Measuring the speed of the Red Queen's Race

Richard Harang and Felipe Ducau
Sophos Data Science Team

SOPHOS

Who we are

- Rich Harang

  @rharang; richard.harang@sophos.com

  - Research Director at Sophos – PhD UCSB; formerly scientist at U.S. Army Research Laboratory; 8 years working at the intersection of machine learning, security, and privacy

- Felipe Ducau

  @fel_d; felipe.ducau@sophos.com

  - Principal Data Scientist at Sophos – MS NYU Center for Data Science; specializes in design and evaluation of deep learning models

Data Science @ Sophos

Josh Saxe

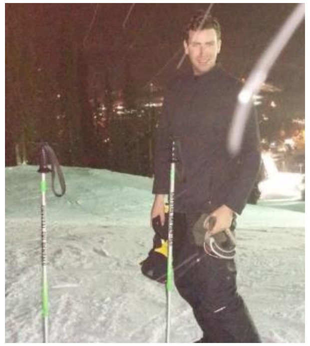
Maddy Schiappa

Rich Harang

Hillary Sanders

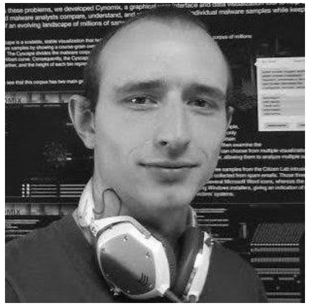
Adarsh Kyadige

Konstantin Berlin



Ethan Rudd

Felipe Ducau



Alex Long

Cody Wild

William Lee

Matt Stec

Matt Burnett

You?

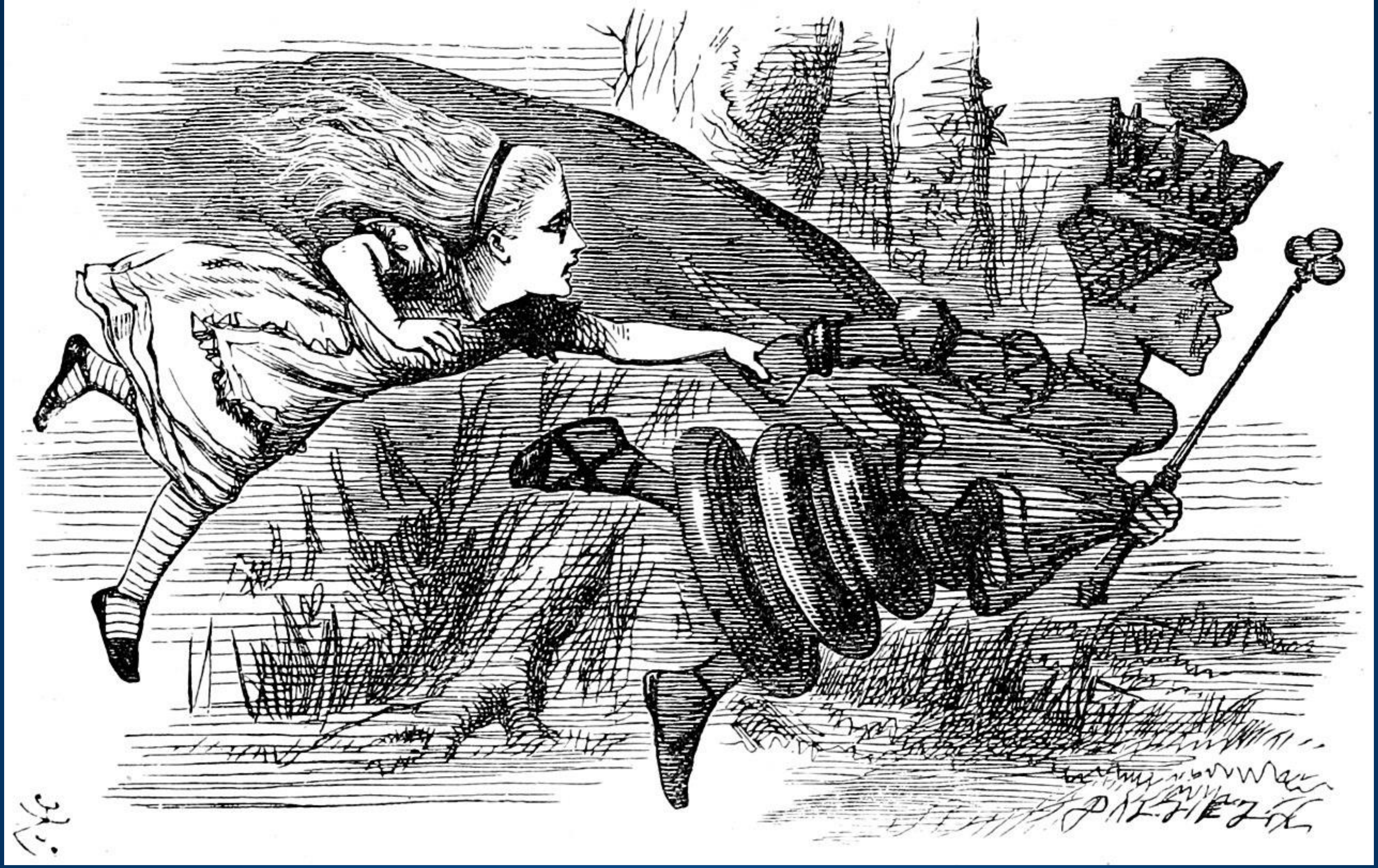
The talk in three bullets

- The threat landscape is constantly changing; detection strategies decay
- Knowing something about how fast and in what way the threat landscape is changing lets us plan for the future
- Machine learning detection strategies decay in *interesting ways* that tell us useful things about these changes

Important caveats

- A lot of details are omitted for time

- We're data scientists first and foremost, so…
  - Advance apologies for any mistakes
  - Our conclusions are machine-learning centric



"Now here, you see, it takes all the running you can do to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!“

(Lewis Carroll, 1871)

Faster and faster…

Vandalism

1986 – Brain virus

1988 – Morris worm

1990 – 1260 polymorphic virus

1991 – Norton Antivirus, EICAR founded, antivirus industry starts in earnest

1995 – Concept virus

RATs, loggers, bots

2002 – Beast RAT

2003 – Blaster worm/DDoS

2004 – MyDoom worm/DDoS

2004 – Cabir: first mobile phone worm

2004 – Nuclear RAT

2005 – Bifrost RAT

2008-2009 – Conficker variants

Crimeware, weapons

2010 – Koobface

2011 – Duqu

2012 – Flame, Shamoon

2013 – Cryptolocker, ZeuS

2014 – Reign

2016 – Locky, Tinba, Mirai

2017 – WannaCry, Petya

Two (static) detection paradigms

Signatures

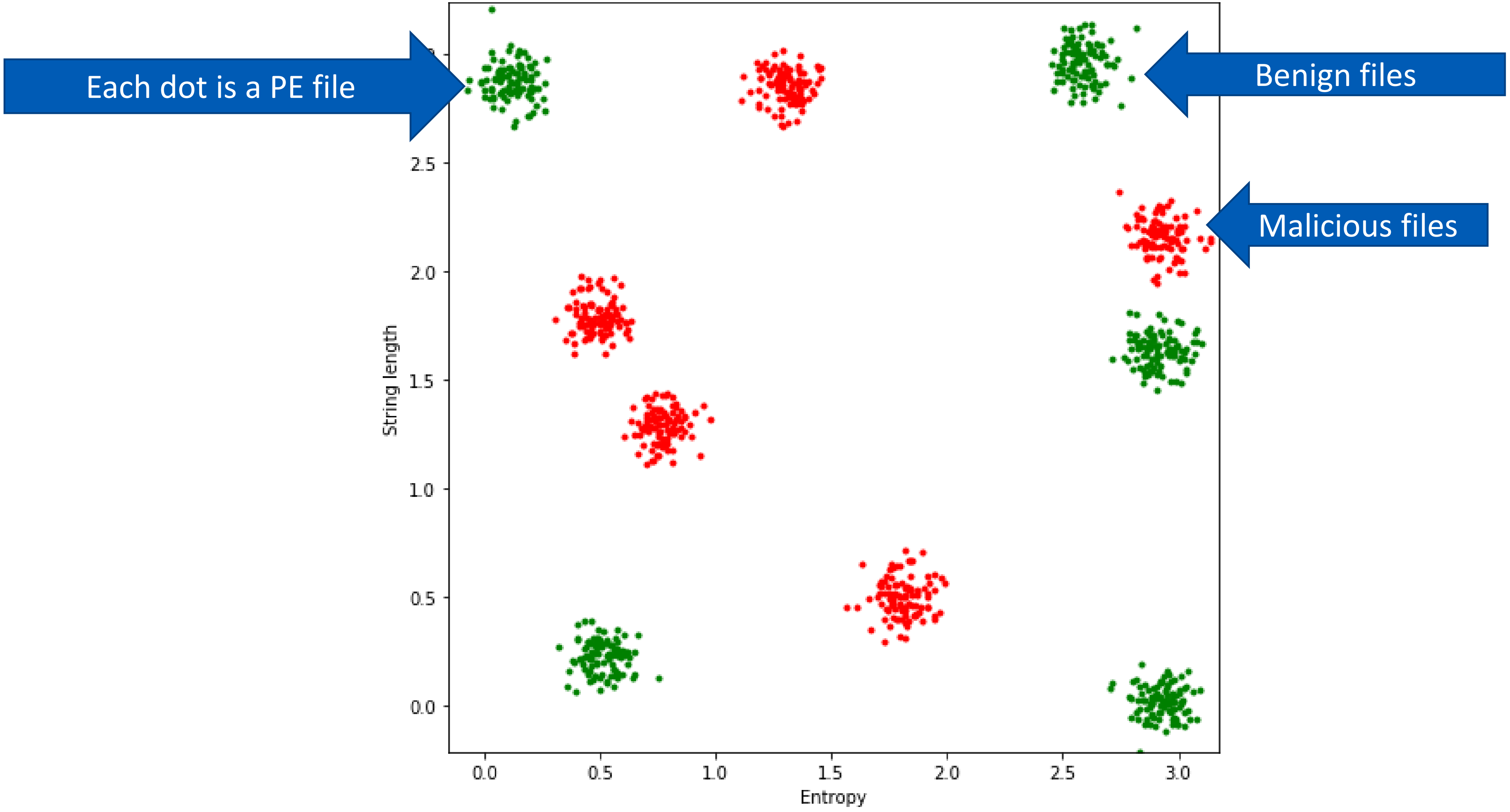
- Highly specific, often to a single family or variant
- Often straightforward to evade
- Low false positive rate
- Often fail on new malware

Machine learning

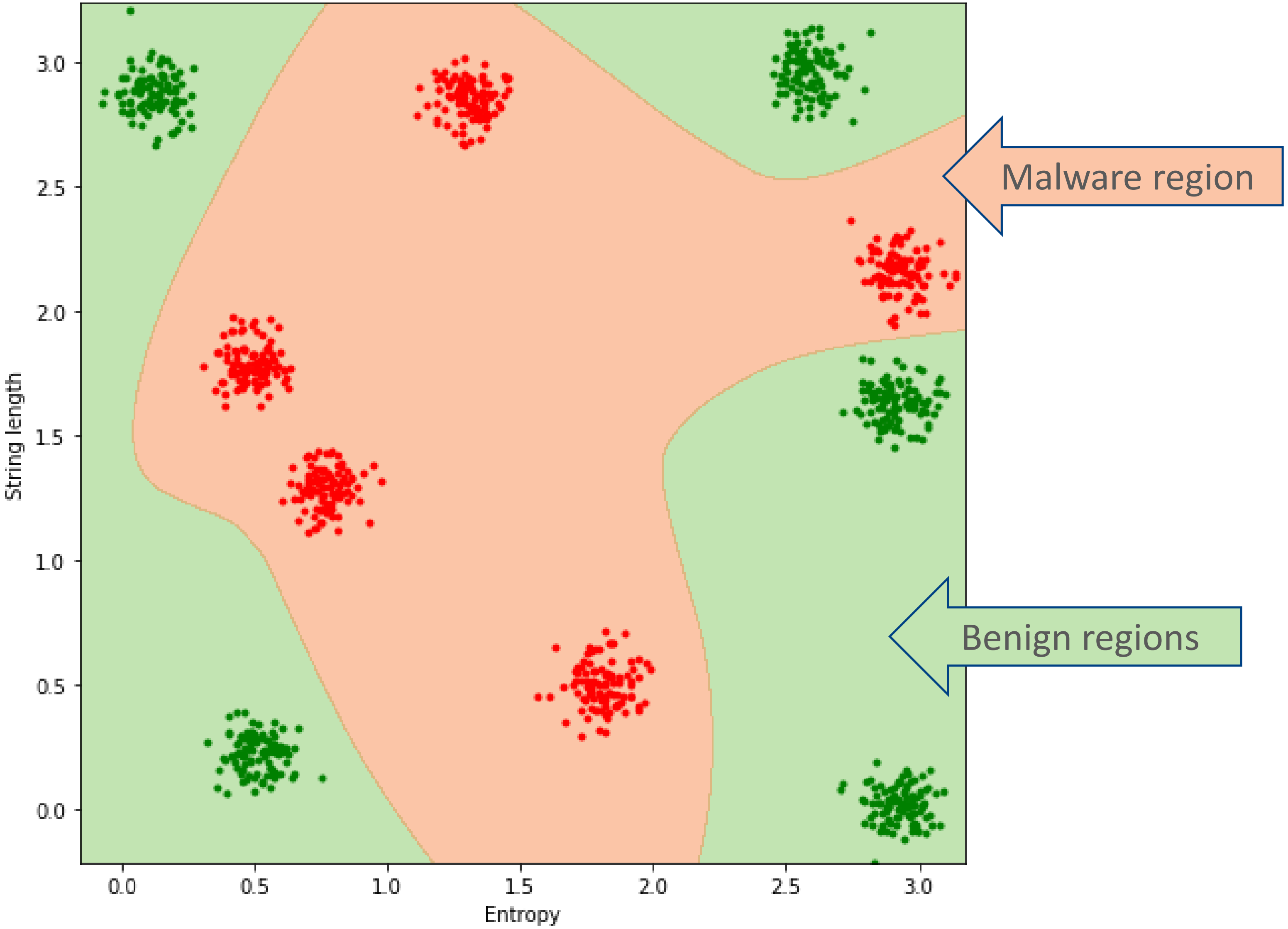
- Looks for statistical patterns that suggest "this is a malicious program"
- Evasive techniques not yet well developed
- Higher false positive rate
- Often does quite well on new malware

A crash primer on deep learning

A toy problem

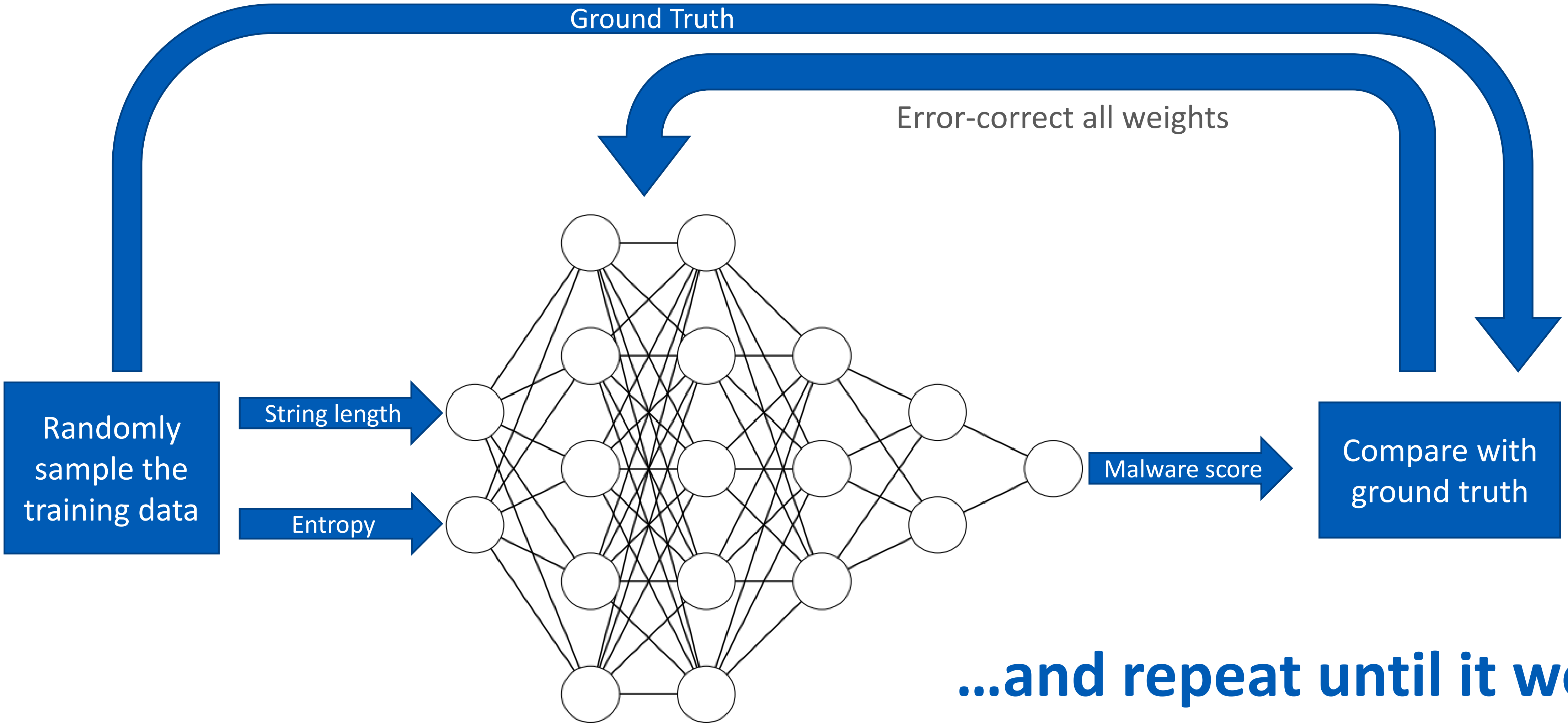


What we want



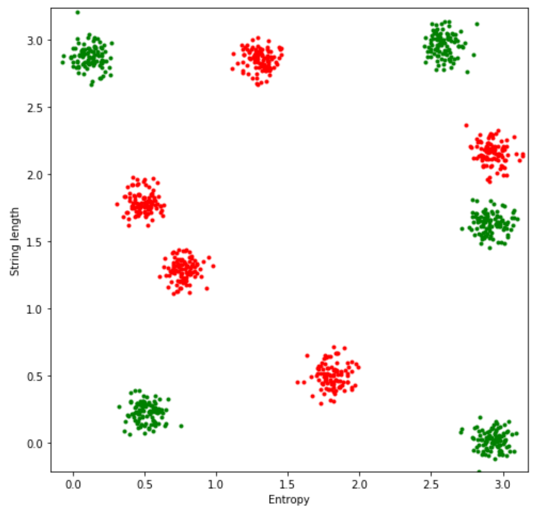
Malware region
Benign regions
String length
Entropy
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

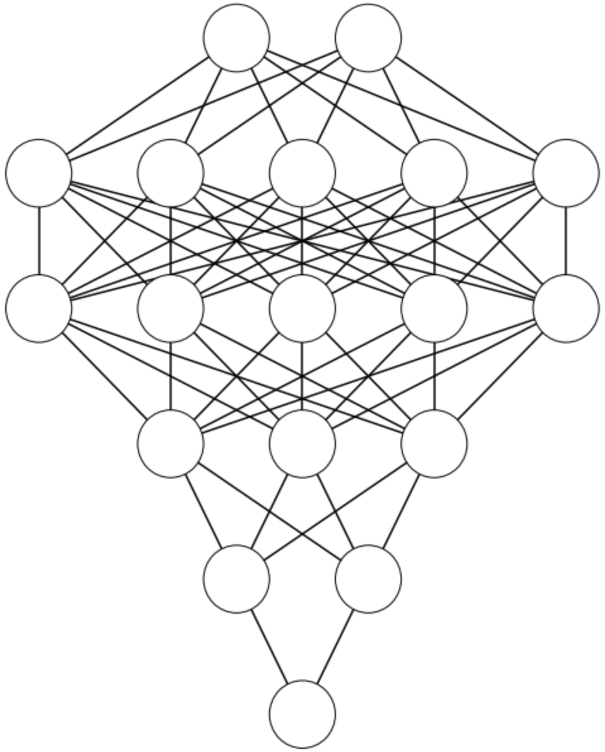
Training the model



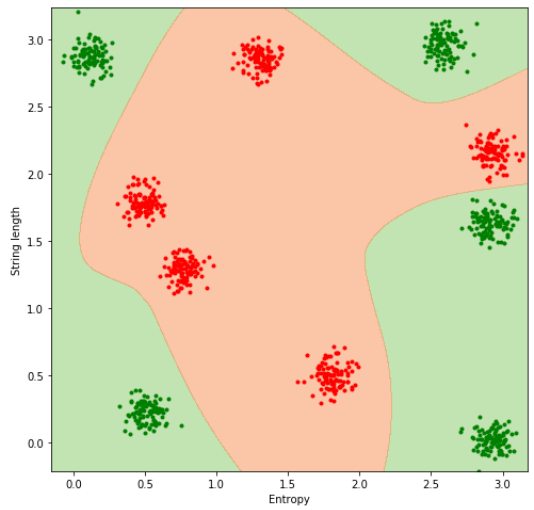
…and repeat until it works

Recipe for an amazing ML classifier

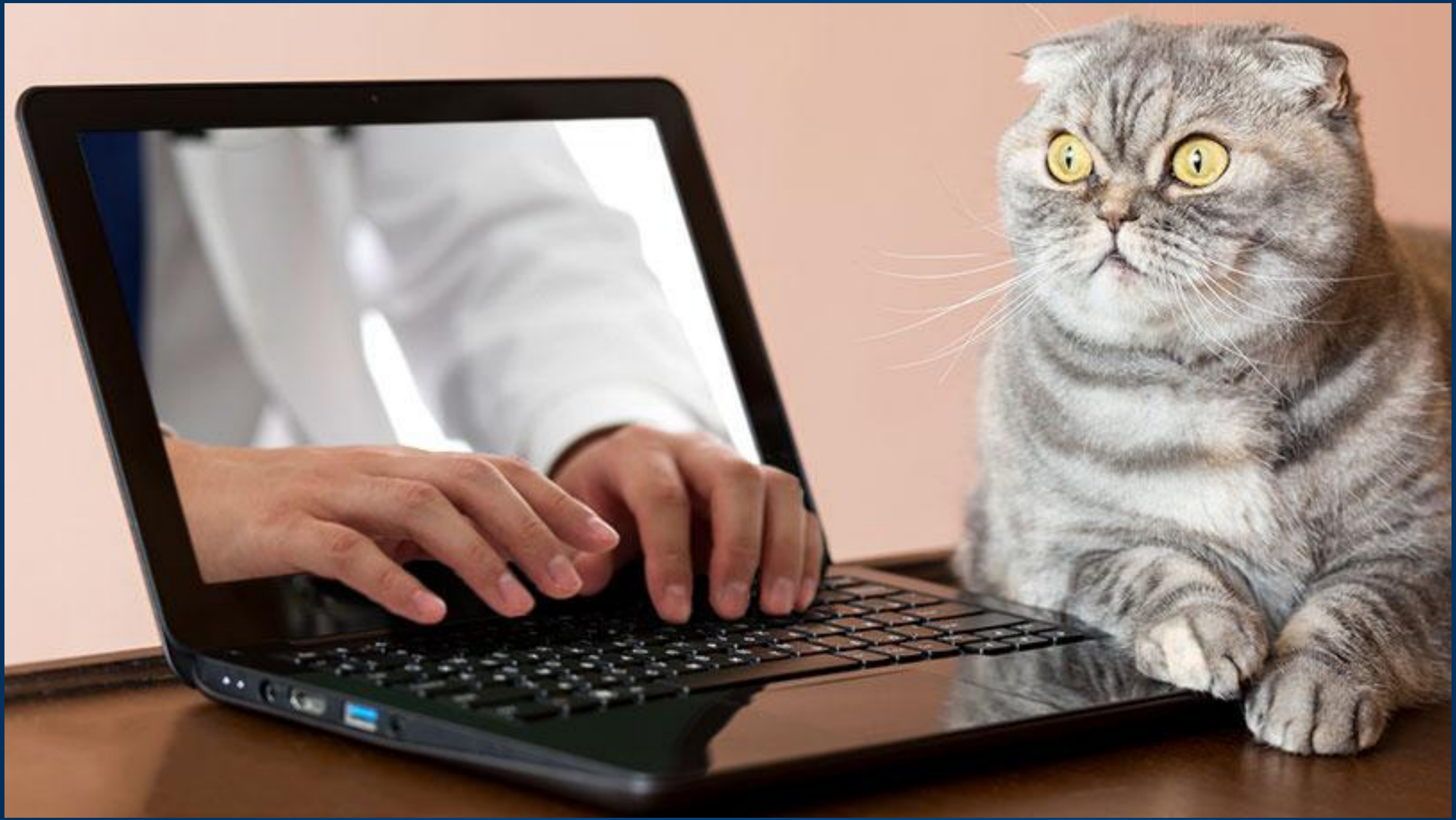




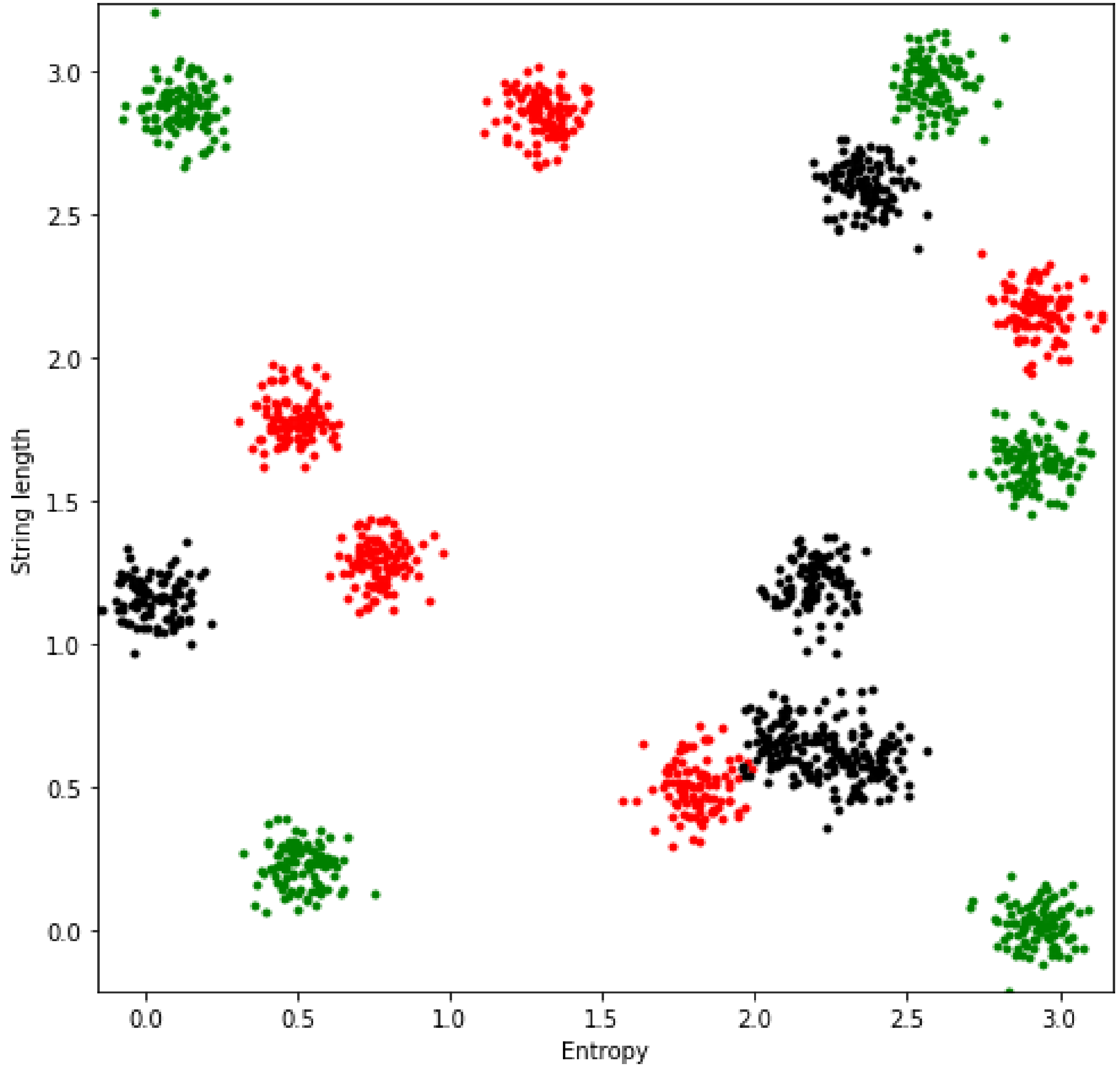
A lot of training time



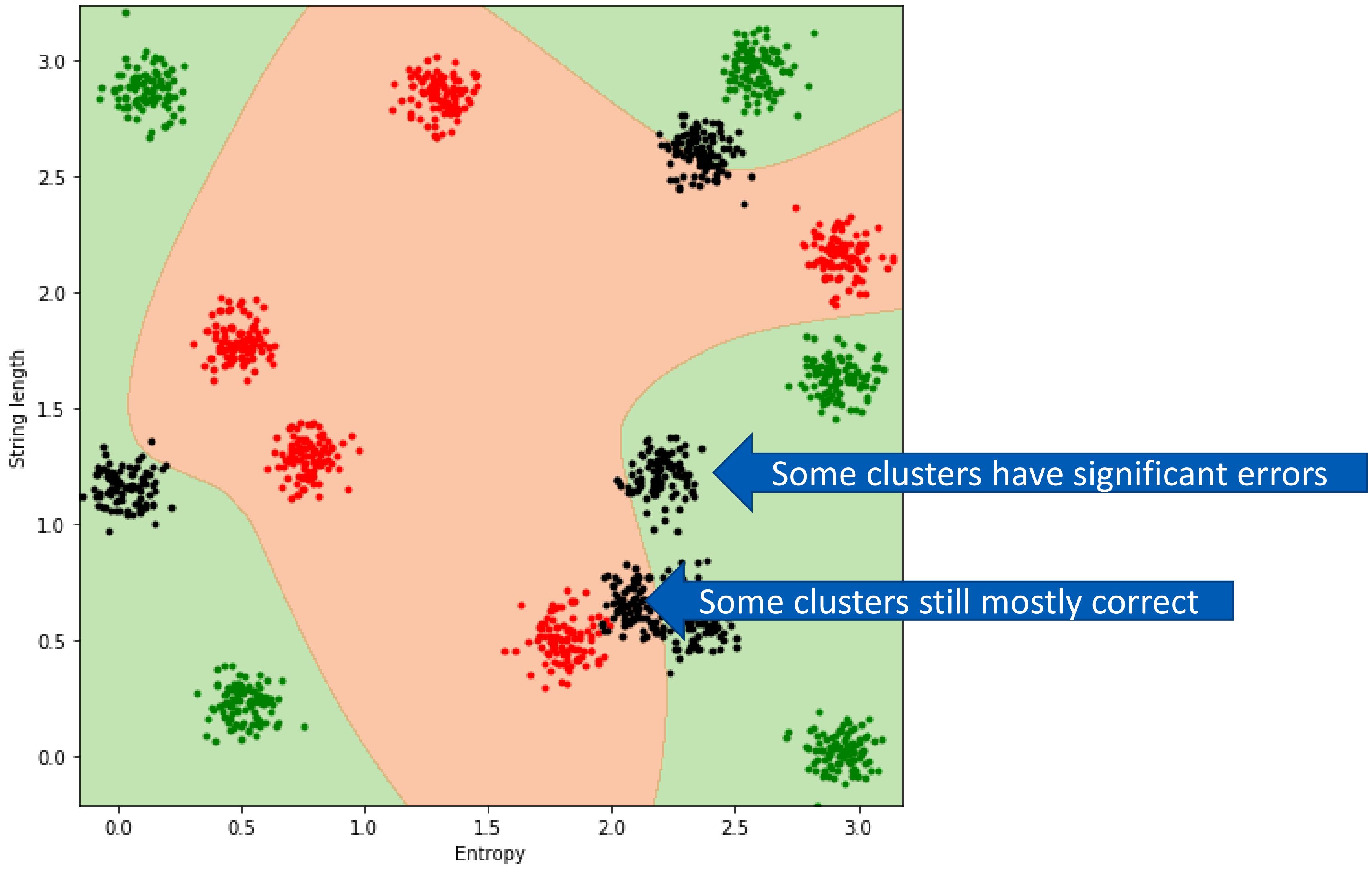
But.



…and six weeks later, we have this.

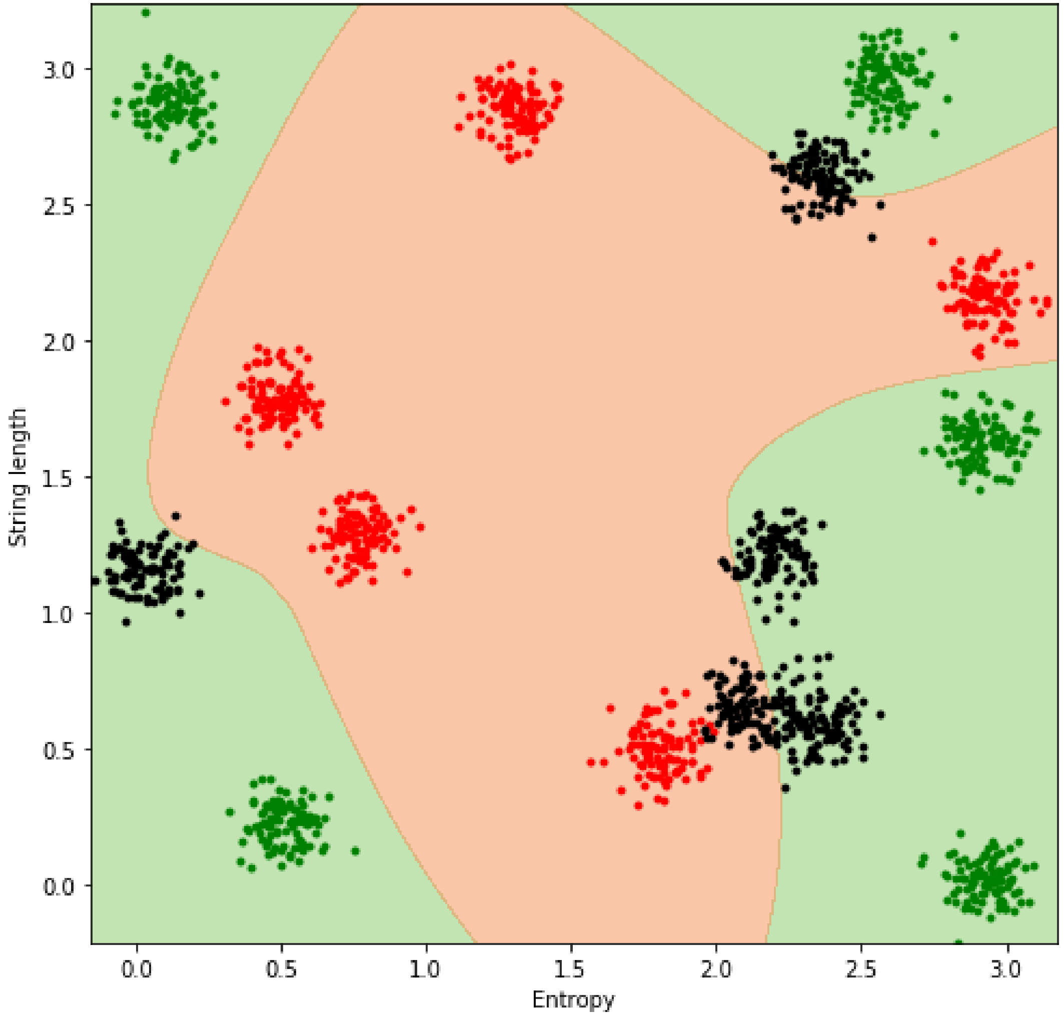


Our model performance begins to decay



Machine learning models decay in *informative ways*

- Decay in performance happens because the data changes
- More decay means larger changes in data

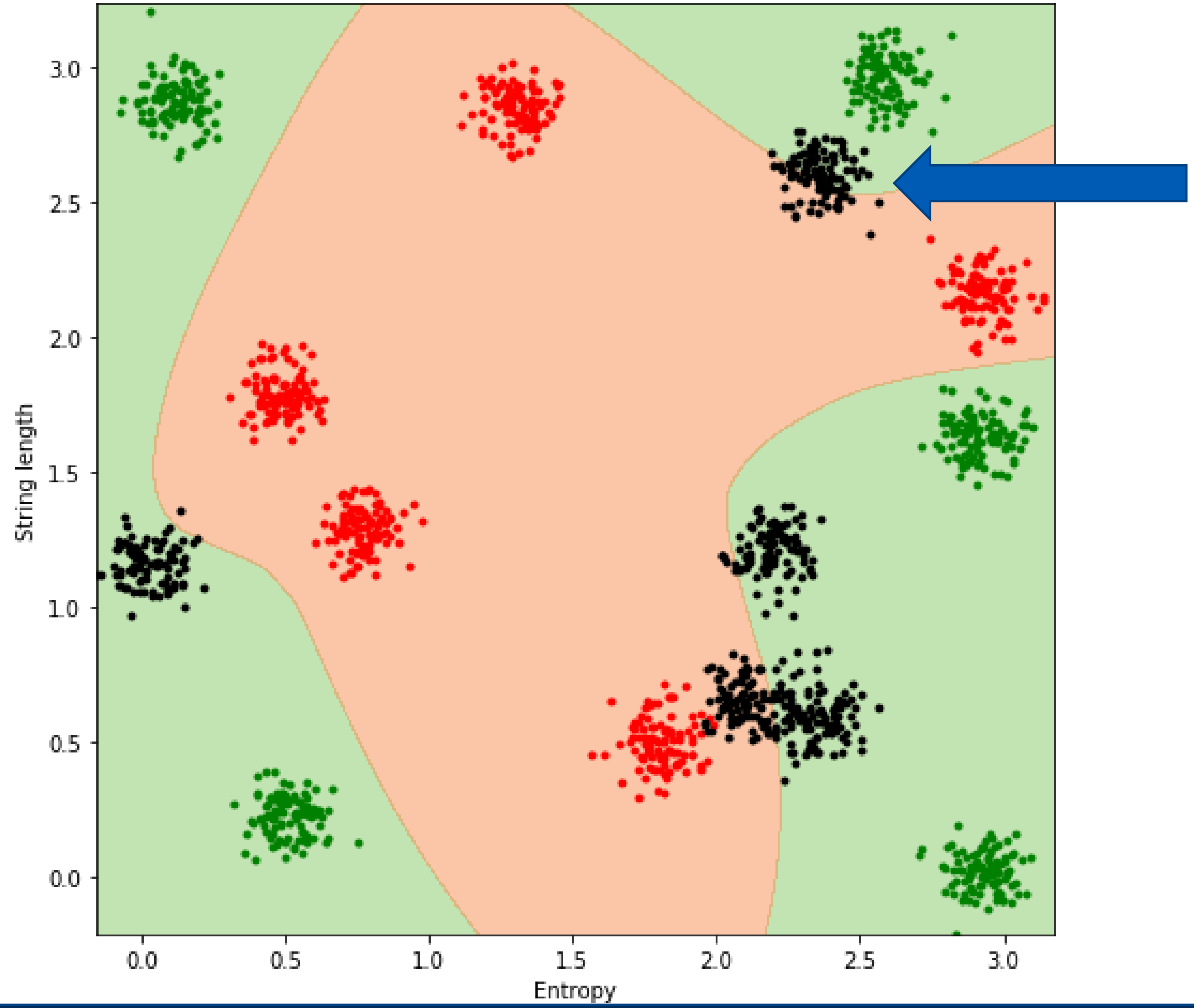


Model confidence

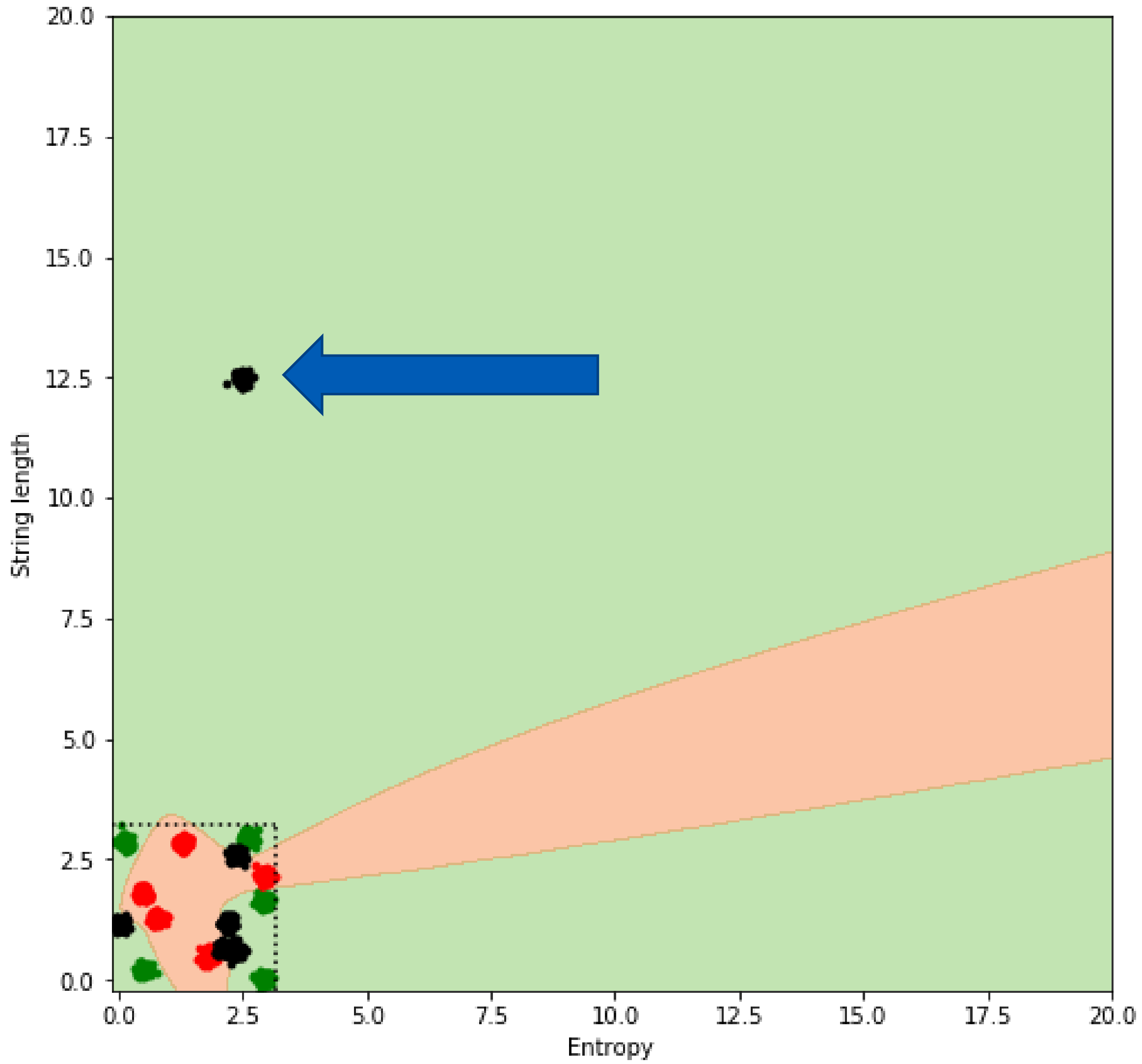
Alice replied: 'what's the answer?'
'I haven't the slightest idea,' said the Hatter.

(Lewis Carroll, 1871)

Intuition: “borderline” files are likely misclassified

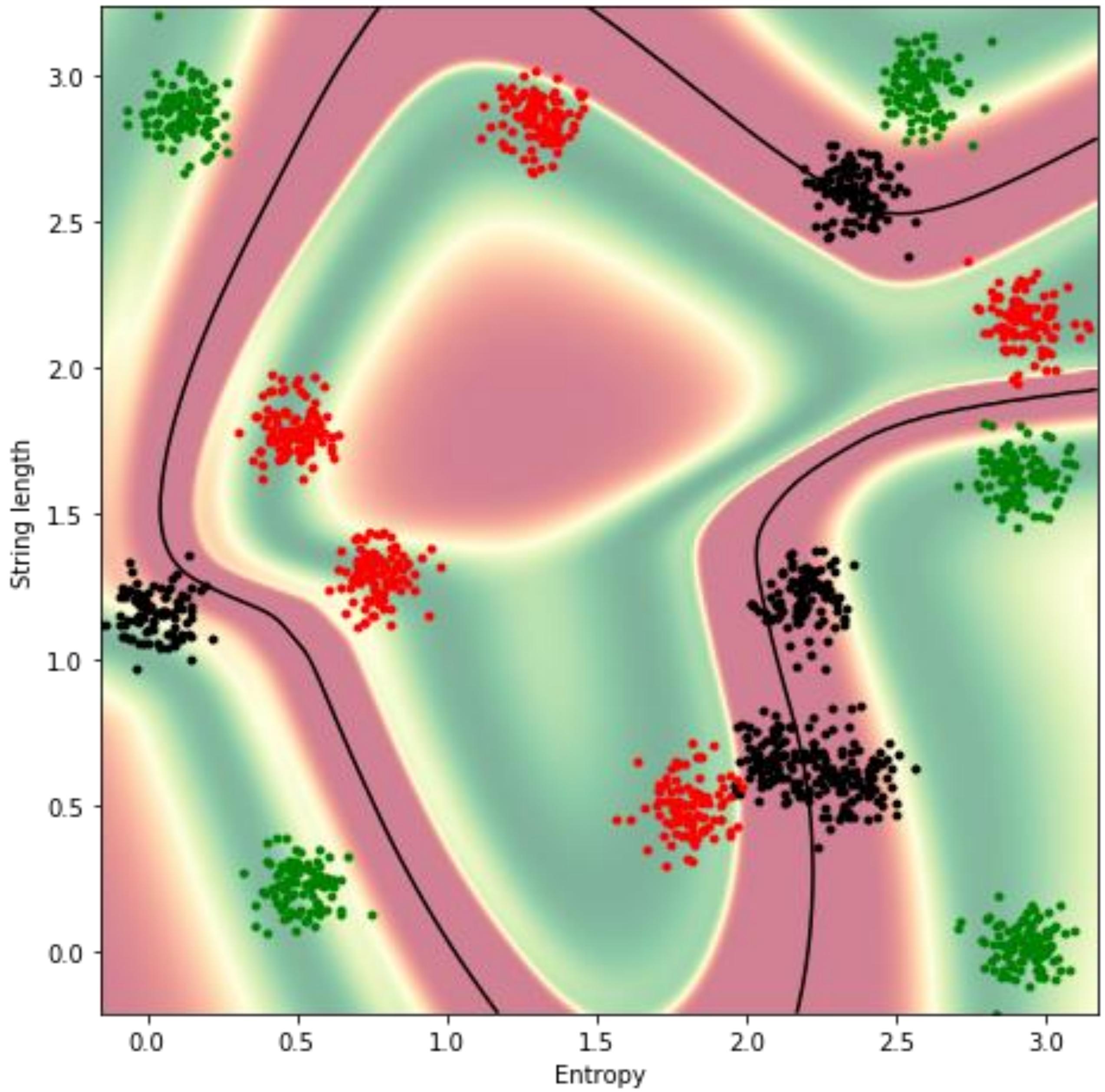


Intuition: “distant” files are likely misclassified



Do it automatically

- “Wiggle the lines” a bit
- Do the resulting classifications agree or disagree on a region?
- Amount of agreement = “Confidence”



https://arxiv.org/pdf/1609.02226.pdf

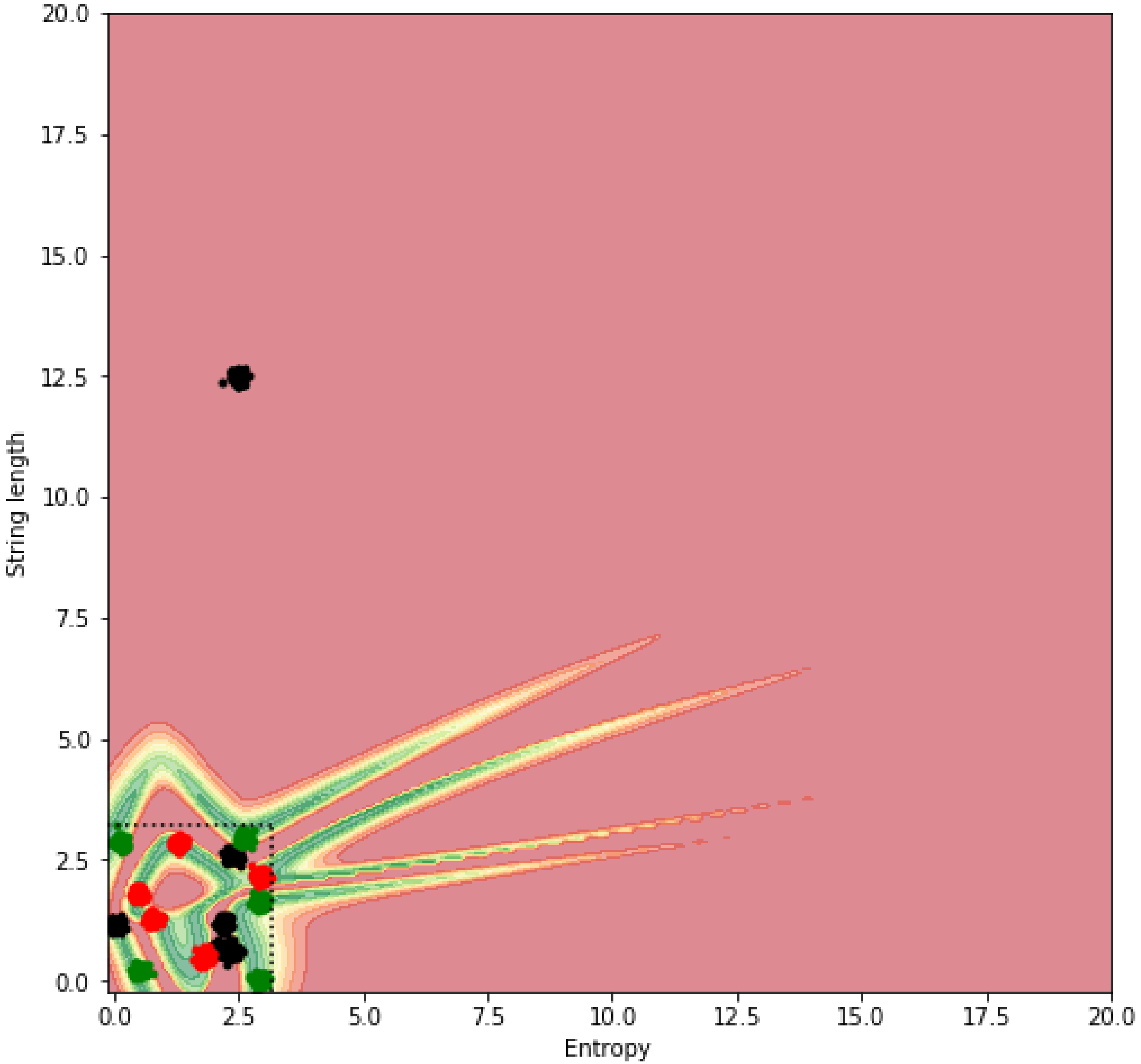
Fitted Learning: Models with Awareness of their Limits

Navid Kardan, Kenneth O. Stanley

Do it automatically

Key takeaway:

- High confidence ≈ Model has seen data like this before!
- Low confidence ≈ This data "looks new"!



https://arxiv.org/pdf/1609.02226.pdf

Fitted Learning: Models with Awareness of their Limits

Navid Kardan, Kenneth O. Stanley

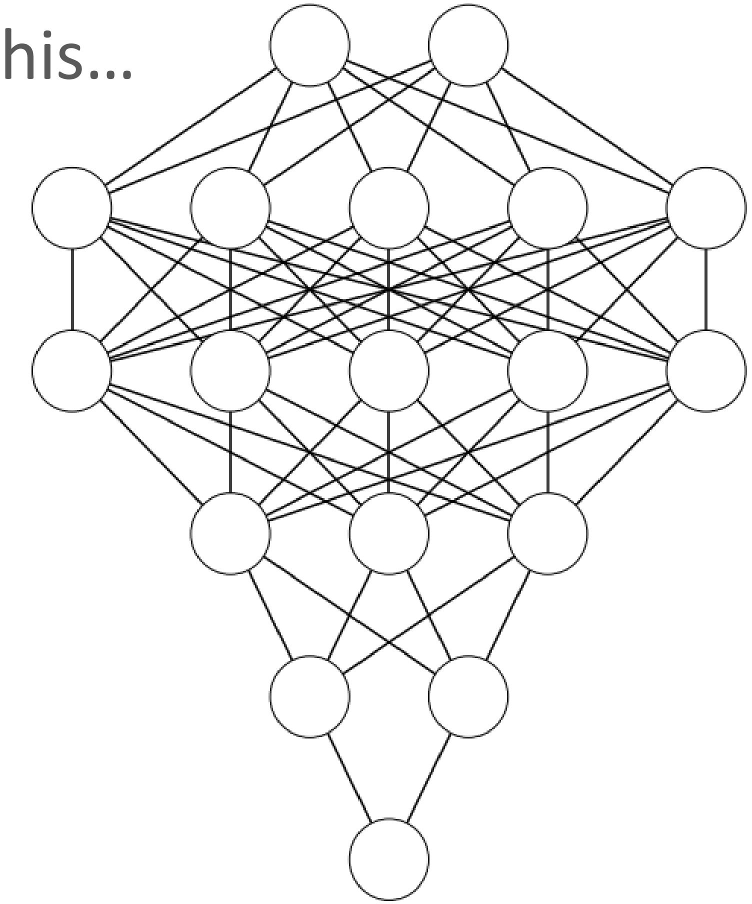
Looking at historical data with confidence

"It's a poor sort of memory that only works backwards," the Queen remarked.

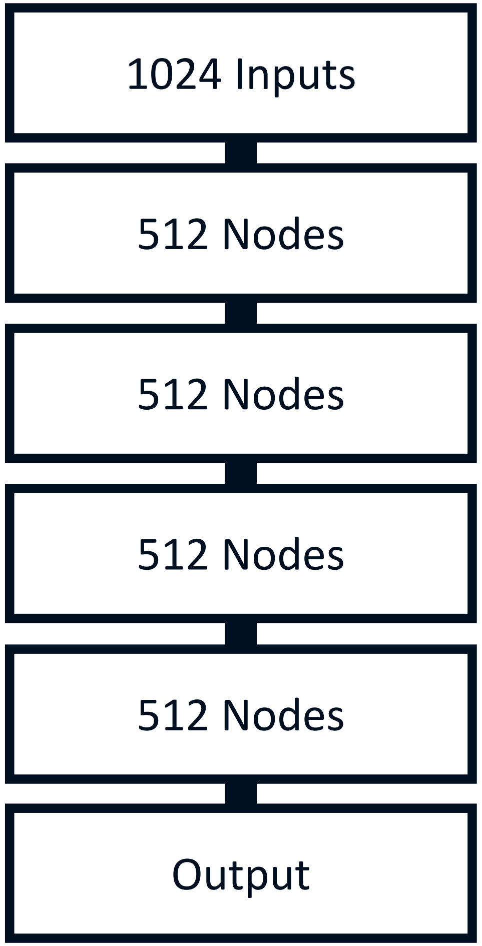
(Lewis Carroll, 1871)

Our model

Go from this…

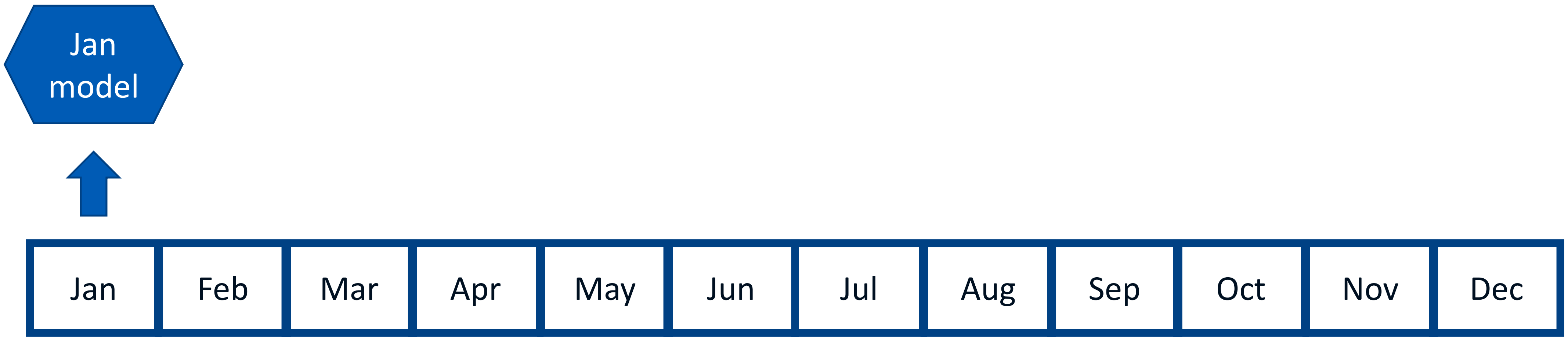


To this…



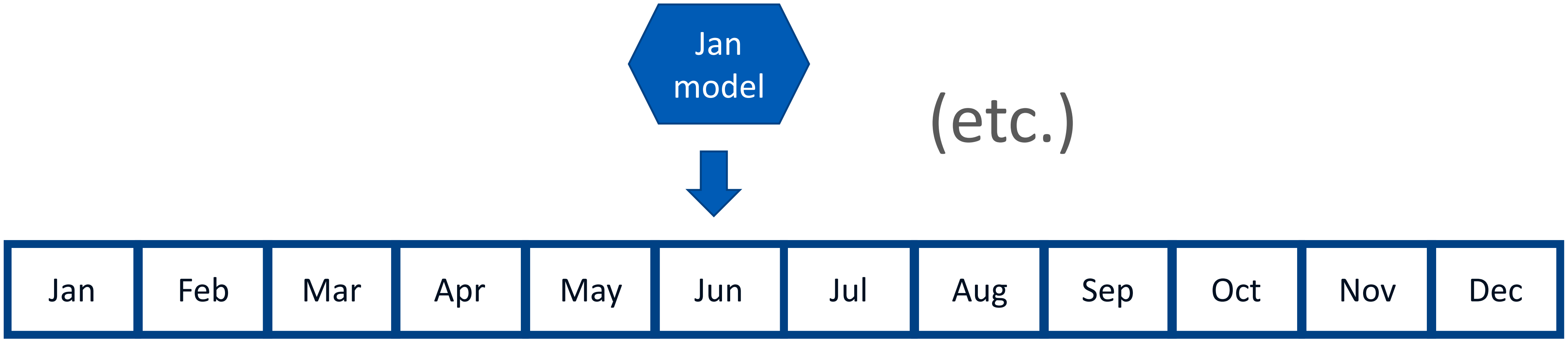
Using confidence to examine changes in malware distribution

- Collect data for each month of 2017 (3M samples, unique sha256 values)
- Train a model on one month (e.g. January)



Using confidence to examine changes in malware distribution

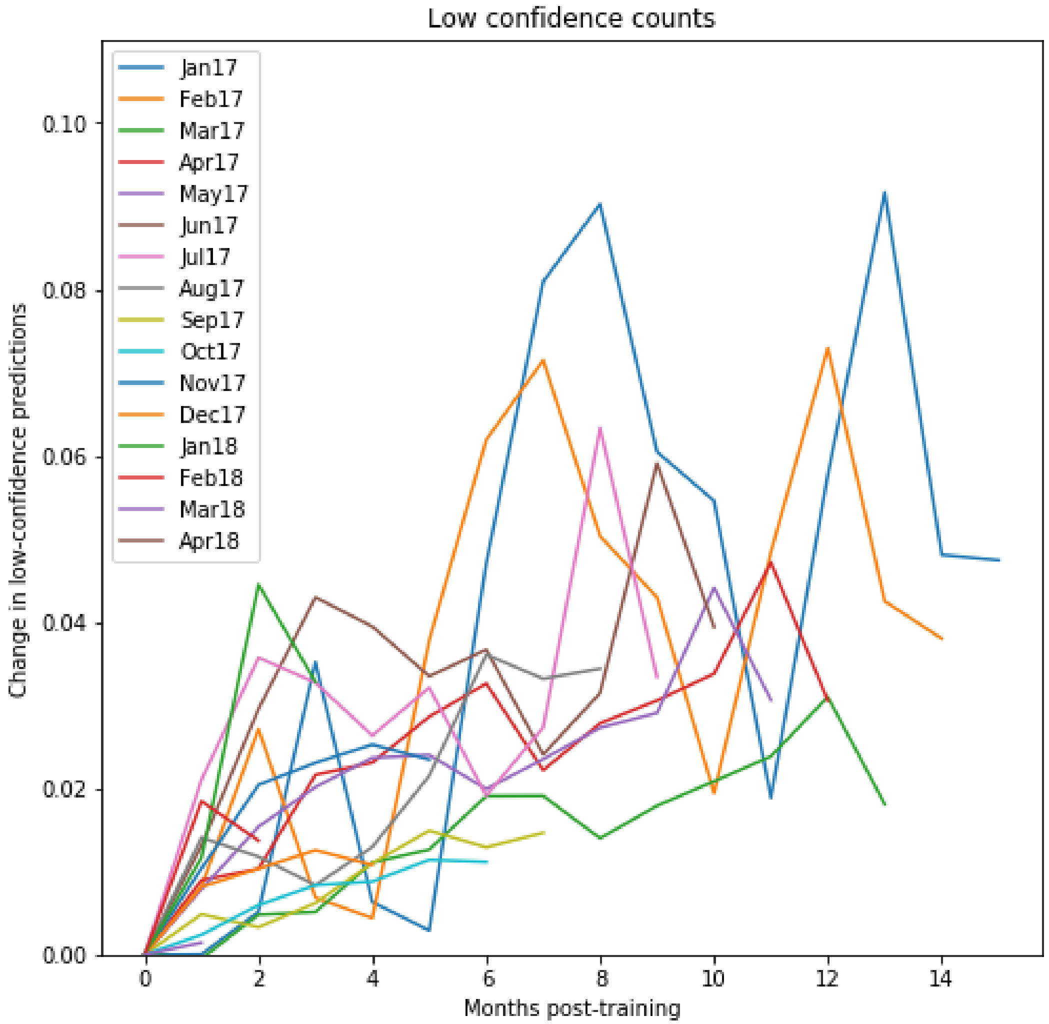
- Collect data for each month of 2017 (3M samples, unique sha256 values)
- Train a model on one month (e.g. January)
- Evaluate it on data from all future months and record the number of high/low confidence samples



Look at change in high/low confidence samples

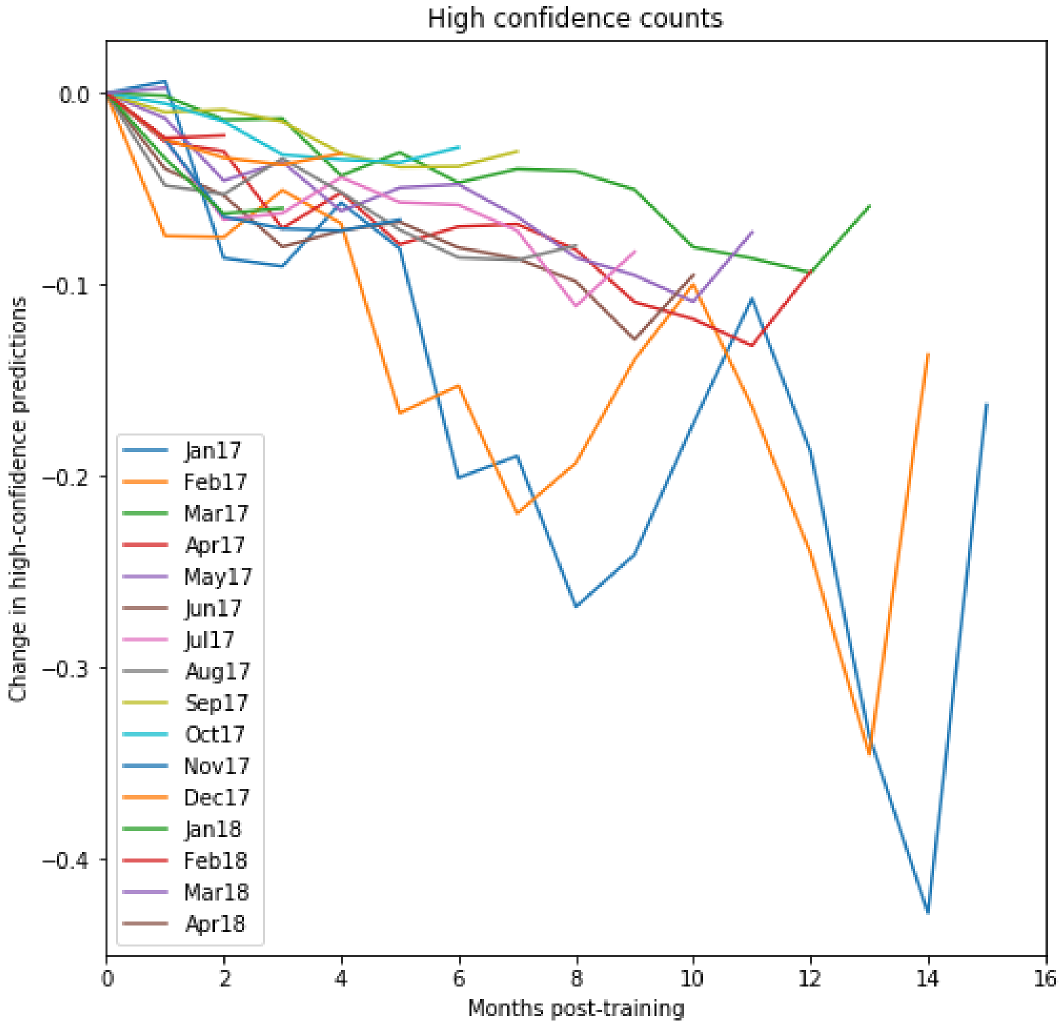
- Train January mode; count low-confidence samples for following months
- And for February
- And so on

- Remember:
  - Low-confidence = “Looks new”

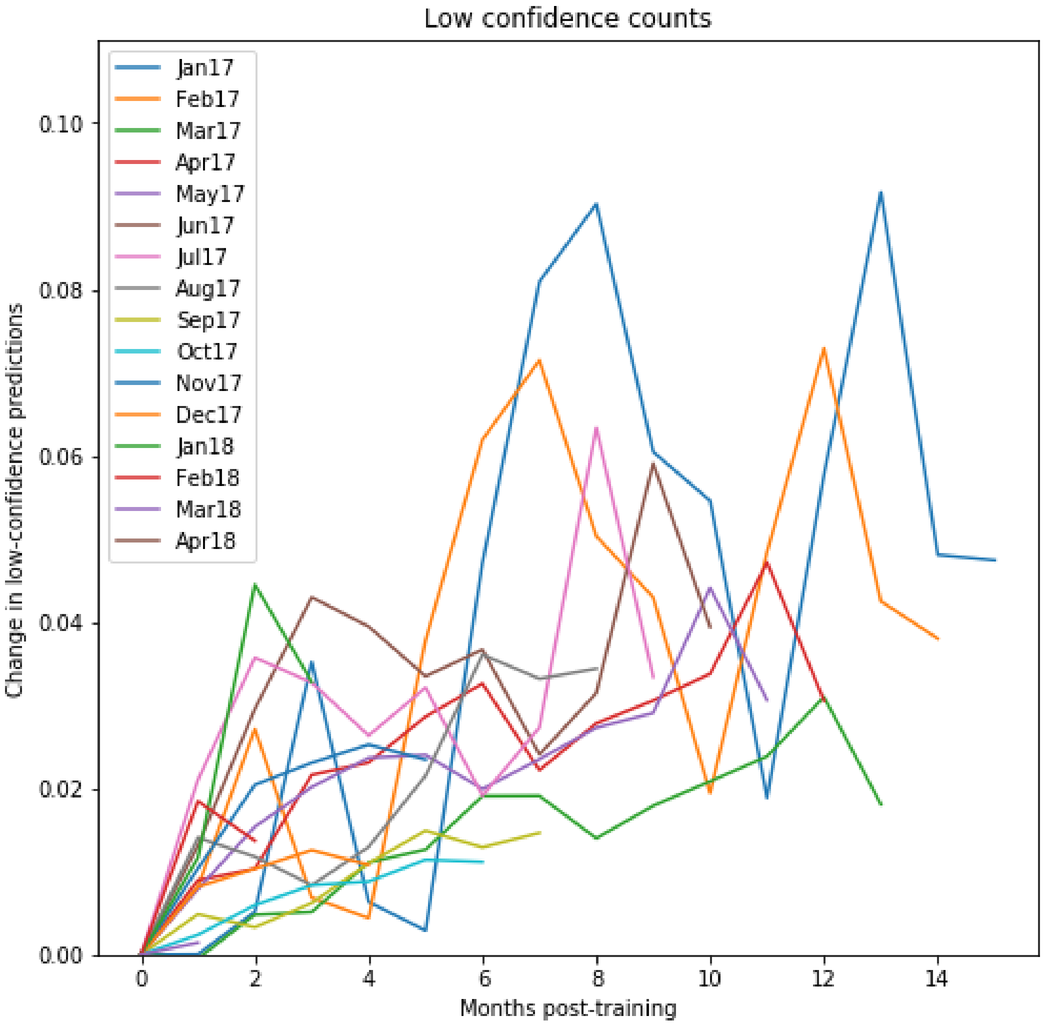


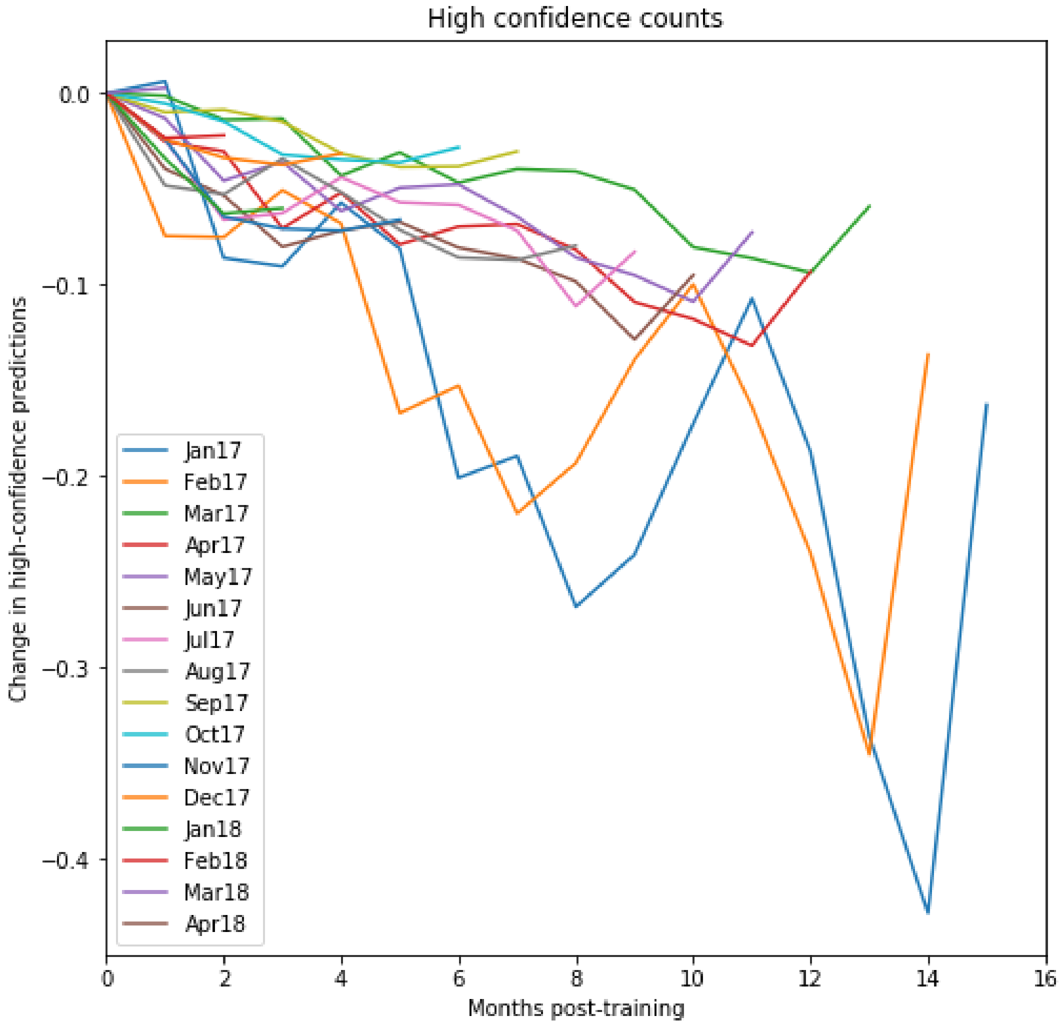
Same thing for high confidence samples

- Remember:
  - High confidence = "Looks like original data"

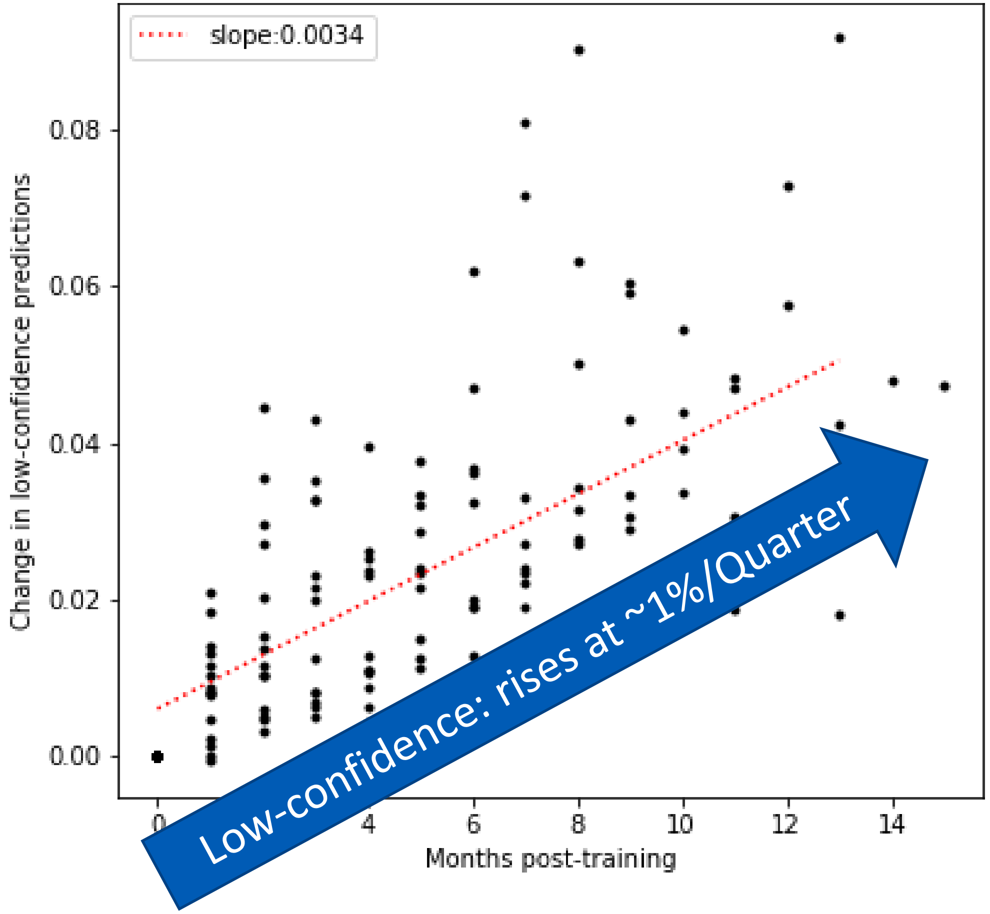


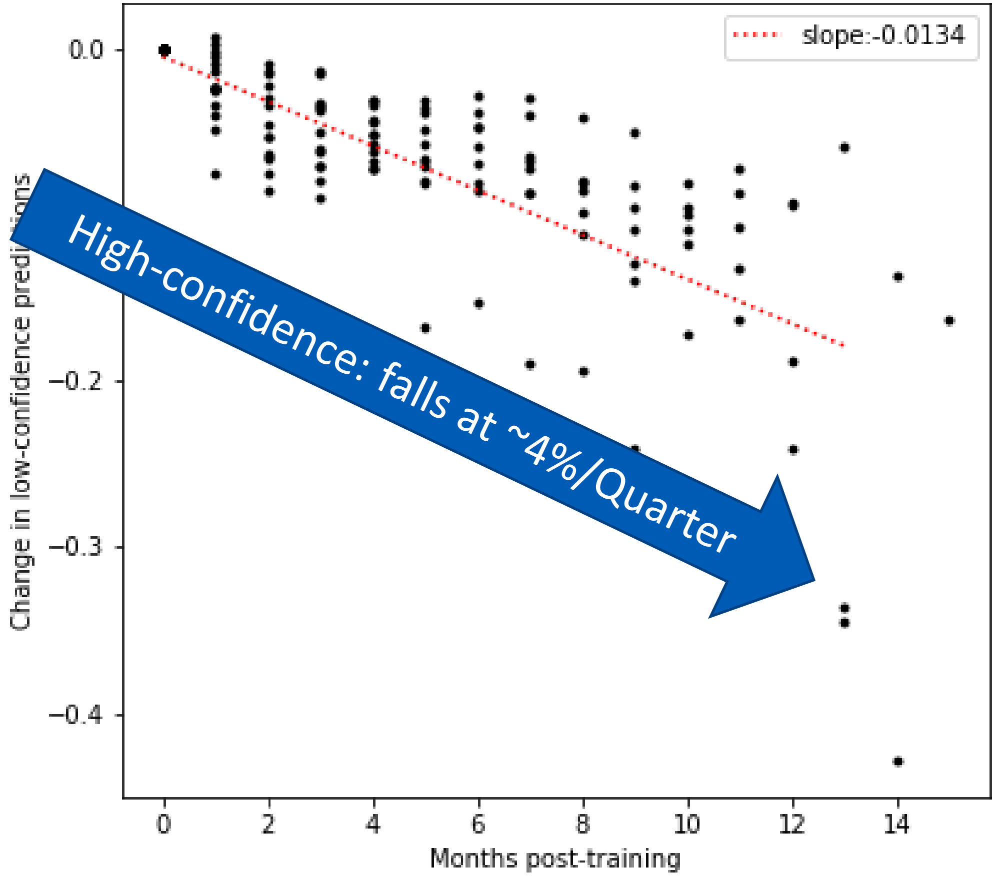
Both forms of decay show noisy but clear trends





Estimate the rates with a best-fit line



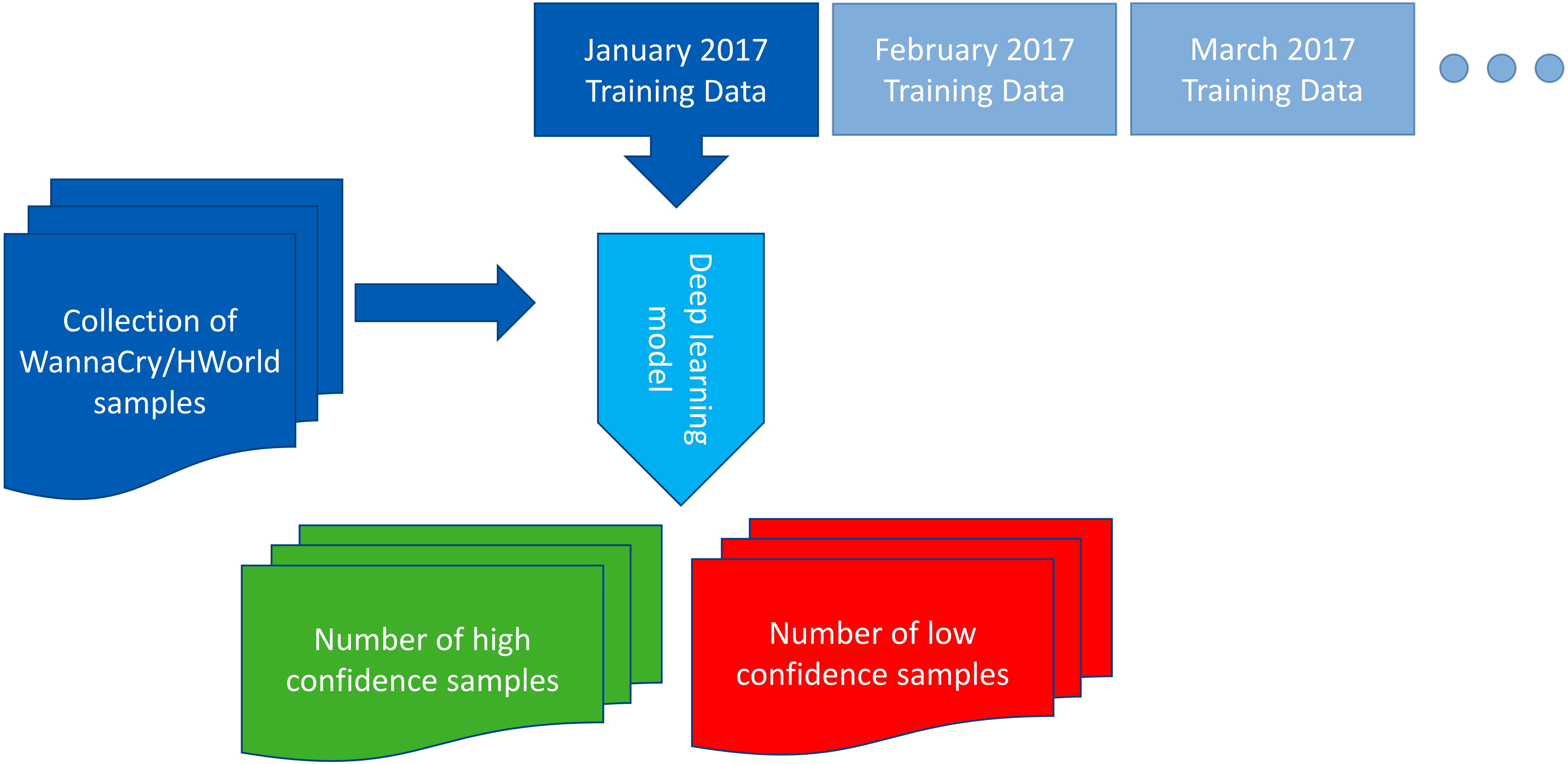


Examining changes within a single family

"I wonder if I've been changed in the night? Let me think. Was I the same when I got up this morning?"

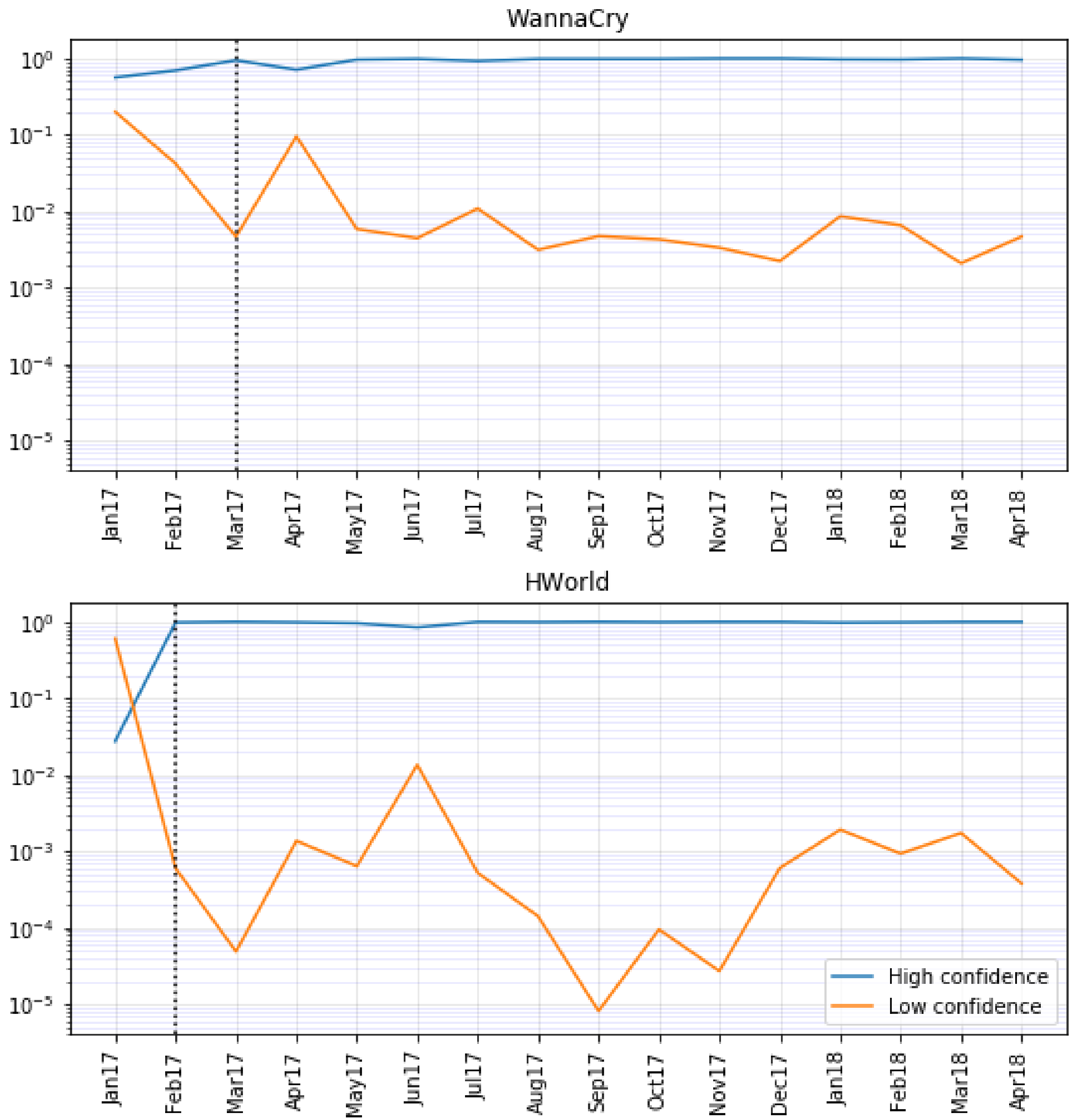
(Lewis Carroll, 1871)

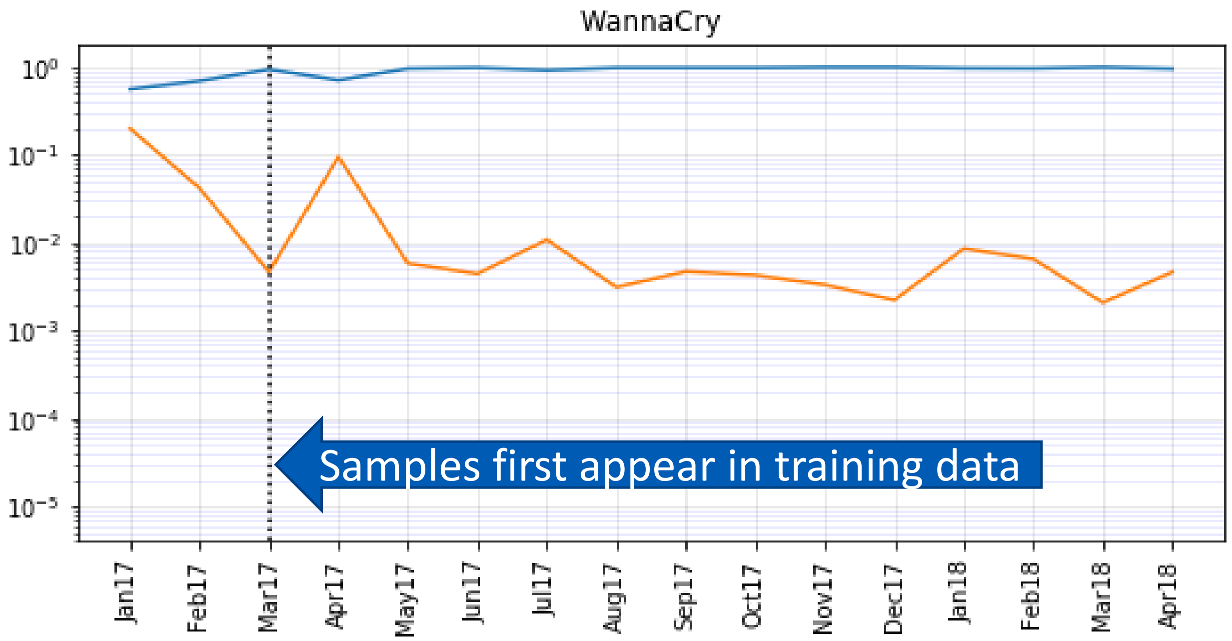
Confidence over time for individual families

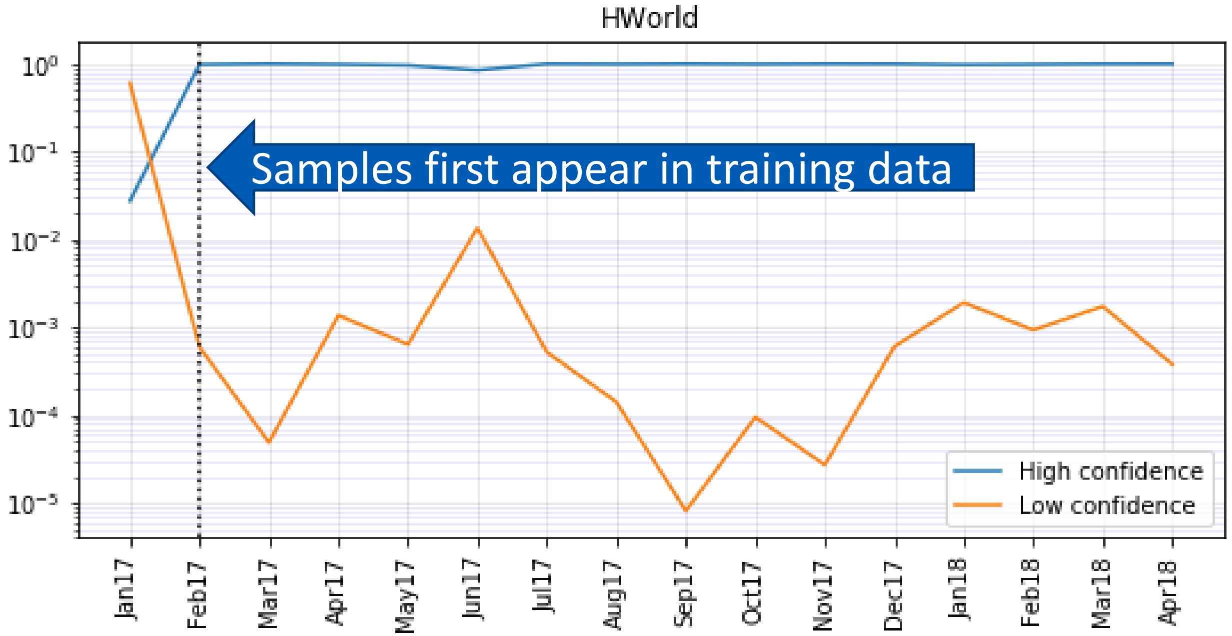


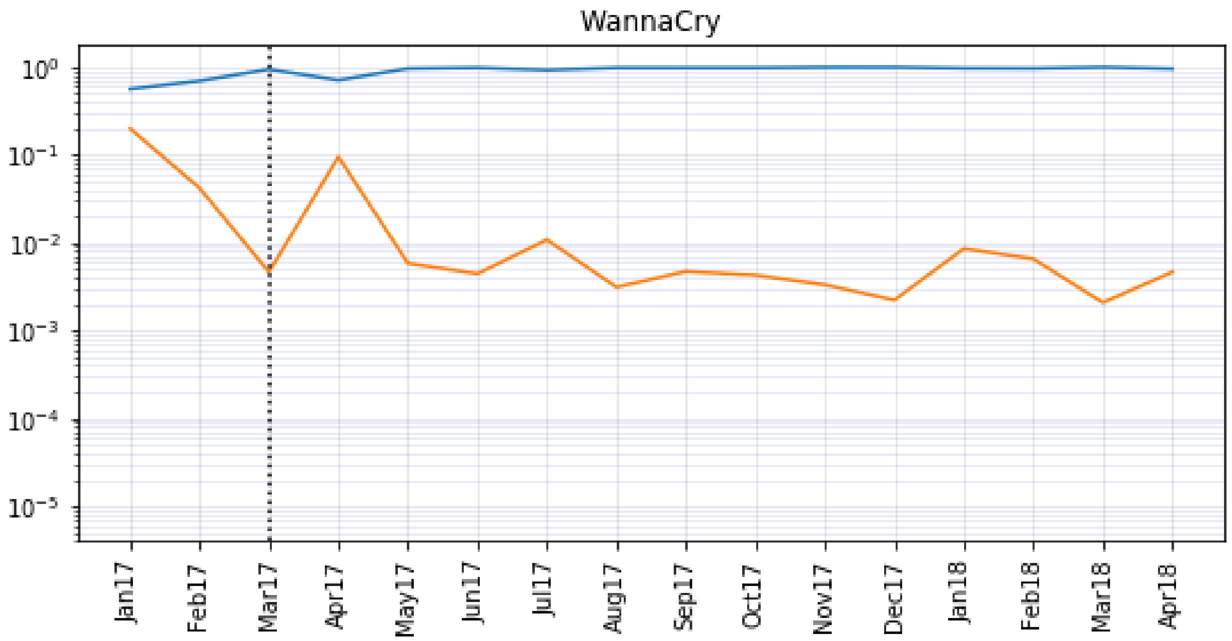
Confidence over time for individual families

- Proportion of samples for family scoring as high/low confidence vs model month

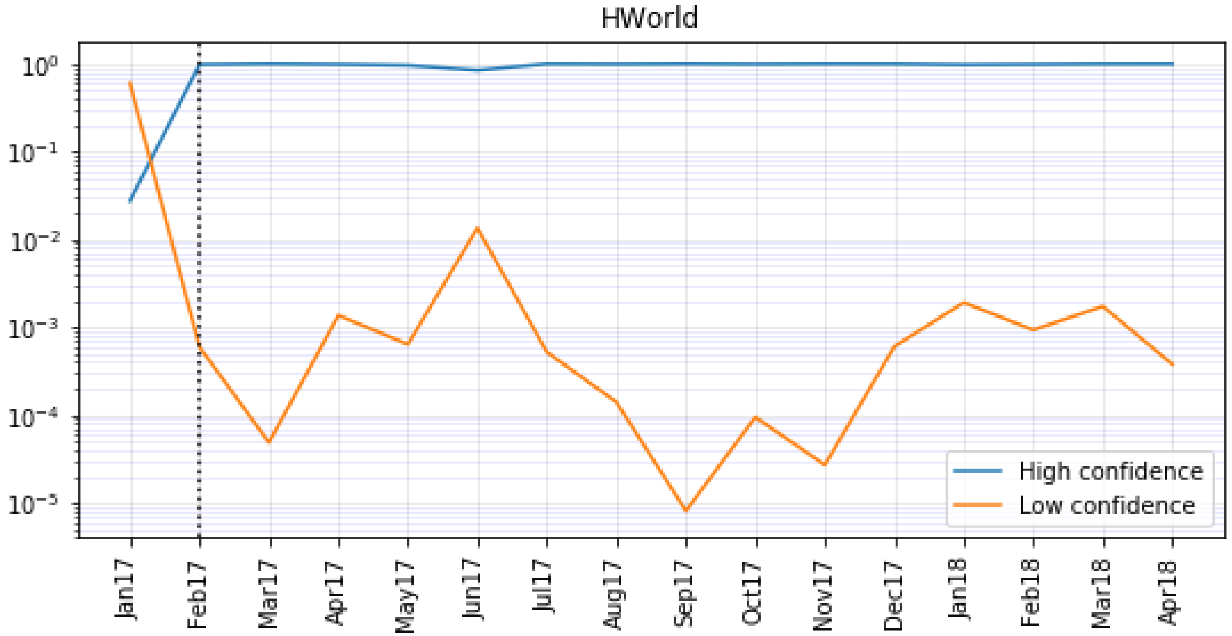



WannaCry
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
Samples first appear in training data
Jan17
Feb17
Mar17
Apr17
May17
Jun17
Jul17
Aug17
Sep17
Oct17
Nov17
Dec17
Jan18
Feb18
Mar18
Apr18

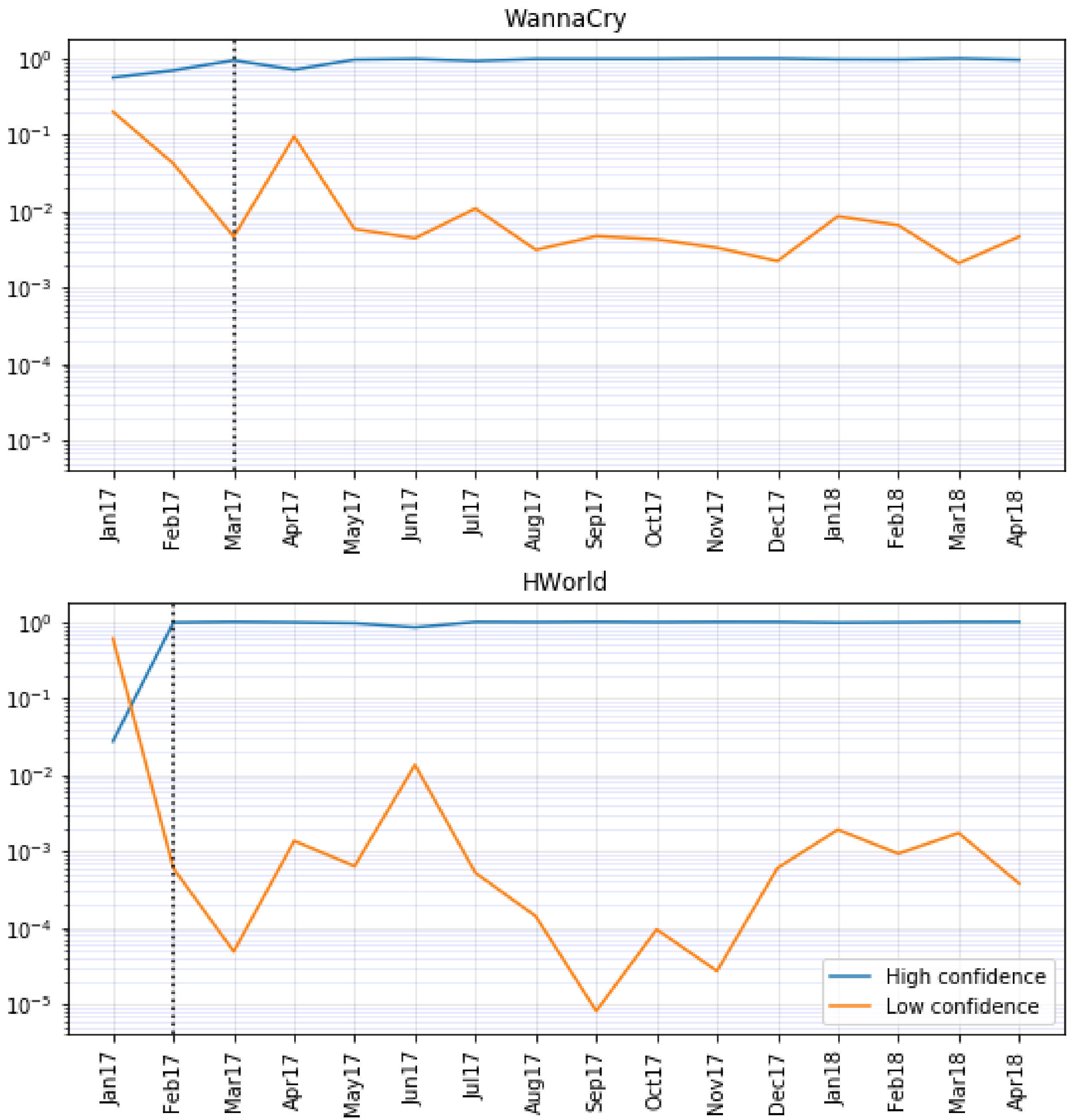

HWorld
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
Samples first appear in training data
High confidence
Low confidence
Jan17
Feb17
Mar17
Apr17
May17
Jun17
Jul17
Aug17
Sep17
Oct17
Nov17
Dec17
Jan18
Feb18
Mar18
Apr18



WannaCry/high confidence: dips as low as 70% after appearing in training data



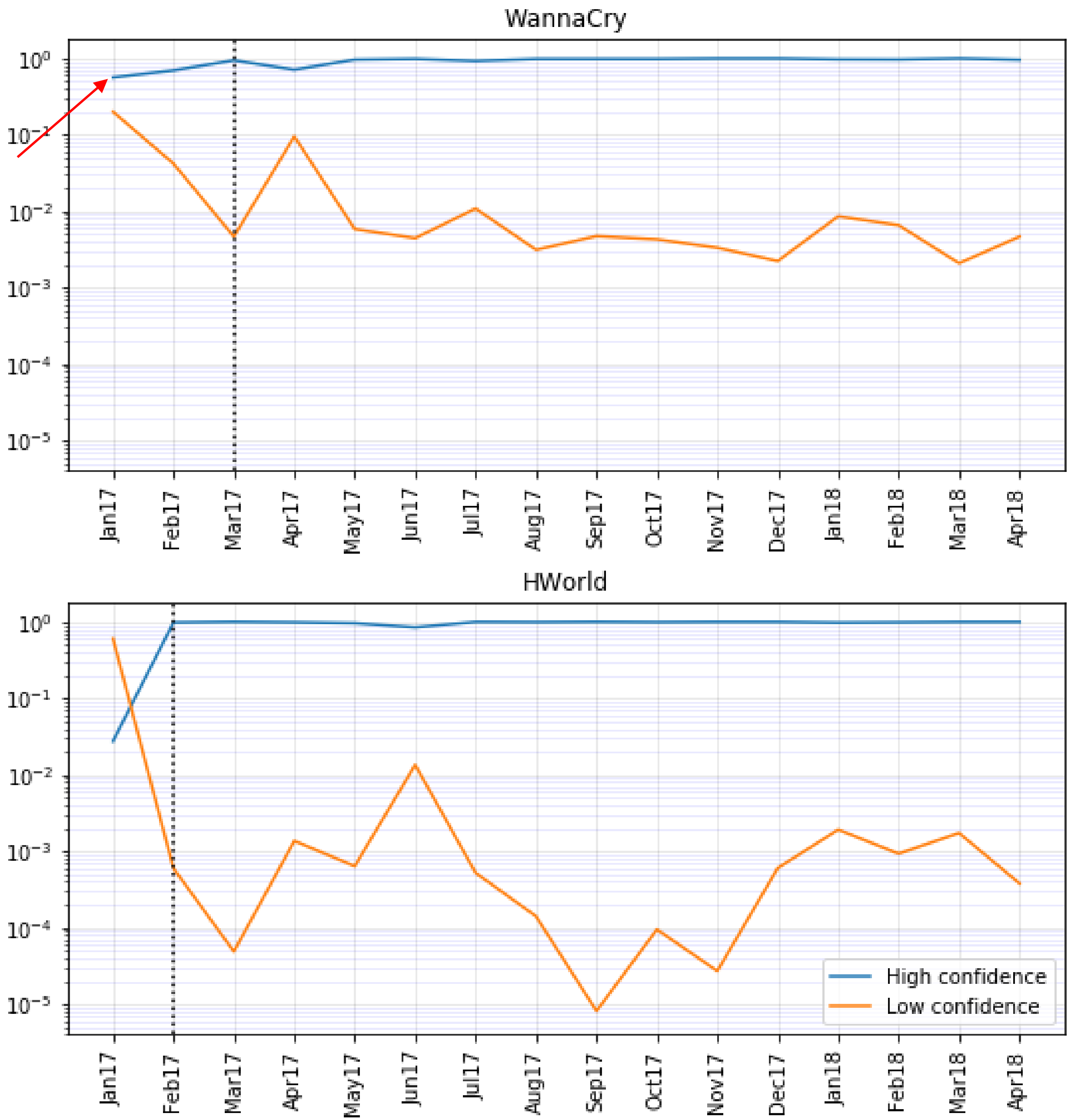
Hworld/high confidence: Never less than 84% after appearing in training data



WannaCry/low confidence: 9% down to 0.2% after appearing in training data

Hworld/low confidence: 1.3% down to 0.0008% after appearing in training data

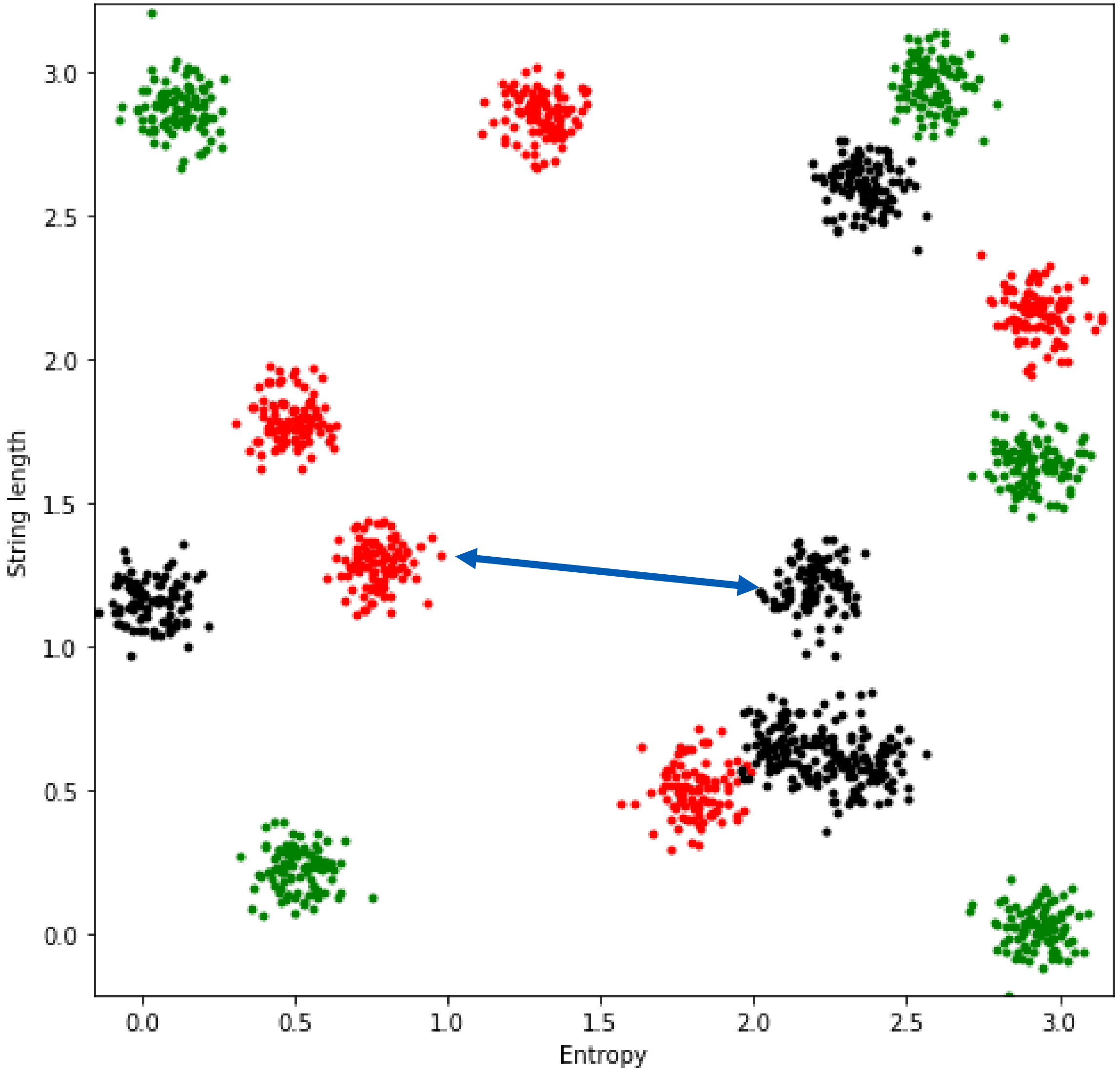
56% of WannaCry samples high-confidence *before* first appearance in training data; 99.98% detection rate in this subset



Distance measures from training data

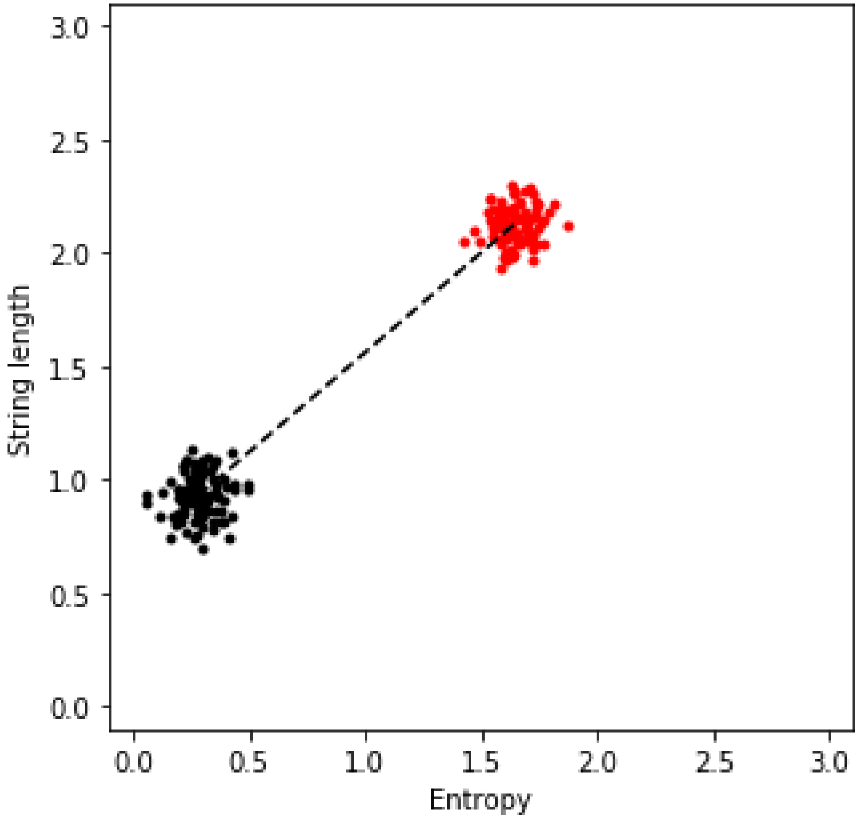
- Large distances = larger change in statistical properties of the sample
  - New family? Significant variant of existing one?

- Look at distances from one month to a later one for samples *from the same family*

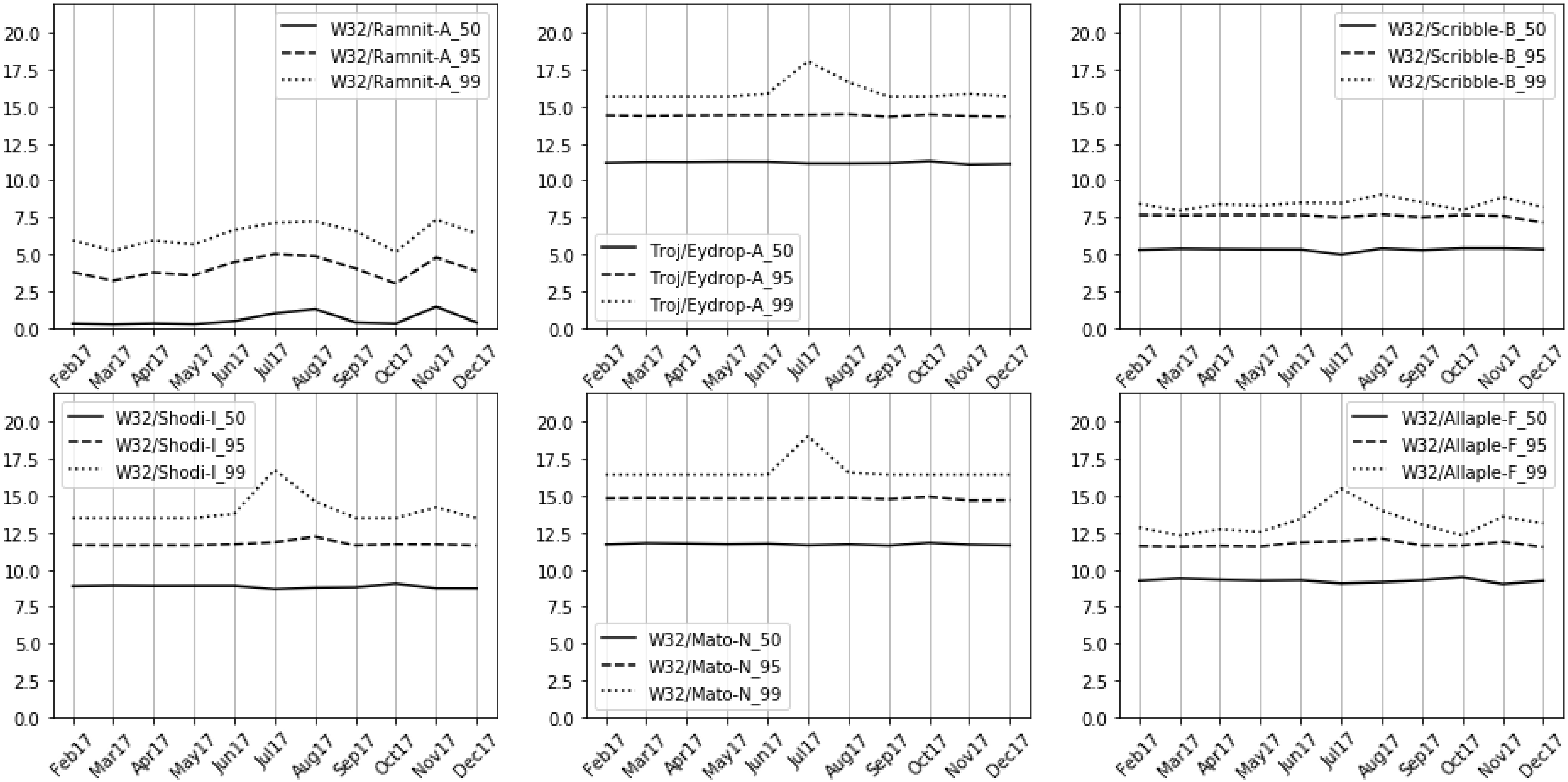


January 2017 to May 2017

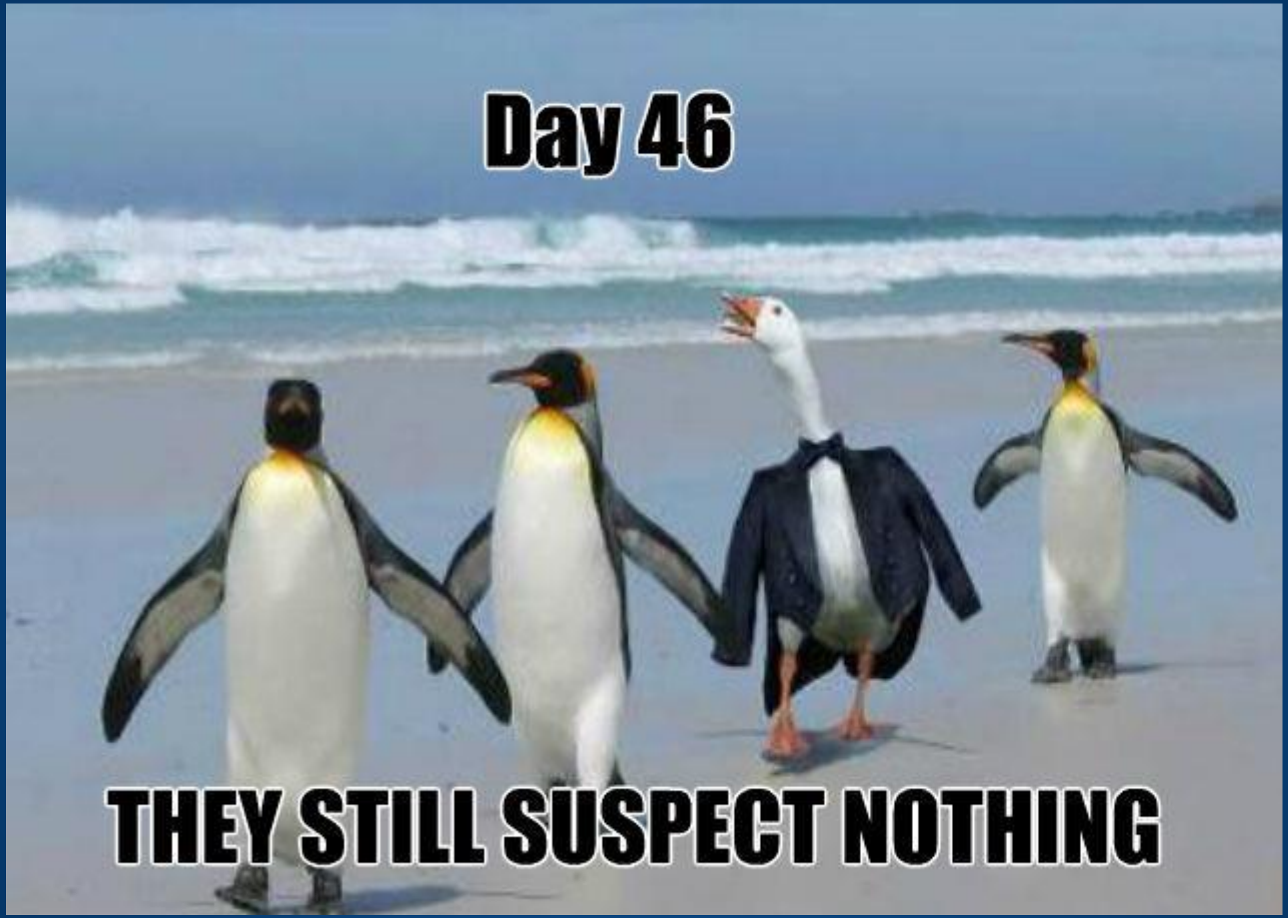
- Changes in the feature representation of samples lead to changes in distance



Distances to closest family member in training data



Distance and new family detection

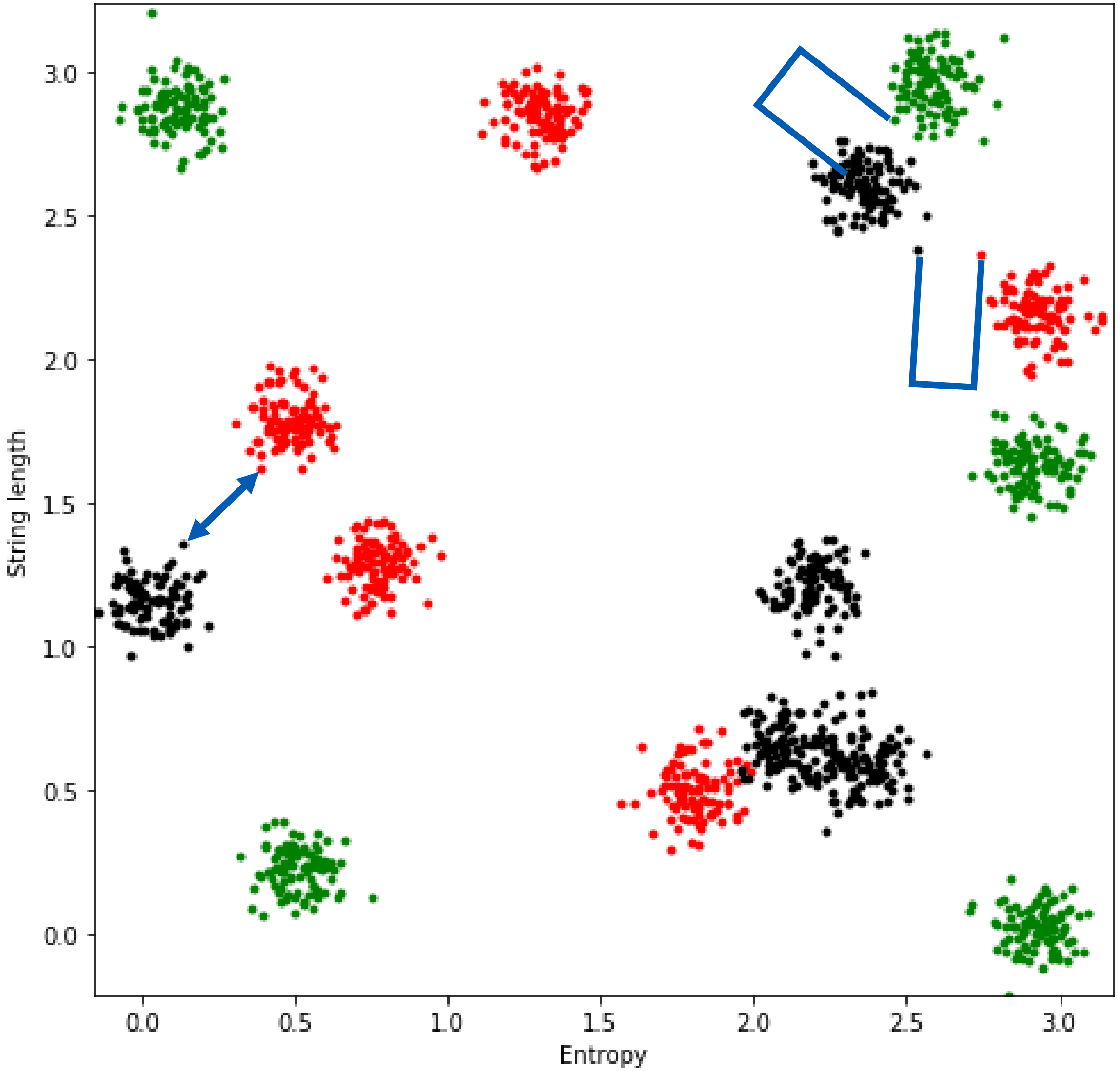


“This thing, what is it in itself, in its own constitution?”

(Marcus Aurelius, Meditations)

Distance measures from training data

- Distance to the nearest point *of any type* in the training data
- Examine against model confidence
- Don’t need labels!

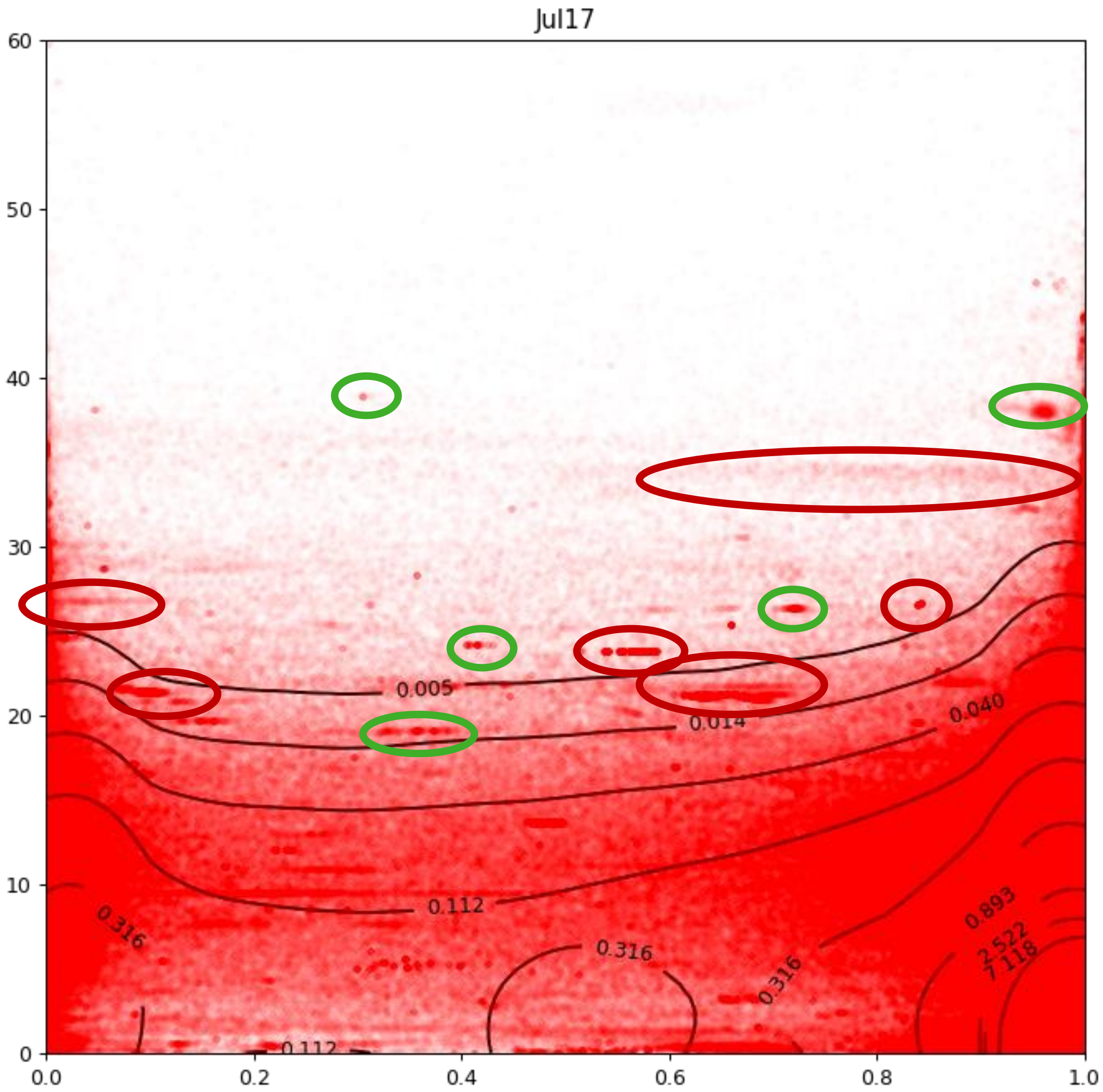


Distances – July data to nearest point in January data

Drill into clusters potentially worth examining further.

- Mal/Behav-238 – 1468 samples
- Mal/VB-ZS – 7236 samples
- Troj/Inject-CMP – 6426 samples
- Mal/Generic-S – 318 samples
- ICLoader PUA – 124 samples

… And several clusters of apparently benign samples



“Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop.”

(Lewis Carroll, 1871)

Conclusion

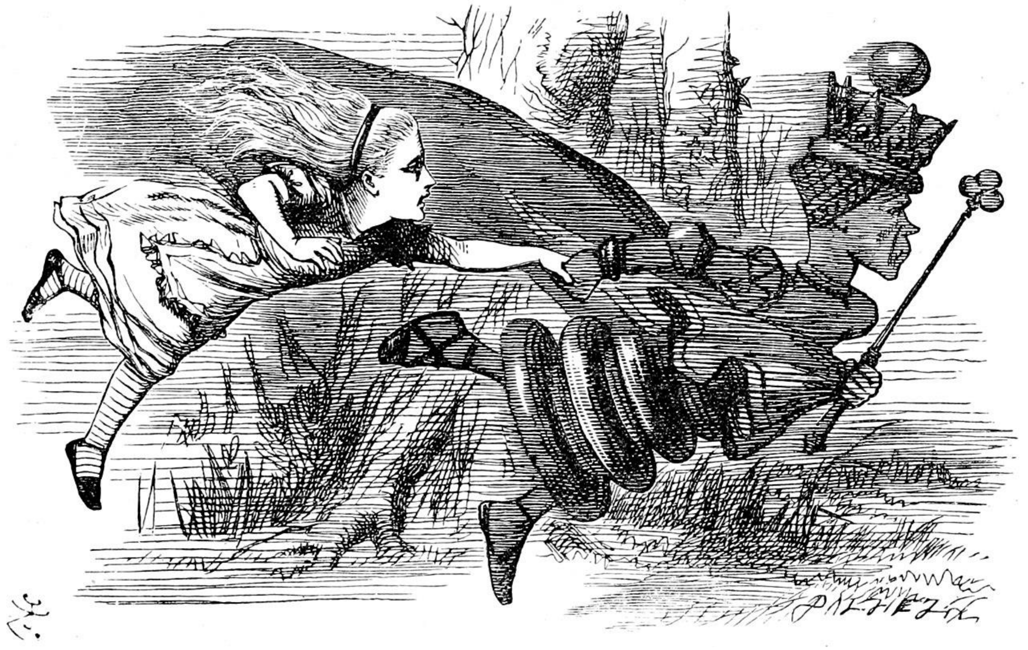
- ML models decay in interesting ways: this makes them useful as analytic tools as well as just simple classifiers
  - Confidence measures – population and family drift
  - Distance metrics – family stability, novel family detection

Practical takeaways

- ML and “old school” malware detection are complementary
  - ML can sometimes detect novel malware; compute and use confidence metrics
- The rate of change of existing malware – from the ML perspective – is slow
  - Retiring seems to be more common than innovation
- There are large error bars on these estimates, and will vary by model and data set, but…
  - Expect to see a turnover of about 1% per quarter of established samples being replaced by novel (from the ML perspective) samples
  - About 4% per quarter of your most identifiable samples will be retired

Additional thanks to…

- Richard Cohen and Sophos Labs
- Josh Saxe and the rest of the DS team
- BlackHat staff and support



- … and John Tenniel for the illustrations

- Code + tools coming soon: https://github.com/inv-ds-research/red_queens_race

SOPHOS

Cybersecurity made simple.