Reducing the “Horseness” of Music Information Retrieval methods

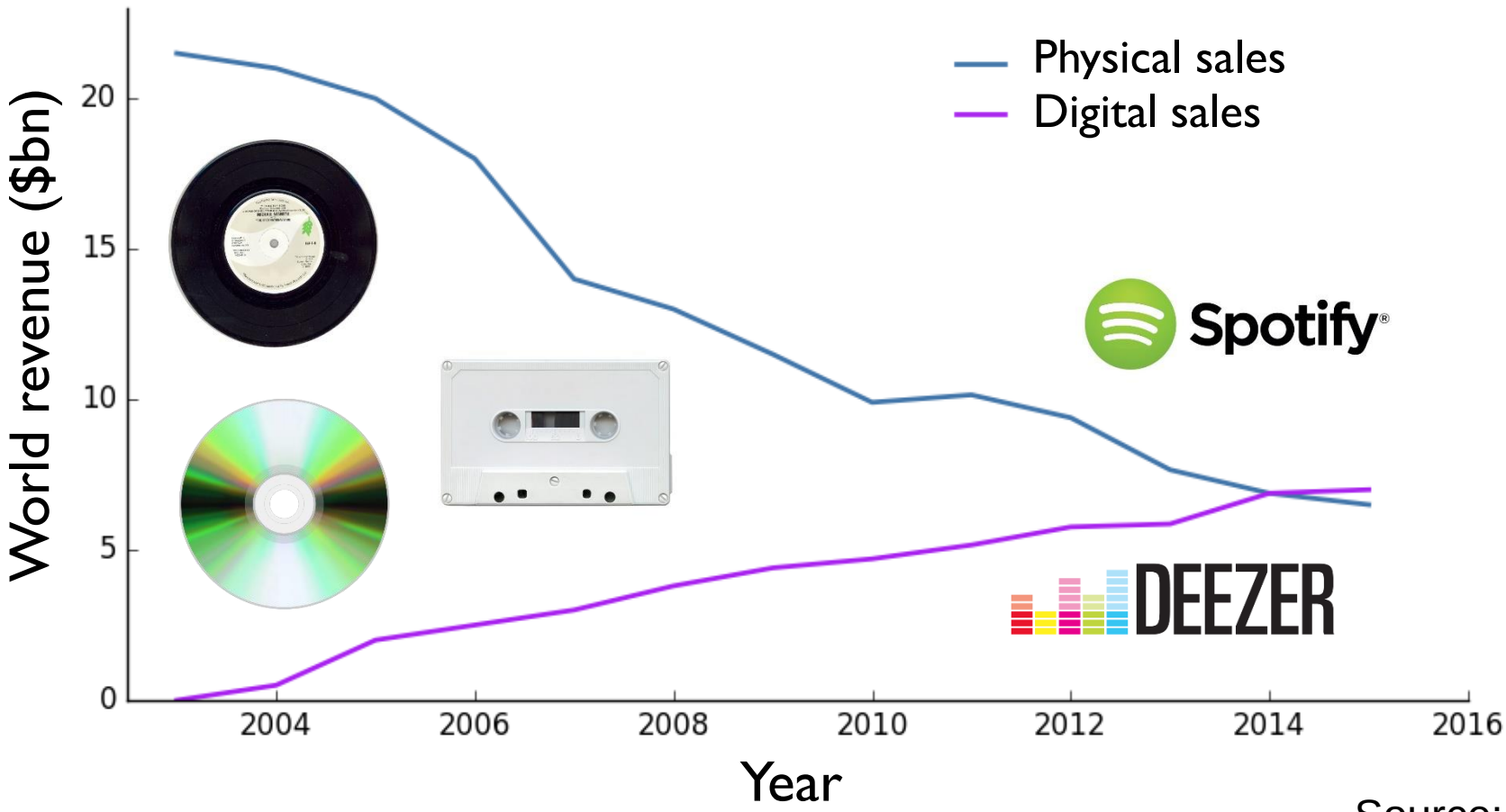
PhD Thesis in I.T. applied to music

Yann Bayle

September 20th 2017

HORSE2017

Musical industry



Source: IFPI

Streaming

Recommendation

Playlist

- ♪ **Genre** (Rock, Blues, …)
- ♪ **Mood** (Joy, Nostalgia, …)
- ♪ **Activity** (Sport, Work, …)
- ♪ **Top 100**
- ♪ **Celebrities** (« Obama », …)

Tag tracks

Music tagging

Methods	Advantages	Drawbacks	Examples
Manual (editor)	Precise	Little	PANDORA
Manual (community)	Plenty	Incorrect Ambiguous Abuse	SOUNDCLOUD last.fm
Automatic (data usage)	Precise	Coverage	Spotify
Automatic (autotagging)	Coverage	Precise	DEEZER

Goal

Enhance autotagging for music recommendations

Focus on Instrumentals and Songs

Tools for development

- ♪ Database Management
- ♪ Signal processing
- ♪ Machine learning
- ♪ Statistical analysis

Test with industrial partners

simbals

DEEZER

RECISIO
sing the world

How to guarantee « Horsefree » methods?

"a *horse* is just a system that is not actually addressing the problem it appears to be solving." (Sturm 2014)

Example

Song/Instrumental classification

Precision on Instrumental detection

*25

Dataset	Algorithm	Precision (%)
1,677 tracks (MSD)	SVMBFF (Gouyon *et al.*, 2014)	82.0
41,491 tracks (SATIN)		12.5
	Random prediction	11.0
	Bayle *et al.*, (2017)	82.5

- ♪ SVMBFF: 68 features per track
- ♪ Proposed algorithm: 39 features per frame

Is bigger better?

Dataset

- Diversified
  - Sources (Cross-dataset comparison)
  - Samples (Representative)
- Deep learning approaches require a lot of data

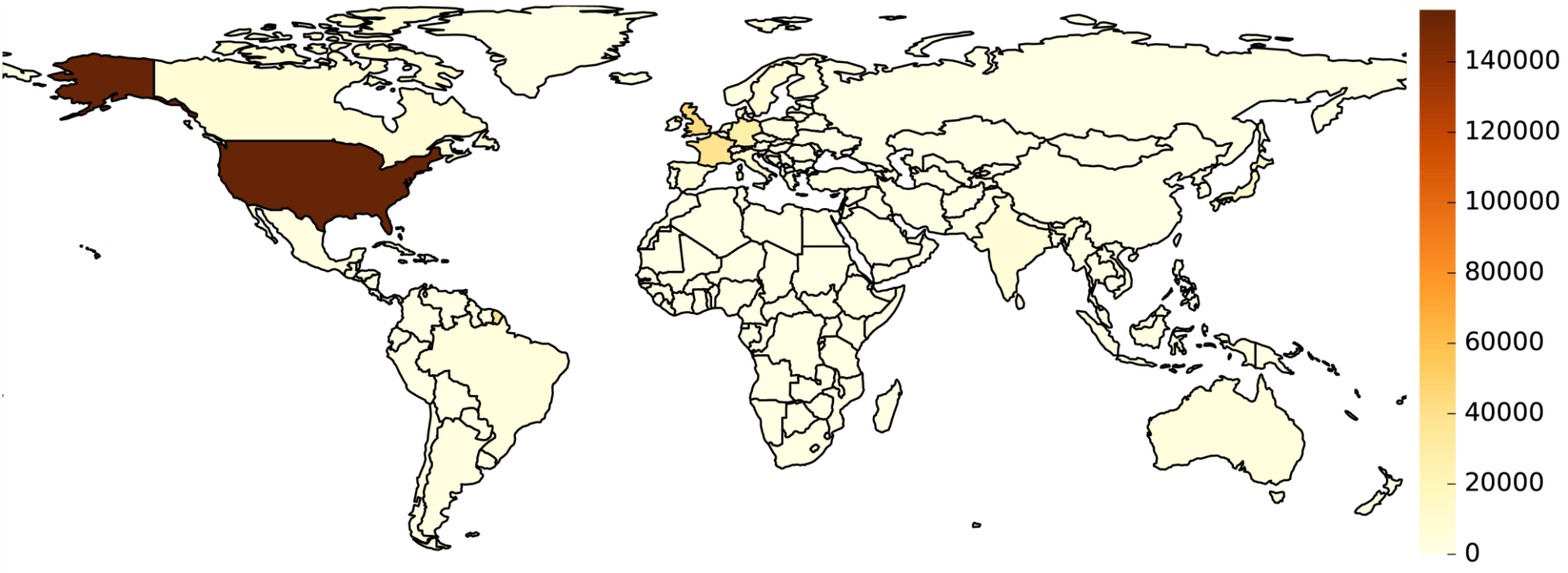
Image research field

- ~2bn images
- *Duplicate Discovery on 2 Billion Internet Images* (Wang *et al.*, 2013)

Music research field

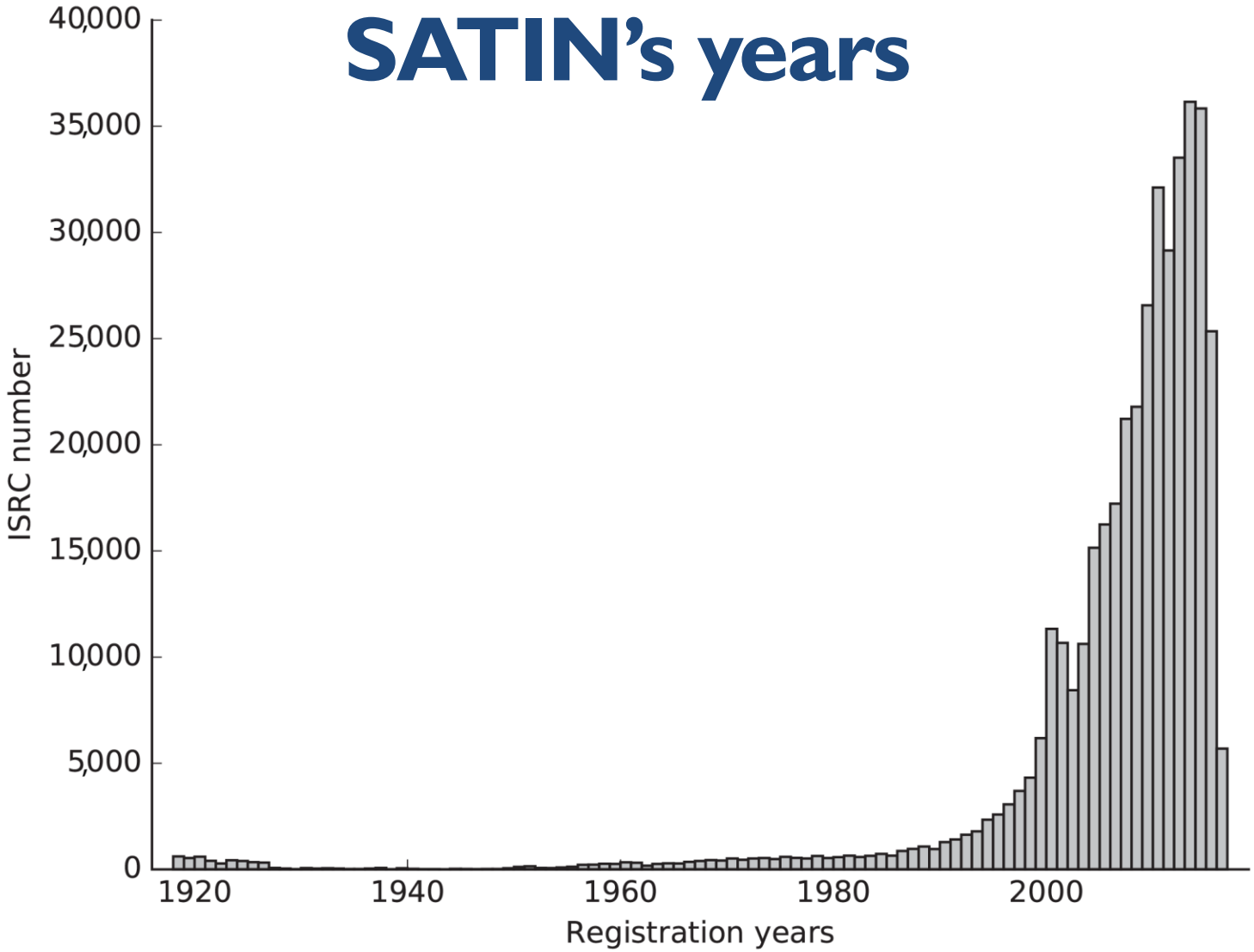
- Deezer: 40M tracks under copyright
- AcousticBrainz: features for 2.7M tracks
- FMA: 106k tracks available for the research community

SATIN's world repartition



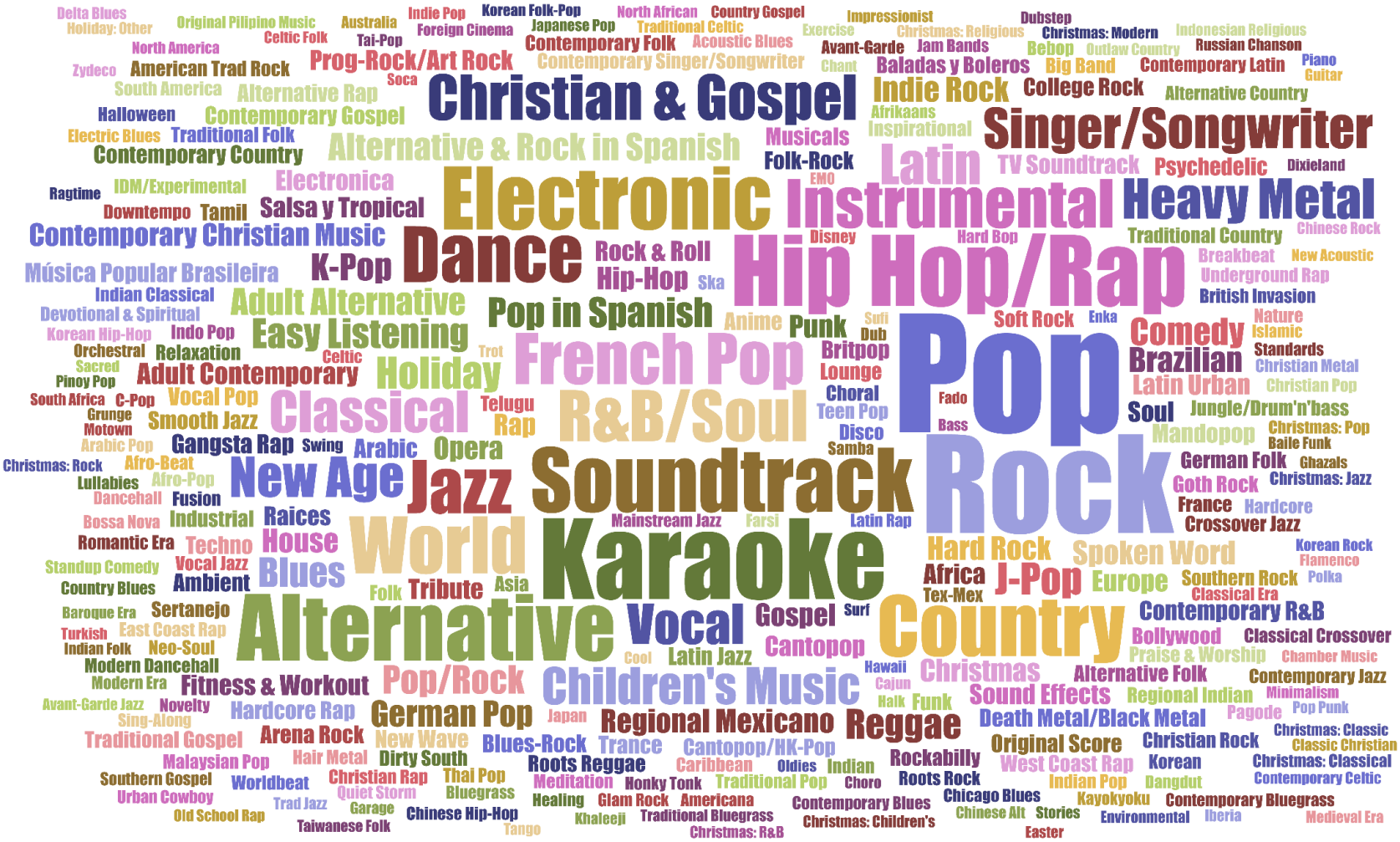
♪ Bias toward western music

SATIN's years



♪ Bias toward 21st century music

SATIN's wordcloud



♪ Reduce genre bias

Bigger but not too big!

Artist and album filtering

- *A closer look on artist filters for musical genre classification* (Flexer 2007)
- Detect studio recording and mastering signature
- Up to which point to filter?
- Human can distinguish song from same artist with 20 albums?
- Filtering reduce the dataset

Data augmentation

- Copyright restriction and filtering reduce the dataset size
- Artificially increase the dataset (pitch, speed, add noise, filter,…)
- A software framework for musical data augmentation (McFee *et al.*, 2015)
- Work in progress: Adding phase-based data augmentation for NN with raw signal as input

Human annotations

Quality

- Track-level (track from 30s to 12m)
- Frame-level (sample precise to seconds)
- *Evaluating Hierarchical Structure in Music Annotations* (McFee *et al.*, 2017)
- From ground truths to L-measure: multi-annotators and multi-level aggregation.

Objective and subjective

- Subjective: Genre, Mood, Activity…
- Objective: Instrumental/Song
- "The tags Vocals and Non-Vocals are well-defined and relatively objective, mutually exclusive, and always relevant." (Gouyon *et al.*, 2014)

Definitions

Oxford dictionary

- **Song**: A short poem or other set of words set to music or meant to be sung
- **Instrumental**: music performed on instruments, with no vocals

Notes

- The voice is an instrument
- What about **humming**?
- **Scat**: Improvised jazz singing in which the voice is used in imitation of an instrument
- A **Song** is a musical piece containing human voice, whereas an **Instrumental** does not.

Examples

- Joe Satriani – Crow chant (cf music excerpt)
- Michael Gregorio (cf video)
- **Objective** definition but **subjective** perception?

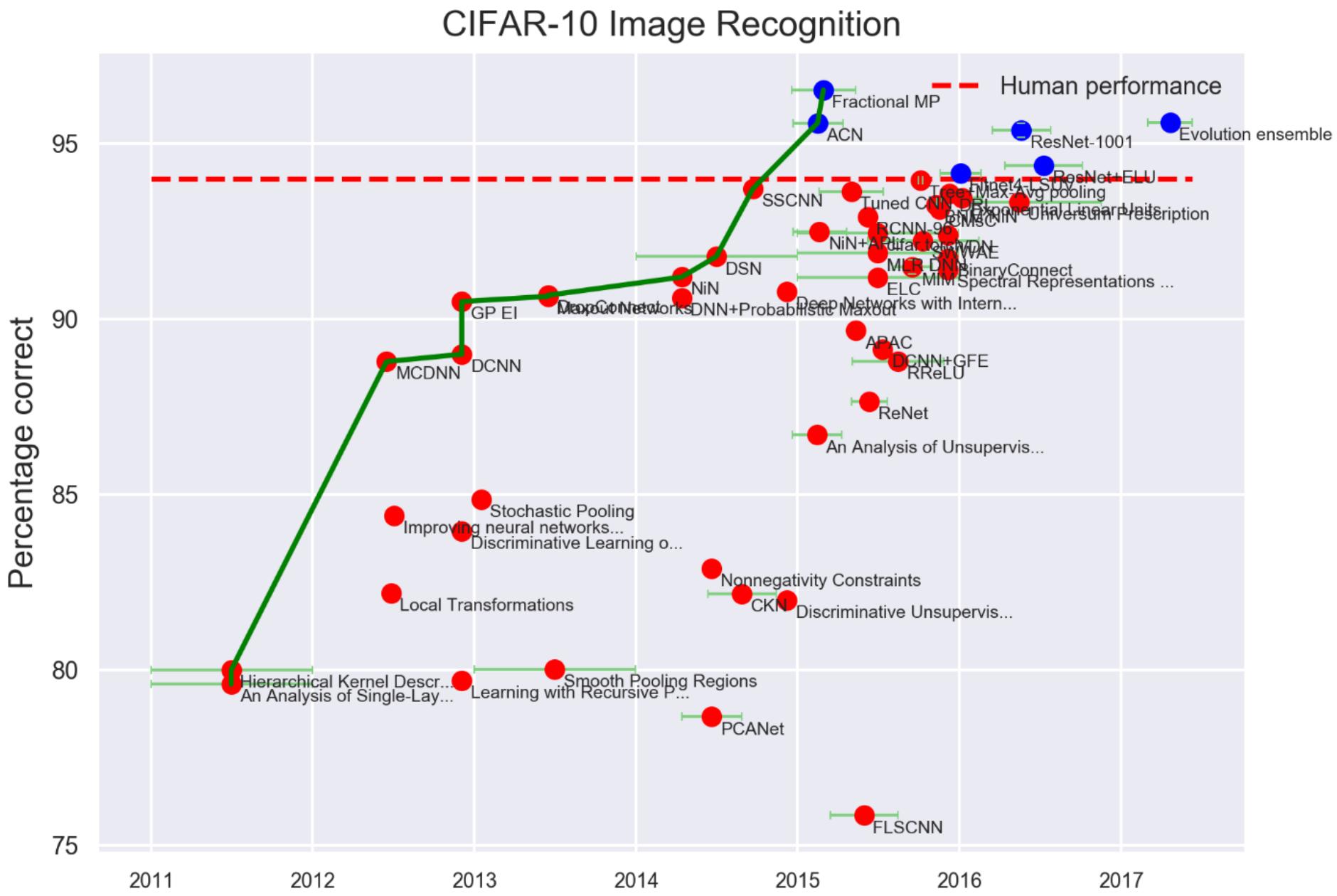
Can we measure "Horseness"?

Comparison to baseline

- **Human** detection performances
- **Random** classification (on the **dataset**)
- **Random** input (in the **system**)

Project « AI Metrics »

- Human detection threshold comparison
- State-of-the-art per task in multiple fields
  - video games, image, video, music,…
- https://github.com/ai-metrics/ai-metrics


CIFAR-10 Image Recognition
Human performance
Percentage correct
95
90
85
80
75
2011
2012
2013
2014
2015
2016
2017
Fractional MP
ACN
ResNet-1001
Evolution ensemble
ResNet+ELU
Tree+Max-Avg pooling
Universum Prescription
SSCNN
Tuned CNN
RCNN-96
NiN+APL
SWWAE
MLR-DNN
BinaryConnect
MIM
Spectral Representations ...
ELC
DSN
NiN
Deep Networks with Intern...
DNN+Probabilistic Maxout
GP EI
APAC
DCNN+GFE
RReLU
MCDNN
DCNN
ReNet
An Analysis of Unsupervis...
Stochastic Pooling
Improving neural networks...
Discriminative Learning o...
Nonnegativity Constraints
CKN
Discriminative Unsupervis...
Local Transformations
Hierarchical Kernel Descr...
An Analysis of Single-Lay...
Smooth Pooling Regions
Learning with Recursive P...
PCANet
FLSCNN

Horse and metrics

	Undetected	Detected
Instru	**TN**	**FP**
Song	**FN**	**TP**

- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- Accuracy, F-Measure,… but:
  - **Medecine**: 0 false negative required
  - **Music recommendation**: minimum of false positive needed

Horse and metrics

Checklist to diminish horseness of a method

- Metric with statistic and math
- User listening experience
  - Subjective
  - Different expectation
  - Time-consuming
  - Too few number of participants

Scientist validation

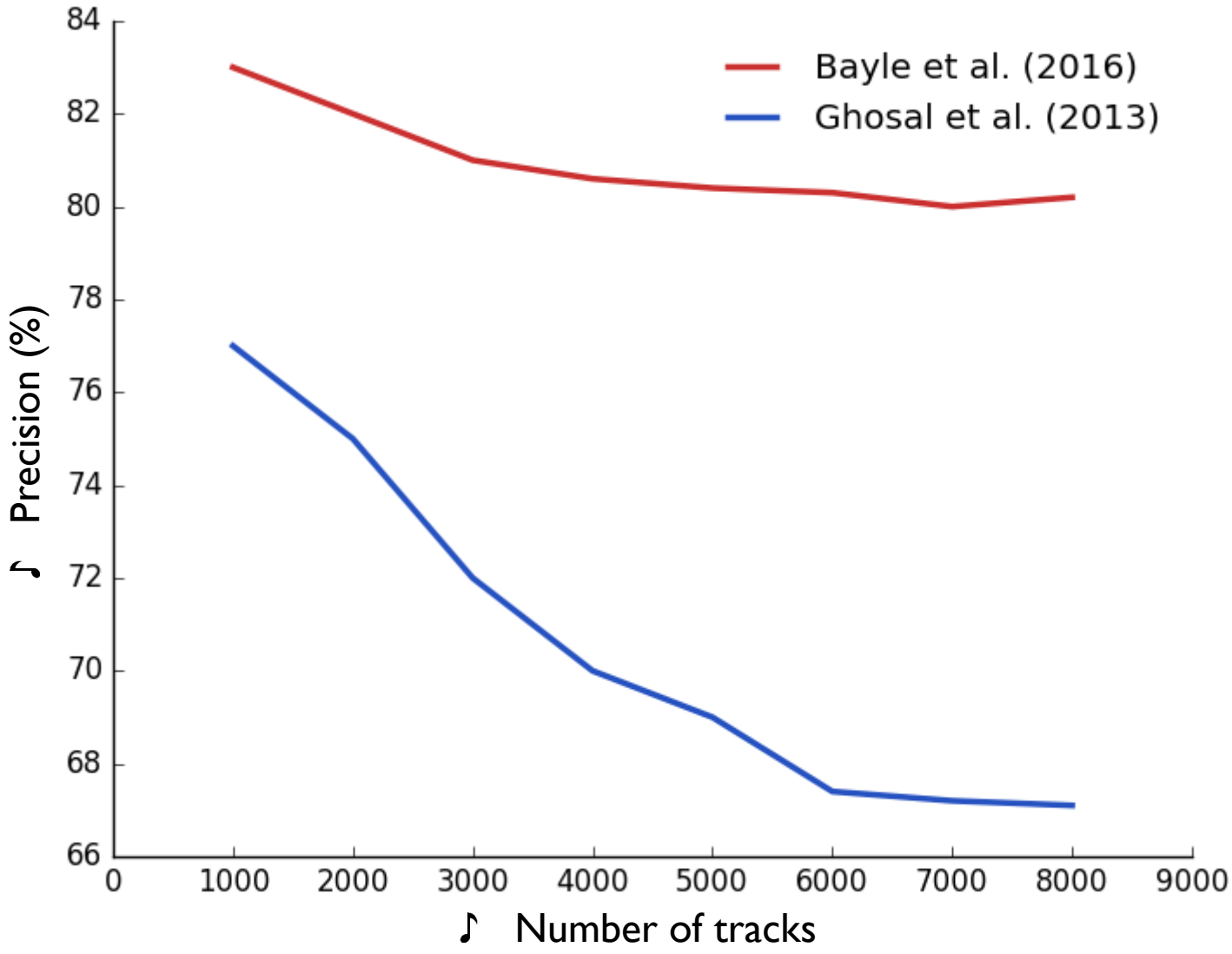
- Check the results or what the ML is learning?
- *Auralisation of deep convolutional neural networks: listening to learned features* (Choi *et al.*, 2015)

Reproducibility and replicability

Examples in Song/Instrumentals classification

- *A hierarchical approach for speech-instrumental-song classification* (Ghosal *et al.*, 2013)
  - Precision @ 95%
  - 540 excerpts of 30s: « inhouse dataset »
- SRCAM (Gouyon *et al.*, 2014)
  - Source code in matlab
  - Crash for more than 1k tracks
  - Cannot run on industrial server with 40k tracks

Reproducibility and replicability



Reproducibility and replicability

Materials

- *Replicability is not reproducibility: nor is it good science* (Drummond 2009)
- https://github.com/audiolabs/APSRR-2016
- https://infoscience.epfl.ch/record/136640
- https://github.com/faroit/reproducible-audio-research
- https://rescience.github.io/
- https://github.com/Cloud-CV/EvalAI

~~Conclusion and solutions~~ Ideas

Checklist to diminish « horseness » of a method

- Definition of the problem/task/goal
- Objective/subjective tag ⇔ objective/subjective solution?
- Dataset
  - Bigger
  - Diversified
    - Sources (Cross-dataset comparison)
    - Samples (representative)
- Data augmentation
- Cross-validation
- Preprocessing
  - Normalise signal/spectrograms
- Comparison to baseline
  - Human performances
  - Random classification (on the dataset)
  - Random input (in the system)
- Auralisation of deep convolutional neural networks: listening to learned features (Choi 2015)
- Reproducible research and replicable code
- User listening experiment for validation?
- Ground truth and L-measure

- Y. Bayle, P. Hanna, and M. Robine, "Large-scale classification of musical tracks according to the presence of singing voice," in *JIM*, 2016, pp. 144–152.
- Y. Bayle, P. Hanna, and M. Robine, "Persistent musical database for music information retrieval," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, 2017, p. 1–4.
- K. Choi, G. Fazekas, M. Sandler, J. Kim, "Auralisation of deep convolutional neural networks: listening to learned features," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015.
- C. Drummond, "Replicability is not reproducibility: nor is it good science," 2009.
- A. Flexer, "A closer look on artist filters for musical genre classification," World, 19(122), 16-7, 2007.
- A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha, "A hierarchical approach for speech-instrumental-song classification," *Springerplus*, vol. 2, no. 526, pp. 1–11, Dec. 2013.
- F. Gouyon, B. L. Sturm, J. L. Oliveira, N. Hespanhol, and T. Langlois, "On evaluation validity in music autotagging," *arXiv*, Sep. 2014.
- B. McFee, E. J. Humphrey, J. P. Bello, "A Software Framework for Musical Data Augmentation," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015, pp. 248-254.
- B. McFee, O. Nieto, M. M.Farbood, J. P. Bello, "Evaluating Hierarchical Structure in Music Annotations," Frontiers in psychology, 8, 2017.
- J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015, pp. 121–126.
- B. L. Sturm, "A Simple Method to Determine if a Music Information Retrieval System is a "Horse"," IEEE Transactions on Multimedia, 16(6), 2014, pp 1636-1644,
- X.-J. Wang, L. Zhang, C. Liu, "Duplicate Discovery on 2 Billion Internet Images," in Proceedings of the IEEE CVPRW, 2013.