



THE COMPUTER SECURITY GROUP AT UC SANTA BARBARA

# MEERKAT

## Detecting Website Defacements through Image-based Object Recognition

---

**Kevin Borgolte**

Christopher Kruegel

Giovanni Vigna

`kevinbo@cs.ucsb.edu`

`chris@cs.ucsb.edu`

`vigna@cs.ucsb.edu`



August 13th, 2015  
USENIX Security 2015

## Your Moment of Zen, Mr Stewart

We are writing you today via Mr Trump's website because, seeming, the only way to get anyone to pay attention any more is to grease a Presidential candidate's website. We agree it is a regrettable state of affairs, we blame big Quinoa mostly.

.. but, we digress..

Mr Stewart, we at [@TelecomixCanada](#) would like to take this opportunity to thank you for the many happy years of quality journalism and entertainment you and your team have undertaken at Comedy Central. While even we, having wired live fire ustreams out of Gaza under Mossad's gaze, are unable to get Comedy Central's website video to work - undaunted we remain your loyal and grateful fans.

Understanding your technical interests remain unexplored you will probably be told of this by one of your most excellent producers. Know, Sir, that your steadfast dedication to the irony and power of Truth has inspired a generation which we ourselves now serve. That our collective thanks appears here will, we hope, amuse you as much as it will them.

We note, with some annoyance, that your natural opponents are beginning to talk smack about you (and presumably your mother).. Labouring, perhaps, under the misapprehension that it is once again business as usual while you enjoy your days tossing ball with the scion and evenings pursuing Mrs Stewart round the sofa with ice tongs.

Be comforted Mr Stewart, as a direct result of [your good work](#) these many years, they labour in false hope.

Our role at Telecomix is largely custodial, demolition we leave to those more talented. In point of fact, this represents the first time our promissory has added a message to a device not our own [in some years](#). It is a measure of the respect with which we hold you and the depth of appreciation we have of your time with us.

Should you ever come to wonder what the stars look like over the North Atlantic on cool clear evening, flag us. Our currency is greatly devalued at the moment, bring a popup if you can, tents aren't as comfortable.

Before Mr Trump's 3 dollar website people (perhaps the [Elephant murderer](#) over at godaddy) figure out how to remove this thank you note, over to you [John Oliver](#).

So far, very promising .. Are you aware we have a federal election coming up between a belligerent cowboy Economist, an angry Irishman with a French passport and a young dad with nice hair? We are unsure if they are registered on Ashley Madison, but fertile ground for an Englishman to properly introduce himself to his colonial brothers and sisters to the north never the less.

Well that's about it, other than to join you all in celebrating America's first openly Asshole Presidential Candidate. Godspeed Mr Trump.

Bests always.. [@TelecomixCanada](#)

[#DataLove](#) [#MMM2015](#)

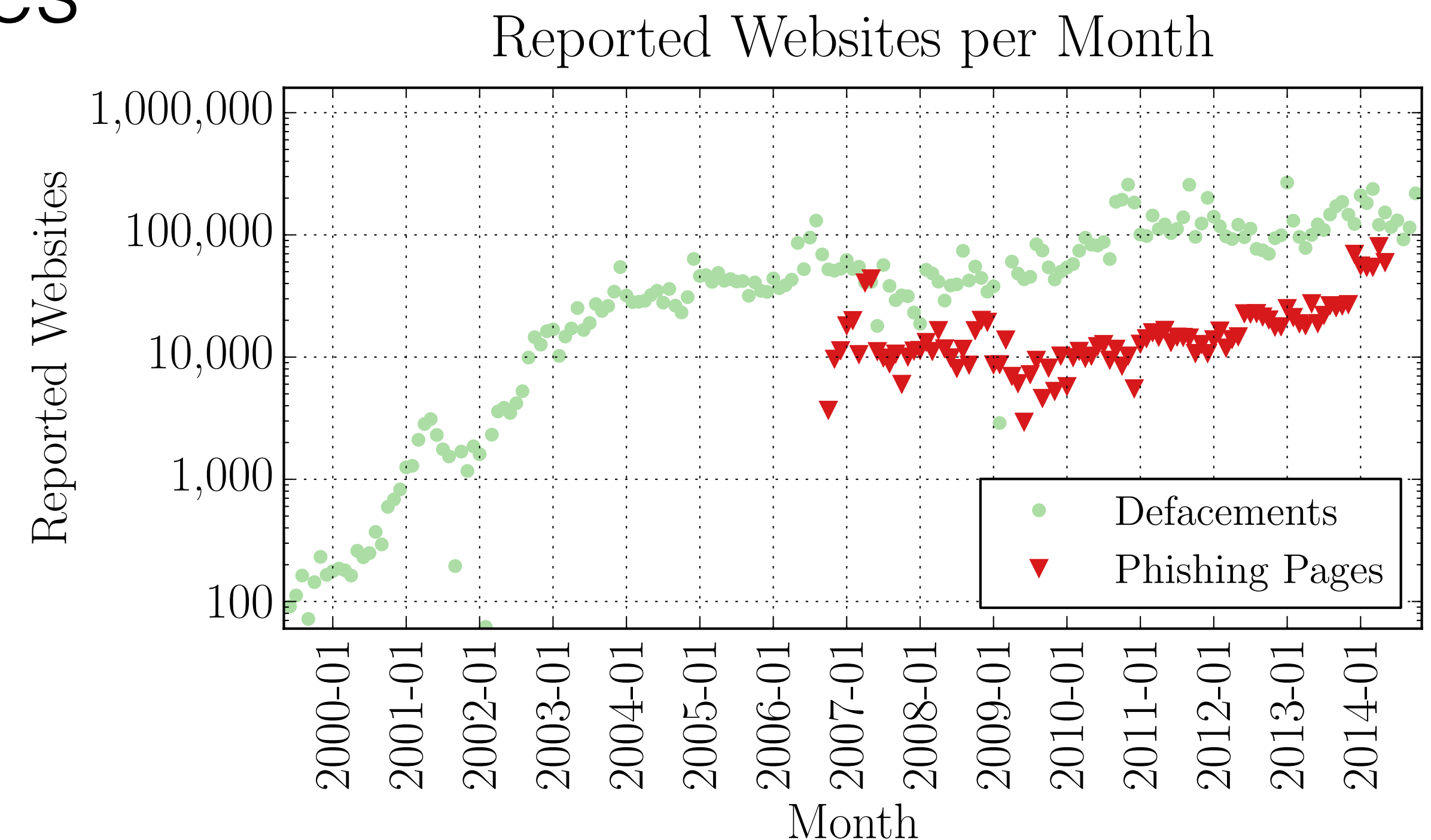


[Archived](#)



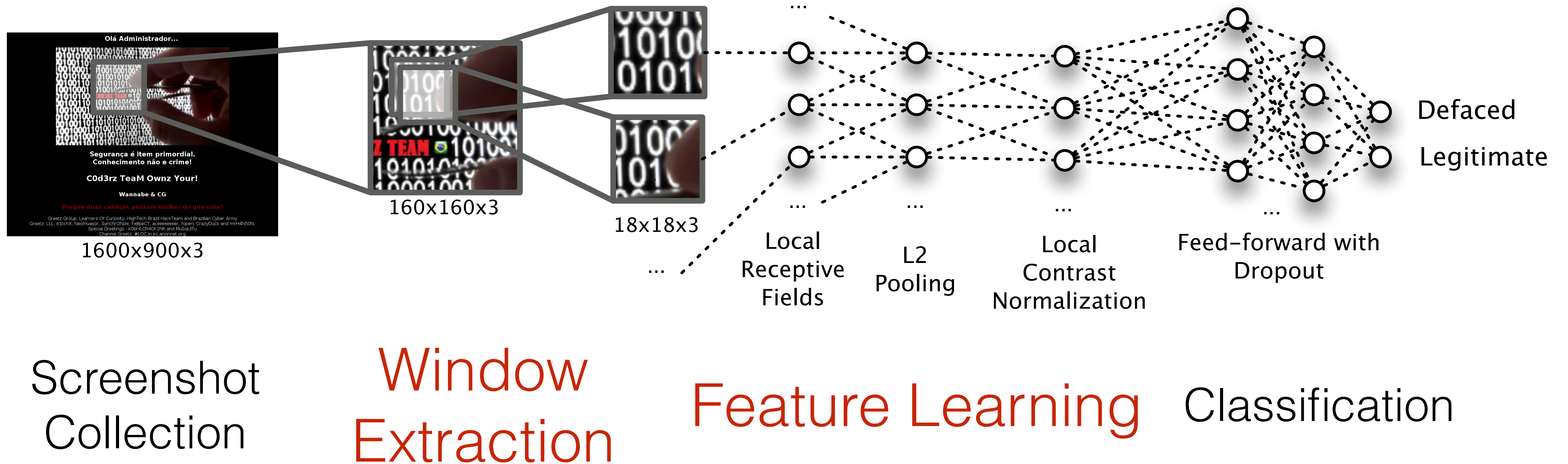
Source: The Register, August 3rd 2015, [http://www.theregister.co.uk/2015/08/03/trump\\_website\\_hacked/](http://www.theregister.co.uk/2015/08/03/trump_website_hacked/)

- Prolific defacers:  
Team System Dz, 2,800 websites in 10 months (~10/day)
- Over 4,700 manually-verified defacements each day (Zone-H)
- Defacements to Phishing Pages  
Average: ~7 to 1  
Maximum: ~33 to 1
- Over 53,000 websites from top 1 million lists were defaced in 2014

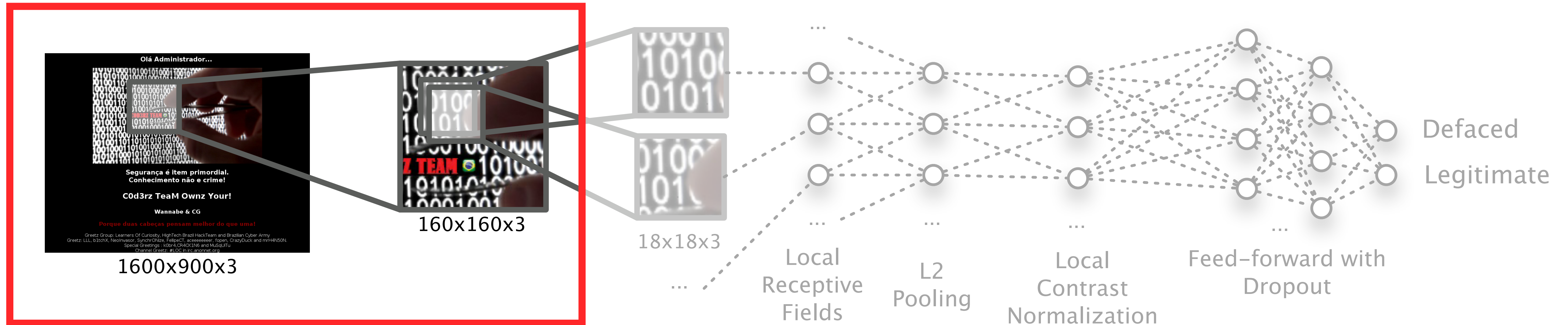


- Prior work looks at websites' code, host-based IDS etc.
  - Compares to prior version / known good state
- MEERKAT: Visually, like a human analyst
  - Render website in browser
  - Take screenshot
  - Does the screenshot look like a defacement?
- No previous version of website needed
- No manual feature engineering

# Approach: Deep Neural Network

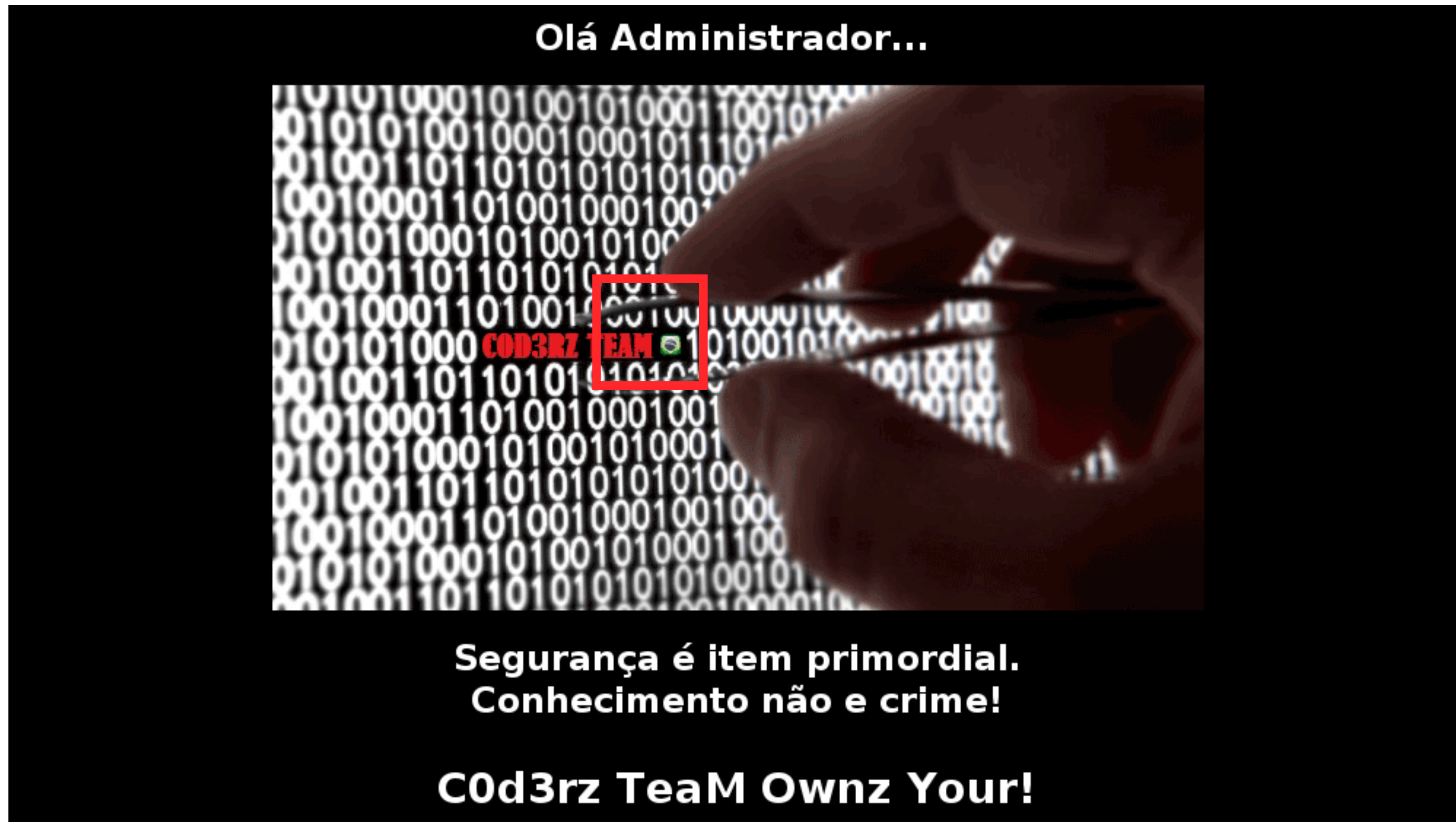


# Approach: A Window “Into” The Defacement



- Full-size screenshots impractical; window “into” defacement instead
- Size of window?
  - Too large  $\Rightarrow$  overfit (high variance)
  - Too small  $\Rightarrow$  underfit (high bias)
- Extract window from which part of the screenshot?

# Approach: Representative Window Extraction (1) **seclab**



# Approach: Representative Window Extraction (2)

---

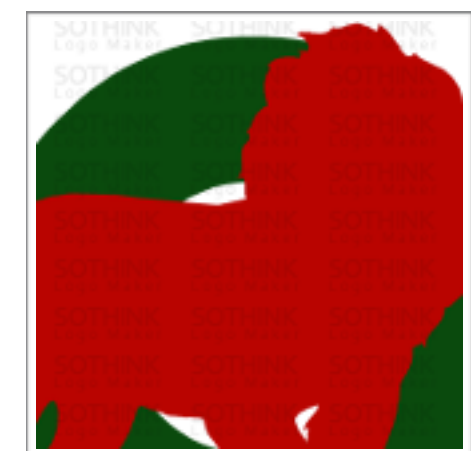
- Sample windows from 2-dimensional Gaussian distribution
  - Bias heavily toward center of page
  - If outside of screenshot, resample
- Why?
  - Center of page is descriptive!
  - Non-trivial to poison training set

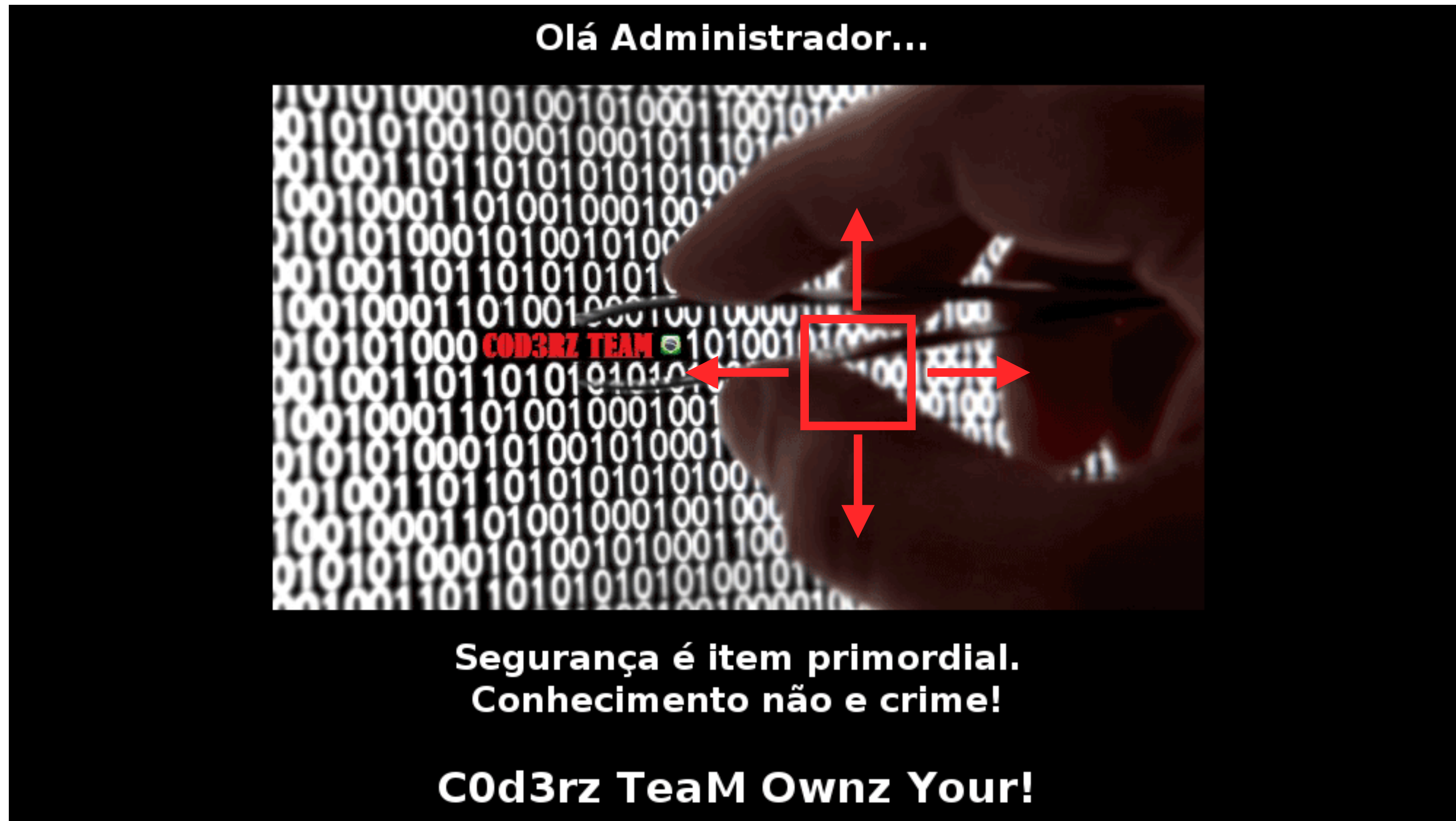


- Feature Learning:  
Stacked Auto-Encoders
- Classification:  
Feed-Forward Neural Network with Dropout
- Implemented on-top of Caffe
- Trained on GPU, training time in weeks

# Approach: Features Learned

- Color combinations
  - Green on black? Black on white/bright gray?
- Letter combinations
  - Broken and mixed encodings
- Leetspeak
  - “pwned” or “h4x0red”
- Typographical and grammatical errors
  - “greats to” or “visit us in our website”
- Defacement group logos





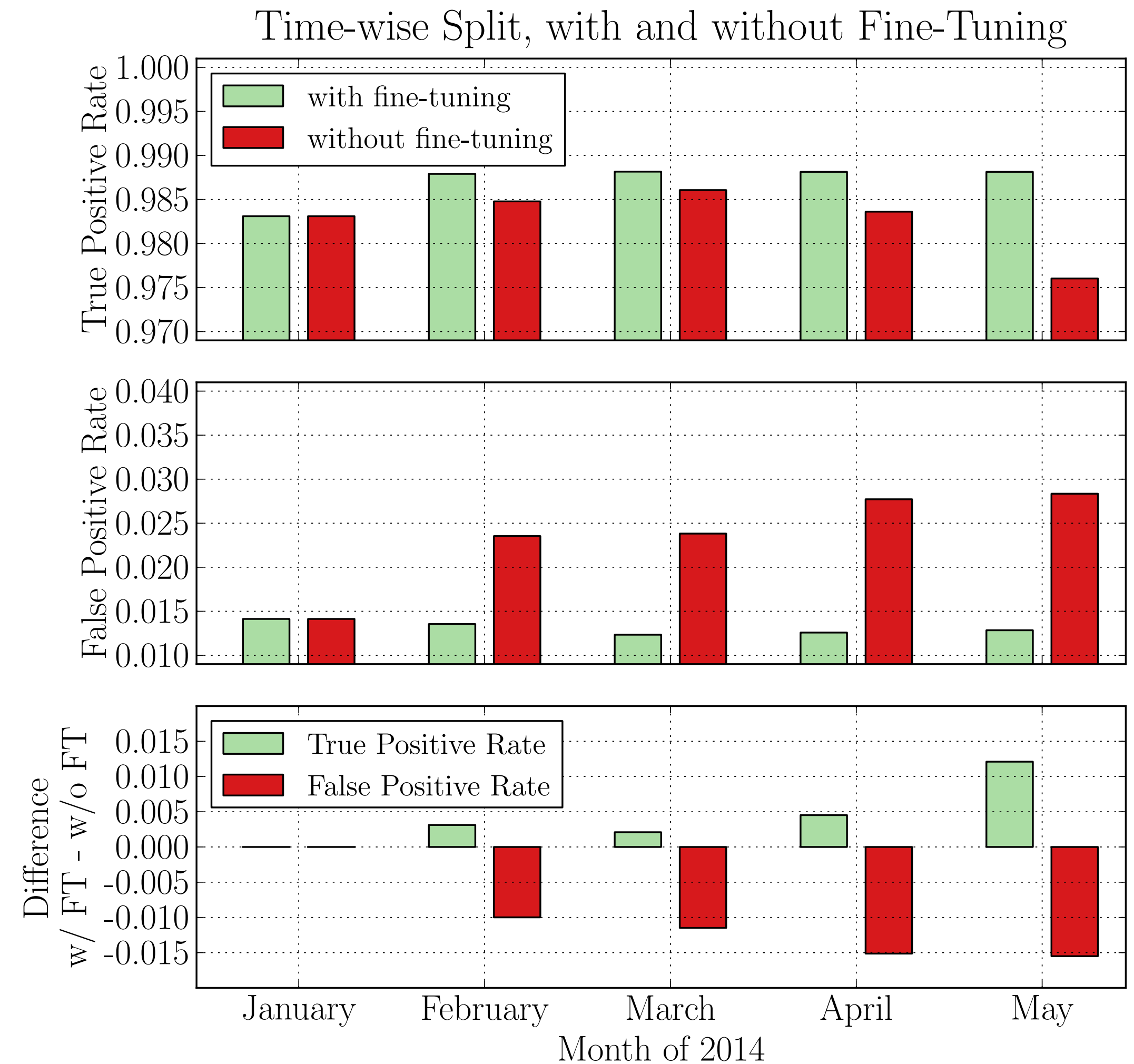
- Dataset
  - 10,053,772 defaced websites = positives
    - Defaced websites manually verified by Zone-H
  - 2,554,905 legitimate websites = negatives
    - Legitimate websites not verified, might be defaced
- Dataset skewed toward defacements
  - Report Bayesian detection rate (BDR):  $P(\text{true positive}|\text{positive})$
- Unverified legitimate websites  $\Rightarrow$  TPR & BDR are lower-bounds!

- 10-fold cross-validation
- Results:
  - TPR: avg. 97.878% [97.422%, 98.375%]
  - FPR: avg. 1.012% [0.547%, 1.419%]
  - BDR: avg. 99.716% [99.603%, 99.845%]
- Traditional evaluation has problems:
  - Same defacement possibly in two bins
  - Defacements from 1998 vs. 2014

- Fingerprinting and delayed defacements
- Tiny defacements
- Huge advertisements
- Concept drift (natural and adversarial)
  - Major: learn new features from new data (no feature engineering)
  - Minor: adjust weights of deeper classification layer

# Limitations: Minor Concept Drift & Fine-Tuning

- Train on Dec 2012 to Dec 2013
  - 1.78 million defacements
  - 1.76 million legitimate pages
- Test on Jan to May 2014
  - 1.54 million samples, 50/50 split
- Fine-tune Jan, Feb, Mar, Apr
- BDR in Jan: 98.583%
  - w/o FT drops to 97.177%
  - w/ FT increases to 98.717%
- Team System Dz started Jan 2014!



- Introduced MEERKAT
- Learns features automatically, match domain knowledge
- Does not require prior version of website
- Outperforms state of the art
- Gracefully tackles minor and major concept drift



Thank you for your attention!

**Questions?**



kevinbo@cs.ucsb.edu  
<http://kevin.borgolte.me>  
twitter: @caovc