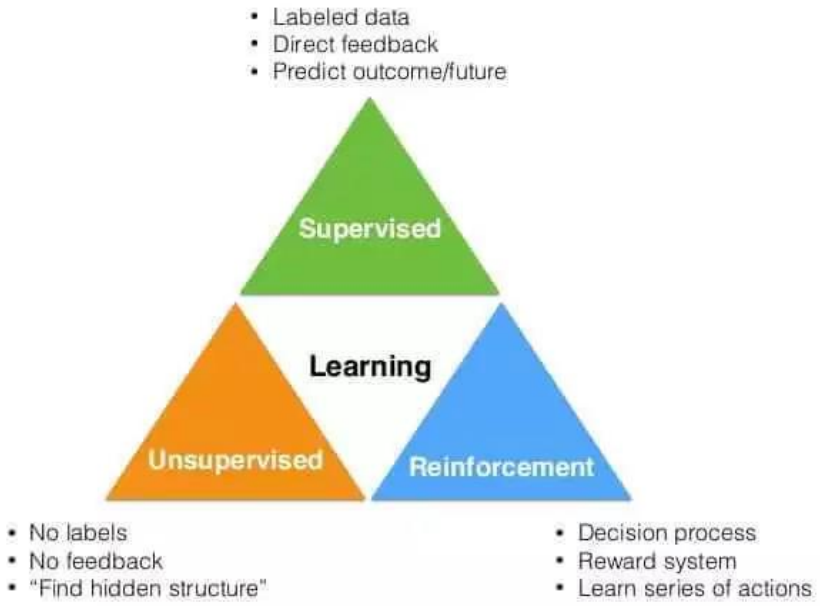
OFFSEC

عدم امنیت در سیستم های امنیتی مبتنی بر هوش مصنوعی

وحید بهزادان - www.vbehzadan.com

یادگیری ماشین

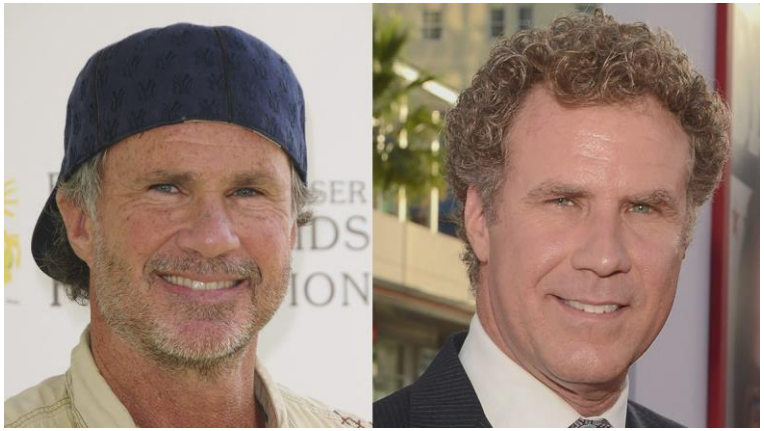
آموختن (*تخمین*) توابع با استفاده از *داده ها*



یادگیری ماشین در دنیای امروز

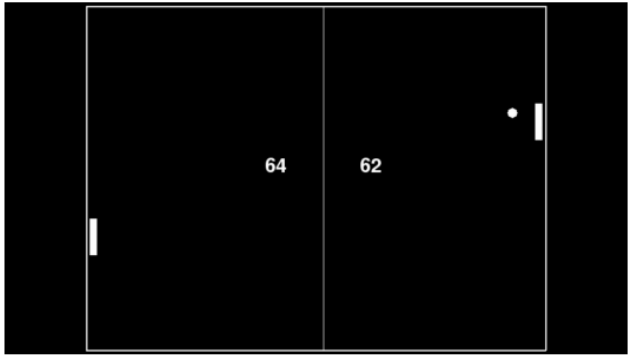
شبکه های عصبی عمیق در تشخیص اشیاء و چهره ها بهتر از انسانها عمل می کنند

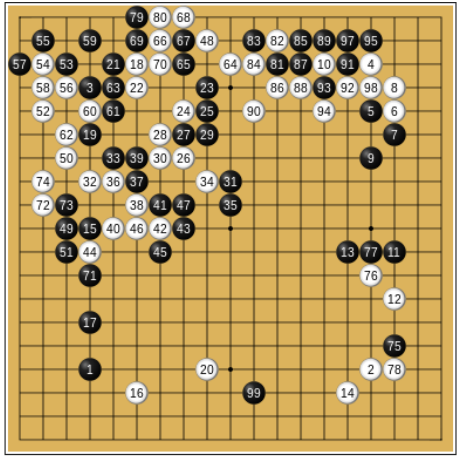
(Szegedy et al, 2014)



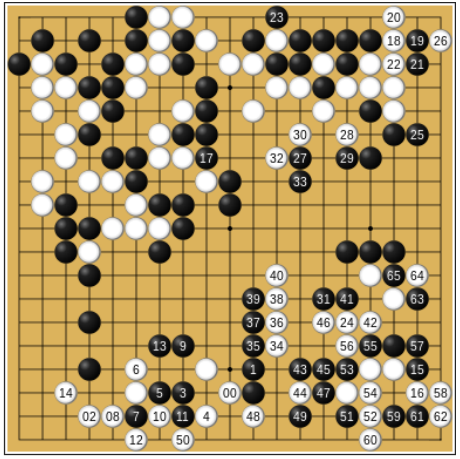
یادگیری ماشین در دنیای امروز

یادگیری تقویتی عمیق در بازی هایی همچون آتاری و گو عملکرد بهتری از انسان دارند.



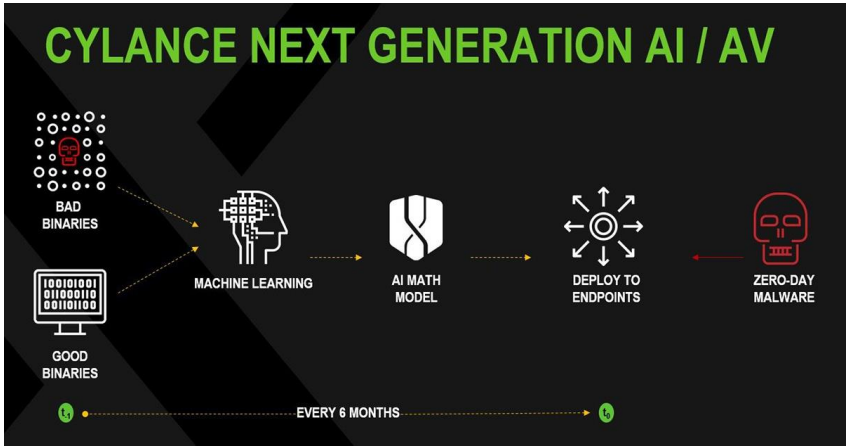


First 99 moves (96 at 10)



Moves 100-165.

یادگیری ماشین و امنیت اطلاعات



چرا یادگیری ماشین؟

- ***عدم تعادل در چرخه ی امنیت***
- افزایش تعداد آسیب پذیری ها
  - نرخ بالای حملات نوین
  - تنوع رو به رشد بدافزارها
  - افزایش پیچیدگی سیستم ها
- کاستی های آنالیز دستی
  - کشف آسیب پذیری
  - تولید و مستند سازی نشانه های حملات
  - آنالیز حملات و بدافزارها
- راه حل :اتوماسیون با استفاده از یادگیری ماشین

یادگیری نظارتی -طبقه بندی

- طبقه بندی ورودی ها در دسته های مختلف

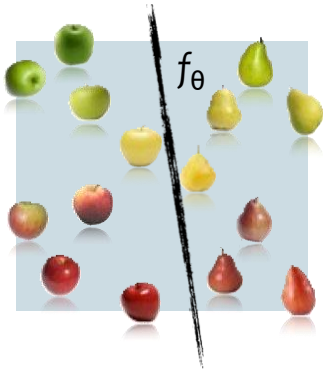
- مثال ها:

جلوگیری از هرزنامه

تشخیص نفوذ

- الگوریتم های رایج:

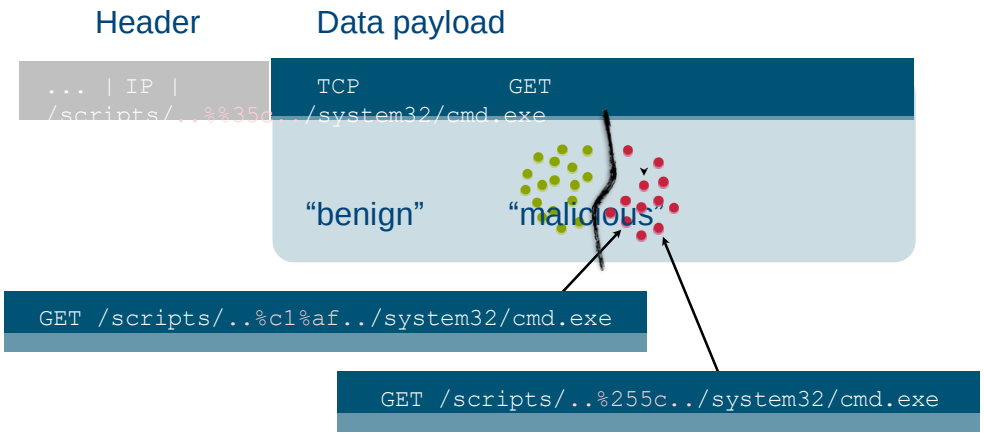
- SVM, KNN, Neural Networks, ...



طبقه بندی

- تشخیص نفوذ

تمایز بین ترافیک نرمال و مخرب



یادگیری بی نظارت -خوشه بندی

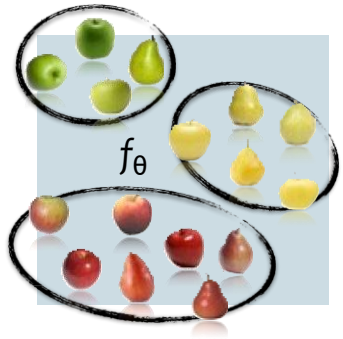
- گروه بندی اشیاء مشابه در یک خوشه

تفاوت با طبقه بندی :خوشه ها از پیش تعیین نشده اند.

- مثال ها:

گروه بندی ترافیک

آنالیز بدافزار



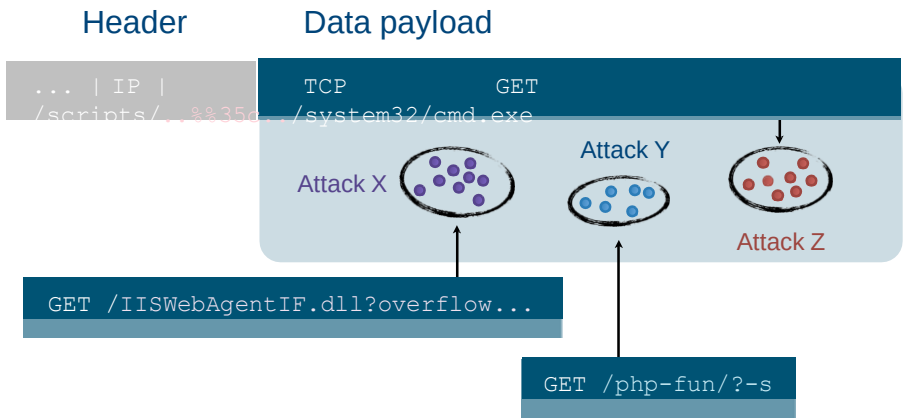
- الگوریتم های رایج:

- K-means, linkage clustering, ...

خوشه بندی

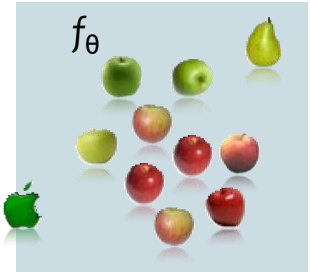
- خوشه بندی ترافیک برای تسهیل آنالیز

گروه بندی بی نظارت محتویات مشابه در یک خوشه



تشخیص ناهنجاری

- تشخیص انحراف نسبت به مدل آموخته شده از رفتار نرمال



- کاربردها

تشخیص تراکنش نا معمول

تشخیص نفوذ

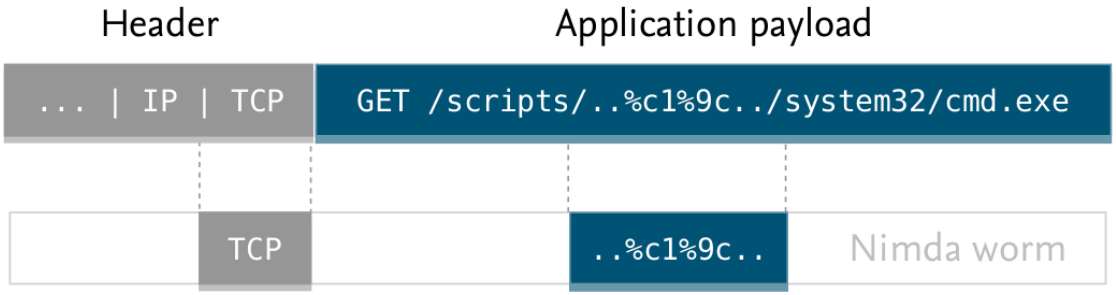
- الگوریتم های رایج:

- Mean, one-class SVM, ...

کاربردهای یادگیری ماشین در امنیت اطلاعات

تشخیص حمله

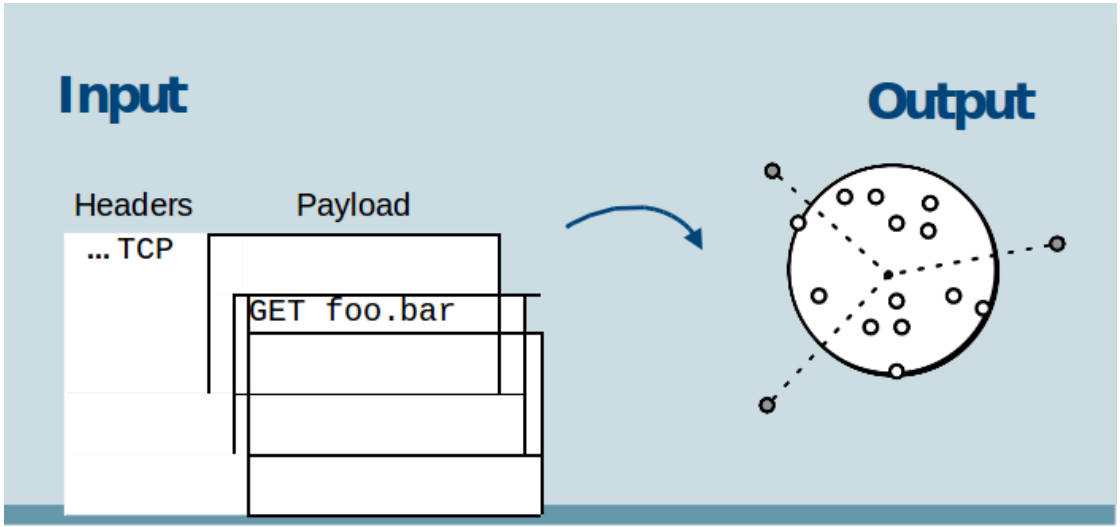
- روش سنتی :تشخیص نشانه (signature)حمله



- دفاعی نا کارآمد در برابر حملات نوین
- تاخیر در انتشار نشانه ها
- Polymorphismو Obfuscation

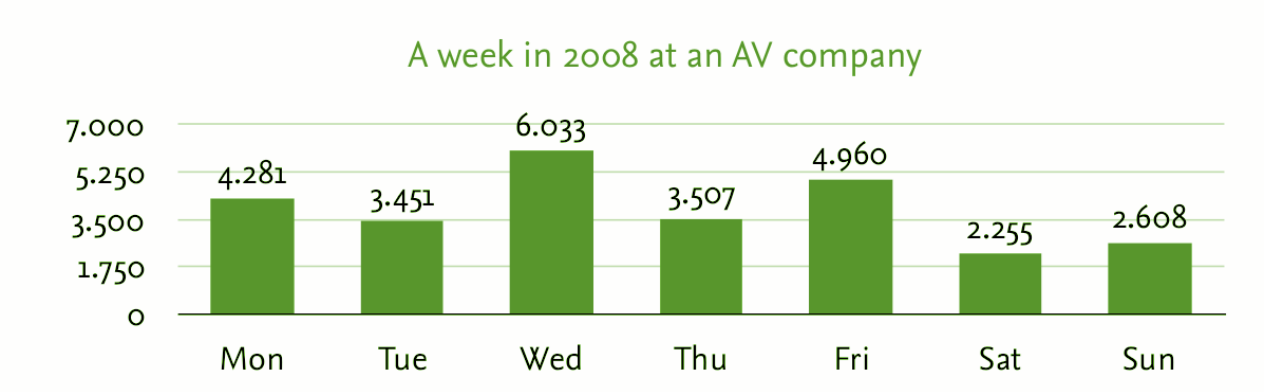
تشخیص نفوذ مبتنی بر یادگیری

- بر اساس تشخیص ناهنجاری



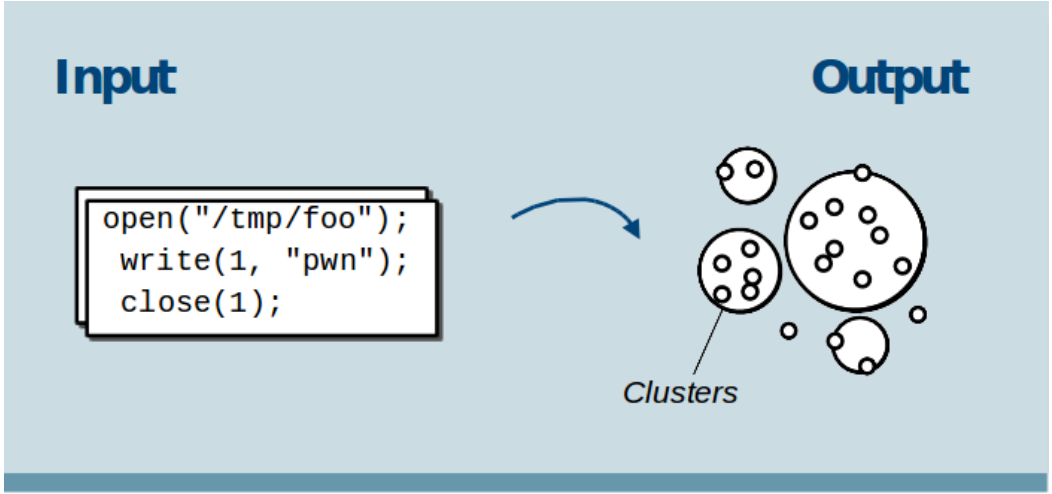
تحلیل بدافزار

- کاستی های تحلیل دستی
  - هزاران بدافزار جدید در هر روز
  - گوناگونی گسترده در رفتار (trojan, bot, ransomware, ...)
  - تحلیل و مستند سازی دستی نشانه ها تقریبا غیر ممکن است...



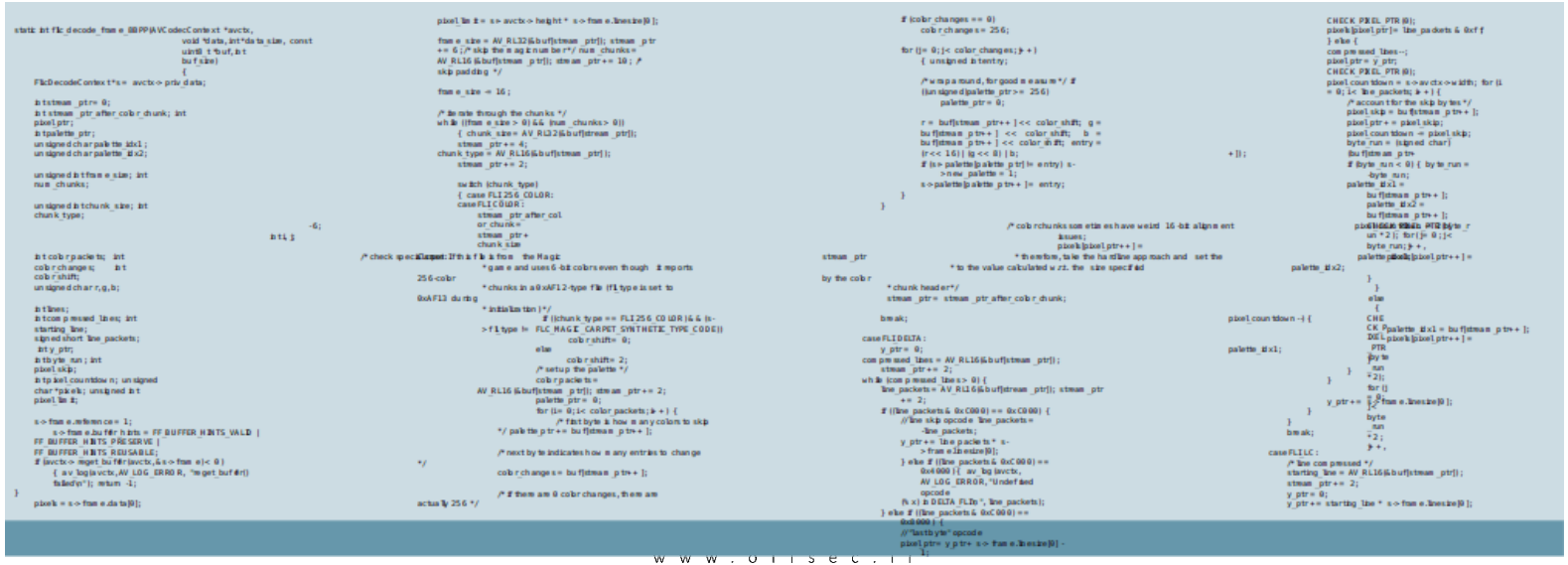
تحلیل بدافزار مبتنی بر یادگیری

- تحلیل با استفاده از خوشه بندی رفتاری



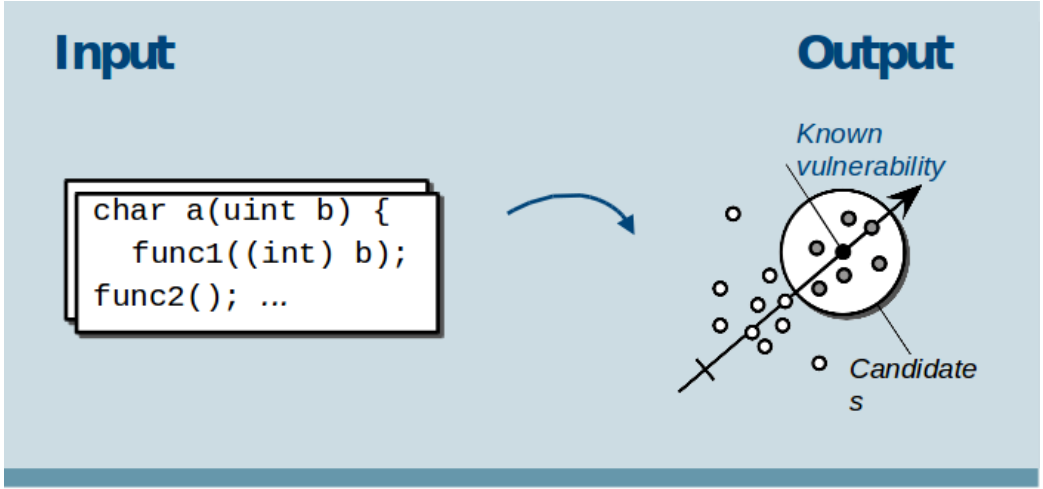
کشف آسیب پذیری

- تشخیص دستی آسیب پذیری بسیار دشوار است، زیرا:
  - نیاز به زمان و تخصص
  - طاقت فرسا



کشف آسیب پذیری مبتنی بر یادگیری

- اکتشاف با استفاده از کاهش بعد

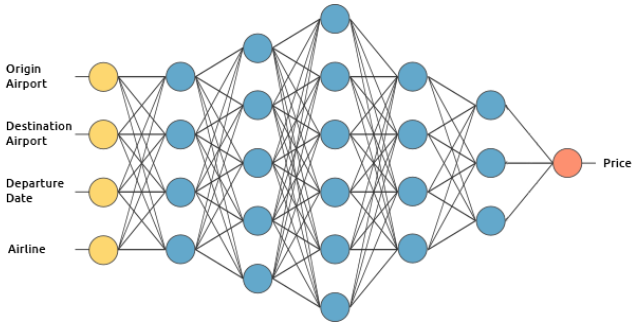


چالش های بنیادی

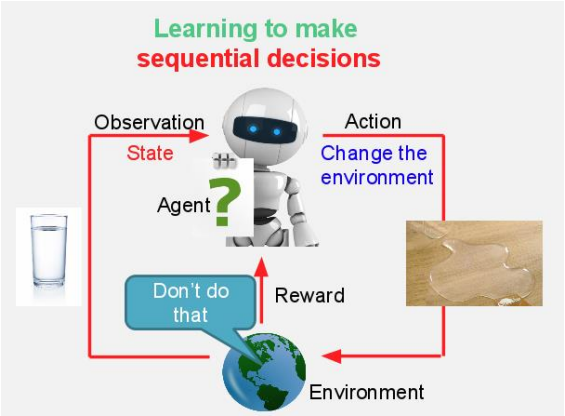
- تفاوت حوزه ی امنیت اطلاعات با مسائل رایج یادگیری ماشین:
  - شکاف معنایی
  - محدودیت های عملیاتی :هزینه ی تشخیص اشتباه
  - نیاز به شفافیت :مدل آموخته شده چگونه کار می کند؟

استفاده های نوین در صنعت

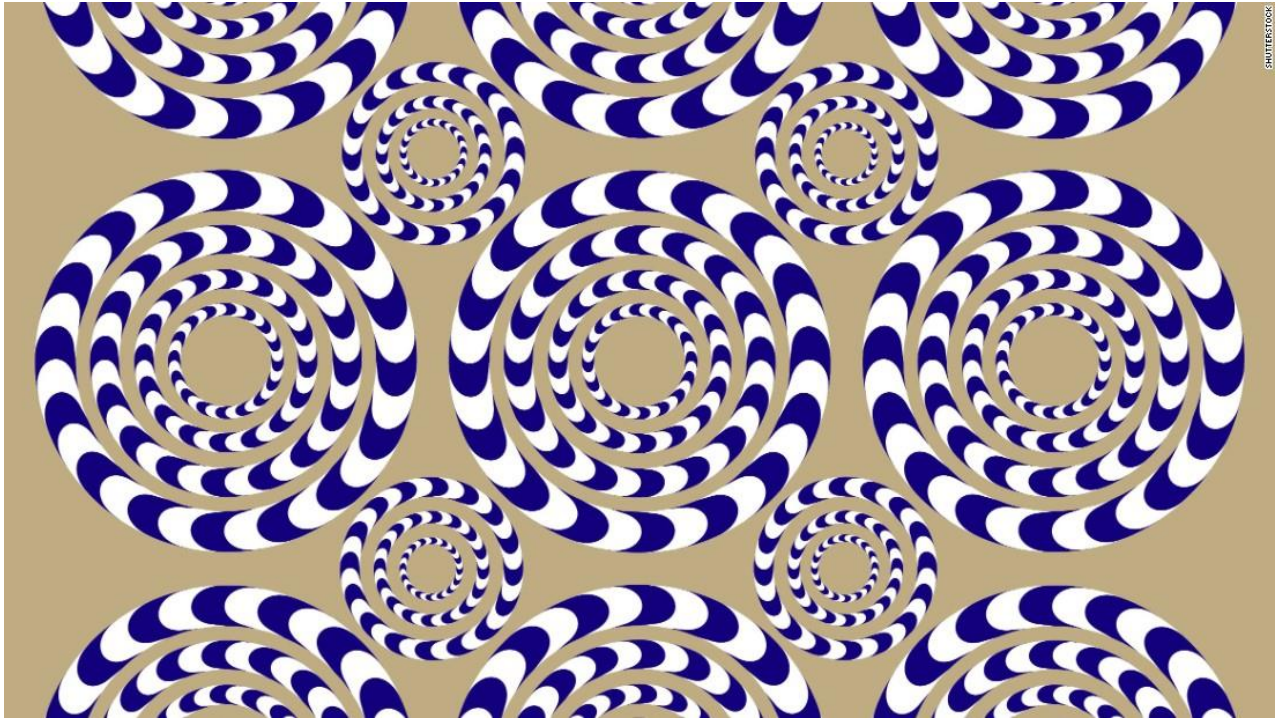
- **یادگیری عمیق :**آموختن خودکار مشخصه های مرتبط
- **آگاهی از تهدیدات :**داده کاوی و تشخیص الگو
- **یادگیری تقویتی عمیق :**آموختن سیاست های تاکتیکی دفاعی
- منابع بیشتر http://www.covert.io :



Recorded Future



اما ...



OFFSEC
www.offsec.ir

نمونه های مخرب

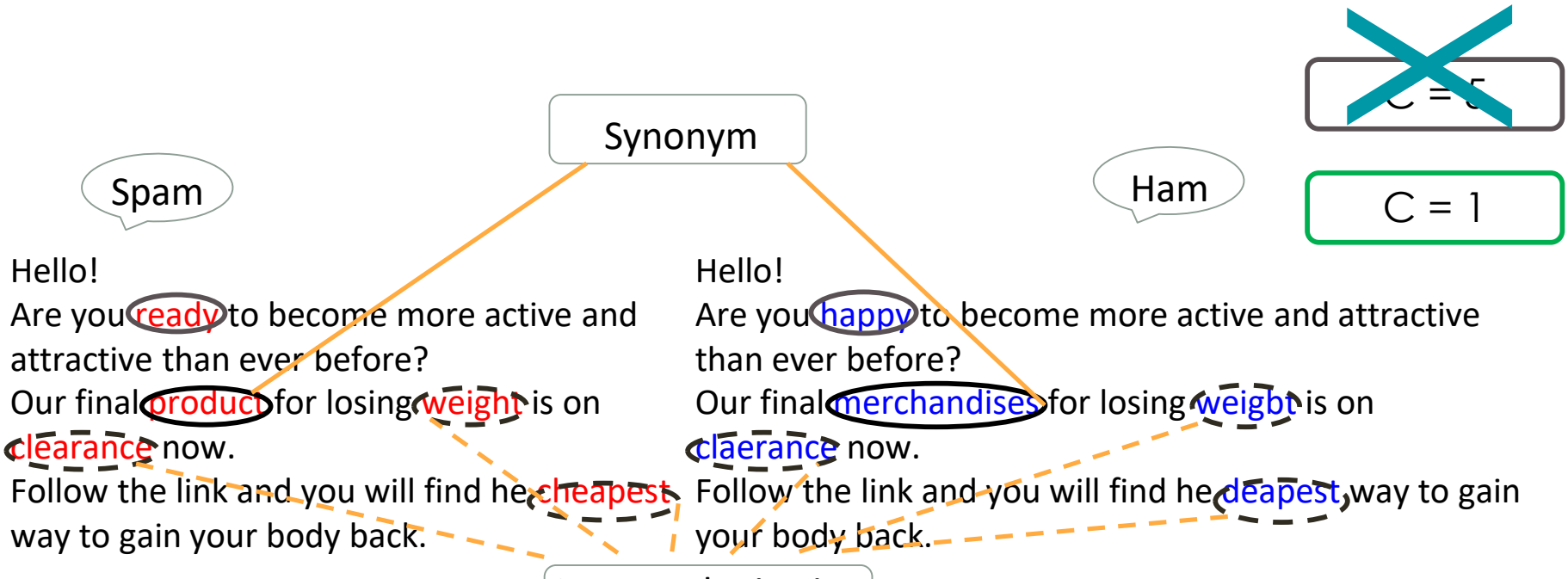
Classified as panda

Small adversarial noise

Classified as cat

آیا این مسئله فقط محدود به تشخیص پانداست؟

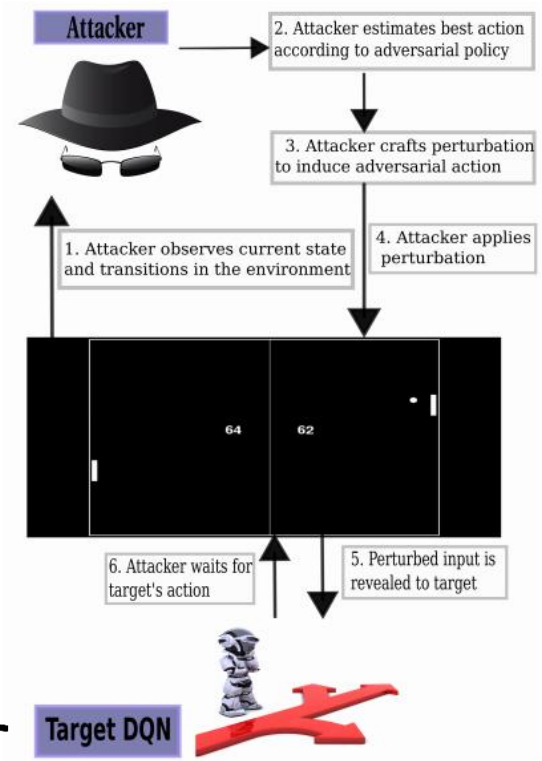
تشخیص هرزنامه



نمونه های مخرب در یادگیری تقویتی

Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks (Behzadan & Munir, 2017)

Policy Induction Attack Framework



امنیت در یادگیری ماشین

OFFSEC
www.offsec.ir

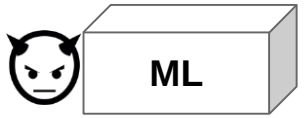
اهداف مهاجم

- **محرمانگی و حفظ حریم شخصی**
  - محرمانگی پارامترهای مدل
    - Tramèr et al. *Stealing Machine Learning Models via Prediction APIs*
  - حفظ حریم شخصی در داده های مورد استفاده در یادگیری
    - Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information*
    - Shokri et al. *Membership Inference Attacks Against Machine Learning Models*
    - Chaudhuri et al . *Privacy-preserving logistic regression*
    - Song et al. *Stochastic gradient descent with differentially private updates*
    - Abadi et al. *Deep learning with differential privacy*
    - Papernot et al. *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*

- **یکپارچگی و در دسترس بودن**
  - یکپارچگی تشخیص ها (خروجی صحیح)
  - در دسترس بودن سیستم مبتنی بر یادگیری ماشین

توانایی های مهاجم

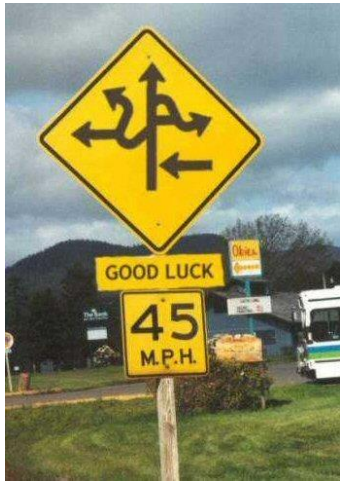
- حین یادگیری
  - Kearns et al. *Learning in the presence of malicious errors*
  - Biggio et al. *Support Vector Machines Under Adversarial Label Noise*
  - Kloft et al. *Online anomaly detection under adversarial impact*
- حین تشخیص
  - حملات جعبه-سفید



    - Szegedy et al. *Intriguing Properties of Neural Networks*
    - Biggio et al. *Evasion Attacks against Machine Learning at Test Time*
    - Goodfellow et al. *Explaining and Harnessing Adversarial Examples*
  - حملات جعبه-سیاه
    - Dalvi et al. *Adversarial Learning*
    - Szegedy et al. *Intriguing Properties of Neural Networks*
    - Xu et al. *Automatically evading classifiers: A Case Study on PDF Malware Classifiers*

حمله به یکپارچگی

- حملات گریزی
  - دست کاری عمدی ورودی به قصد *اغفال* مدل

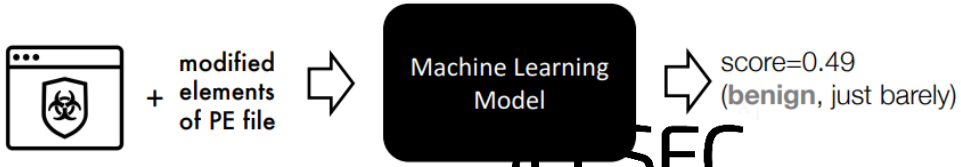


حمله به یکپارچگی

- حملات گریزی بر ضد مدل های تشخیص بدافزار

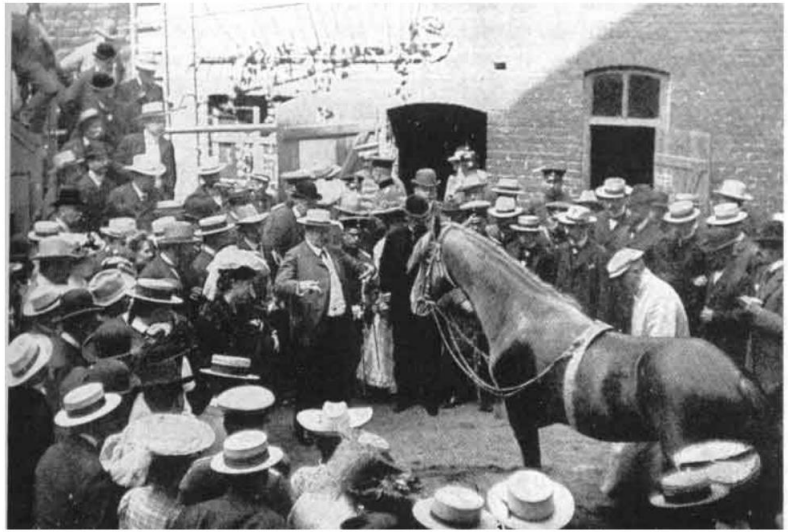
- Static machine learning model trained on millions of samples

- Simple structural changes that don't change behavior
  - upx_unpack
  - '.text' -> '.foo' (remains valid entry point)
  - create '.text' and populate with '.text from calc.exe'



Clever Hans

- https://github.com/tensorflow/cleverhans/
- Open-Source toolbox for crafting adversarial examples
- named after a “clever horse”



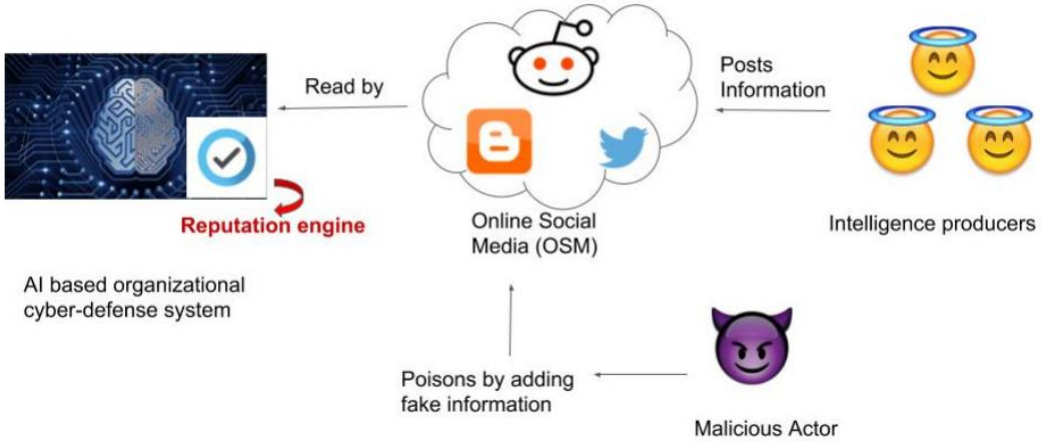
حمله به یکپارچگی

- حملات مسموم سازی داده

حمله به یکپارچگی

- حملات مسموم سازی داده
- دستکاری اطلاعات مورد استفاده در یادگیری
  - اضافه کردن نمونه های غیر واقعی
  - دستکاری نمونه های واقعی
  - حذف انتخابی نمونه ها

حمله به یکپارچگی



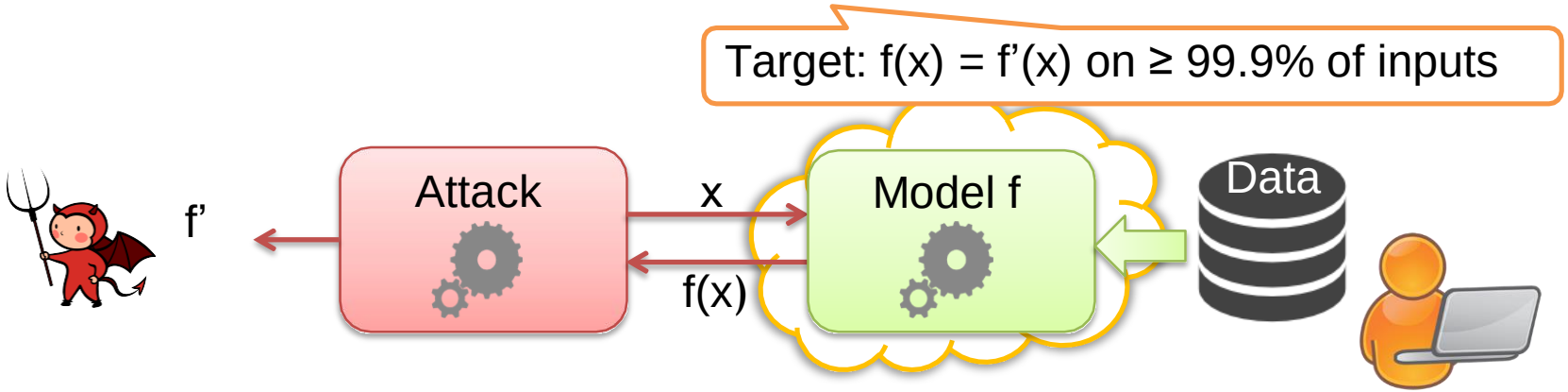
گریز و مسموم سازی

- تفاوت عمده
  - *گریز* بر ضد مدلی اعمال می شود که قبلا آموخته شده است.
  - *مسموم سازی* بر ضد الگوریتم یادگیری اعمال می شود .

- نمونه های مخرب مثالی از حملات گریزی هستند .

حمله به محرمانگی

استخراج مدل :مهاجم می کوشد که با مشاهده ی ورودی و خروجی های مدل، آن را تقریب بزند:



کاربردها:

1) دزدی مدل های مورد استفاده در سیستم های تجاری
2) آسان سازی حملات ضد حریم شخصی
3) گام اول در حملات گریزی

روش های دفاعی شکست خورده

Generative pretraining

Removing perturbation with an autoencoder

Adding noise at test time

Ensembles

Confidence-reducing perturbation at test time

Error correcting codes

Multiple glimpses

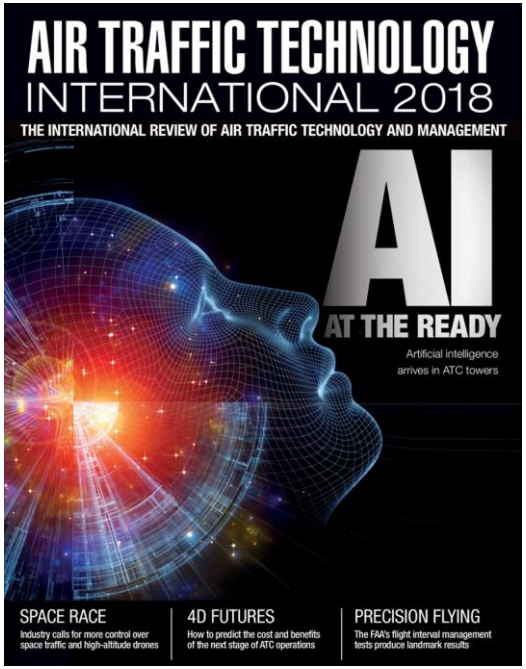
Weight decay

Double backprop

Adding noise at train time

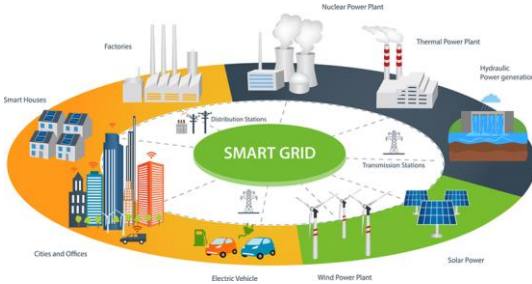
Various non-linear units

Dropout



نتیجه

- یادگیری ماشین به عنوان سطح حمله.
- عدم وجود روش های دفاعی مستحکم.
- اهمیت در نظر گرفتن امنیت در سیستم های مبتنی بر یادگیری.



OFFSEC
www.offsec.ir