
1904-2016

HORSE 2016

On "Horses" and "Potemkin Villages" in Applied Machine Learning

Queen Mary
University of London

centre for digital music

FIRE SAFETY

Queen Mary
University of London

centre for digital music

WELCOME

* Premier edition!
* Acknowledgments to the EPSRC *Platform Grant on Digital Music (EP/K009559/1)*
* HORSE 2016 is co-organised by
  * Applied Machine Learning Lab
    http://machinelearning.eecs.qmul.ac.uk/
  * Machine Listening Lab
    http://machine-listening.eecs.qmul.ac.uk/

WELCOME

* Thank you for attending!
  * It *is* a weird idea for a workshop.
  * And you get free coffee and lunch!
* Thank you to all the contributors.

Queen Mary
University of London

centre for digital music

SCHEDULE, part 1

10h00	Bob L. Sturm	Horse taxonomy and taxidermy
10h30	Roisin Loughran	When The Means Justifies the End: Why We Must Evaluate on More than Mere Output
11h00	Mathieu Lagrange	Computational experiments in Science: Horse wrangling in the digital age
11h30	Tim Hospedales	Gated Neural Networks for Option Pricing: Enforcing Sanity in a Black Box Model
12h00	Geraint Wiggins	TBA (keynote)
13h00	Lunch	

Queen Mary University of London

centre for digital music

SCHEDULE, part 2

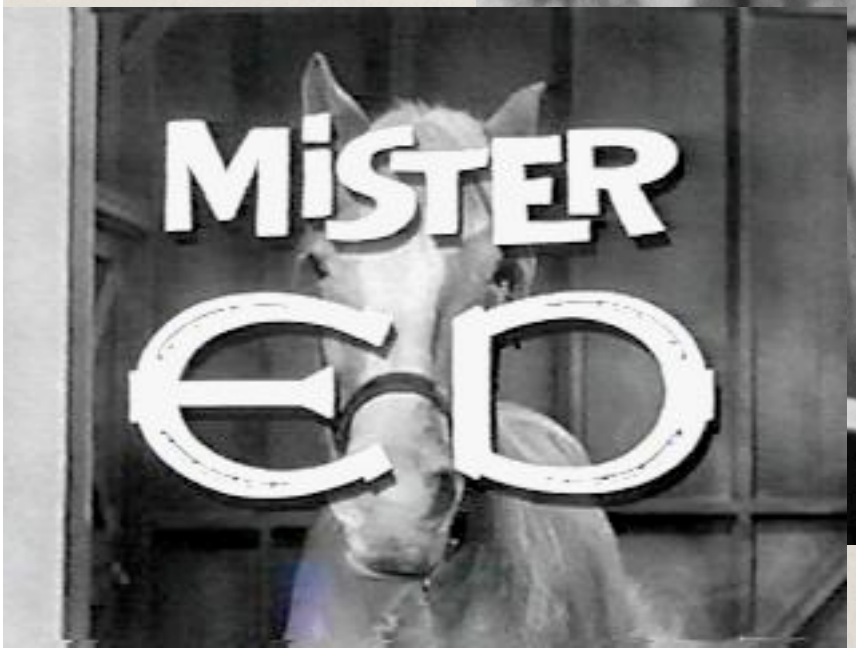
14h00	Sacha Krstulovic	Avoiding deadly horses in Automatic Environmental Sound Recognition
14h30	Francisco Rodríguez Algarra	You don't hear a thing... but my Horse knows it's Rock!
15h00	Jeff Clune (via skype)	How much do deep neural networks understand about the images they recognize?
15h30	Ricardo Silva	The role of causal inference in machine learning
16h00	Ian Goodfellow (via skype)	Adversarial Examples and Adversarial Training

16h30 : Make our way to Half Moon Pub (Mile End Wetherspoons)

Queen Mary University of London

centre for digital music

In memoriam, Ed Young
(1919-2016)



1958-1966

centre for digital music

"Horse" taxonomy and taxidermy

Bob L. Sturm

Queen Mary University of London

centre for digital music

What is a "horse"?

"a 'horse' is just a system that is not actually addressing the problem it appears to be solving."

Sturm, "A simple method to determine if a music information retrieval system is a 'horse'," *IEEE Transactions Multimedia* 16(6): 1636–1644, 2014.

Is a system giving the right answers for the right reasons?

Queen Mary University of London

centre for digital music

Why is this important?

* Applying machine learning and artificial intelligence within many domains requires transparency and responsibility!
  * health care
  * finance
  * surveillance
  * autonomous vehicles
  * government
* "Why exactly was my loan application rejected?"
* "What could I have done differently so that my application would not have been rejected?"

Queen Mary University of London

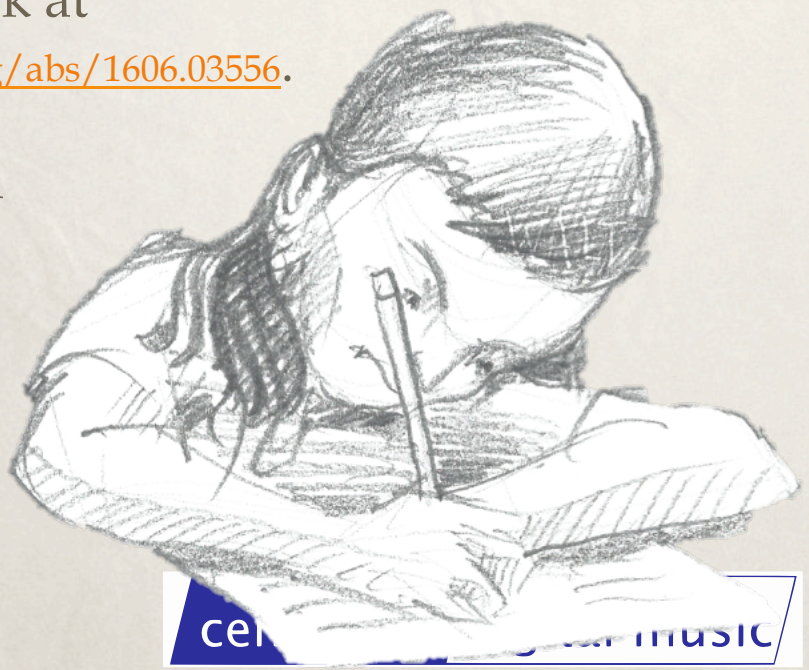
centre for digital music

Why is this important?

* J. Vincent, "First Click: **Deep learning is creating computer systems we don't fully understand**", *The Verge* (July 12, 2016) http://www.theverge.com/2016/7/12/12158238/first-click-deep-learning-algorithmic-black-boxes

* A. Das et al., "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?", 2016 http://arxiv.org/abs/1606.03556.

* "Like good students, computers need to be able to show their working."



Queen Mary University of London

Why is this important?

* L. Alexander, "**Is an algorithm any less racist than a human?**", *The Guardian* (Aug 3, 2016)
  https://www.theguardian.com/technology/2016/aug/03/algorithm-racist-human-employers-work

* "Any algorithm can – and often does – simply reproduce the biases inherent in its creator, in the data it's using, or in society at large."

  * https://socialmediacollective.org/reading-lists/critical-algorithm-studies/#1.1

  * http://algorithmwatch.org/das-adm-manifest-the-adm-manifesto/#English

Why is this important?

* T. O'Reilly, "**The great question of the 21st century: Whose black box do you trust?**", *linkedin.com* (Sep 13, 2016)
  https://www.linkedin.com/pulse/great-question-21st-century-whose-black-box-do-you-trust-tim-o-reilly

* "Understanding how to evaluate algorithms without knowing the exact rules they follow is a key discipline in today's world."
  * Goodfellow et al., "Explaining and harnessing adversarial examples," in Proc. ICLR, 2015.
  * Yosinski, Clune, et al., "Understanding Neural Networks Through Deep Visualization," ArXiv e-prints, June 2015.
  * Ribeiro et al., "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016.

Queen Mary University of London

centre for digital music

Why is this important?

* K. Wagstaff, "**Machine learning that matters**", in *Proc. ICML* , 2012.
* "Many machine learning problems are phrased in terms of an objective function to be optimized. It is time for us to ask a question of larger scope: what is the field's objective function? Do we seek to maximize performance on isolated data sets? Or can we characterize progress in a more meaningful way that measures the concrete impact of machine learning innovations?"
* http://www.wkiri.com/mlimpact/

On Taxonomy

On taxonomy

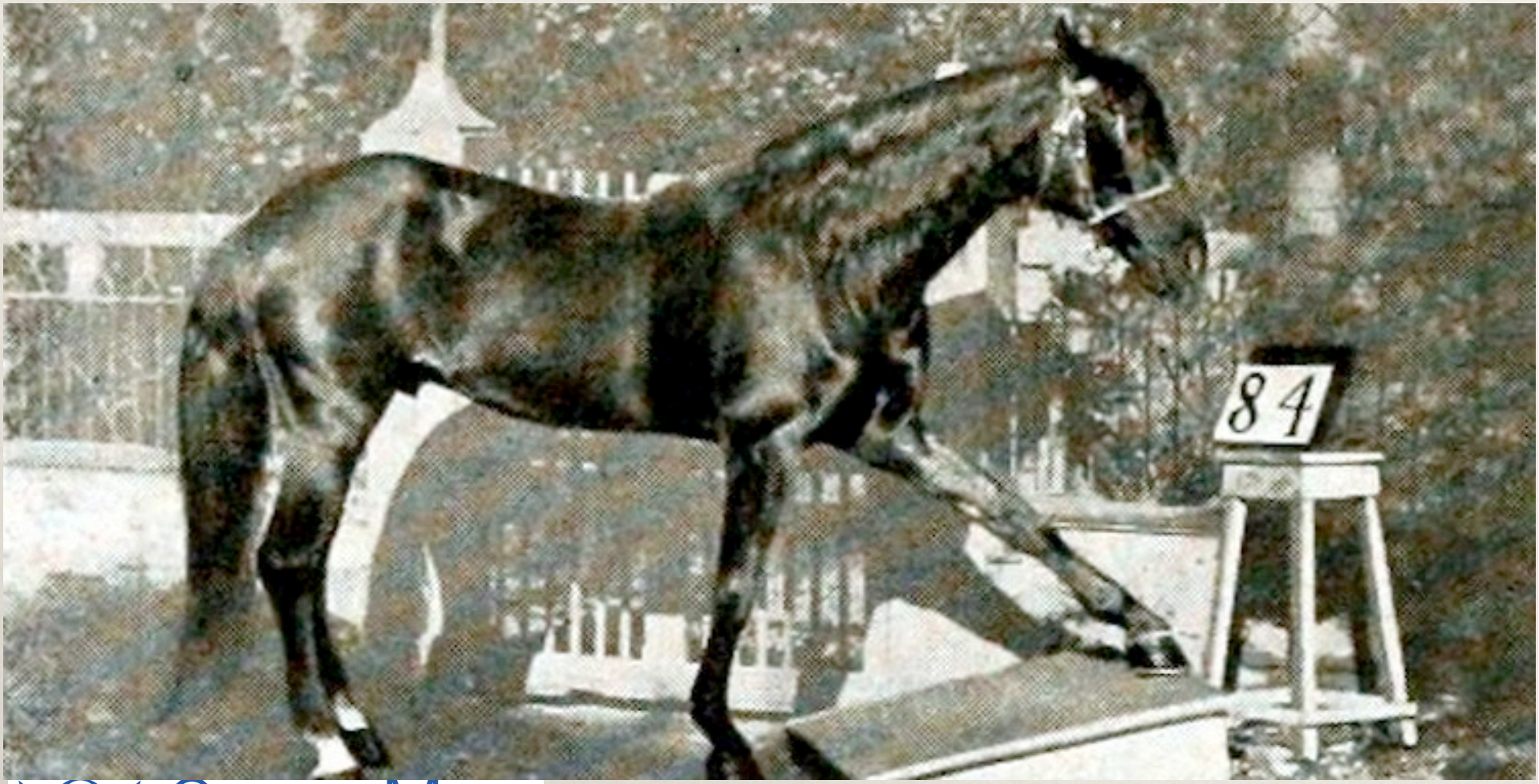
* A "horse" is just a system that is not actually addressing the problem it appears to be solving.
* A system is a "horse" only in relation to a specific problem.
* A "horse" for one problem may not be a "horse" for another
  * "Reproduce ground truth by XYZ" vs.
    "Reproduce ground truth by any means"
* It is *us humans* who infer a relationship between a system and a problem.
* Why, "horse?"

Clever Hans: *A Model Metaphor*

… but also a true life story!

Clever Hans: *A Model Metaphor*

… but also a true life story!



Queen Mary
University of London

centre for digital music

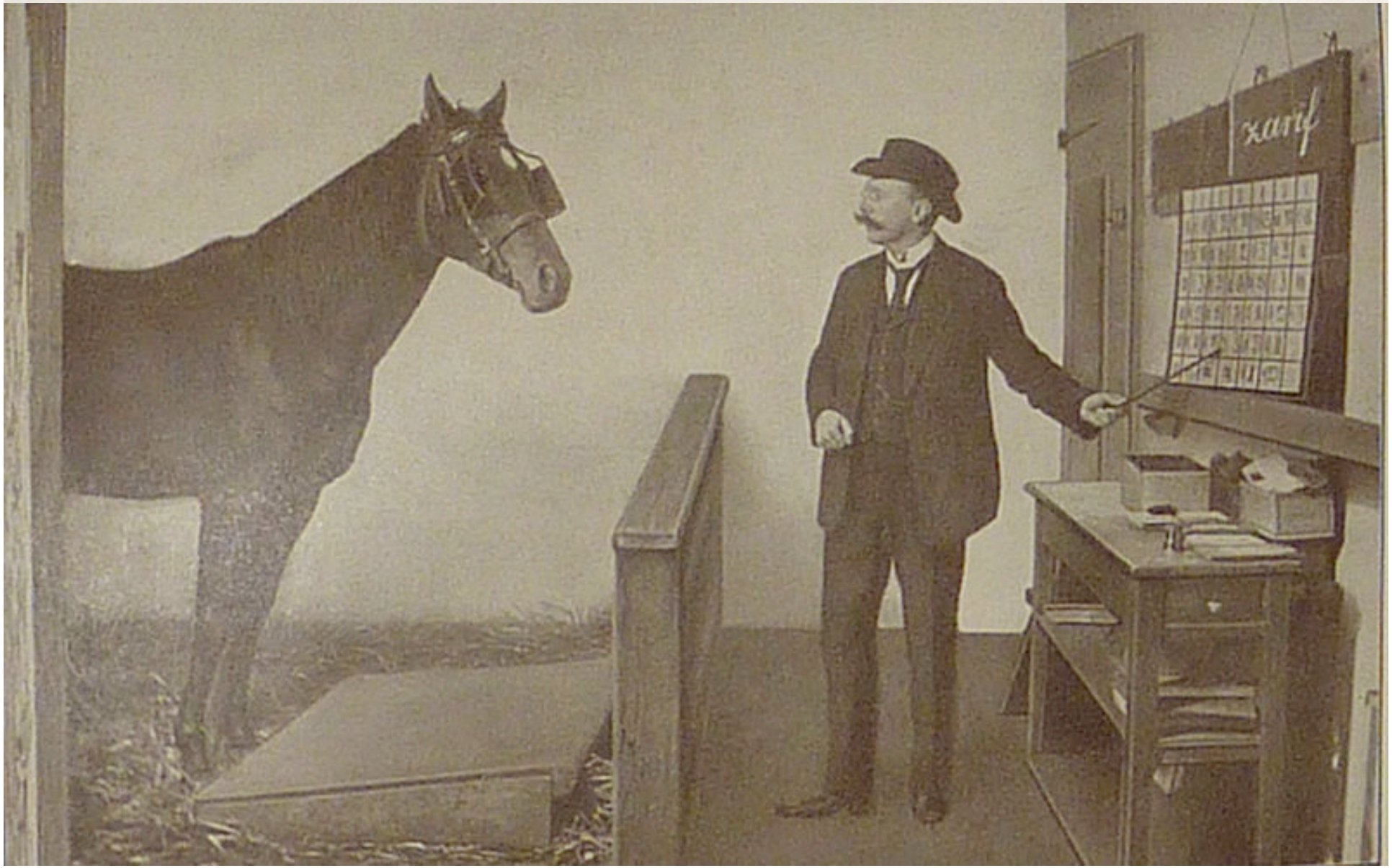
Clever Hans: *A Model Metaphor*

… but also a true life story!

Queen Mary
University of London

centre for digital music

Clever Hans: *A Model Metaphor*



Other possible metaphors

* "Potemkin village"
  * Goodfellow et al., "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.
  * Good figures of merit can be an illusion.
* "Golem" (*not* Smeagol)
  * R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, CRC Press, 2016.
  * Statistical tools are powerful but lack wisdom.
* I argue that Clever Hans is a very appropriate metaphor, *and also provides a useful methodology*.

Similarities

* von Osten believed he had taught Hans the concept of numbers
  * Used a strict pedagogical method
  * Hans was also allowed time in the pasture to think, and "discover many things for himself"
  * Hans could correctly answer unique questions
* Many sincerely believe that they have taught a machine a particular concept because …
  * The dataset is big, carefully constructed, etc.
  * It performs significantly better than random for unique problems
  * Machine learning is well-established, theoretically grounded, successful in many domains

Similarities

* Observers of Clever Hans did not think that each question could be asking something other than what was intended.
  * "How much is 2+3?" -> "Carrot for tapping your hoof until I raise."
* Many do not think that their dataset observations could be encoding something other than what they intend.
  * *There can be many ways to reproduce the ground truth of a dataset.*
  * Just because system performance is inconsistent with that expected of a random one does not mean it is likely to have learned the intended concept! (Today we will see several examples.)

Clever Hans presents a methodology too!

* Pfungst recongised that asking Hans more of the same kind of questions does not measure his arithmetic abilities.
  * Instead, he controlled the experimental conditions to answer:
  1. Does Hans really possess arithmetic acumen?
  2. If not, then how does he appear to?

* I will discuss two kinds of experiments inspired by Pfungst that get to the matter of determining system success.
  1. Does a system really possess an understanding of the concept? (Is it giving the right answer for the right reasons?)
  2. If not, then how does it appear to?

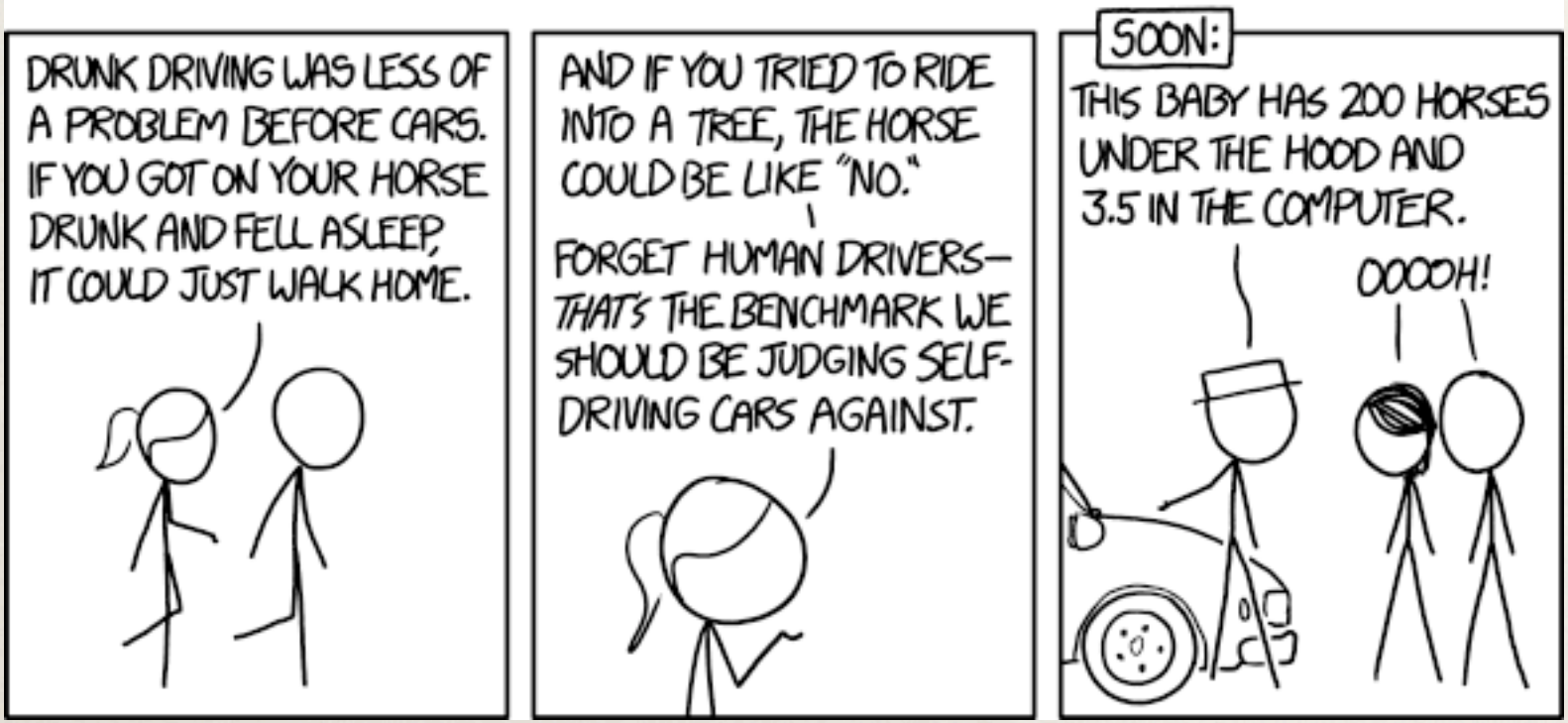
Queen Mary University of London

centre for digital music

Is "horse" appropriate?

* Some have said that the "horse" metaphor misunderstands the fundamentals of supervised machine learning
  * E.g., I am expecting too much, I am fighting straw men, etc.
* What is the supervised machine learning doing?
  * Estimating a joint probability distribution from a labeled dataset
  * It is not equipped to tease out causal relationships, or judge relevance
* *Yet we see these kind of claims all the time:*
  * "System ABC performs significantly better than random, and so recognises QRS."
  * "We find that feature XYZ is important for recognising QRS."

http://xkcd.com/1720/



Queen Mary University of London

centre for digital music

"Horse" is appropriate

* The metaphor of Clever Hans combines system and researcher.
  * von Osten et al. claimed remarkable things of his horse
  * Pfungst tested the claims in properly controlled and relevant ways
* Clever Hans also shows why bigger data doesn't necessarily solve the problem. *It is about designing relevant experiments.*
* Seeking to know why a system behaves the way it does also addresses questions about its generalisation for a given problem --- *a principal aim of machine learning*.

An example “horse”

A music rhythm recognition system

- Amplitude envelope periodicities over 10 s
- Deep neural network
- 91% train/test accuracy in 7 rhythm classes

Queen Mary University of London

Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

centre for digital music

Does this system really know how to identify and discriminate between:

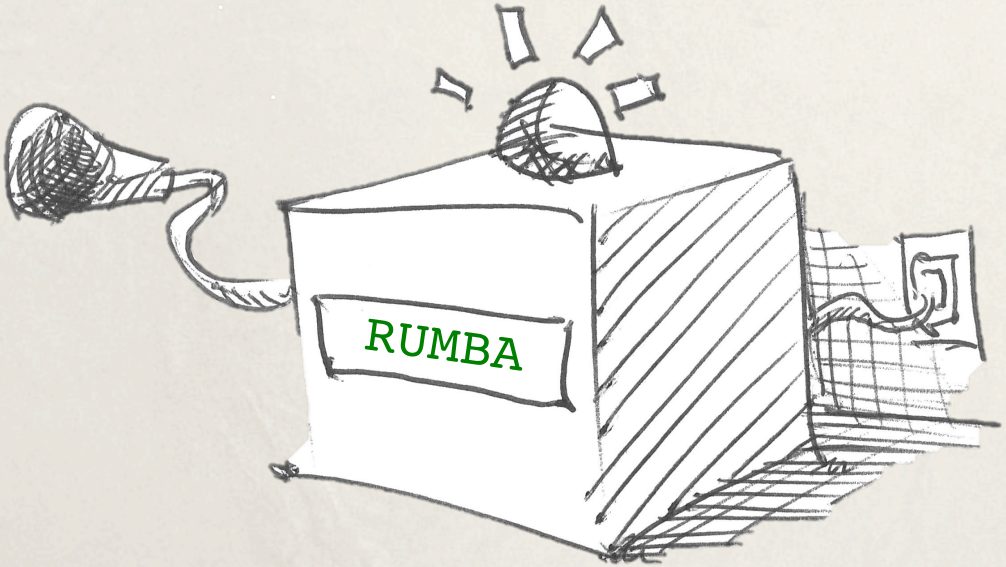
1. Cha cha cha
2. Jive
3. Quickstep
4. Rumba
5. Samba
6. Tango
7. Waltz

Will this system be useful to someone seeking rhythm information about a music recording?

Queen Mary
University of London

centre for digital music

An intervention experiment

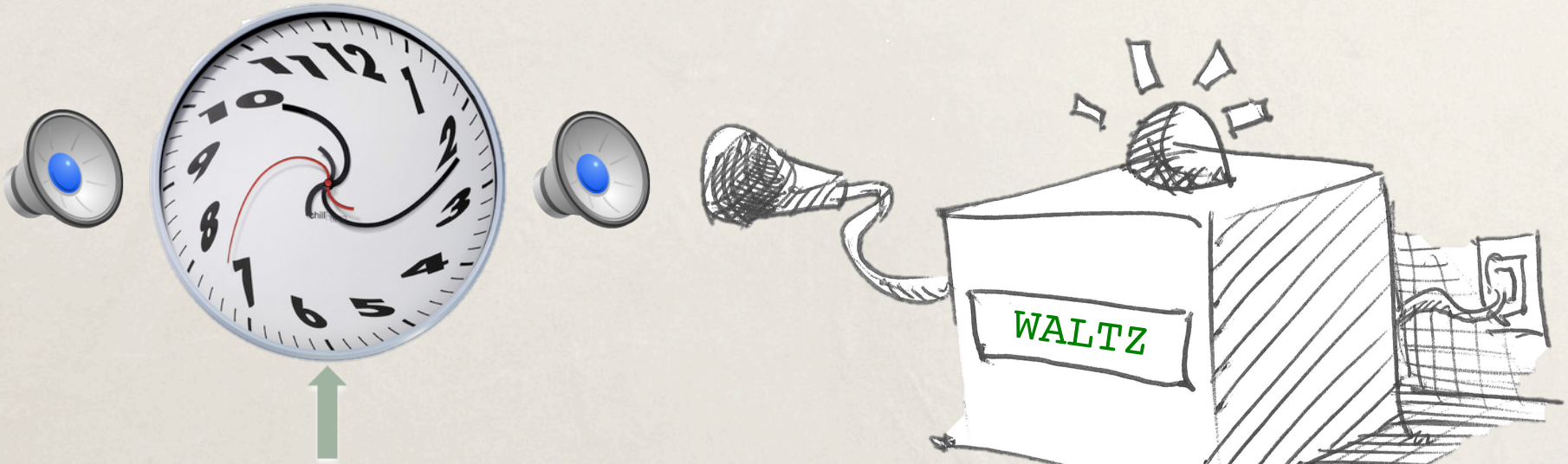


Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

Queen Mary University of London

centre for digital music

An intervention experiment



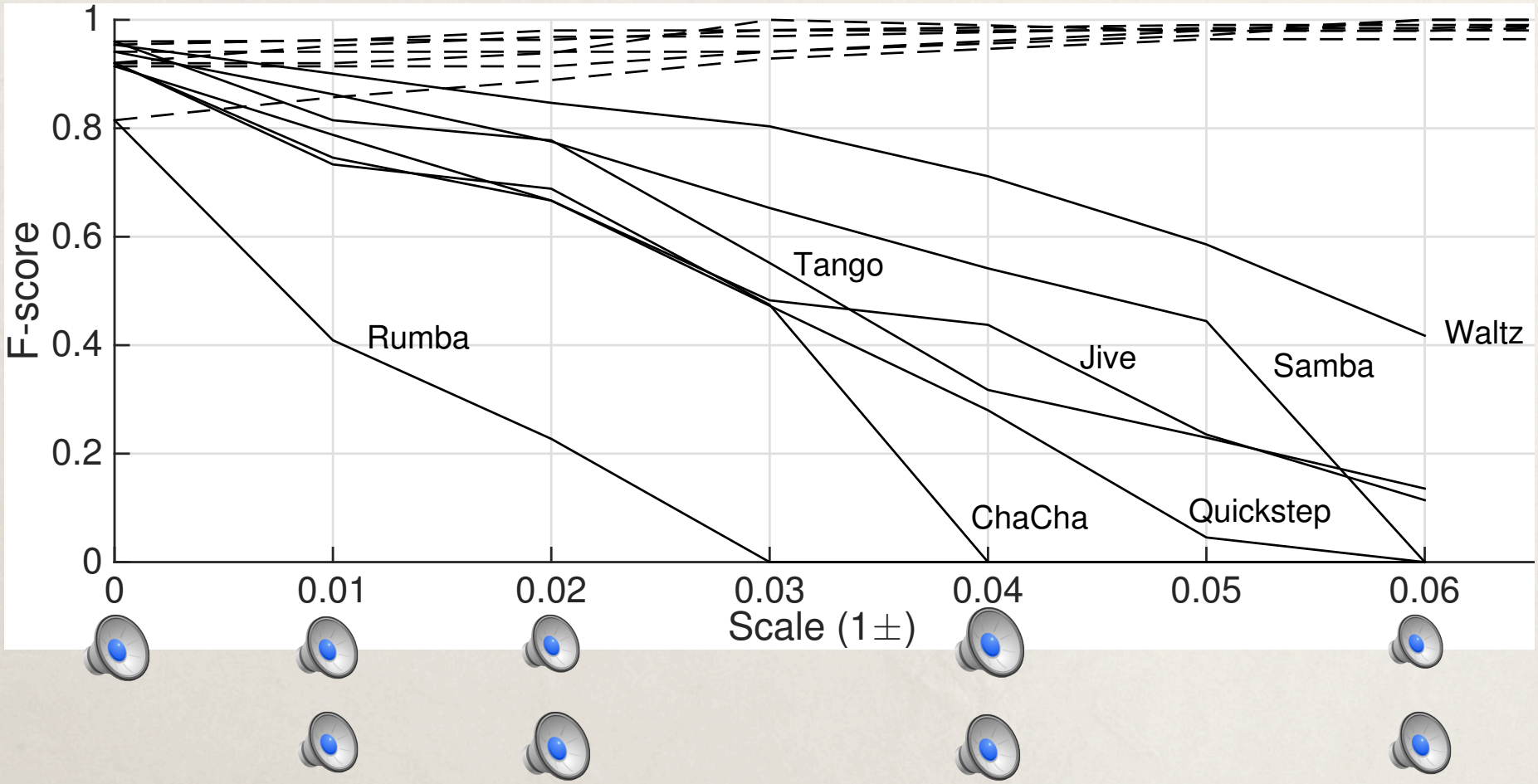
TIME DILATION
INTERVENTION

Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

Queen Mary
University of London

centre for digital music

Intervention Results

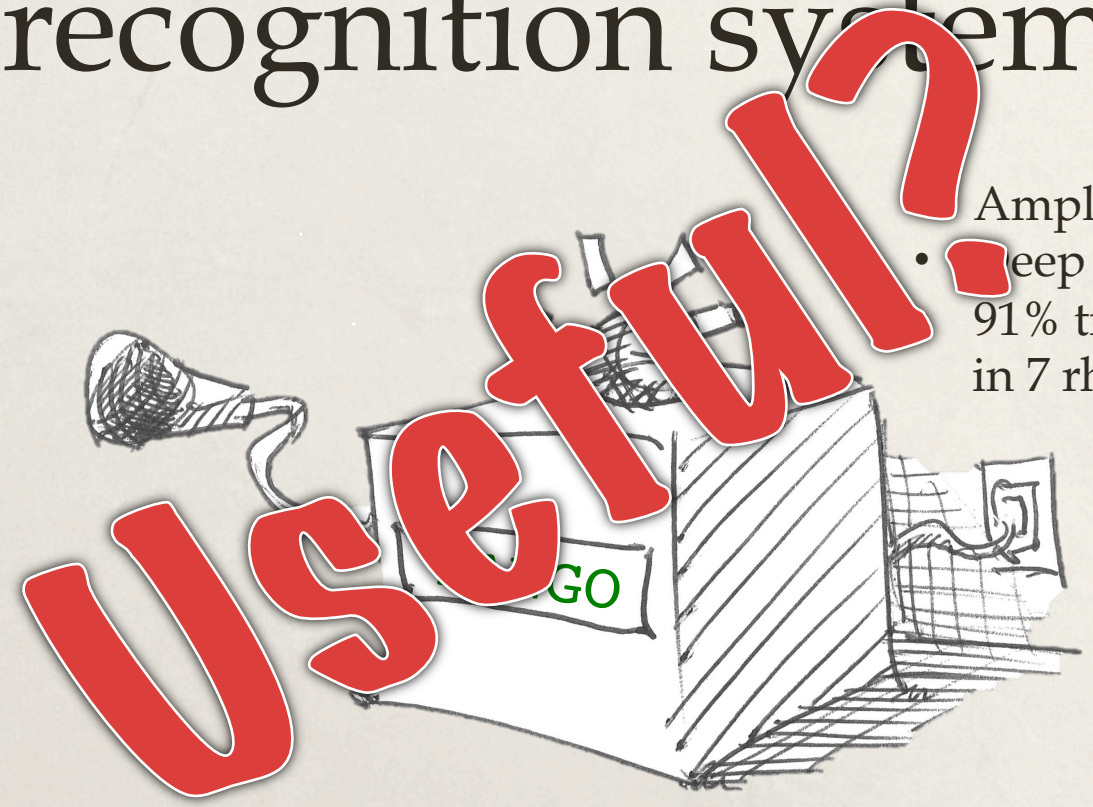


Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

Queen Mary University of London

centre for digital music

A music rhythm recognition system



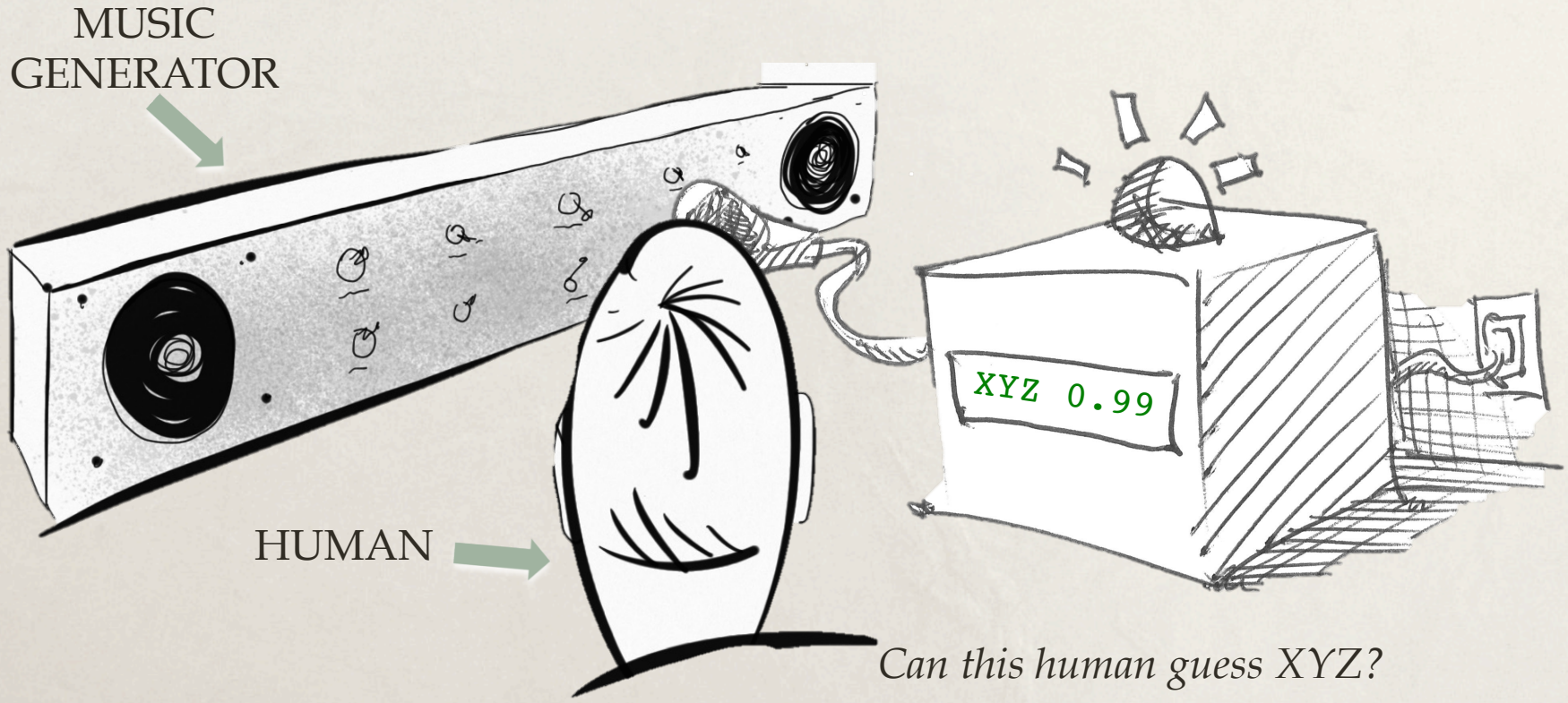
- Amplitude periodicities
- eep neural network
- 91% train/test accuracy in 7 rhythm classes

Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

Queen Mary University of London

centre for digital music

A generation experiment



Can this human guess XYZ?

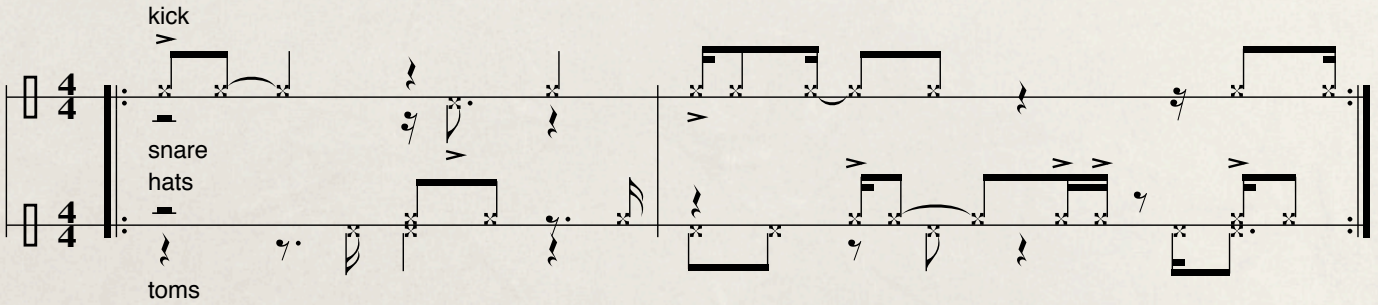
Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?," in Proc. ACM MIRUM Workshop, 2012.

Queen Mary
University of London

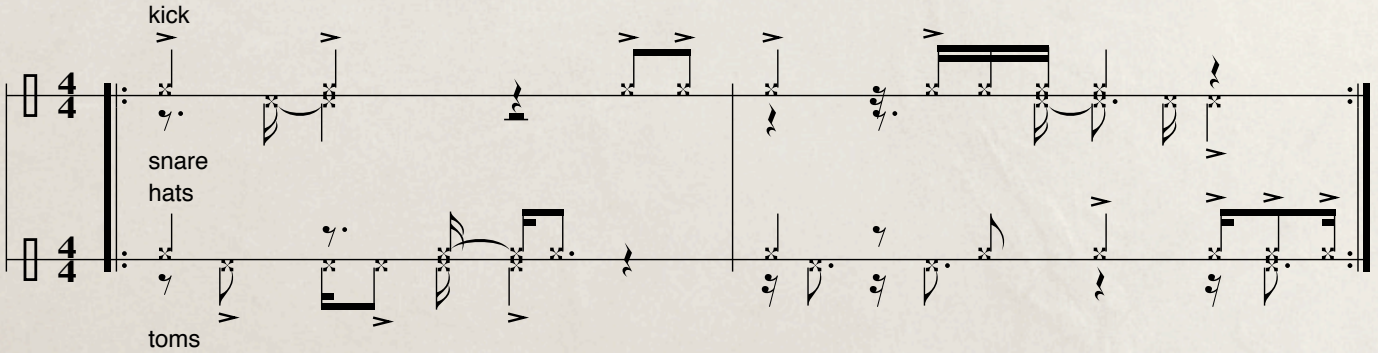
centre for digital music

Guess the rhythm

Cha cha cha
Jive
Quickstep
Rumba
Samba
Tango
Waltz



"???"



"???"

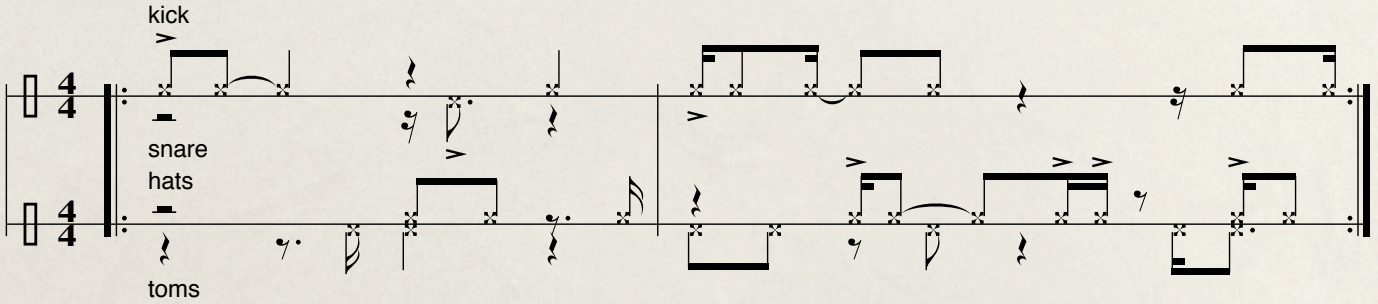
Queen Mary University of London

Sturm, "The "horse" inside: Seeking causes of the behaviours of music content analysis systems," ACM Computers in Entertainment (accepted), 2016.

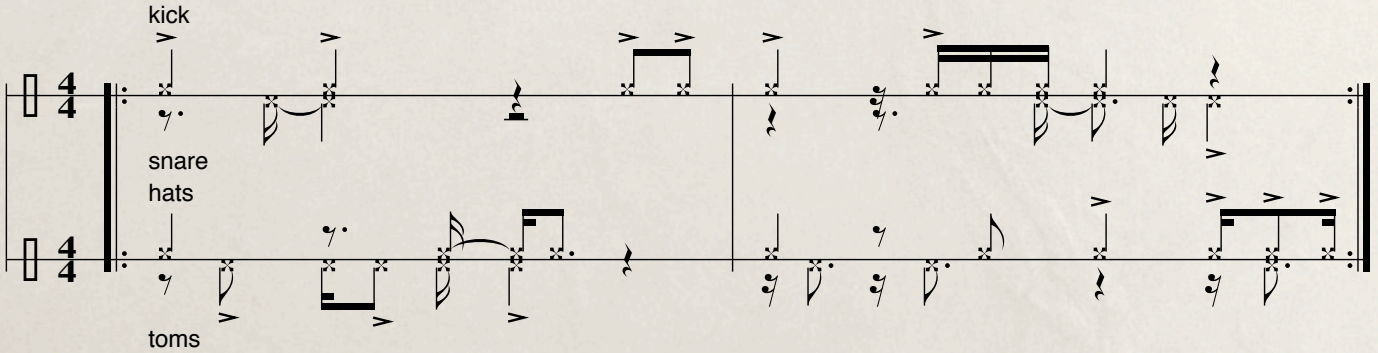
centre for digital music

Guess the rhythm

Cha cha cha
Jive
Quickstep
Rumba
Samba
Tango
Waltz



"Tango"



"Waltz"

Sturm, "The "horse" inside: Seeking causes of the behaviours of music content analysis systems," ACM Computers in Entertainment (accepted), 2016.

Queen Mary University of London

centre for digital music

A music rhythm recognition system

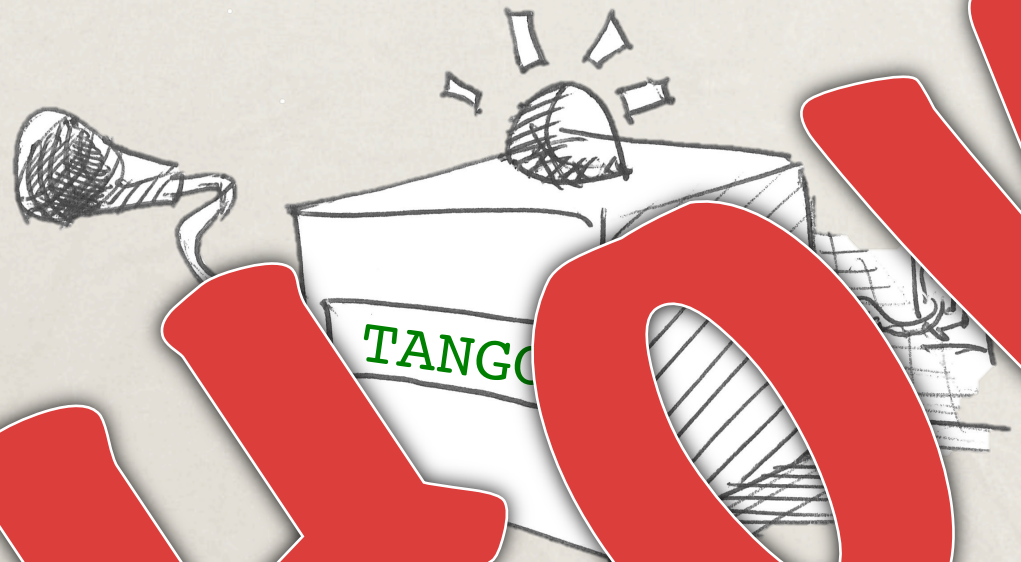
- Amplitude envelope periodicities over 10 s
- Deep neural network
- 91% train/test accuracy in 7 rhythm classes

Queen Mary University of London

Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

centre for digital music

A music rhyth recognition system



de en e
ties over 10 s
al network
91% train/test accuracy
7 rhythm classes

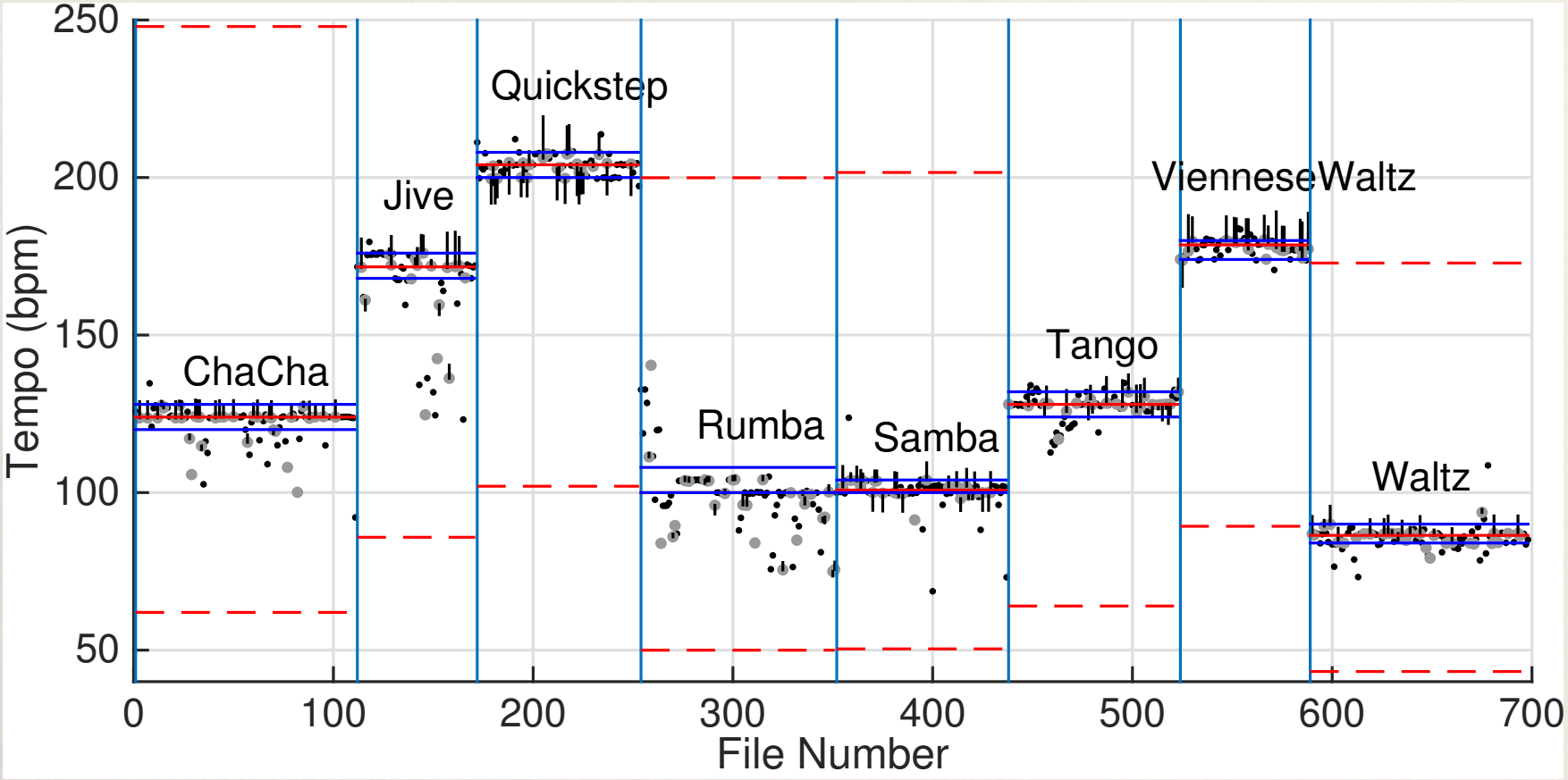
WOW?

Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

Que Mary
University

centre for digital music

One lurking cue

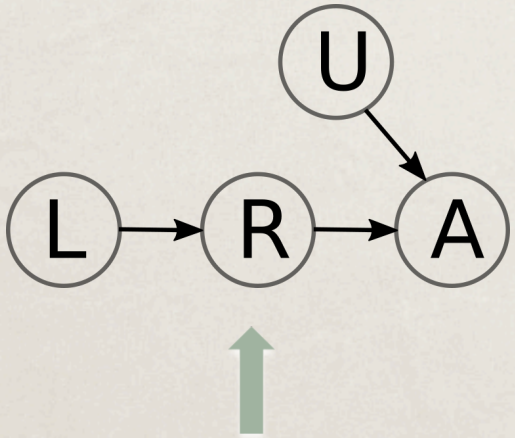


Sturm, Kereliuk, and Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in Proc. Int. Workshop on Cognitive Info. Process., 2014.

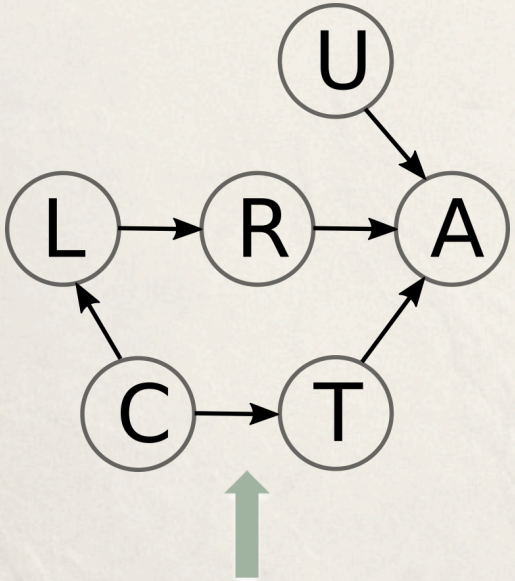
Queen Mary University of London

centre for digital music

Machines learn the darndest things!



INTENDED PROBLEM

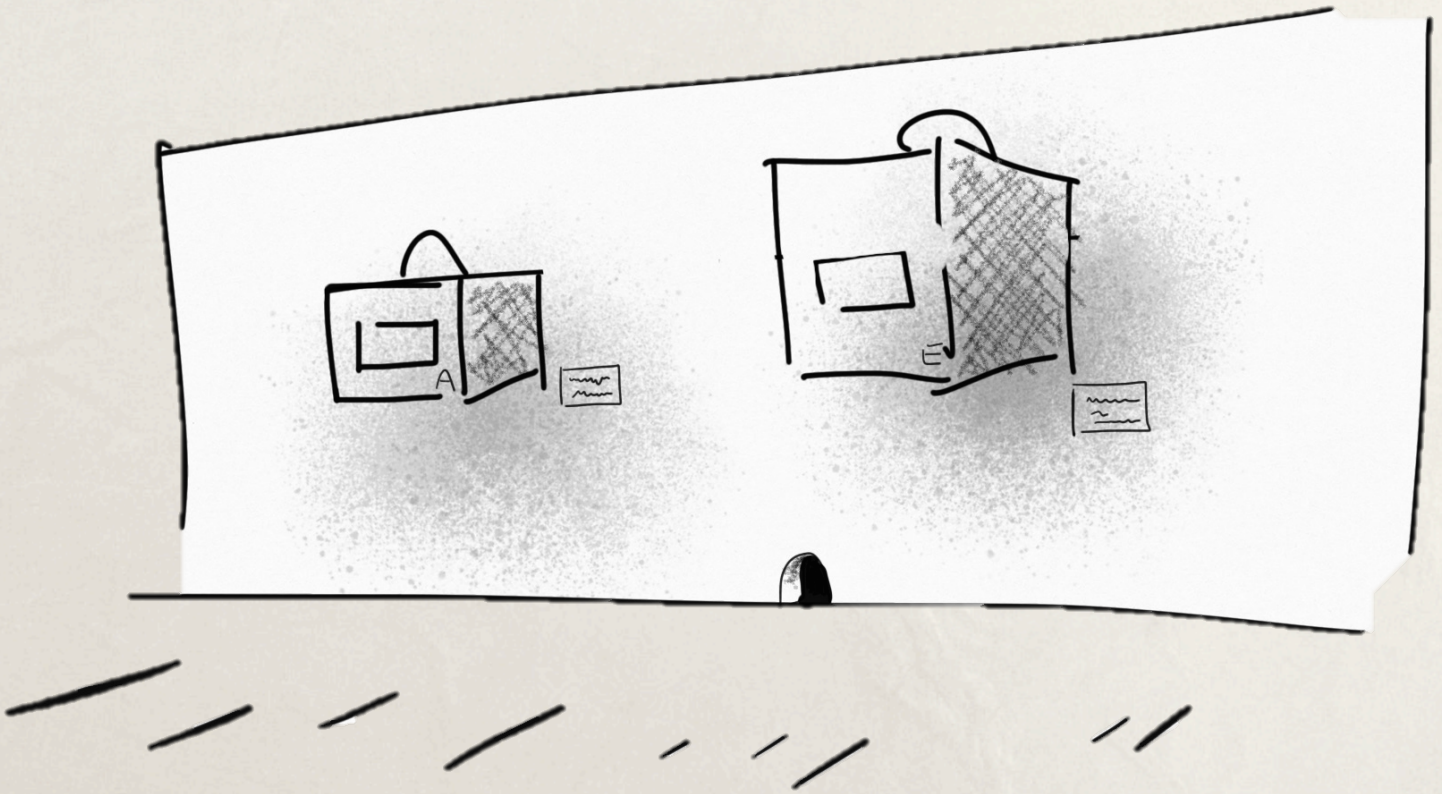


ALTERNATE ROUTE

Queen Mary University of London

centre for digital music

On Taxidermy



Queen Mary
University of London

centre for digital music

On Taxidermy

* What should you do upon discovering a system to be giving the right answers for the wrong reasons?
  * You have uncovered valuable information to improve a system, its training, our datasets, etc.
  * Publish your findings *with reproducible code*.
  * Examples: http://www.eecs.qmul.ac.uk/~sturm/software/
* A pessimist would call Pfungst's work a "negative result", but it is one of the greatest contributions to modern scientific methodology.
  * O. Pfungst, Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology. New York: Henry Holt, 1911.

Queen Mary University of London

centre for digital music

On Taxidermy

* What should you do when someone discovers *your* system to be a "horse"?
  * Don't take it personally
  * Celebrate! They have uncovered useful information to improve your system, its training, our datasets, and moving forward.
  * Collaborate to build a better system.
* The default position for any machine learning system is, "horse" until proven otherwise.
  * Don't stop at the cross-validation accuracies, confusions, etc.
  * *Work to explain what the system has actually learned to do.*


Clever Hans revealed to be a “horse”
1st year of HORSE!
1904-2016