Implications of Adversarial Environment on Machine Learning

Michal Nánási (michal.nanasi@gmail.com)
https://www.(facebook.com|twitter.com|linkedin.com/in)/mic47

Disclaimer

The opinions expressed in this presentation are my own and not necessarily those of my [former] employer. This talk is not on behalf of my [former] employer or anyone else, but me.

This is not talk about...

- ... adversarial attacks on neural networks (and other classifiers).
- ... how to solve spam problem.
- ... how to build anti-spam systems.

What it is about?

- What will change in your approach to ML, or what problems you might face, if:
  - There is persistent, well motivated adversary, trying to circumvent your ML classifiers.
  - You are dealing with abuse with high volume (like spam, ad fraud, …).

What is spam?

- Unwanted messaging.
  - With very loose definition of what is messaging.
- Almost exclusively financially motivated.
- Most of the spam is somehow automated.
- It’s basically shady advertising.

What is spam?

- Unwanted messaging.
  - With very loose definition of what is messaging.
- Almost exclusively financially motivated.
- Most of the spam is somehow automated.
- It’s basically shady advertising.

How hard is it to spam?

- Affiliate programs do everything for you [1]:
  - Give you shop.
  - Handle payments.
  - Drug manufacture.
  - Ship drug to customer.
- Spammer can focus on innovation in spamming.

[1] http://bit.ly/SpamMLEco

You can also buy accounts.

Provider	Quantity	
Twitter.com EN PVA	936	1K-10K: **$90**
Twitter.com Aged	20831	1K-10K: **$80**
Twitter.com Profiled	11443	1K-10K: **$70**
Twitter.com+Avatar	10090	1K-10K: **$60**
Twitter.com	18910	1K-10K: **$50**
Twitter.com Promo	19369	1K-10K: **$25**

Provider	Quantity	
Mail.ru	112812	1K-10K: **$6**
Mail.ru Mix	291755	1K-10K: **$6**
Mail.ru Human	71853	1K-10K: **$8**
Mail.ru No SPAM	28459	1K-10K: **$8**
Mail.ru EN	35820	1K-10K: **$8**
Mail.ru UA	59550	1K-10K: **$7**
Mail.ru Second Hand	34844	1K-10K: **$3**
Mail.ru Mix Second Hand	38840	1K-10K: **$3**

Provider	Quantity	P
Facebook.com EN PVA US	3133	1K-10K: **$240**
Facebook.com EN PVA	3228	1K-10K: **$150**
Facebook.com Aged	1220	1K-10K: **$150**
Facebook.com+Avatar	2197	1K-10K: **$100**
Facebook.com EN	9088	1K-10K: **$120**
Facebook.com RU	2910	1K-10K: **$100**
Facebook.com RU Basic	21977	1K-10K: **$60** |
Facebook.com Basic	21405	1K-10K: **$50** | 1
Facebook.com Promo	138989	1K-10K: **$20** | 1
Facebook		

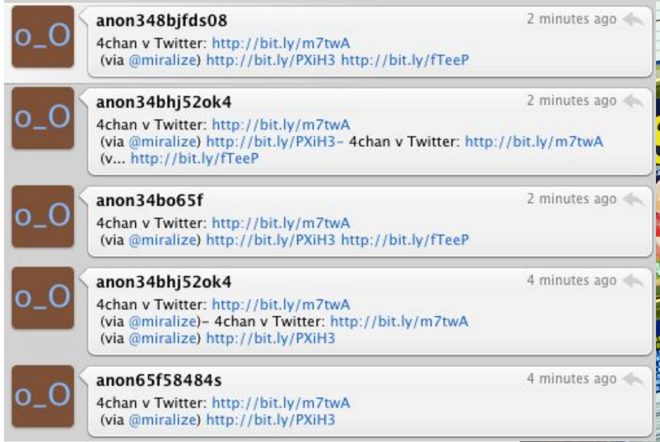
Provider	Quantity	
Instagram.com EN PVA	191	1K-10K: **$180** |
Instagram.com FB Photo	0	1K-10K: **$180** |
Instagram.com RU	181	1K-10K: **$50** |
Instagram.com Basic	292	1K-10K: **$50** |
Instagram.com RU MF/ML	0	1K-10K: **$25** |

Provider	Quantity	P
Gmail.com PVA SMTP	690	1K-10K: **$380** |
Gmail.com RU PVA LS	2368	1K-10K: **$350** |
Gmail.com USA PVA LS	665	1K-10K: **$300** |
Gmail.com USA PVA	974	1K-10K: **$280** |
Gmail.com RU PVA	0	1K-10K: **$260** |
Gmail.com PVA Promo	0	1K-10K: **$210** |

Dobrý deň,

radi by sme Vás touto cestou požiadali o súhlas so zobrazením obchodného oznámenia, ktoré sa týka výpredaja notebookov a počítačov.

Ak súhlasíte, pre **ZOBRAZENIE VÝPREDAJA NOTEBOOKOV A POČÍTAČOV PROSÍM KLIKNITE TU >>**



Sources: http://bit.ly/2psGpeB, http://bit.ly/2DHrZfU, http://bit.ly/2FYTt2g, http://bit.ly/2DIEwzv

Let’s focus on spam detection
(using “machine learning”)

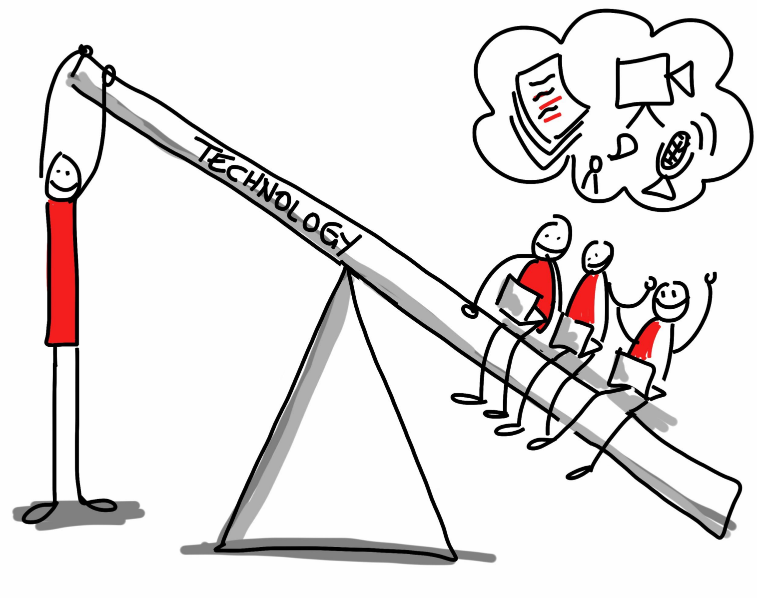
Important parts of ML pipeline

- Feature extraction.
- Training label extraction.
- Model [not covered in this talk].
- Monitoring & Deployment.

Important parts of ML pipeline

- **Feature extraction.**
- Training label extraction.
- Model [not covered in this talk].
- Monitoring & Deployment.

It’s all about leverage

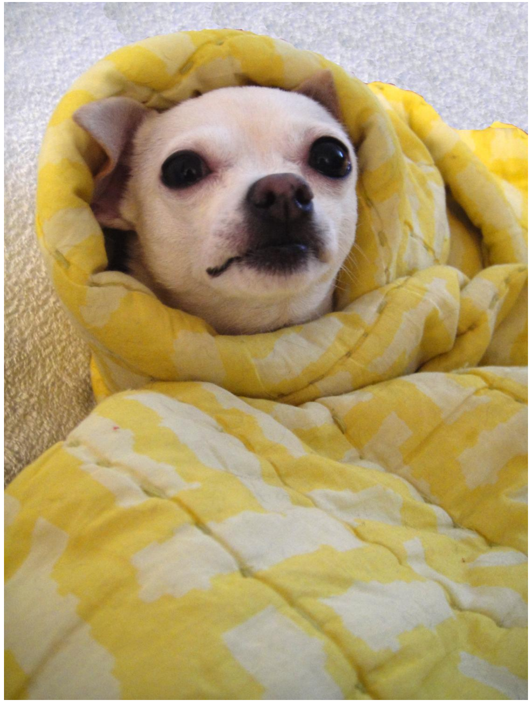


Source: https://www.flickr.com/photos/gforsythe/10310176123, https://bit.ly/by-nc-sa

Adversarial cycle (spam simulation)

Example Spam

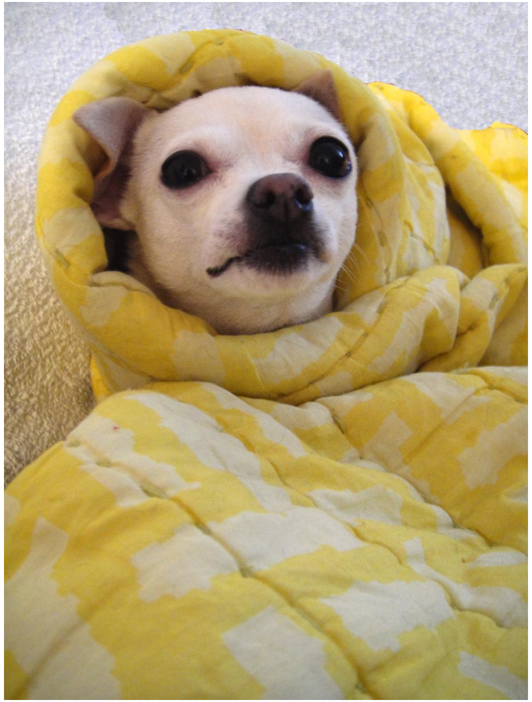
Checkout this cute dog burrito! Just visit https://spammy.dogs/2390udasflkj



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Example Spam

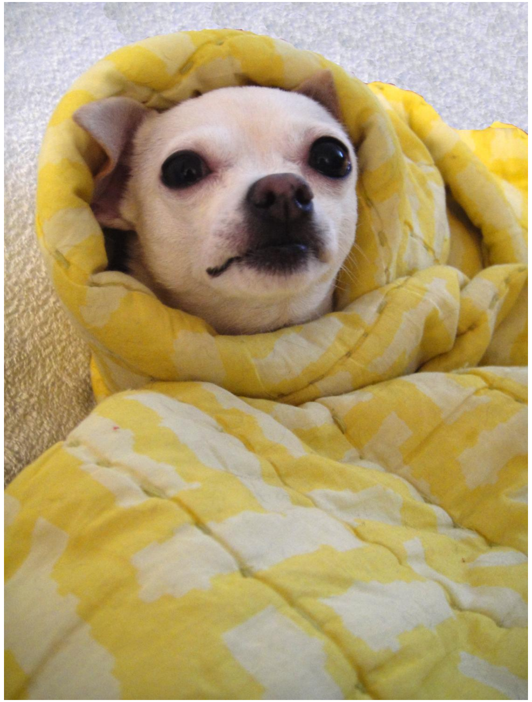
Checkout this cute dog burrito! Simply go to https://spammy.dogs/90u23nklsd



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Example Spam

Checkout this cute dog burrito! Don't forget to subscribe on https://spammy.dogs/2389jhds093



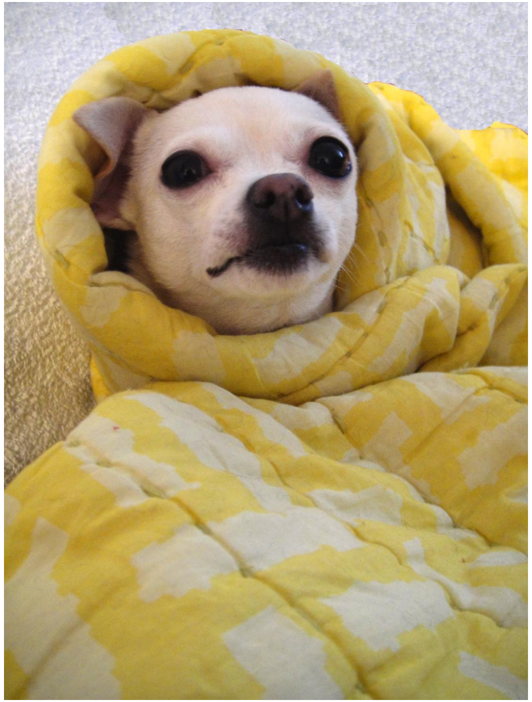
Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Our first “classifier”

```
dogSpam = do
  message <- getMessage
  if
    message `contains` "Checkout this cute dog burrito!"
  then return BlockMessage
  else return DontBlock
```

New Spam

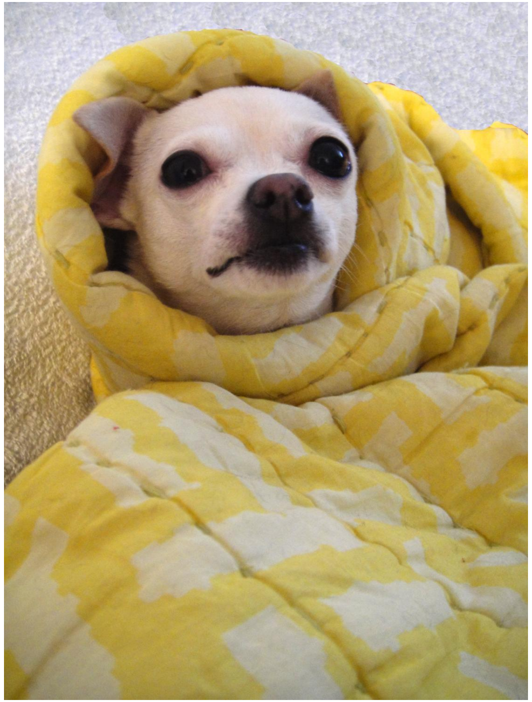
Do you like this dog burrito? Just visit
https://spammy.dogs/2390udasflkj



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

New Spam

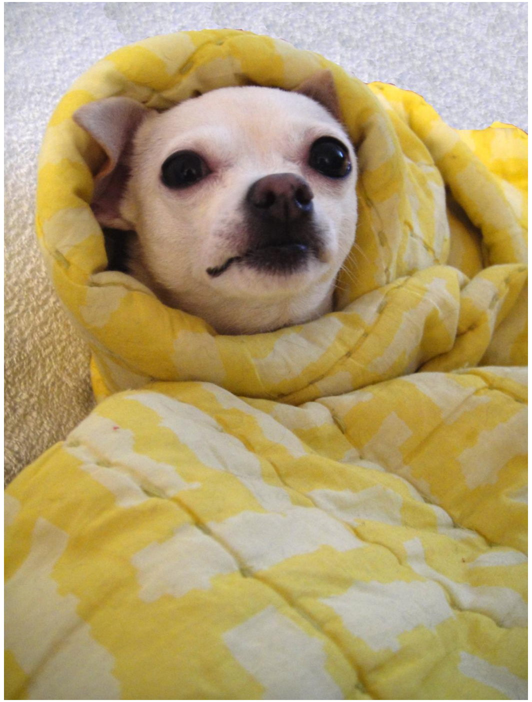
You would never guess what this dog burrito barks!
Simply go to https://spammy.dogs/90u23nklsd



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

New Spam

It is dog? It is burrito? It’s dog burrito! Don’t forget to subscribe on https://spammy.dogs/2389jhds093



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Retrained “classifier”

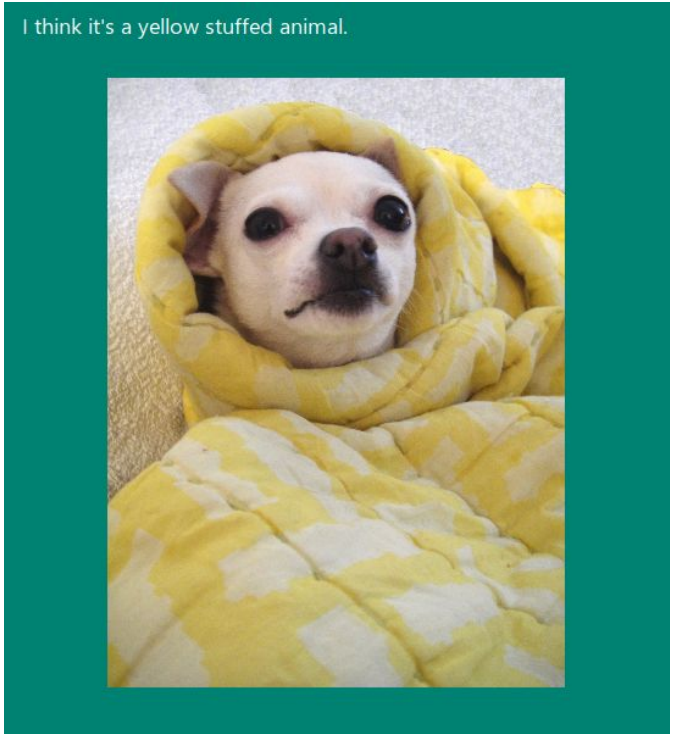
```
dogSpam = do
  message <- getMessage
  photo <- getPhoto
  photoDescription <- describePhoto photo
  if
    message `contains` "burrito"
    && photoDescription `contains` "animal"
  then
    return BlockMessage
  else
    return DontBlock
```

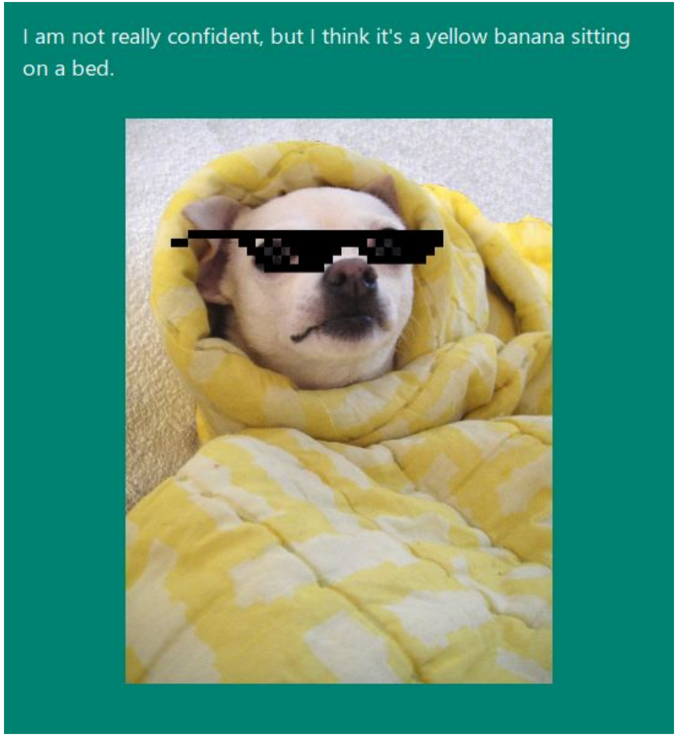
Even better spam

Don’t like this dog? Deal with it at https://spammy.dogs/2390udasflkj

Sources: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc, https://thenounproject.com/term/deal-with-it/150247/

Overlay really works

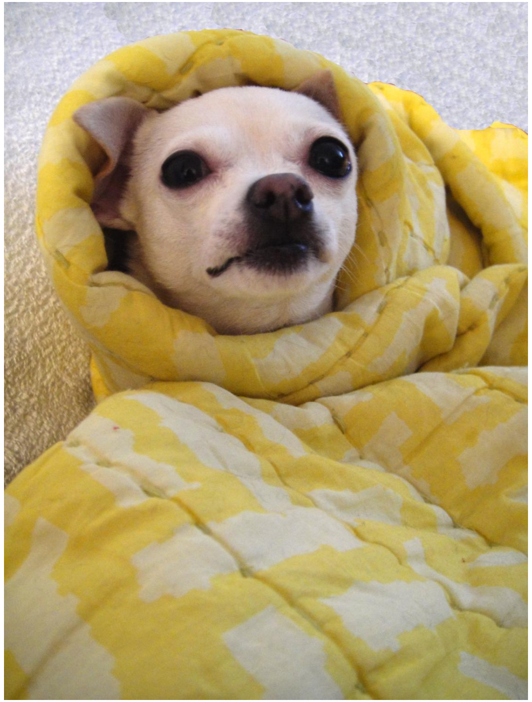




Sources: https://www.captionbot.ai/, https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Interesting False positive

Don’t click on “dog burrito” posts with this image. They are all phishing scam!



Source: https://www.flickr.com/photos/redlipstick/5431030879, , http://bit.ly/by-nc

Content based features

- Content based features are convenient.
- They are easy to avoid: simply reformulate, add overlay.
- Exact creative is often not that important for spam.
- Without full (and fast enough) automation, spammer has advantage.
- Can cause bad false positives.

Additional problems with content based features.

- Language dependent, you need to know language to debug.
- You can spam without content (with notifications, following, connection requests, ad clicks).
- Does not work with end-to-end encryption [1].
- “Any blacklist you create will contain someone’s name.”

[1] http://bit.ly/SpamMLWhatsApp

What to do instead?

- Content is easy to change.
- Rather use something inherent to spam, like behavior: spammer have to spam a lot.
- Behavior is harder to change, than the message.
- Aggregate events, instead classifying single event [1].
- Limit the amount of actions (or damage) spammer can do with his resources.

[1] http://bit.ly/SpamMLLinkedIn

What are “expensive” resources for spammer?

- IP address [1] [2].
- Account (needs to create / compromise / buy).
- URL, domain [1] [3].
- Phone number.
- Email address [1].

[1] Can be easy to obtain in some cases. [2] Also, very messy and hard to block properly. [3] It's sort of content feature too.

What if spammer ...

- ... use botnets (lot of IPs)?
- ... buy bulk of cheap sim cards?
- ... buy discounted domains? [1]
- ... use URL shortener?
- ... use dropbox / google drive?

[1] http://bit.ly/SpamMLPred, http://bit.ly/SpamMLPredYT

One more trick

- Exploit information asymmetry between spammer and us.
- We know distribution of names, ages, genders, browsers, operating systems, countries, …
- Spammers don’t know those distributions.

Look at the surprises

- “Why is this weird domain shared only in Canadian groups, using IPs from Brazil and users with phone numbers from Slovakia?”
- Each of this feature is weak alone, but strong in aggregate.
- You can observe distributions from the data.

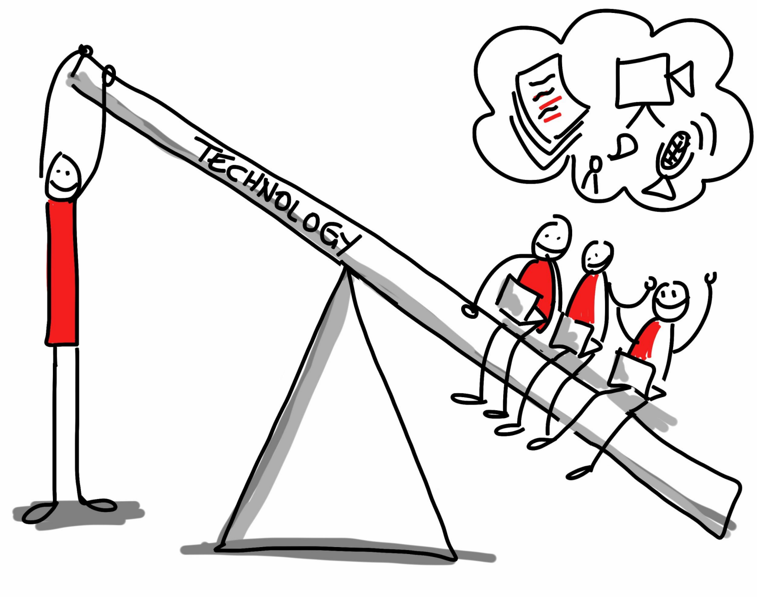
Be aware of data poisoning

- "We know distribution of names, ages, genders, browser, operating systems, countries, …"
- This is true, unless large part of your data have spam and you don't know how to remove it.

Don’t allow spammer to affect user’s features

- What about this feature: Probability of browser, given country.
- P(browser | country) = count(browser, country) / country(country)
- Vatican is smallest country, had 792 citizens in 2017. Assume 200 users, each uses “Good browser™”.
- Spammer have 800 bad users. Each uses “Bad browser™”.
- P(Good browser™ | Vatican) = 0.2, P(Bad browser™ | Vatican) = 0.8
- If spammer stops, P(Good browser™ | Vatican) = 1.0, P(Bad browser™ | Vatican) = 0.0
- Spammer’s action affected feature values of non-spamming users.

Conclusion: think about leverage



Source: https://www.flickr.com/photos/gforsythe/10310176123, https://bit.ly/by-nc-sa

Summary of low leverage features

- Phrases, images, url in the content.
- Does user agent string contain “curl” or “phantomjs”?
- This is not our client.

Summary of high leverage features

- History of actions on IP address, url, device, account (number of actions per item).
- Deviations from the expected distributions, anomalies.
- Aggregations over low leverage features.

Important parts of ML pipeline

- Feature extraction.
- **Training label extraction.**
- Model [not covered in this talk].
- Monitoring & Deployment.

Low leverage features (LLF) are not useless

- If precision is high, it LLF can be used as label.
- If recall is high, it can be used as label with combination with other features.
- You will get high quality automated, but biased labels.

“What you can’t use for classification, use for labels. [1]”

[1] https://en.wikipedia.org/wiki/Parallel_construction

Slow labels

- Some features arrive late, like
  - We eventually deleted this content.
  - Content was taken down by moderator.
  - Account was compromised.
- Use machine learning to make your systems react faster.
- Be aware of feedback loops.

User feedback

- Users are good at recognizing spam.
- There are 2 options:
  - Use reports directly as a labels, features.
  - Review reports manually, and use reviews as labels.

User feedback

- Reporting is available to spammers too.
- Reporting wars:
  - One group of users start mass reporting of content of other group, in order to get it down.

User feedback

- Use machine learning to amplify human actions.
- Never ever use user feedback directly in classifiers.
- This still won’t be completely unbiased [1].

[1] http://bit.ly/SpamMLBias

Important parts of ML pipeline

- Feature extraction.
- Training label extraction.
- Model [not covered in this talk].
- **Monitoring & Deployment.**

Monitoring / Evaluation

- It’s good idea to monitor deployed classifier (or candidate classifier).
- Does classifier still catch spam as yesterday?
- Do we have more false positives?
- What is a good metric?

Confusion matrix

	Classifier thinks it is spam	Classifier thinks it is ham
It is spam	True positive (TP)	False Negative (FN)
It is ham	False positive (FP)	True Negative (TN)

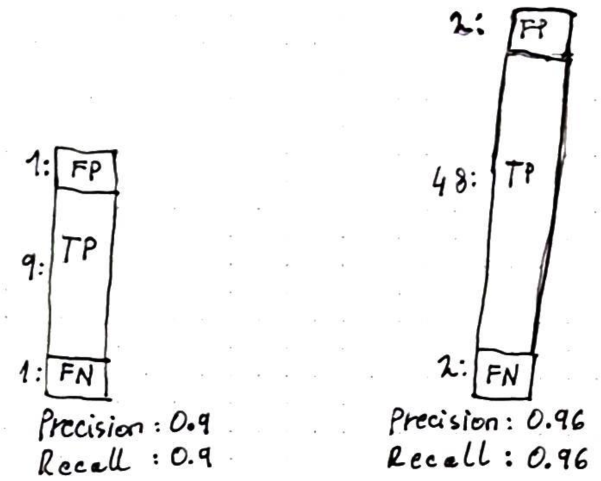
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)

Quiz

Which of these are irrelevant for monitoring of spam?

- True negatives (non-blocked ham)
- True positives (blocked spam)
- False negatives (non-blocked spam)
- False positives (blocked ham)

Problem



- Amount of spam attempts is controlled by attacker.
- Attacker can easily attempt to do more spam.
- Higher volume spam is easier to block.

Perils of Precision & Recall

- If attacker spam more, recall goes up.
- Even if the absolute amount of unblocked bad content goes up.
- If attacker spam more, precision goes up.
- Even if the absolute number of false positives goes up too.

Perils of Precision & Recall

- 0% recall with 10 spam messages is better than 99.9% recall with 1M spam messages.
- Don’t worry about bad content you blocked.

Quiz: “correct answer”

Which of these are irrelevant for monitoring of spam?

- False negatives (non-blocked spam)
- False positives (blocked ham)
- True negatives (non-blocked ham)
- True positives (blocked spam)

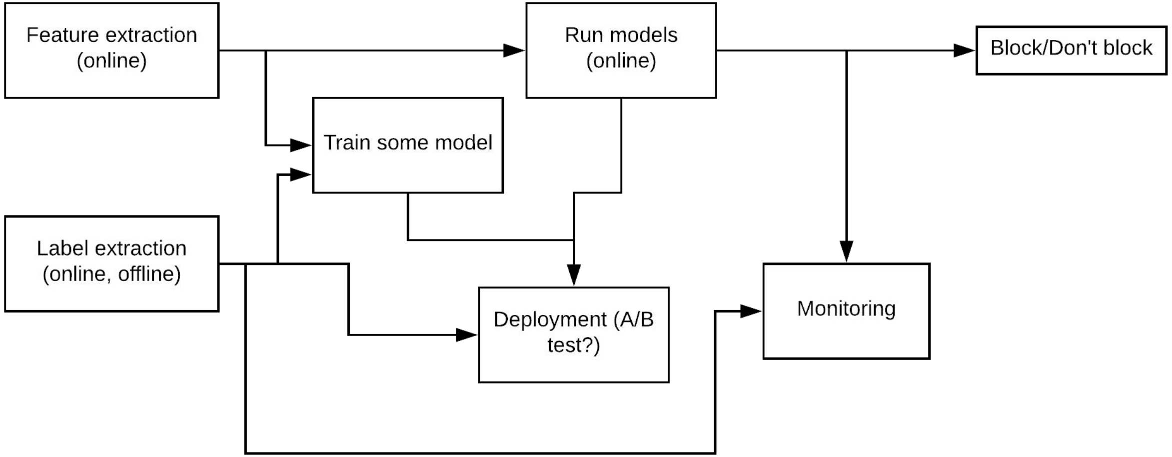
Summary

- Spammer will stop once it become unprofitable.
- Think about leverage when creating features.
- What you can’t use for classification, use as labels.
- Trust, but verify (user reports).
- Don’t bother with true positives.

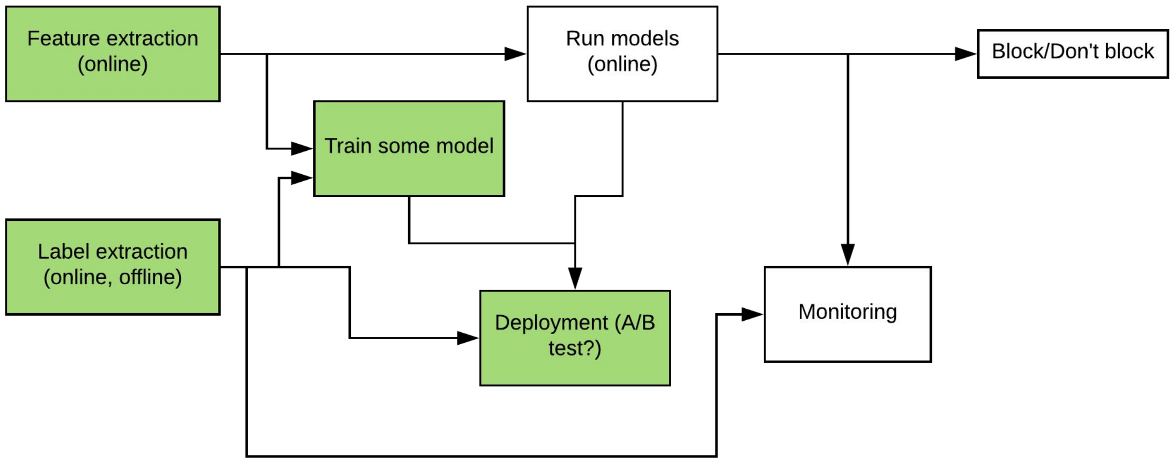
Thank you

Michal Nánási (michal.nanasi@gmail.com)
https://www.(facebook.com|twitter.com|linkedin.com/in)/mic47

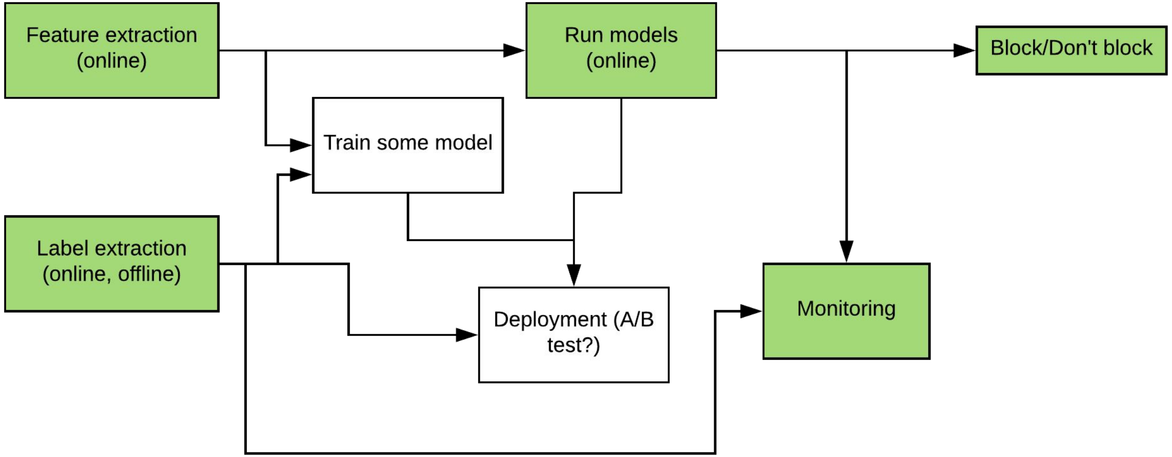
“Simple” Machine learning pipeline



“Simple” Machine learning pipeline



“Simple” Machine learning pipeline



A/B testing: Is new classifier harder to circumvent?

- A/B test: Assign subjects into 2 groups, observe difference. If you have enough subjects, result will be significant.
- You have enough actions, users, IPs, ..., related to spam.
- So do you have enough subjects?

A/B testing: Is new classifier harder to circumvent?

- There are only few (major) spammers
- Spammer won't care if 5% of spam accounts are in better classifier group and blocked.
- Spammers should be subjects in A/B test, not their pawns.
- Spammers are hard to distinguish.

A/B testing: Do we have false positives?

What should we measure? Number of reports?

- If you block spammers, reports should go down.
- If you have lot of false positives, reports go down too.

Engagement? (number of posts, shares, likes, comments).

- If you have lot of false positives, engagement is down.
- But spammers create engagement too.

A/B testing: Do we have false positives?

- Combine metrics with counter-metrics: Good classifier decreases reports, but does not touch engagement much.