Defending Networks with Incomplete Information: A Machine Learning Approach

Alexandre Pinto
alexcp@mlsecproject.org
@alexcpsec
@MLSecProject

Agenda

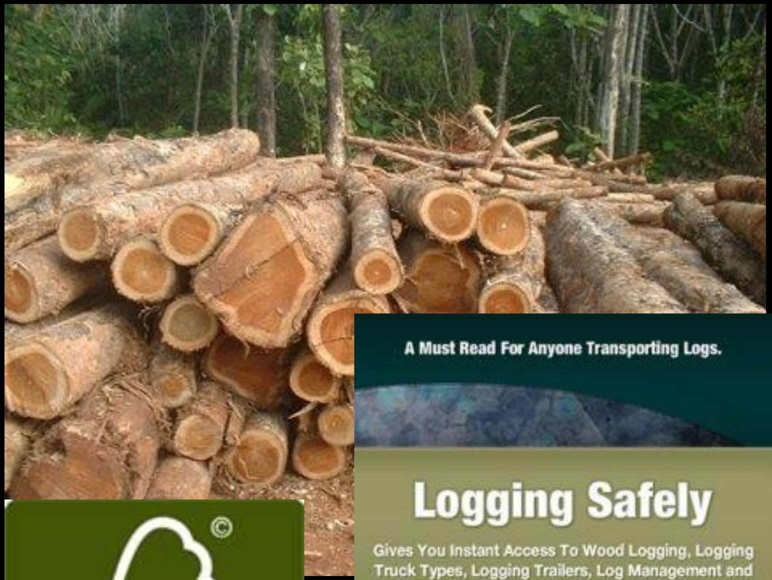
- Security Monitoring: We are doing it wrong
- Machine Learning and the Robot Uprising
- More attacks = more data = better defenses
- Case study: Model to detect malicious agents
- MLSec Project
- Acknowledgments and thanks

Who's this guy?

- 12 years in Information Security, done a little bit of everything.
- Past 7 or so years leading security consultancy and monitoring teams in Brazil, London and the US.
  - If there is any way a SIEM can hurt you, it did to me.
- Researching machine learning and data science in general for the past year or so. Active competitor in Kaggle machine learning competitions.

The Monitoring Problem

- Logs, logs everywhere
- Where?
  - Log management
  - SIEM solutions
- Why?
  - Compliance
  - Incident Response



Monitoring / Log Management is Hard

- Gartner Magic Quadrant for Security Information and Event Management 2013.
  - “Organizations are failing at early breach detection, with more than 92% of breaches undetected by the breached organization”
  - “We continue to see large companies that are re-evaluating SIEM vendors to replace SIEM technology associated with partial, marginal or failed deployments.”
- Are these the right tools for the job?

Monitoring / Log Management is Hard

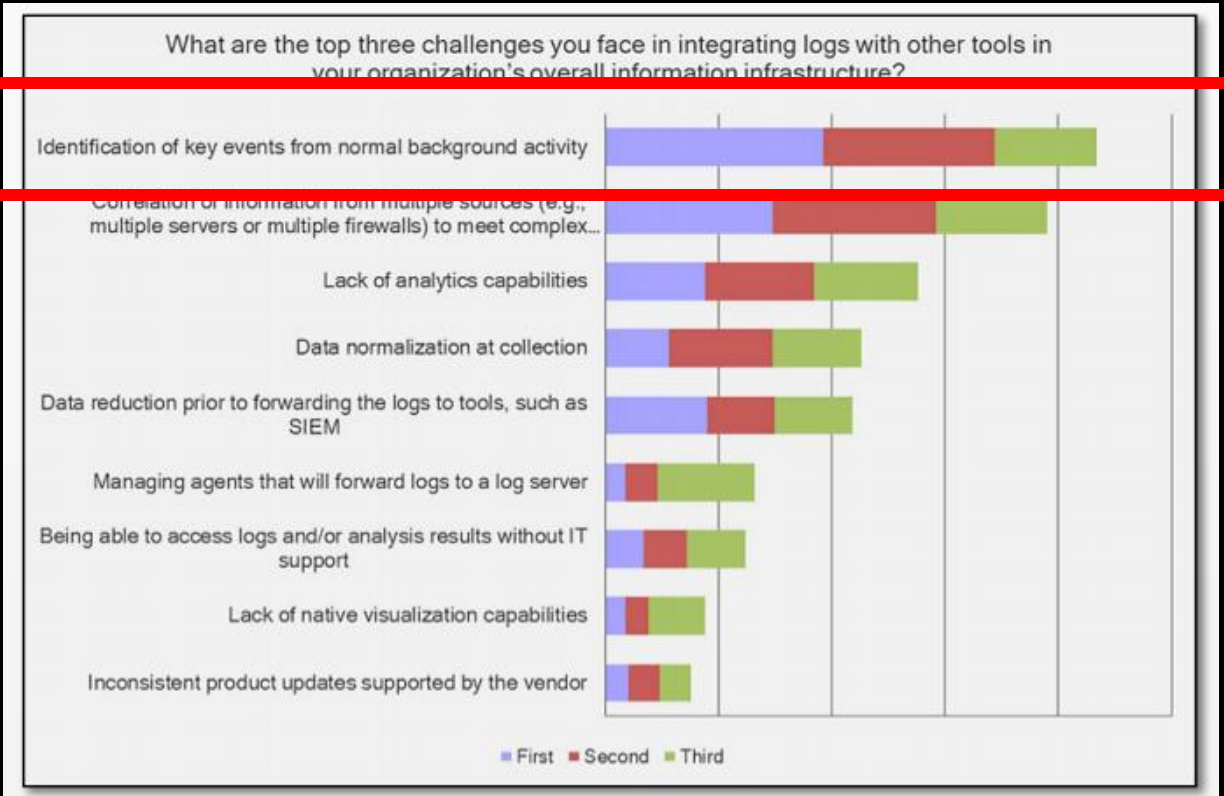


Figure 3. First, Second and Third Most Challenging Aspects of Log Management and Integration

- SANS Eighth Annual 2012 Log and Event Management Survey Results (http://www.sans.org/reading_room/analysts_program/SortingThruNoise.pdf)

Not exclusively a tool problem

- However, there are individuals who will do a good job
- How many do you know?
- DAM hard (ouch!) to find these capable professionals

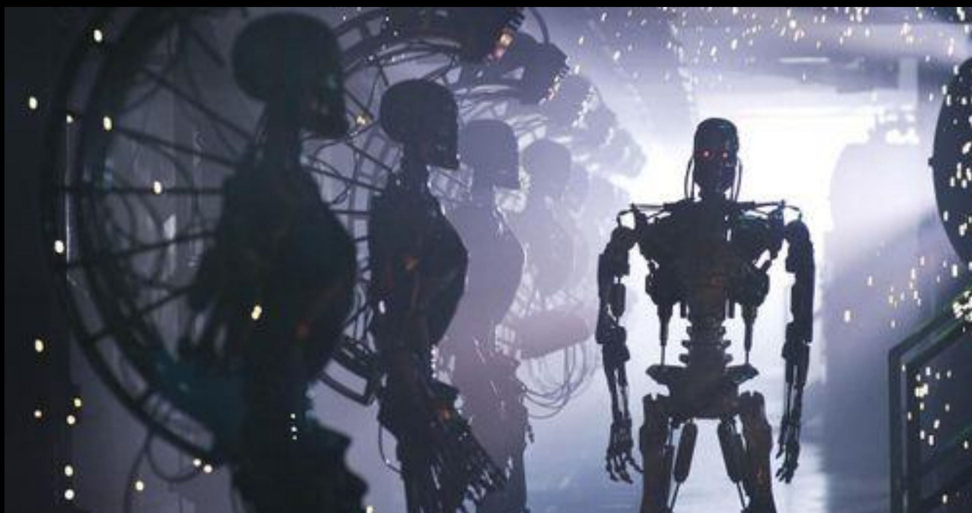
Next up: Big Data Technologies

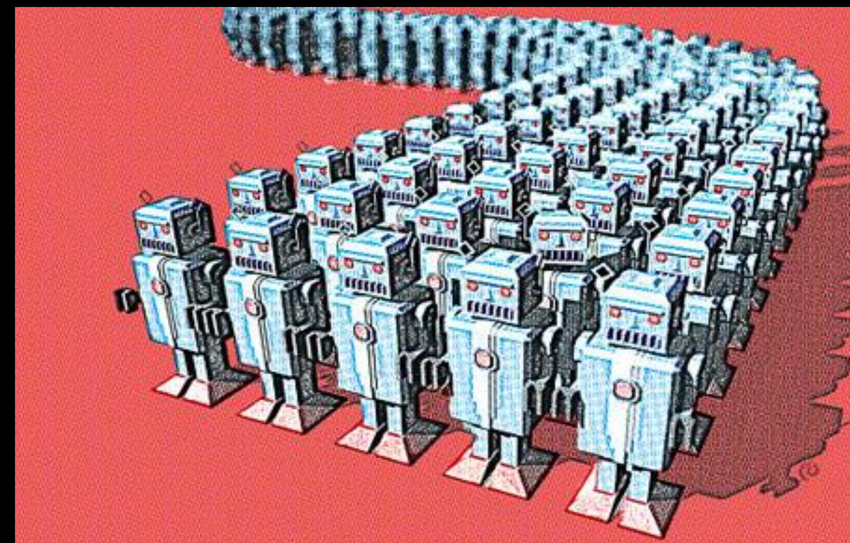
- How many of these very qualified professionals will we need?
- How many know/ will learn statistics, data analysis, data science?

Next up: Big Data Technologies

- How many of these very qualified professionals will we need?
- How many know/ will learn statistics, data analysis, data science?

We need an Army! Of ROBOTS!





Enter Machine Learning

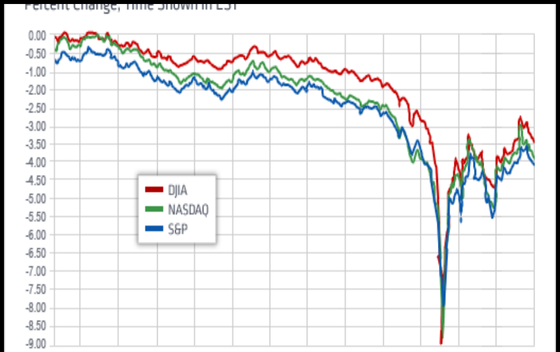
- "Machine learning systems automatically learn programs from data" (*)
- You don't really code the program, but it is inferred from data.
- Intuition of trying to mimic the way the brain learns: that's where terms like *artificial intelligence* come from.

(*) CACM 55(10) - A Few Useful Things to Know about Machine Learning (Domingos 2012)

Applications of Machine Learning

- Sales

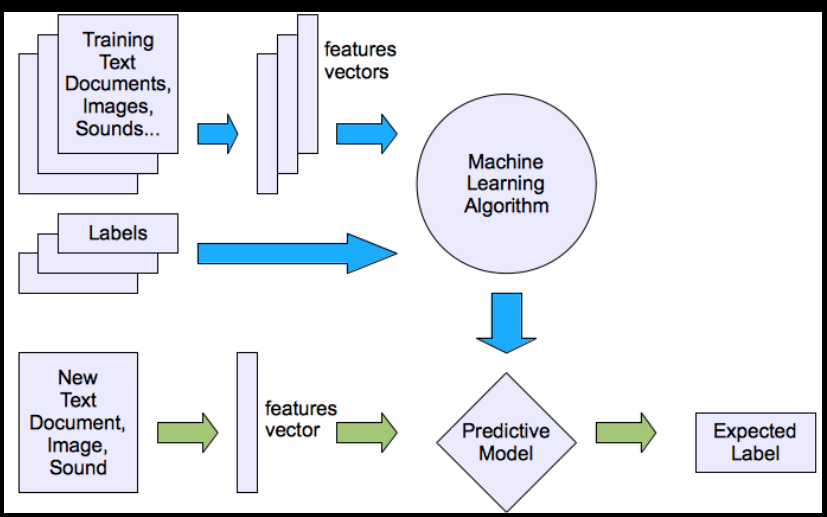
- Trading



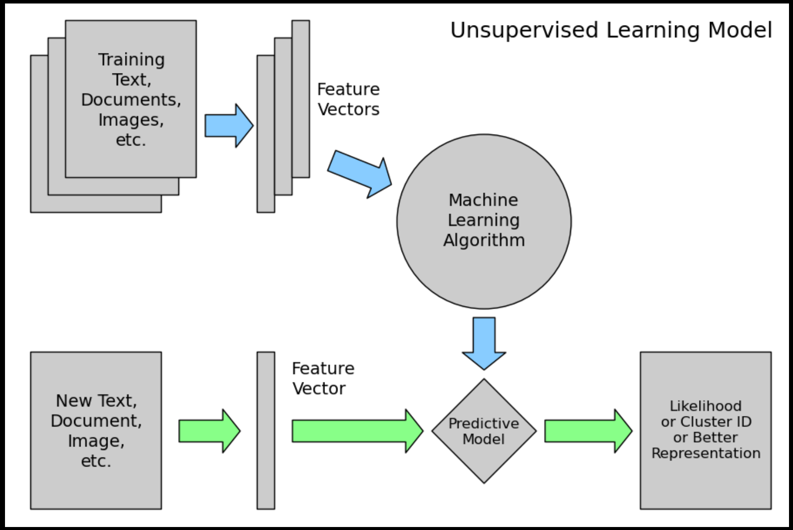
- Image and Voice Recognition

Kinds of Machine Learning

- Supervised Learning:
  - Classification (NN, SVM, Naïve Bayes)
  - Regression (linear, logistic)



- Unsupervised Learning :
  - Clustering (k-means)
  - Decomposition (PCA, SVD)



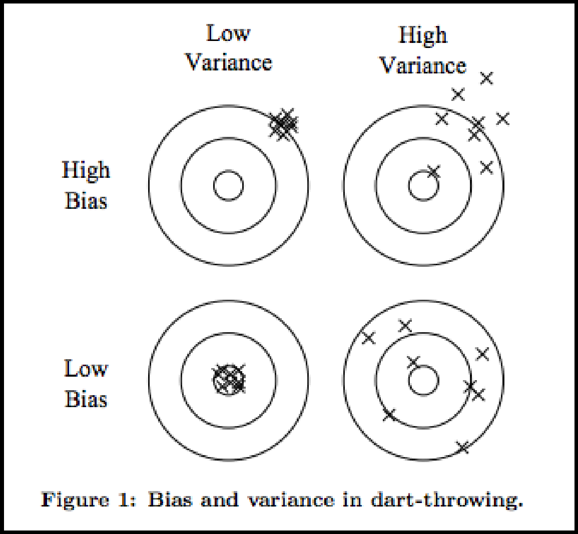
Source – scikit-learn.github.io/scikit-learn-tutorial/general_concepts.html

Remember SPAM filters?

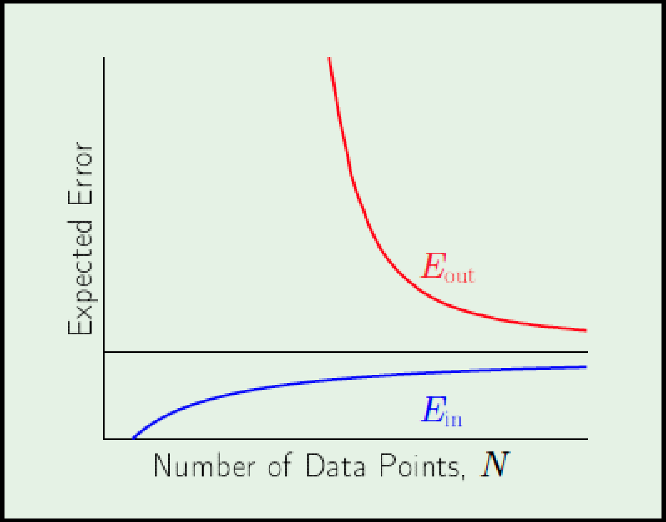
- The original use case for ML in Information Security
- Remember the “Bayesian filters”? There you go.
- How many talks have you been hearing about SPAM filtering lately? ;)

So what is the fuss?

- Models will get better with more data
  - We always have to consider bias and variance as we select our data points
- “I’ve got 99 problems, but data ain’t one”



Domingos, 2012



Abu-Mostafa, Caltech, 2012

Designing a model to detect external agents with malicious behavior

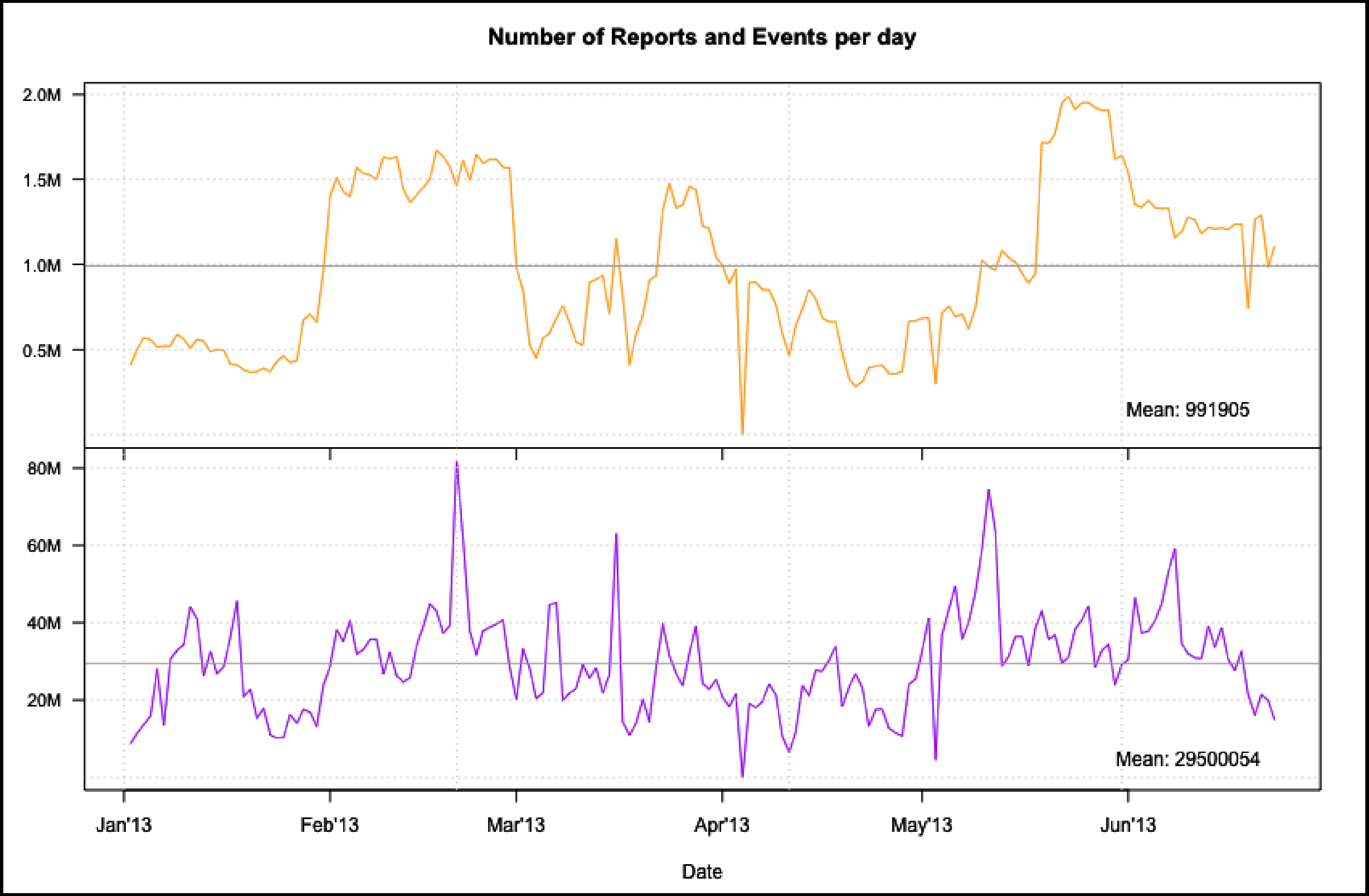
- We've got all that log data anyway, let's dig into it
- Most important thing is the "feature engineering"

Model: Data Collection

- Firewall block data from SANS DShield (per day)
- Firewalls, really? Yes, but could be anything.
- We get summarized “malicious” data per port

```
> sans
             date              ip targetPort protocol reports targets firstSeen lastSeen
      1: 20130622   89.248.171.125         80      TCP   64853   64775  00:14:14 17:51:54
      2: 20130622    93.174.93.179         80      TCP   59580   58487  05:11:15 22:21:41
      3: 20130622    213.186.60.63         80      TCP   58429   58429  00:15:41 21:42:28
      4: 20130622  202.121.166.203         22      TCP  106621   53328  05:18:26 10:10:33
      5: 20130622  218.207.176.125         80      TCP   53241   53241  21:16:09 21:56:07
     ---
1107159: 20130622     65.55.37.104      16766      TCP       2       1  12:31:06 12:31:12
1107160: 20130622     65.55.37.104      16765      TCP       1       1  00:45:24 00:45:24
1107161: 20130622     65.55.37.104      16761      TCP       3       1  09:47:49 09:48:39
1107162: 20130622     65.55.37.104      16759      TCP       2       1  03:29:51 03:30:37
1107163: 20130622     65.55.37.104      16721      TCP       1       1  20:29:24 20:29:24
```

Not quite "Big Data", but enough to play around



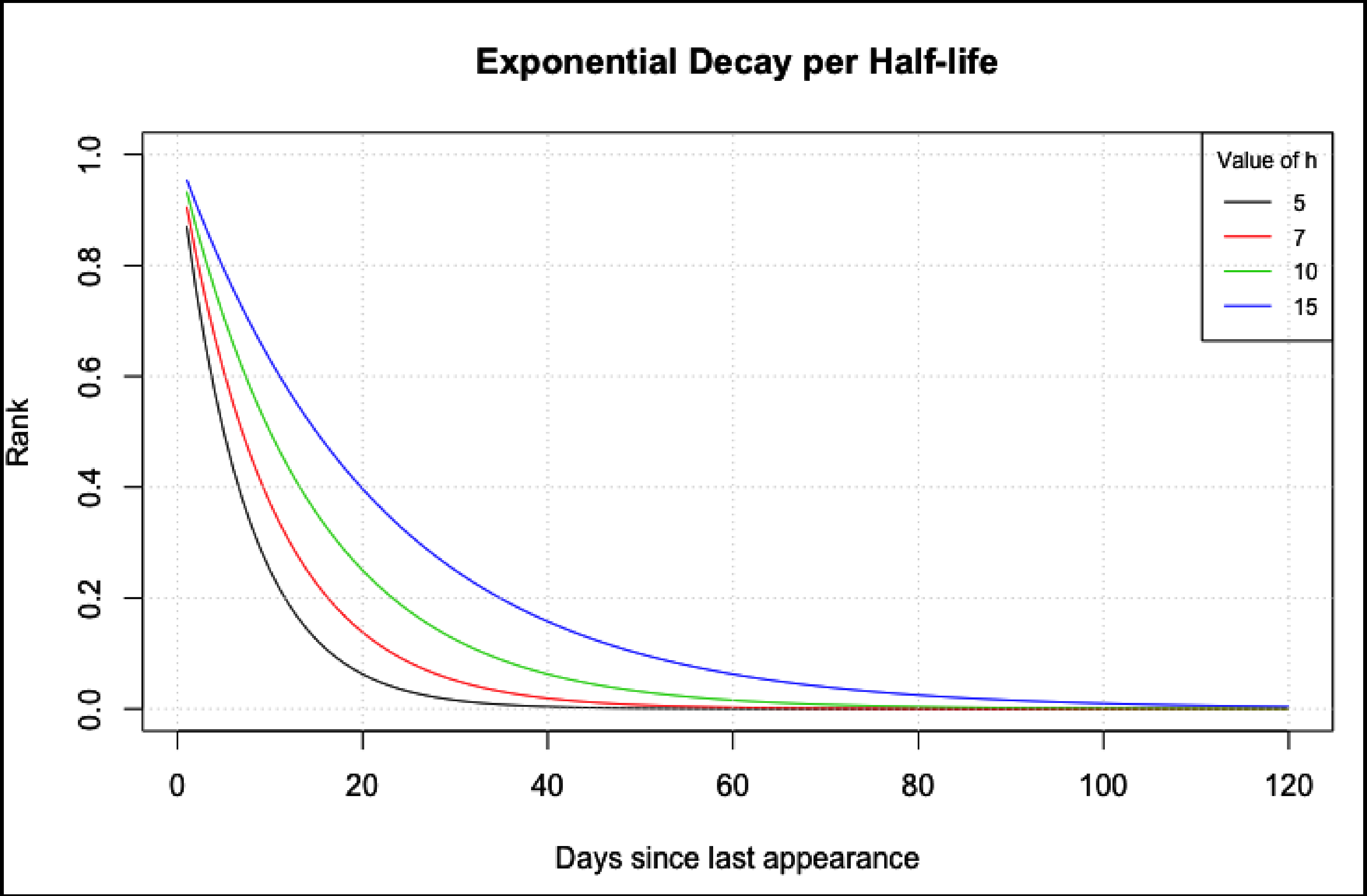
Model Intuition: Proximity

- Assumptions to aggregate the data
- Correlation / proximity / similarity BY BEHAVIOUR
- “Bad Neighborhoods” concept:
  - Spamhaus x CyberBunker
  - Google Report (June 2013)
  - Moura 2013
- Group by Netblock
- Group by ASN (thanks, TC)

Model Intuition: Temporal Decay

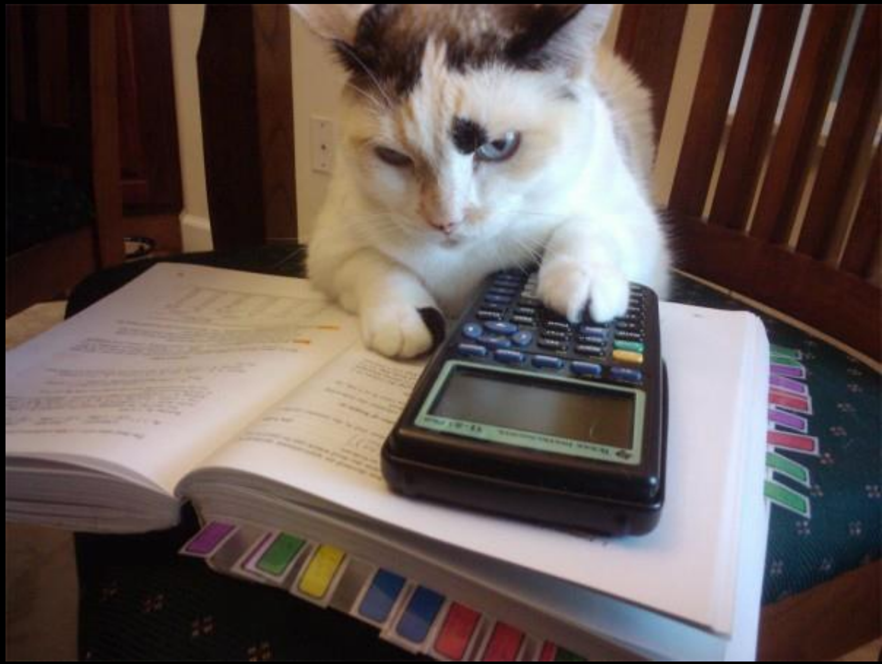
- Even bad neighborhoods renovate:
  - Agents may change ISP, Botnets may be shut down
  - Paranoia can be ok, but not EVERYONE is out to get you
- As days pass, let’s forget, bit by bit, who attacked
- A Half-Life decay function will do just fine

Model Intuition: Temporal Decay



Model: Calculate Features

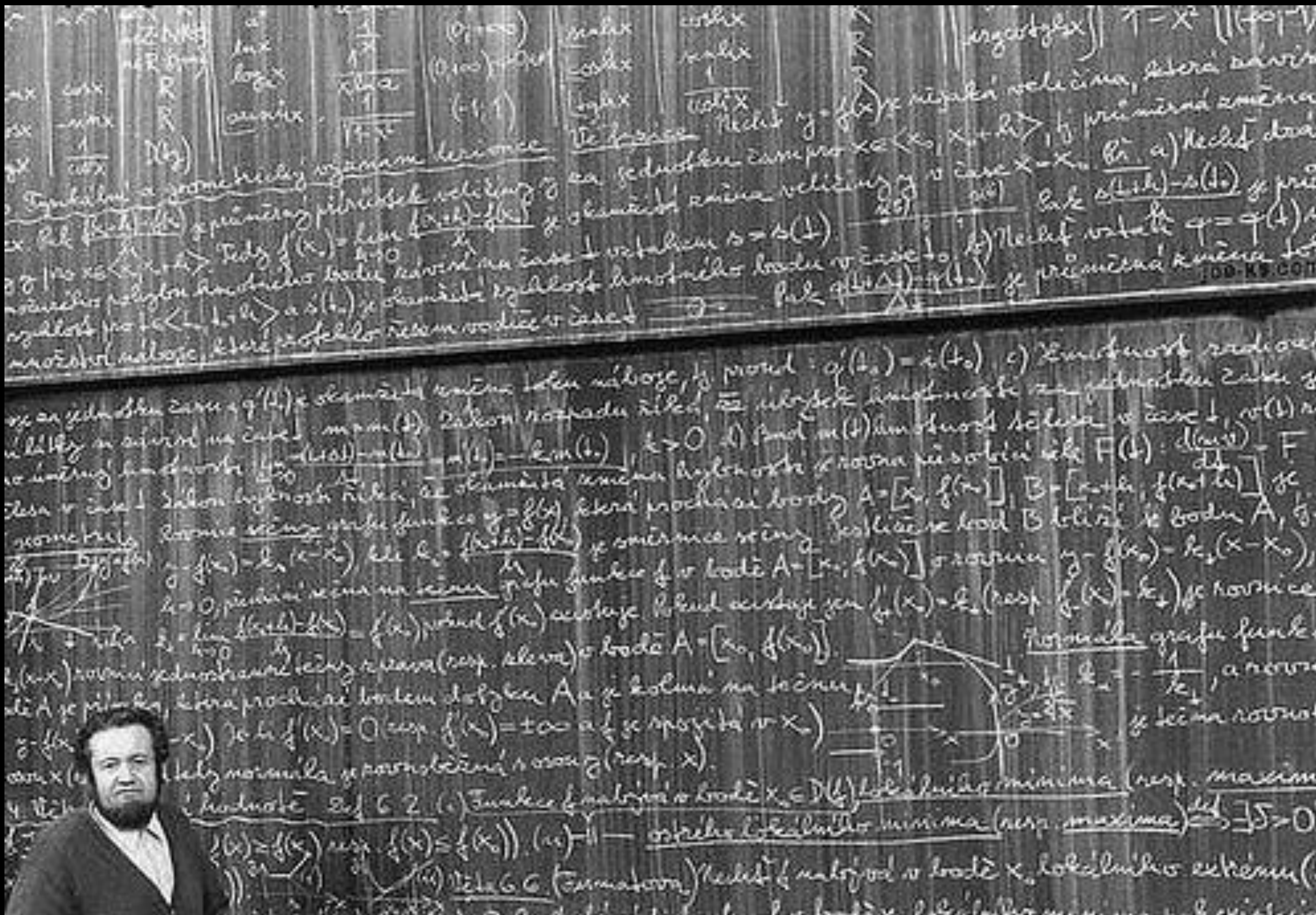
- Cluster your data: what behavior are you trying to predict?
- Create "Badness" Rank = lwRank (just because)
- Calculate normalized ranks by IP, Netblock (16, 24) and ASN
- Missing ASNs and Bogons (we still have those) handled separately, get higher ranks.



Model: Calculate Features

- We will have a rank calculation per day
  - Each “day-rank” will accumulate all the knowledge we gathered on that IP, Netblock and ASN to that day
- We NEED different days for the training data
- Each entry will have its date:
  - Use that “day-rank”
  - NO cheating
  - Survivorship bias issues!

How are we doing so far?



Training the Model

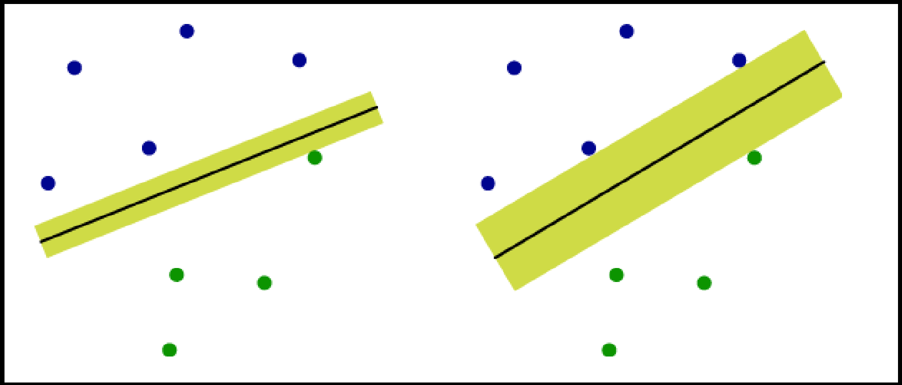
- YAY! We have a bunch of numbers per IP address!
  - How can I use this?
- We get the latest blocked log files (SANS or not):
  - We have "badness" data on IP Addresses - <u>features</u>
  - If they are blocked, they are "malicious" - <u>label</u>
- Sounds familiar?
- Now, for each behavior to predict:
  - Create a dataset with "enough" observations:
  - ROT of 50k - 60k because of empirical dimensionality.

Negative and Positive Observations

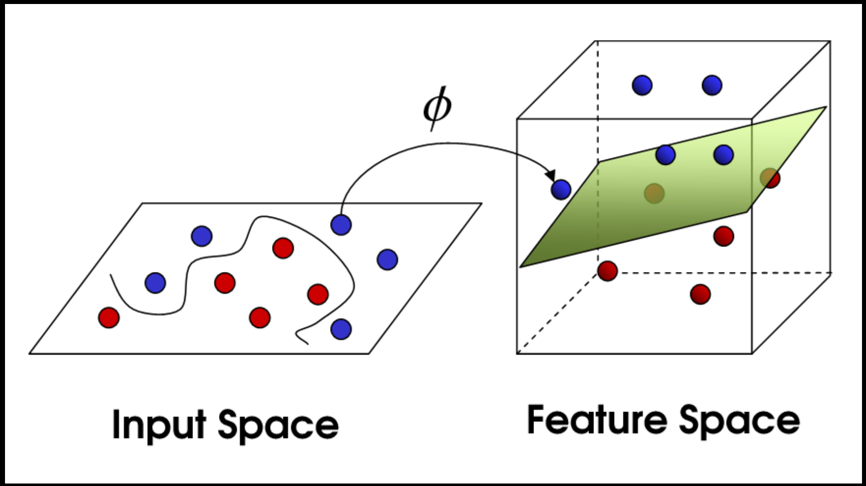
- We also require “non-malicious” IPs!
- If we just feed the algorithms with one label, they will get lazy.
- CHEAP TRICK: Everything is “malicious”
- Gather “non-malicious” IP addresses from Alexa and Chromium Top 1m Sites.

SVM FTW!

- Use your favorite algorithm! YMMV.
- I chose Support Vector Machines (SVM):
  - Good for classification problems with numeric features
  - Not a lot of features, so it helps control overfitting, built in regularization in the model, usually robust
  - Also <u>awesome</u>: hyperplane separation on an unknown infinite dimension.



Jesse Johnson – shapeofdata.wordpress.com



No idea... Everyone copies this one

Results: Training Data

- Cross-Validation: method to test the data against itself

- On the training data itself, 85 to 95% accuracy
- Accuracy = (things we got right) / (everything we had)
- Some behaviors are much more predictable than others:
  - Port 3389 is close to the 95%
  - Port 22 is close to the 85%
  - SANS has much more data on port 3389. Hmmm……

Results: New Data

- And what about new data?
- With new data we know the labels, we find:
  - 80 – 85% true positive rate (sensitivity)
  - 85 – 90% true negative rate (specificity)

$$LR+ = \frac{\Pr(T+|D+)}{\Pr(T+|D-)}$$

- This means that:
  - If the model says something is "bad", it is <u>5.3 to 8.5 times MORE LIKELY to be bad</u>.
- Think about this. Our statistical intuition is bad.
- Wouldn't you rather have your analysts look at these?

Results: Really New Data

Final Remarks

- These and other algorithms are being developed in a personal project of mine: MLSec Project
- Sign up, send logs, receive reports generated by models!
  - FREE! I need the data! Please help! ;)
- Looking for contributors, ideas, skeptics to support project as well.
- Please visit **http://mlsecproject.org** or just e-mail me.

Thanks!

- Q&A?
- Don't forget your feedback forms!

Alexandre Pinto
alexcp@mlsecproject.org
@alexcpsec
@MLSecProject