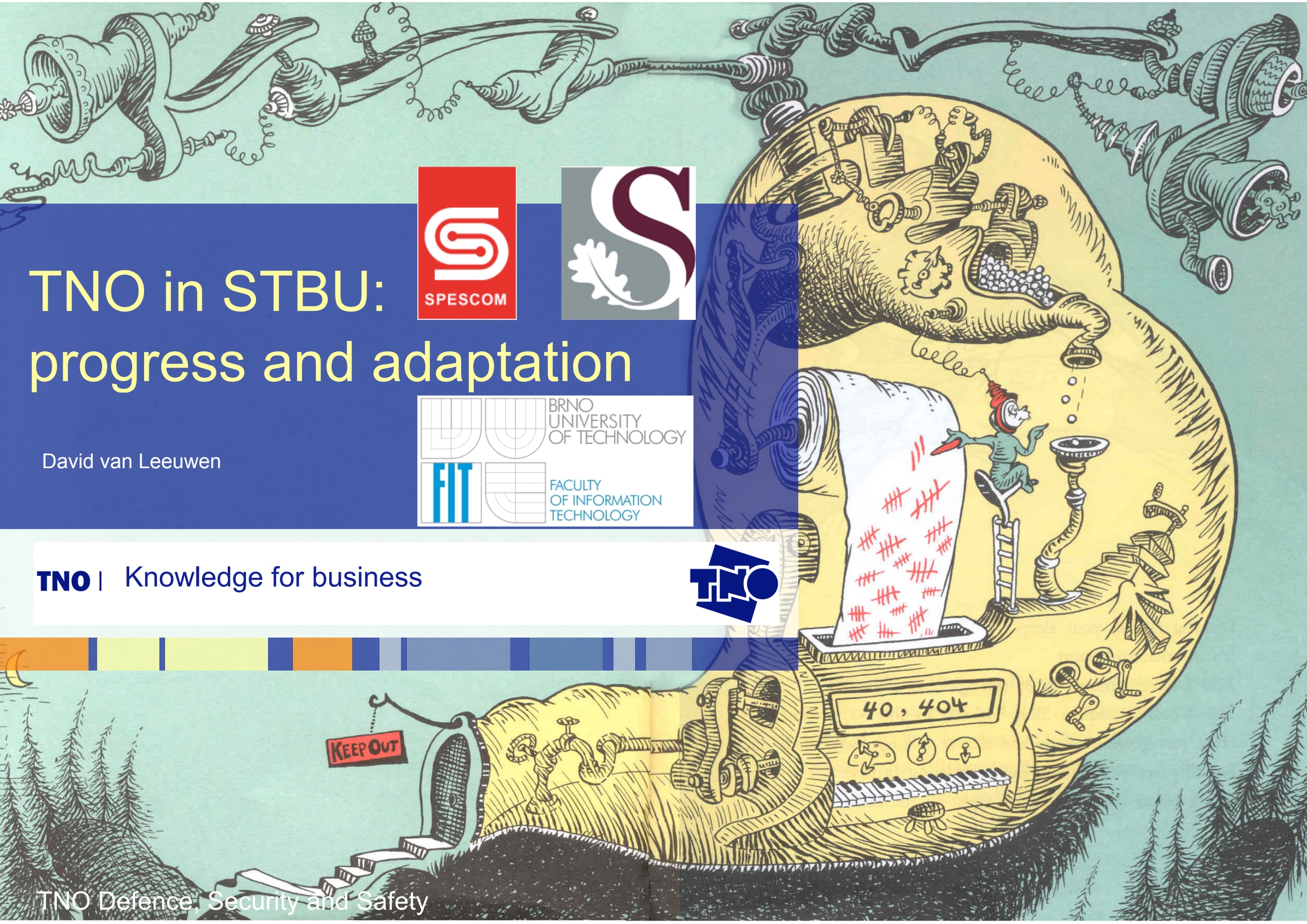
TNO in STBU: progress and adaptation

David van Leeuwen

BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

TNO | Knowledge for business

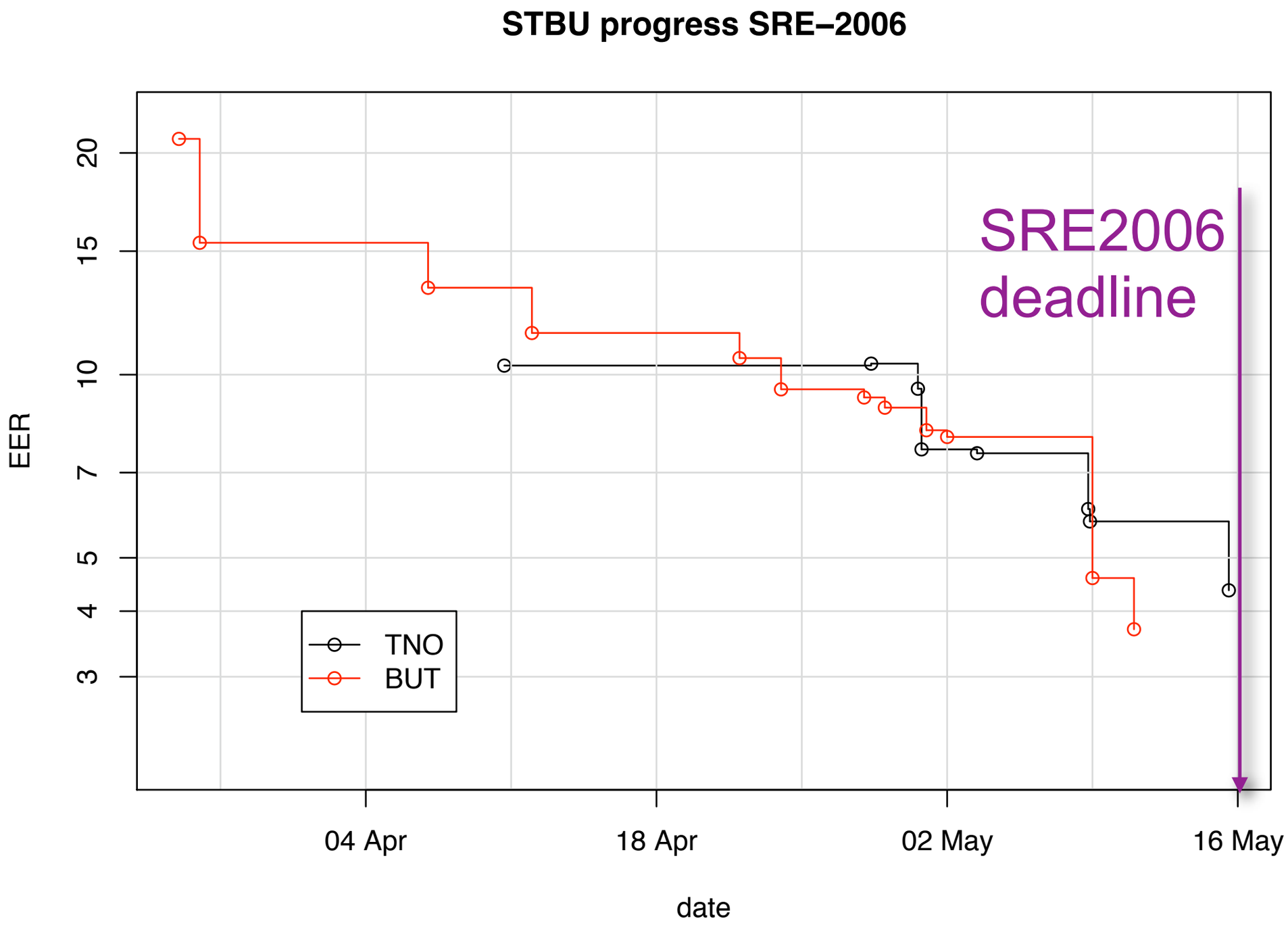


TNO Defence, Security and Safety

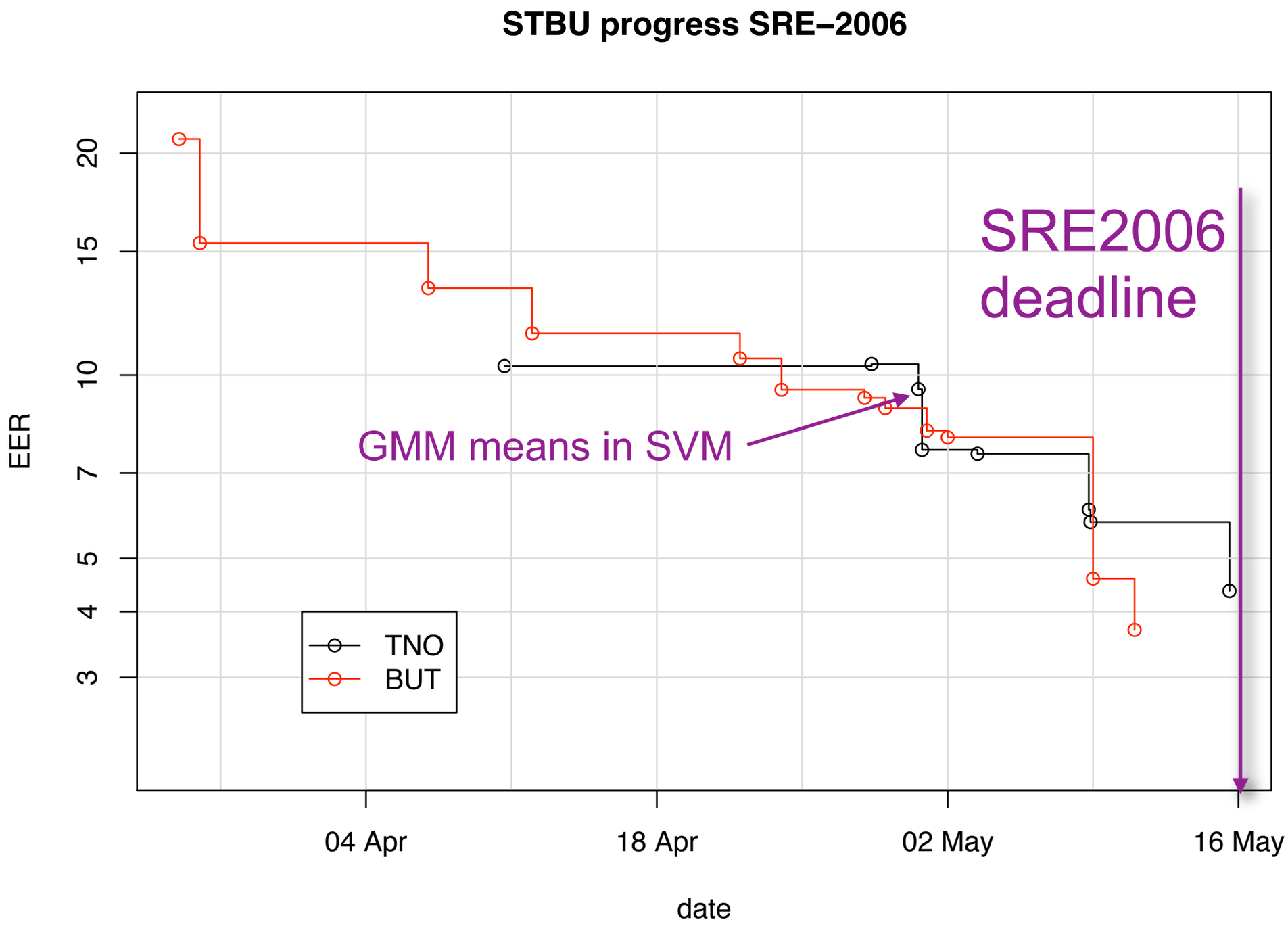
The bottom line: devtest (SRE-2005) results, evaluation (SRE-2006) results

	EER	min DCF
sys2005: GMM-2048, UBM 591s, FM 8c, T-norm	10.32%	0.0390
GMM-512, UBM 1640s, FM 16c, T-norm	10.39%	0.0388
+ GMM means in SVM, no T-norm	7.64%	0.0304
+ T-norm	7.53%	0.0260
+ channel NAP	6.08%	0.0214
+ T-norm	5.79%	0.0189
+ unsupervised adaptation	4.37%	0.0124
SRE-2006, all trials	5.48%	0.0290
core condition (English trials, det3)	4.06%	0.0204

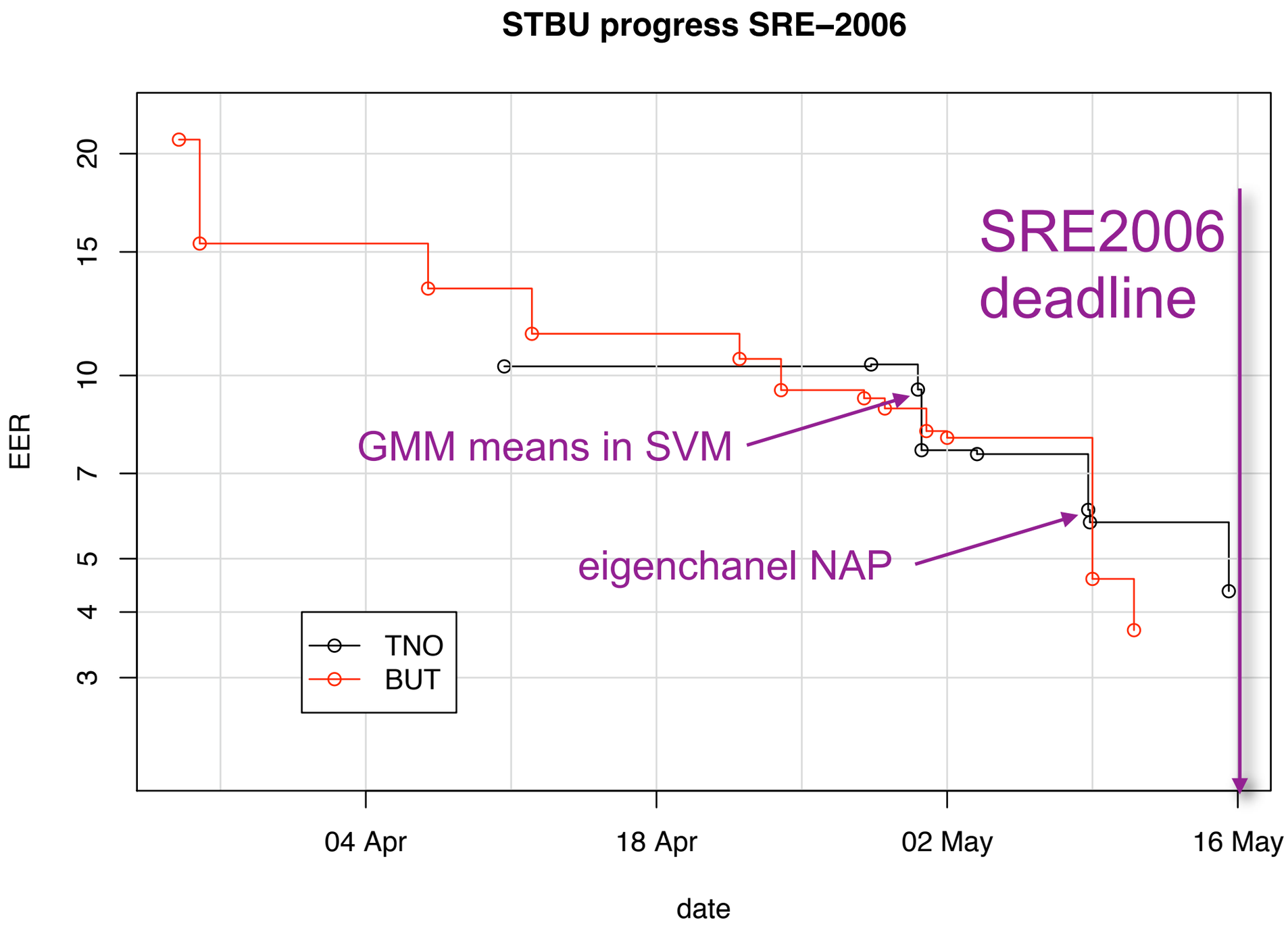
STBU interaction process



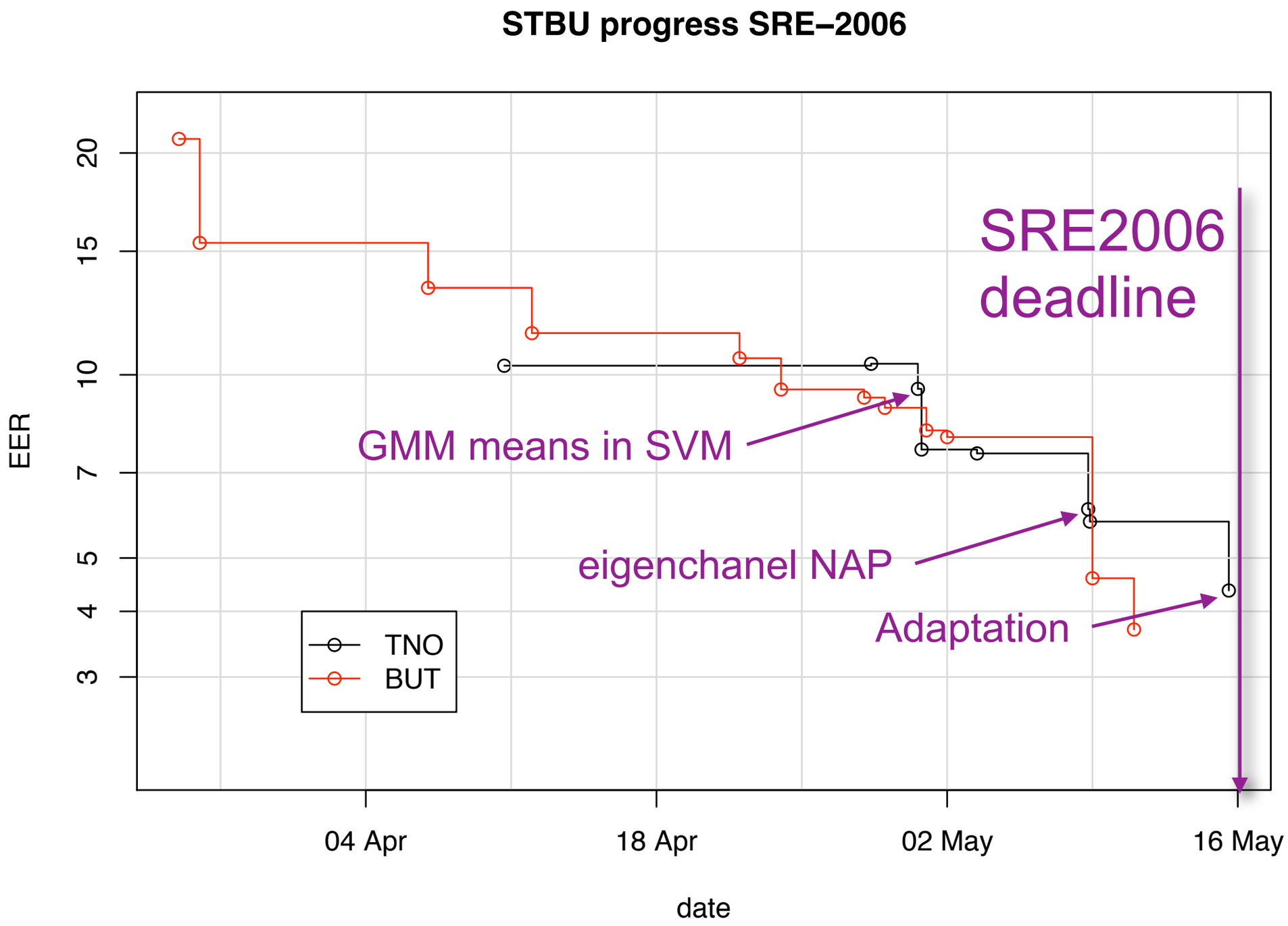
STBU interaction process



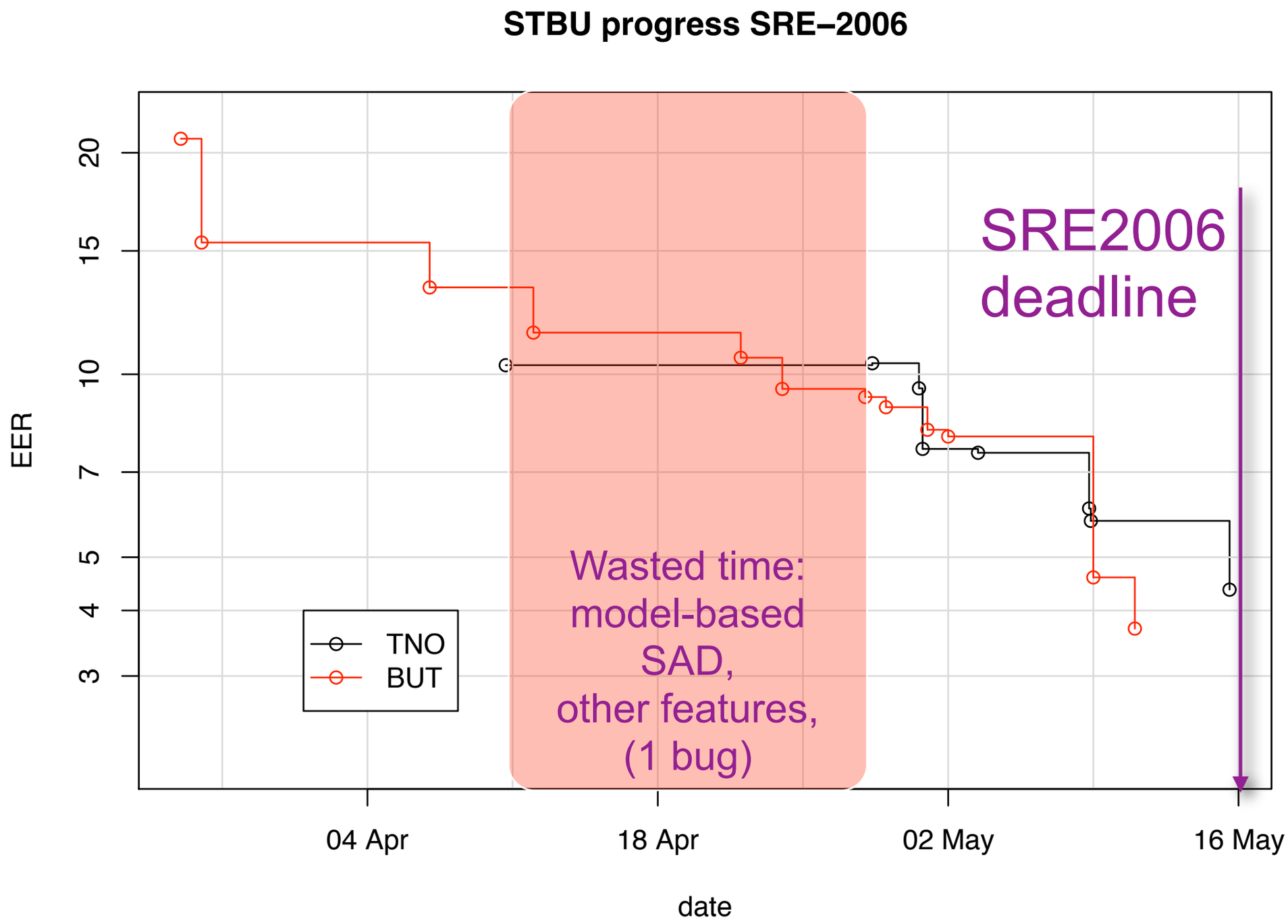
STBU interaction process



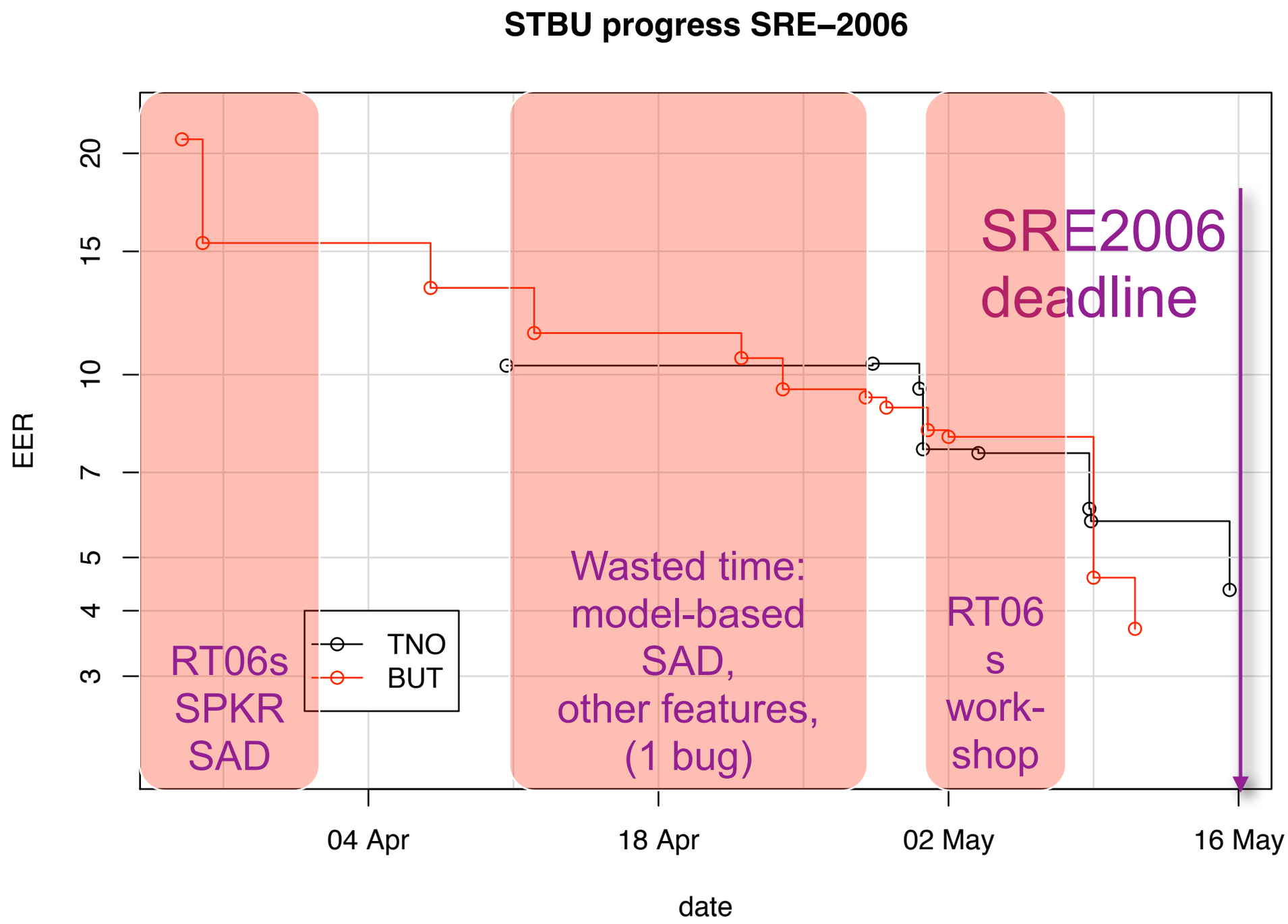
STBU interaction process



STBU interaction process



STBU interaction process



Speakers

LDC Switchboard II phase 3

NIST SRE 2001–2003

NIST SRE 2004

LDC Fisher English PIN>10000

NIST SRE 2005

Speakers

UBM, FM, SVM background

LDC Switchboard II phase 3

NIST SRE 2001–2003

NIST SRE 2004

LDC Fisher English PIN>10000

NIST SRE 2005

Speakers

UBM, FM, SVM background

LDC Switchboard II phase 3

NIST SRE 2001–2003

LDC Fisher English PIN>10000

t-norm

NIST SRE 2004

NIST SRE 2005

Speakers

UBM, FM, SVM background

LDC
Switchboard II
phase 3

NIST SRE
2001–2003

LDC Fisher English PIN>10000

t-norm

NAP training

NIST SRE
2004

NIST SRE
2005

Speakers

UBM, FM, SVM background

LDC Switchboard II phase 3

NIST SRE 2001–2003

LDC Fisher English PIN>10000

t-norm

NIST SRE 2004

NAP training

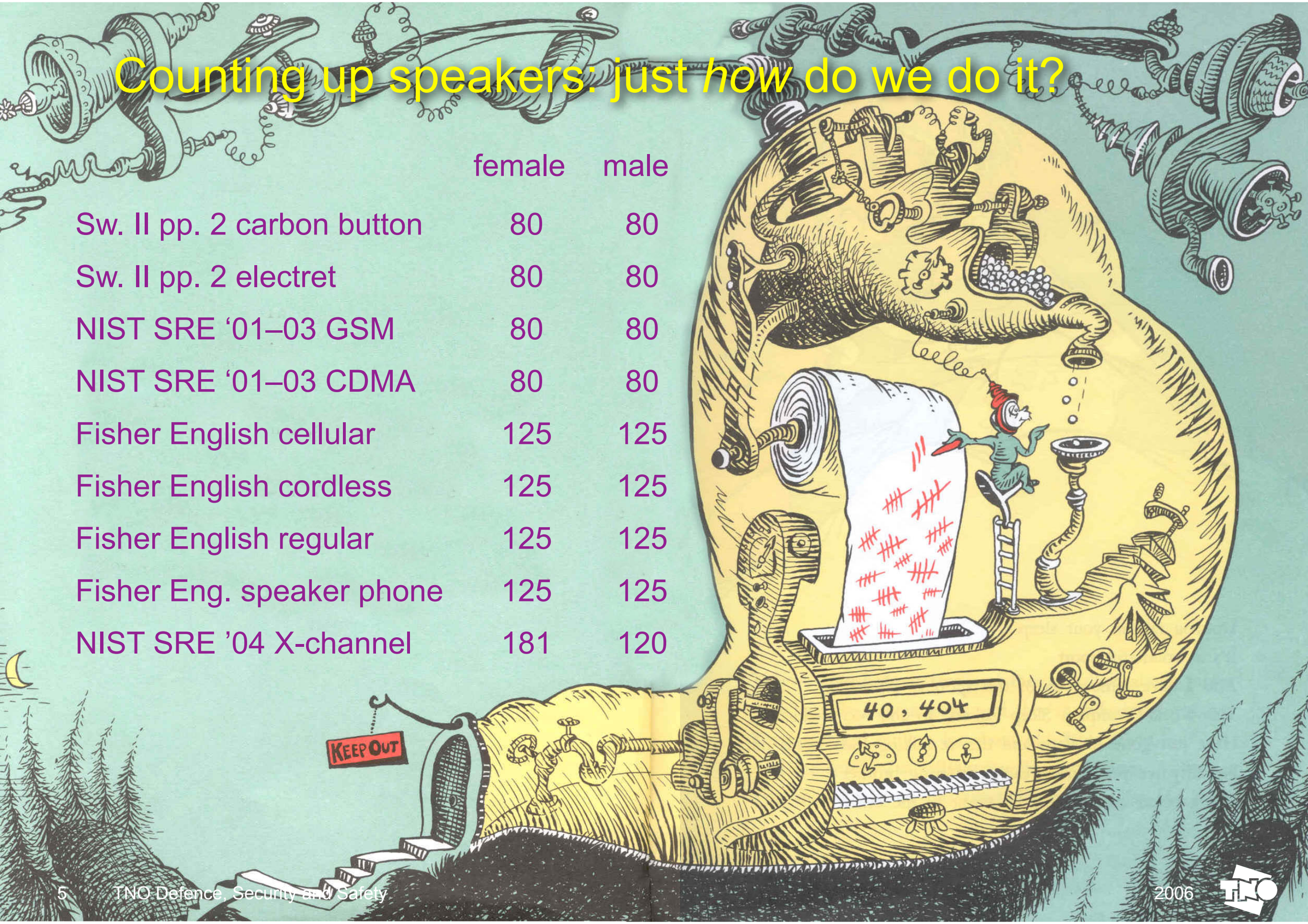
NIST SRE 2004

Calibration

NIST SRE 2005

Counting up speakers: just *how* do we do it?

	female	male
Sw. II pp. 2 carbon button	80	80
Sw. II pp. 2 electret	80	80
NIST SRE ‘01–03 GSM	80	80
NIST SRE ‘01–03 CDMA	80	80
Fisher English cellular	125	125
Fisher English cordless	125	125
Fisher English regular	125	125
Fisher Eng. speaker phone	125	125
NIST SRE ’04 X-channel	181	120



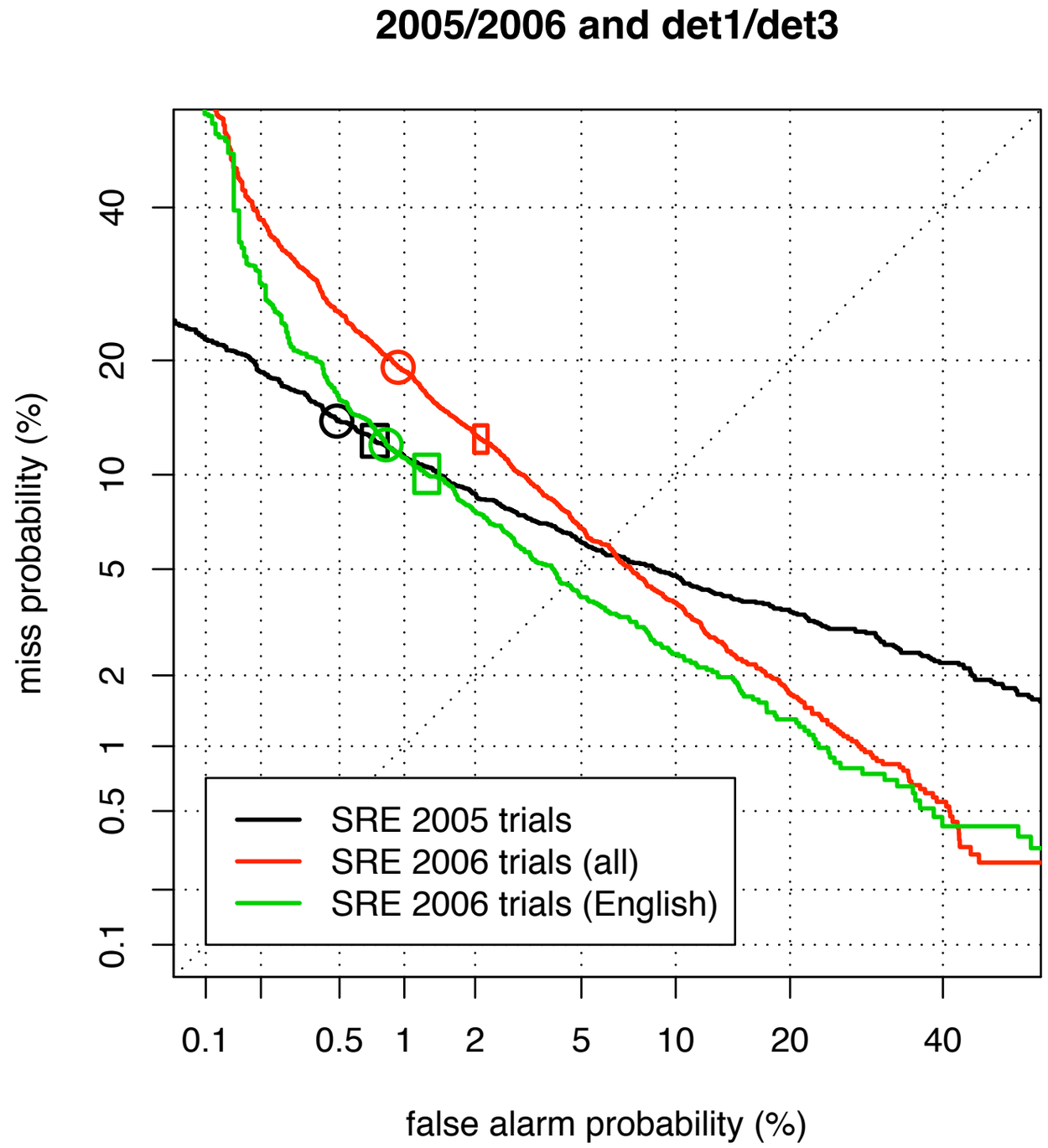
Main observations: TNO N-mode

- SRE 2006 rotated clockwise
  - same EER
  - higher C_{DET}
- English only: easier

Why?

- English UBM?
- X-language dependence?
- Effect of NAP?

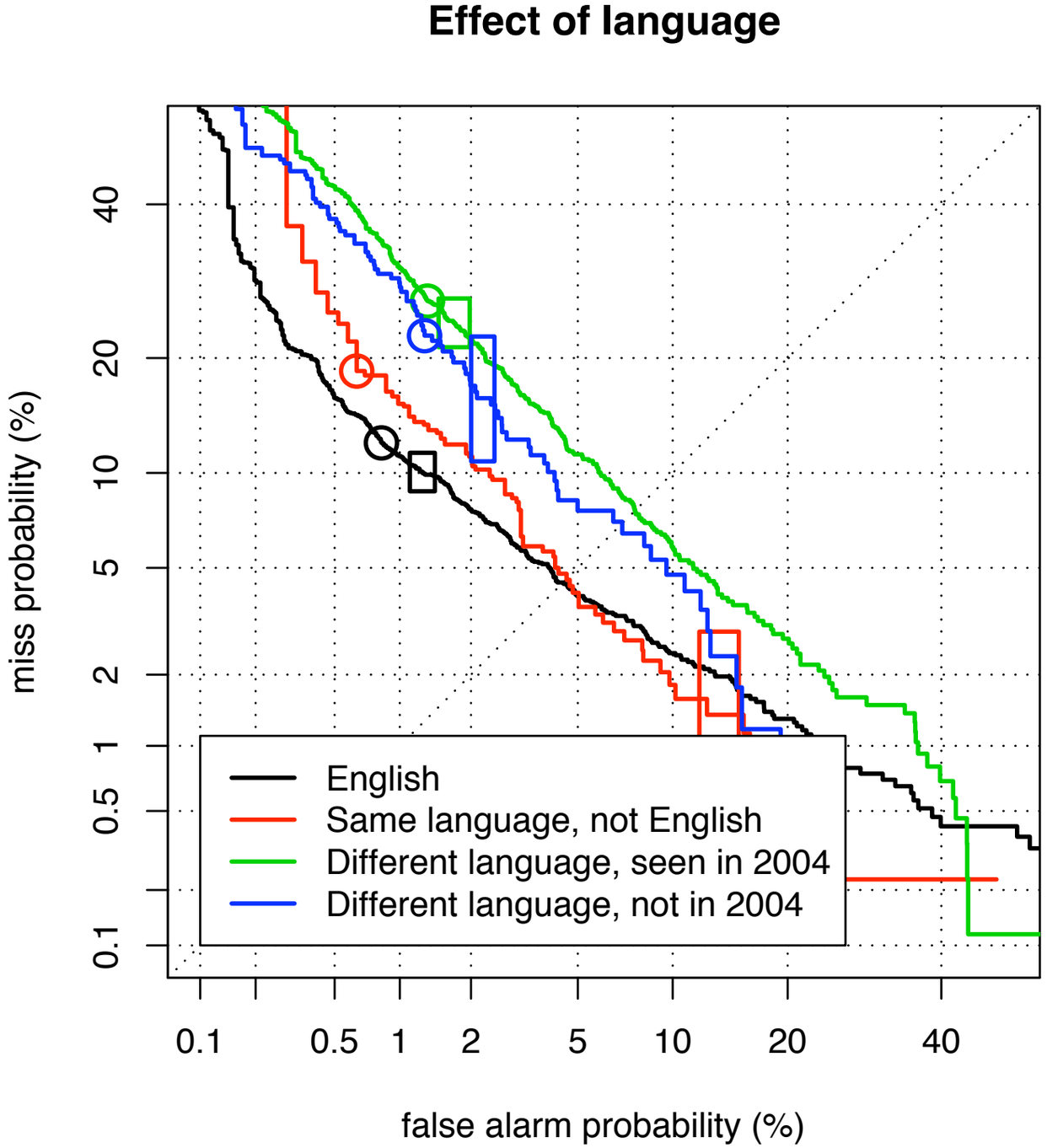
➡2004 languages



Language dependence

- English *not* more easily detected
- Calibration of non-English is off!
- 2004 X-languages *not* easier than new ones

➡NAP did not project X-language effect away



Obtain yourself an EER < 6% within 24 hours

- Collect speakers and data, and 2 key papers William Campbell
- Extract your favorite features
  - Train UBM, obtain UBM-indices for all speakers
- Do (fast) MAP-adaptation of all speakers, stack GMM means into super-vector (SV) with some scaling
  - For all SRE-2004 speakers
    - collect all conversation sides
    - subtract mean super-vector over speaker
  - combine into matrix Δ, compute ‘top 40’ eigenvectors *S* of $\Delta\Delta^T$
- Project all SV’s along *S* using operator $I - SS^T$
- Build SVM for each model speaker, fold model into one vector
  - T-norm models
  - train models
- Score test segments, T-norm
- Perform score to LLR conversion

Some notes on efficiency (courtesy of Niko Brümmer)

- UBM index is essential
  - top-*N* scoring (✘), FM (✘/✔), fast-MAP (✔)
  - no need to evaluate `exp()`
- Fast MAP-adaptation of UBM
  - like top-*N* scoring
  - in ‘E-step’ only compute posterior per component of top-*N* Gaussians

- Calculation of NAP eigenvectors
  - Covariance matrix $\Delta\Delta^T$ is large $(N_{fea} \times N_{gauss})^2 \approx 13k^2$
  - top *M* e.v. $\Delta\Delta^T \approx \Delta$ (top *M* e.v. $\Delta^T\Delta$)
  - ARPACK or Matlab `eigs()` only needs function $f(x) = \Delta^T\Delta x$
  - calculate $\Delta^T\Delta x$ as $\{ (\Delta x)^T \Delta \}^T$

- Calculate projection $(I - SS^T)x$ as $x' = x - S(S^T x)$

The continuing story of unsupervised adaptation
(aka U-mode)

- History:
  - 2003: proposed by Claude Barras (LIMSI) at workshop
  - 2004: 3 sites tried, hardly any positive effect
    - setting threshold was difficult (new data collection)
  - 2005: 1 site tried, clear positive effect
    - in discussion proposal to allow U-mode as primary system
  - 2006: 5 sites tried, 2 (STBU and TNO) designated as *primary*
    - risky, because of calibration issue

- Method still the same
  - process trials in order
  - if T-normed score exceeds threshold *a*
    - 1conv: MAP adapt means using *test* segment, relevance *r*, new SVM
    - 8conv: add *test* segment to *train* list, new SVM

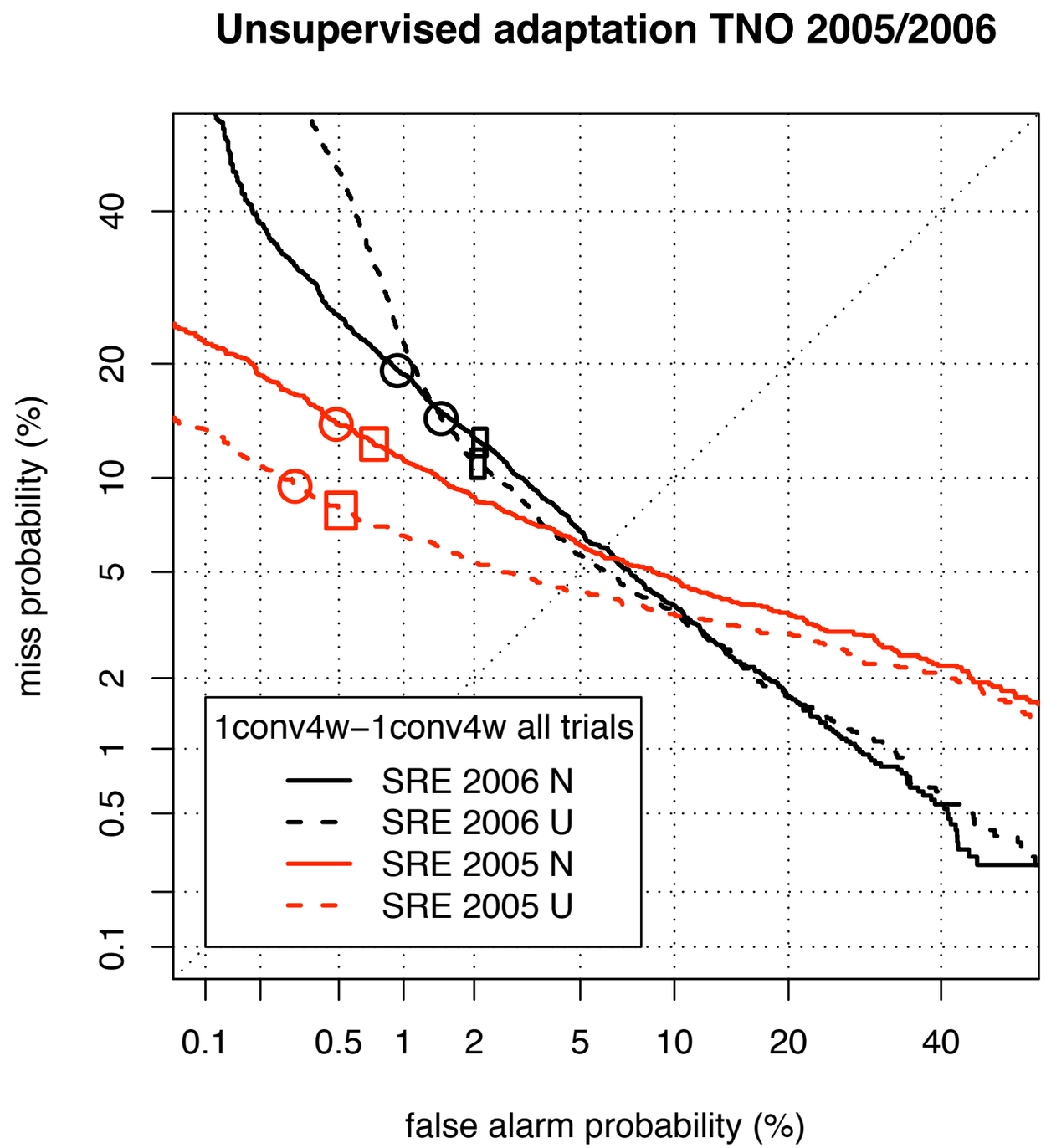
This year's challenge: pathological files
(courtesy of QUT)

- Any form of interaction with the data is *not* allowed
  - People started complaining about (almost) empty, zeroed, identical, files
  - Some GMM-means became *NaN* (bug?), SVM training did not finish
- For adaptation, a pathological file can ruin the model
  - identical files: too much weight to conversation
  - empty files: tend to give very high scores when trained on

- Algorithm
  - File is *pathological* if either
    - all frames have energy > max energy – 30dB
    - occurs in list sent out by QUT
    - raw SVM score > 0.95
  - Then: no adaptation, LR = 1

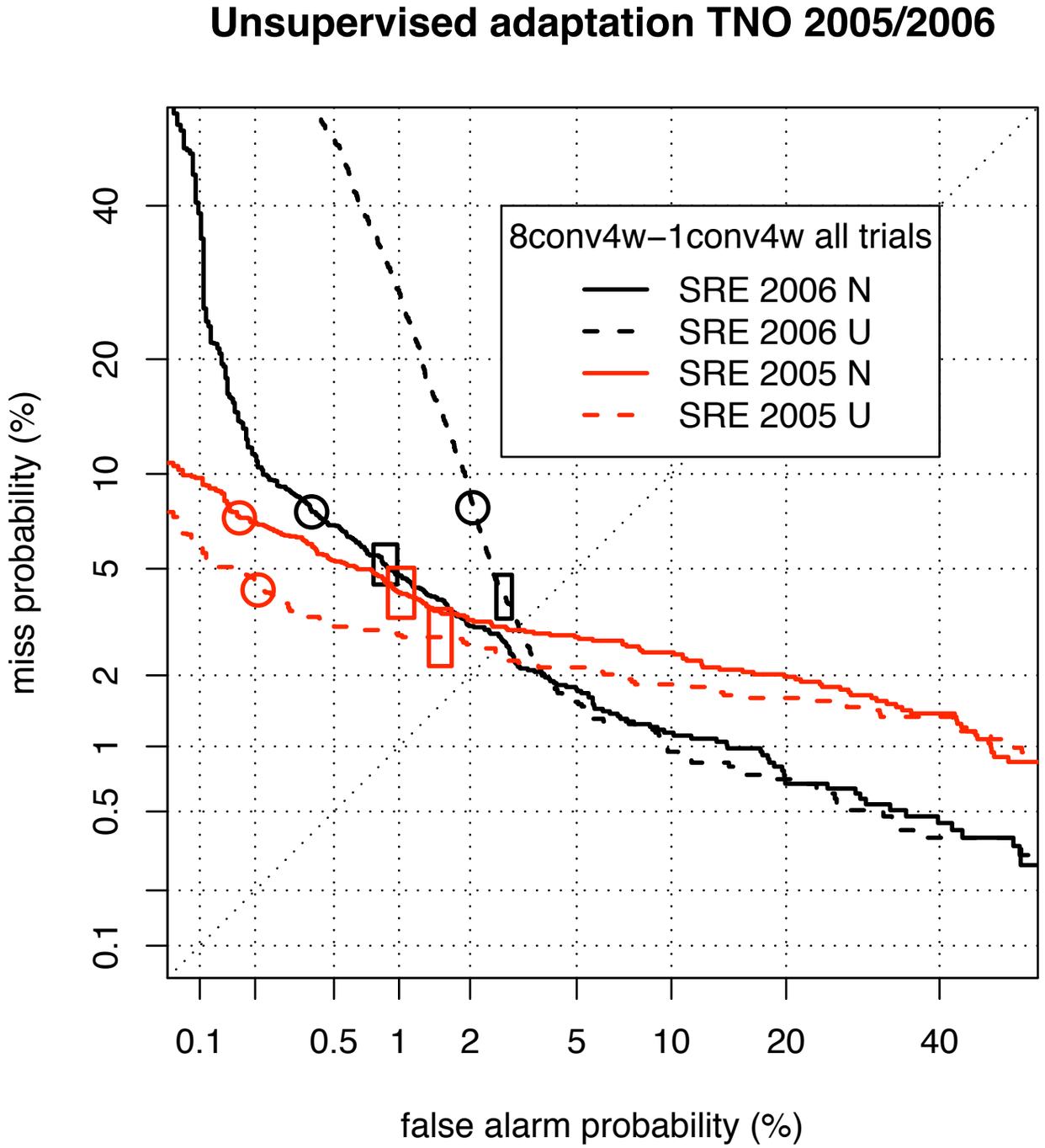
So again, it worked a little bit

- Calibration threshold *a* was OK
- Effect smaller in evaluation
- Did not help/hurt in STBU fusion at C_{DET}



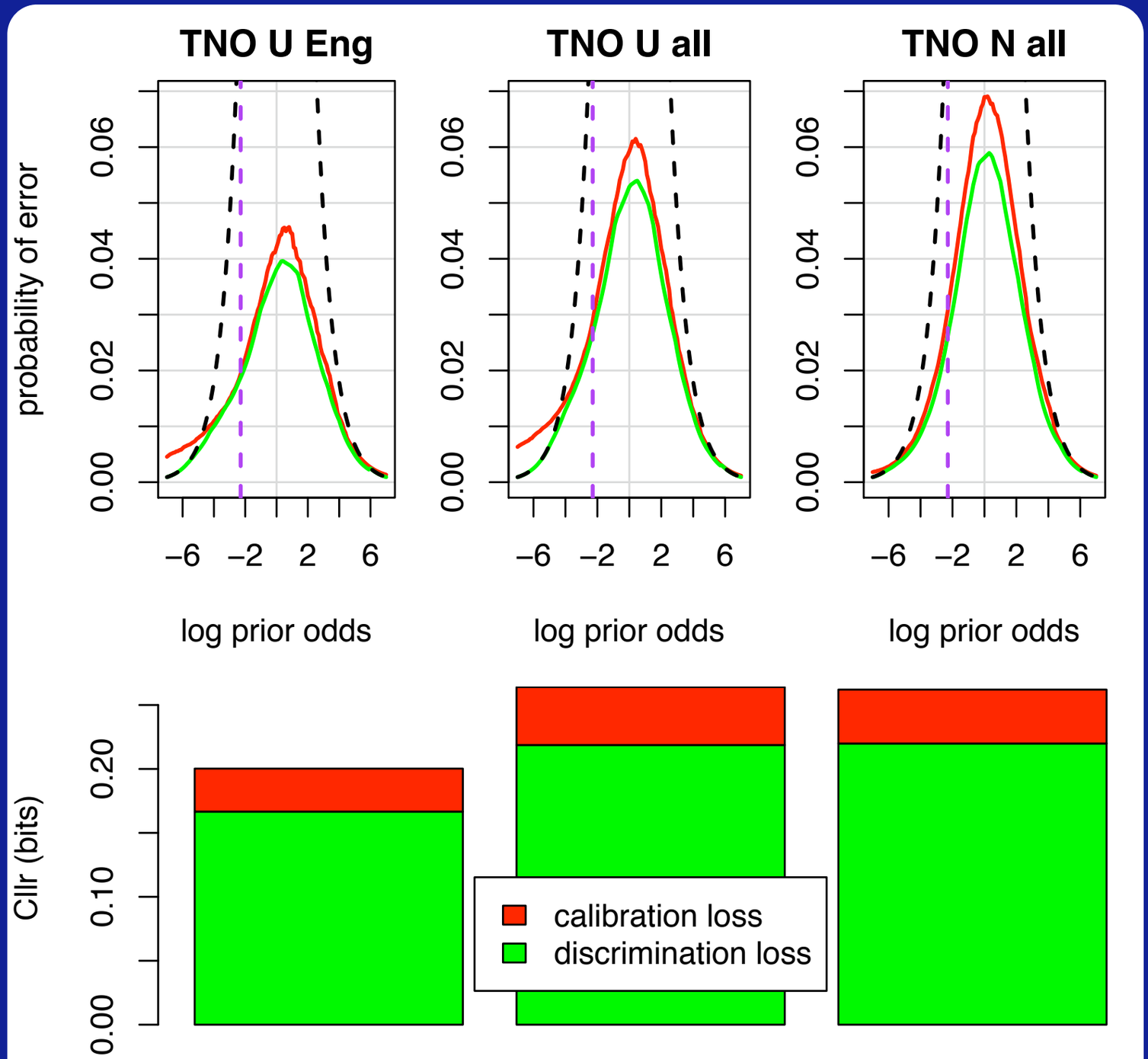
And again, there were problems

- Calibration threshold *a* was OK
- Effect smaller in evaluation
- Did not help/hurt in STBU fusion
- But it didn't work for 8conv4w training in the evaluation!



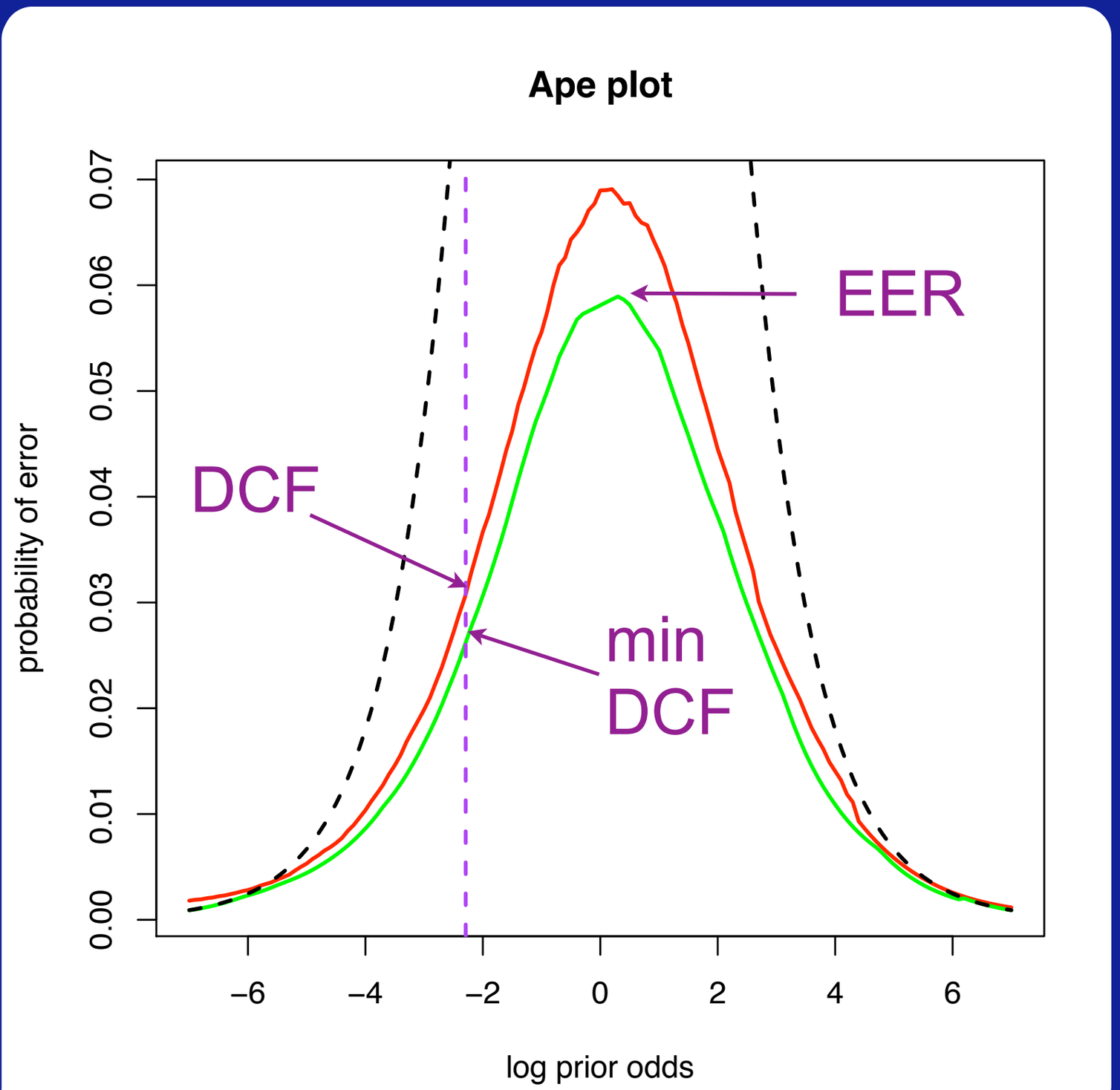
Calibration

Applied Probability of Error shows:



Calibration

Applied Probability of Error shows:

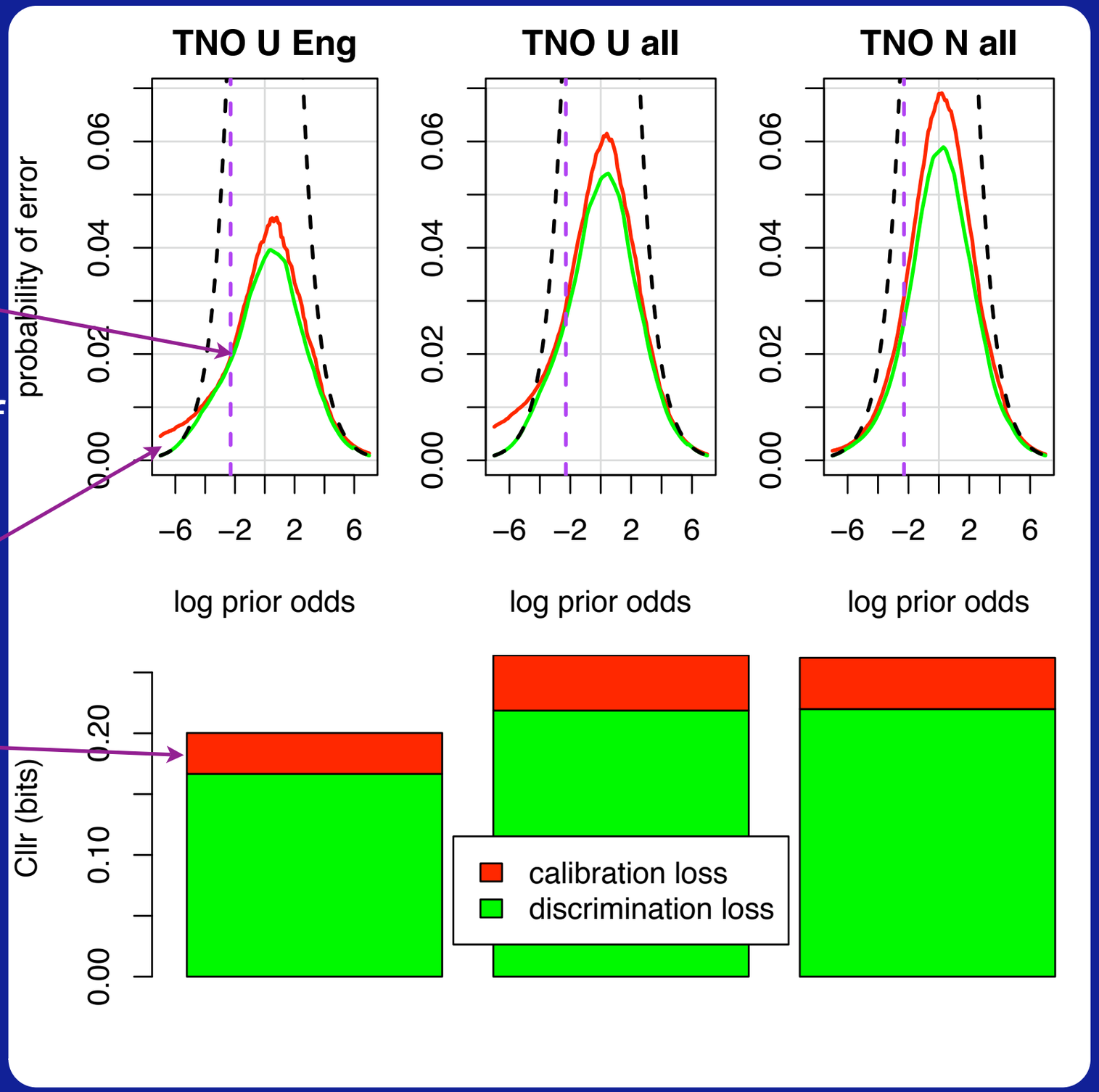


Calibration

Applied Probability of Error shows:

- good calibration around DCF
- fair calibration over wide range of priors
- U-mode in low odds range miscalibrated
- Overall little calibration loss

- we used *FoCal* with one source



Conclusions

- It is very useful to work in a team and share
  - tedious preparation work
  - papers, ideas, understanding, results
- even more than when just sharing *scores*

- MIT's GMM means in SVM is great
- CRIM/SDV/QUT/MIT's eigenchannel/NAP is great

- Choice of speakers for background, T-norm, NAP is important
- Unsupervised adaptation still has interesting challenges
  - calibration
  - algorithm

- FoCal calibration seems fairly robust, calibration over *range* priors
- Is NAP robust against *data collection*?