# Knock It Off:

## Profiling the Online Storefronts of Counterfeit Merchandise

Matt Der
Lawrence Saul
Stefan Savage
Geoff Voelker

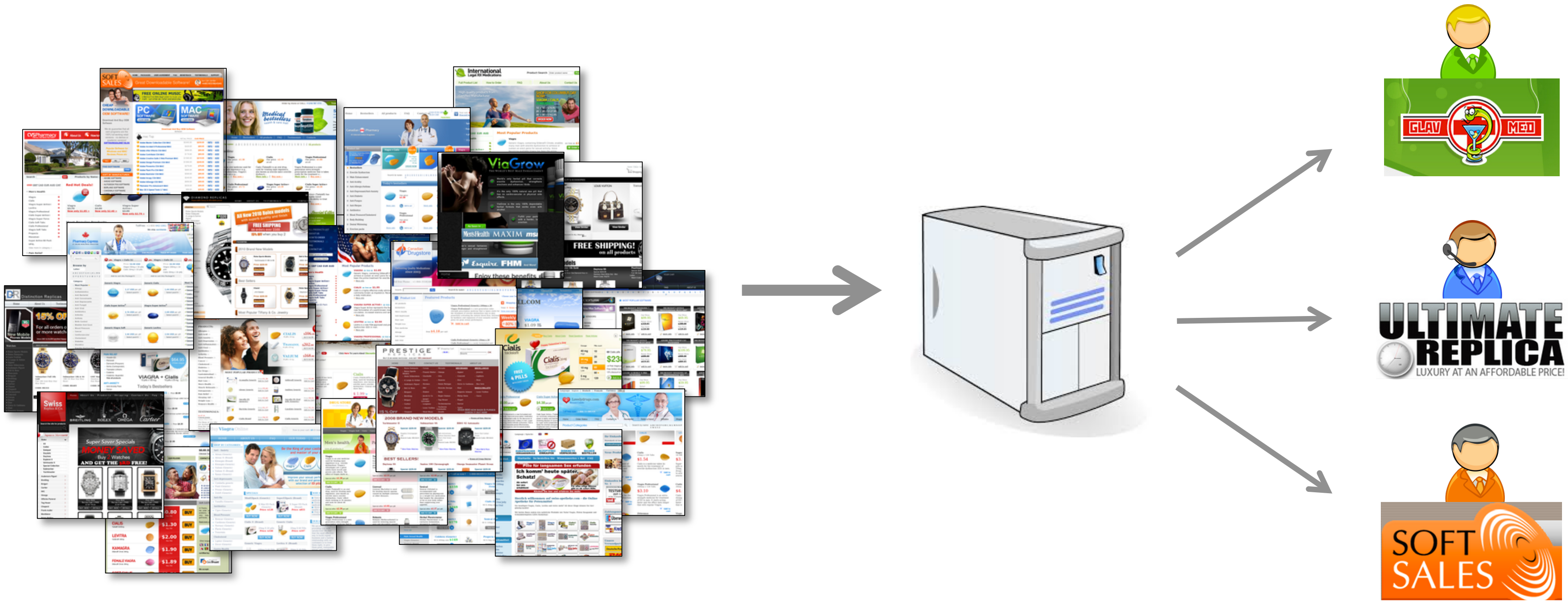
UCSDCSE
Computer Science and Engineering

cns
center for networked systems

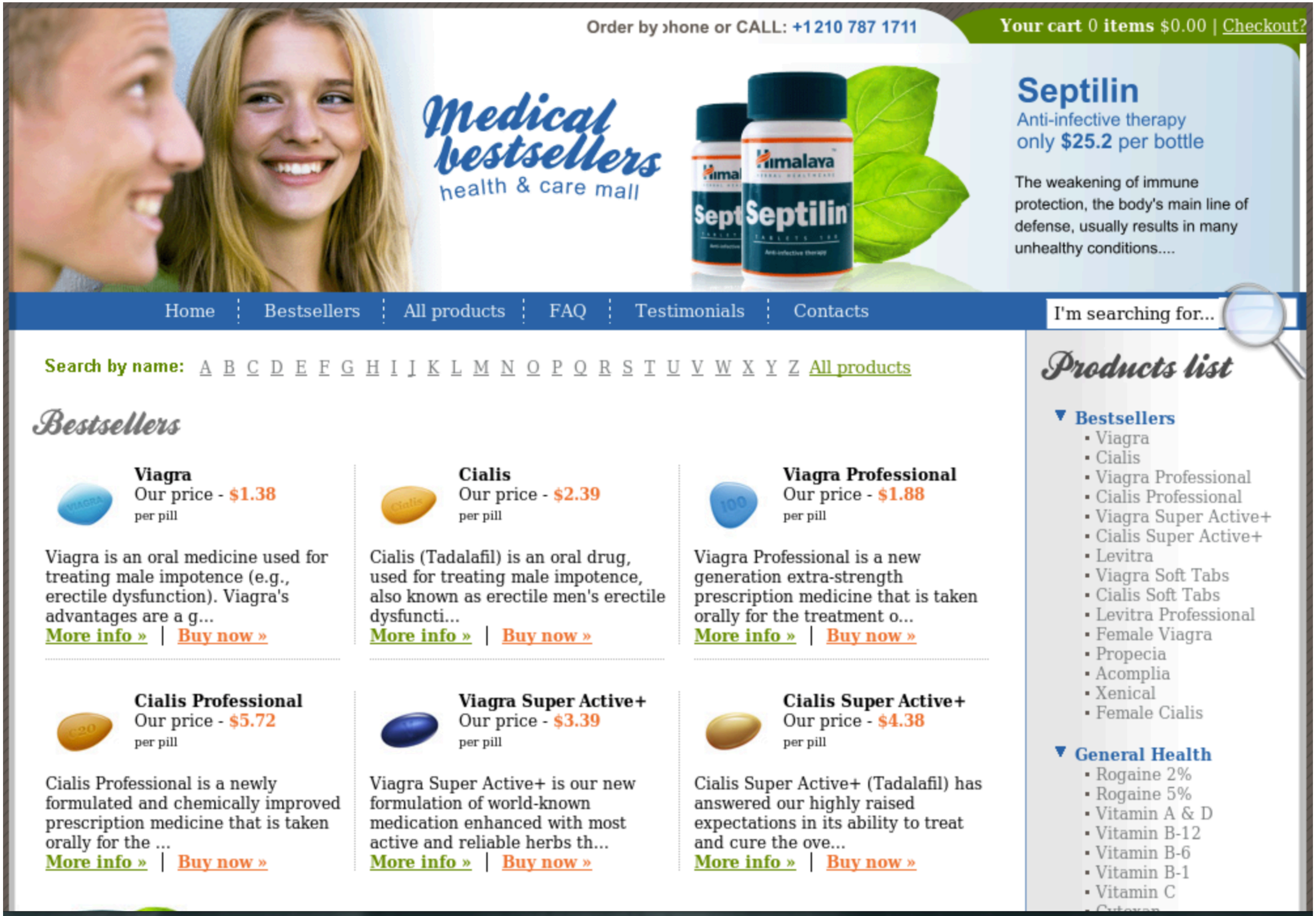
# Problem in a nutshell

Behind the many online storefronts for counterfeit goods lurk a small handful of sophisticated criminal operations.

How can automated, data-driven methods help to identify and target them?



# Counterfeit online storefronts



Order by ɔhone or CALL: +1 210 787 1711

Your cart 0 items $0.00 | Checkout?

**Septilin**

Anti-infective therapy
only **$25.2** per bottle

The weakening of immune protection, the body's main line of defense, usually results in many unhealthy conditions....

Home | Bestsellers | All products | FAQ | Testimonials | Contacts

I'm searching for...

Search by name: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All products

## Bestsellers

**Viagra**
Our price - **$1.38**
per pill

Viagra is an oral medicine used for treating male impotence (e.g., erectile dysfunction). Viagra's advantages are a g...
More info » | Buy now »

**Cialis**
Our price - **$2.39**
per pill

Cialis (Tadalafil) is an oral drug, used for treating male impotence, also known as erectile men's erectile dysfuncti...
More info » | Buy now »

**Viagra Professional**
Our price - **$1.88**
per pill

Viagra Professional is a new generation extra-strength prescription medicine that is taken orally for the treatment o...
More info » | Buy now »

**Cialis Professional**
Our price - **$5.72**
per pill

Cialis Professional is a newly formulated and chemically improved prescription medicine that is taken orally for the ...
More info » | Buy now »

**Viagra Super Active+**
Our price - **$3.39**
per pill

Viagra Super Active+ is our new formulation of world-known medication enhanced with most active and reliable herbs th...
More info » | Buy now »

**Cialis Super Active+**
Our price - **$4.38**
per pill

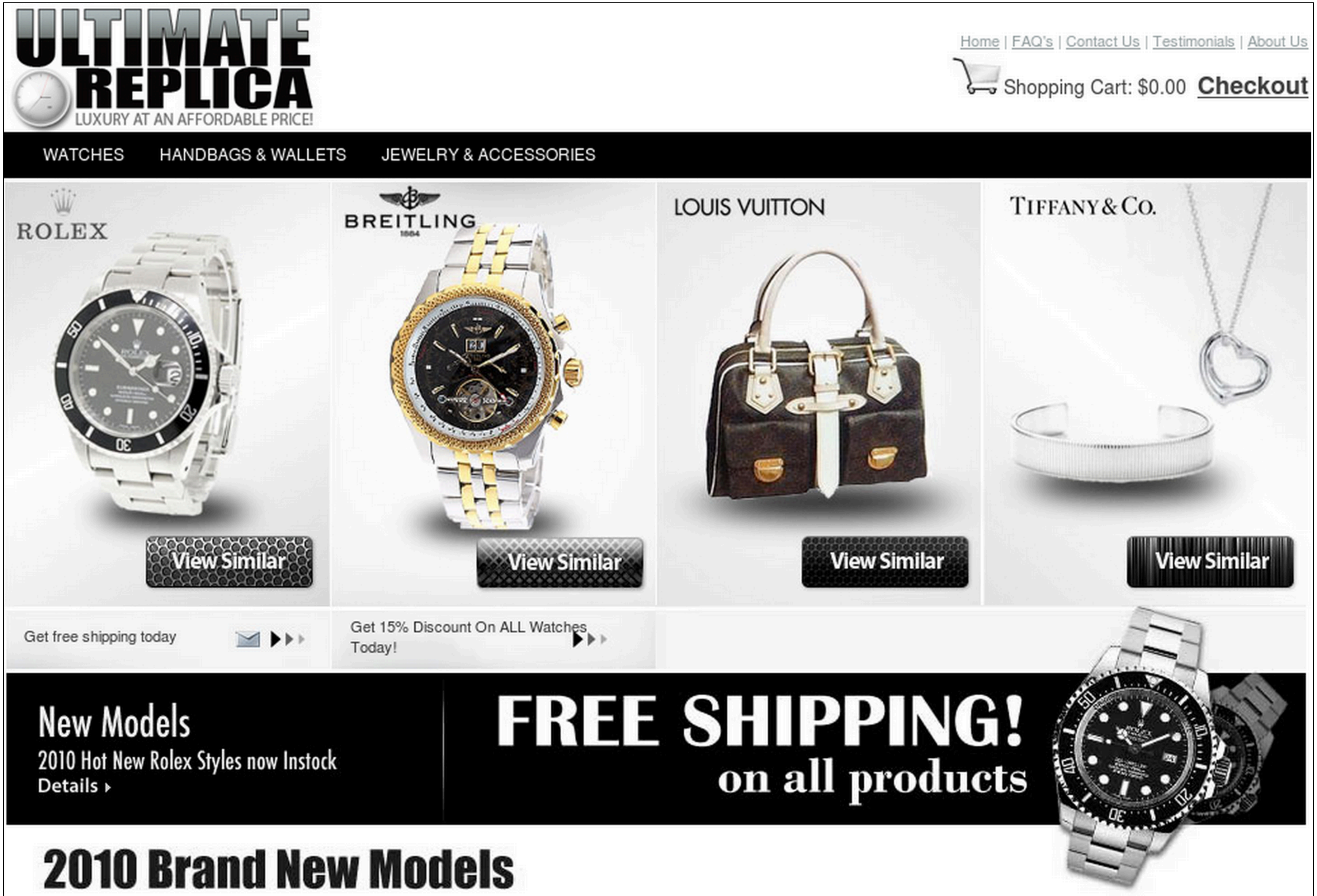
Cialis Super Active+ (Tadalafil) has answered our highly raised expectations in its ability to treat and cure the ove...
More info » | Buy now »

## Products list

**Bestsellers**
- Viagra
- Cialis
- Viagra Professional
- Cialis Professional
- Viagra Super Active+
- Cialis Super Active+
- Levitra
- Viagra Soft Tabs
- Cialis Soft Tabs
- Levitra Professional
- Female Viagra
- Propecia
- Acomplia
- Xenical
- Female Cialis

**General Health**
- Rogaine 2%
- Rogaine 5%
- Vitamin A & D
- Vitamin B-12
- Vitamin B-6
- Vitamin B-1
- Vitamin C

# Counterfeit online storefronts



Get free shipping today

Get 15% Discount On ALL Watches Today!

New Models

2010 Hot New Rolex Styles now Instock

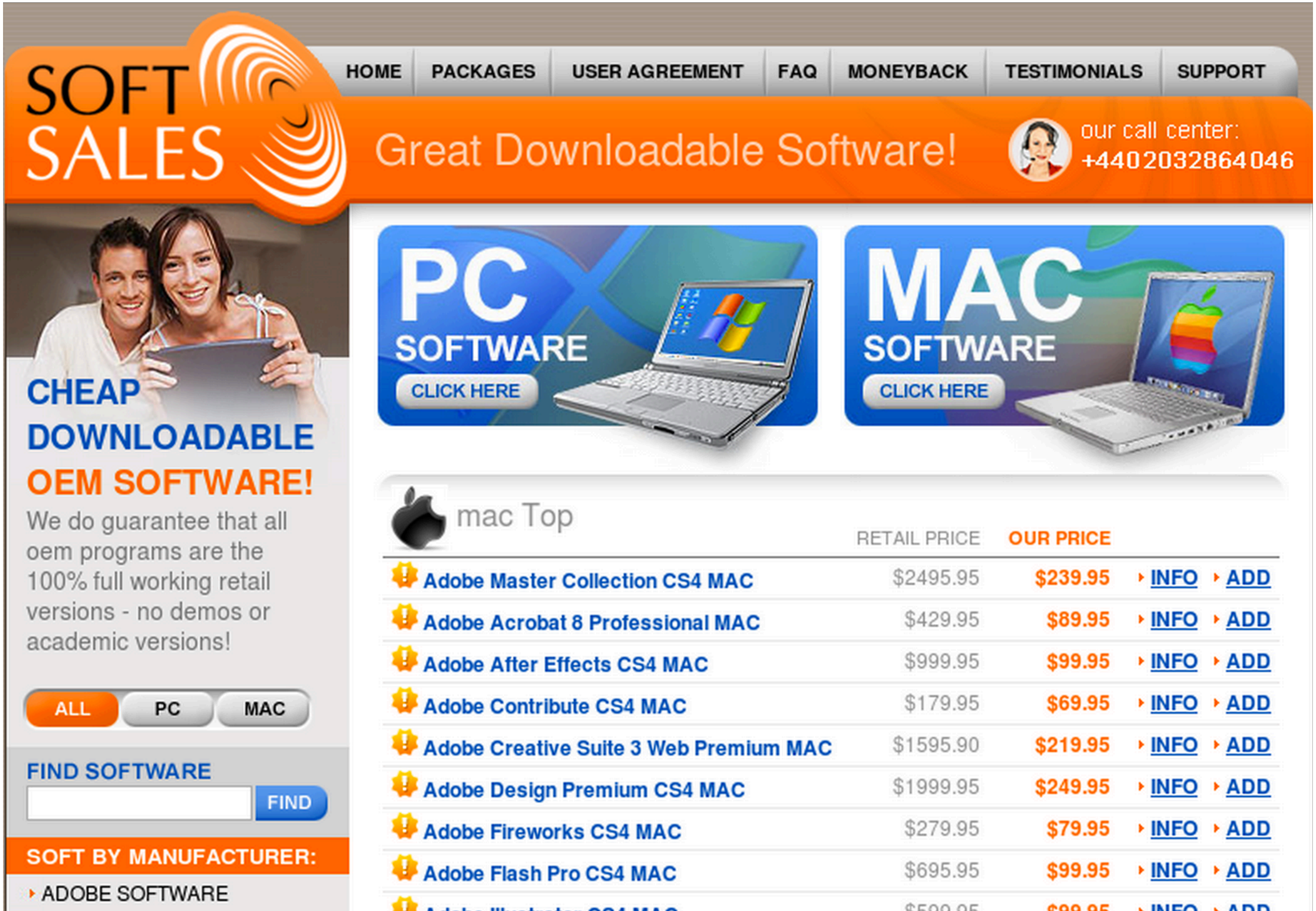
Details ▸

FREE SHIPPING!

on all products

2010 Brand New Models

# Counterfeit online storefronts



Great Downloadable Software!

our call center: +442032864046

CHEAP DOWNLOADABLE OEM SOFTWARE!

We do guarantee that all oem programs are the 100% full working retail versions - no demos or academic versions!

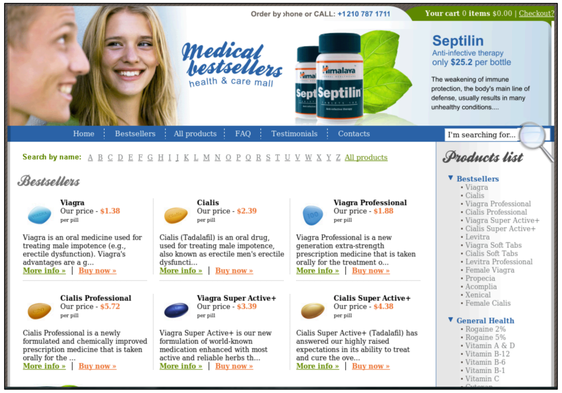
SOFT BY MANUFACTURER:

- ADOBE SOFTWARE

mac Top

| | RETAIL PRICE | OUR PRICE | | |
|---|---|---|---|---|
| Adobe Master Collection CS4 MAC | $2495.95 | $239.95 | INFO | ADD |
| Adobe Acrobat 8 Professional MAC | $429.95 | $89.95 | INFO | ADD |
| Adobe After Effects CS4 MAC | $999.95 | $99.95 | INFO | ADD |
| Adobe Contribute CS4 MAC | $179.95 | $69.95 | INFO | ADD |
| Adobe Creative Suite 3 Web Premium MAC | $1595.90 | $219.95 | INFO | ADD |
| Adobe Design Premium CS4 MAC | $1999.95 | $249.95 | INFO | ADD |
| Adobe Fireworks CS4 MAC | $279.95 | $79.95 | INFO | ADD |
| Adobe Flash Pro CS4 MAC | $695.95 | $99.95 | INFO | ADD |
| Adobe Illustrator CS4 MAC | $599.95 | $99.95 | INFO | ADD |

# Who is running the store?


Order by phone or CALL: +1 210 787 1711
Your cart 0 items $0.00 | Checkout?
Medical bestsellers
health & care mall
Home
Bestsellers
All products
FAQ
Testimonials
Contacts
I'm searching for...
Search by name: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All products
More info »
Buy now »

Septilin
Anti-infective therapy
only $25.2 per bottle

The weakening of immune protection, the body's main line of defense, usually results in many unhealthy conditions....

Bestsellers

**Viagra**
Our price - $1.38
per pill

Viagra is an oral medicine used for treating male impotence (e.g., erectile dysfunction). Viagra's advantages are a g...

**Cialis**
Our price - $2.39
per pill

Cialis (Tadalafil) is an oral drug, used for treating male impotence, also known as erectile men's erectile dysfuncti...

**Viagra Professional**
Our price - $1.88
per pill

Viagra Professional is a new generation extra-strength prescription medicine that is taken orally for the treatment o...

**Cialis Professional**
Our price - $5.72
per pill

Cialis Professional is a newly formulated and chemically improved prescription medicine that is taken orally for the ...

**Viagra Super Active+**
Our price - $3.39
per pill

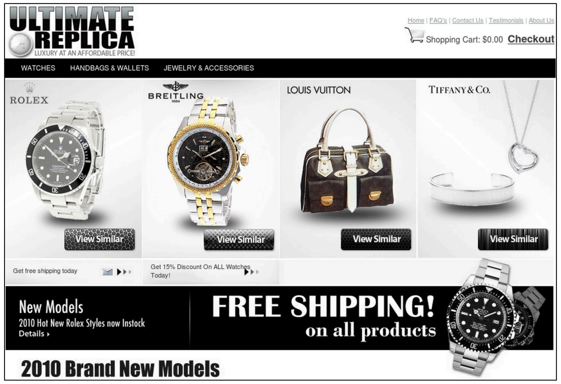
Viagra Super Active+ is our new formulation of world-known medication enhanced with most active and reliable herbs th...

**Cialis Super Active+**
Our price - $4.38
per pill

Cialis Super Active+ (Tadalafil) has answered our highly raised expectations in its ability to treat and cure the ove...

Products list

- Bestsellers
  - Viagra
  - Cialis
  - Viagra Professional
  - Cialis Professional
  - Viagra Super Active+
  - Cialis Super Active+
  - Levitra
  - Viagra Soft Tabs
  - Cialis Soft Tabs
  - Levitra Professional
  - Female Viagra
  - Propecia
  - Acomplia
  - Xenical
  - Female Cialis
- General Health
  - Rogaine 2%
  - Rogaine 5%
  - Vitamin A & D
  - Vitamin B-12
  - Vitamin B-6
  - Vitamin B-1
  - Vitamin C



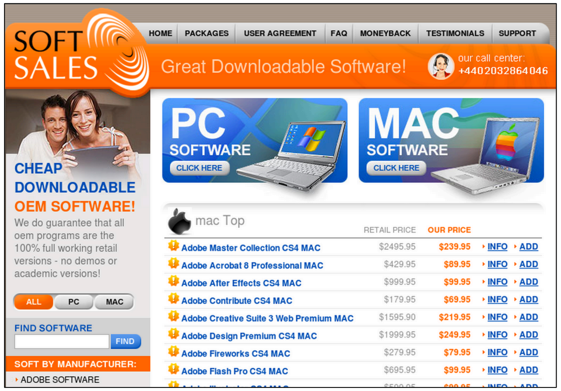
Get free shipping today

Get 15% Discount On ALL Watches Today!

New Models
2010 Hot New Rolex Styles now Instock
Details ›

FREE SHIPPING!
on all products

2010 Brand New Models



Great Downloadable Software!

our call center:
+44 0 2032864046

CHEAP DOWNLOADABLE OEM SOFTWARE!

We do guarantee that all oem programs are the 100% full working retail versions - no demos or academic versions!

mac Top

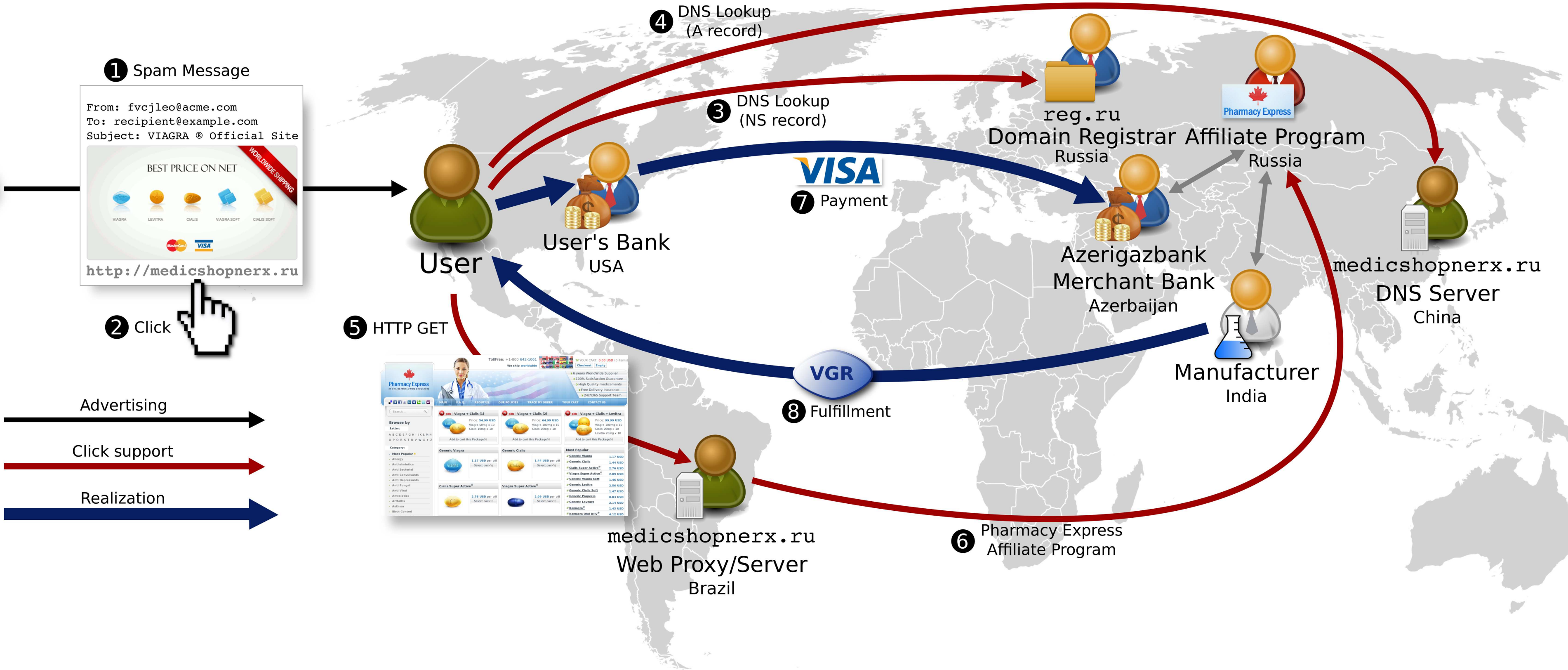
| | RETAIL PRICE | OUR PRICE | | |
|---|---|---|---|---|
| Adobe Master Collection CS4 MAC | $2495.95 | $239.95 | INFO | ADD |
| Adobe Acrobat 8 Professional MAC | $429.95 | $89.95 | INFO | ADD |
| Adobe After Effects CS4 MAC | $999.95 | $99.95 | INFO | ADD |
| Adobe Contribute CS4 MAC | $179.95 | $69.95 | INFO | ADD |
| Adobe Creative Suite 3 Web Premium MAC | $1595.90 | $219.95 | INFO | ADD |
| Adobe Design Premium CS4 MAC | $1999.95 | $249.95 | INFO | ADD |
| Adobe Fireworks CS4 MAC | $279.95 | $79.95 | INFO | ADD |
| Adobe Flash Pro CS4 MAC | $695.95 | $99.95 | INFO | ADD |

“affiliate programs”

# What is an affiliate program?

- Illegitimate business that sells counterfeit goods: millions of $$$ of revenue per month
- Manage Web sites that serve as online storefronts
- Enlist spammers to advertise their storefronts via bulk email
- Contract out payment & fulfillment services

# Click Trajectories



- Bottleneck is payment processing

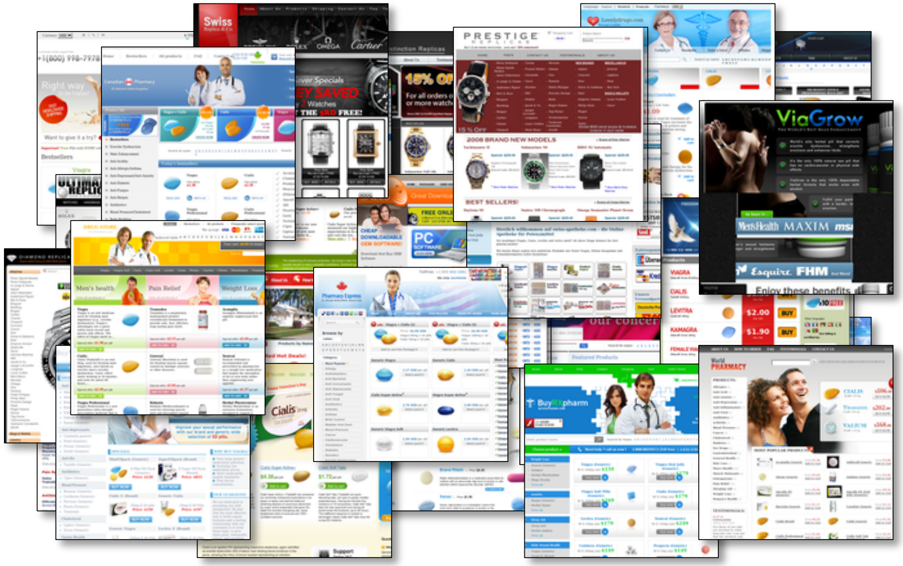
  95% of spam-advertised pharmaceutical, replica, and software products are monetized through only a handful of merchant banks

- If affiliate program can't process credit cards, business collapses

Figure from Levchenko et al, 2011

# Key insight

## 100s of thousands of storefronts



# Key insight

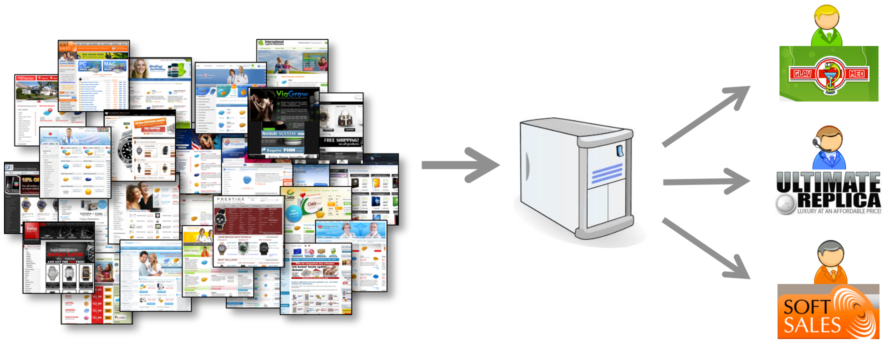
## 100s of thousands of storefronts

## *dozens* of affiliate programs

PRESTIGE
LUXURY FOOTWEAR AND BAGS
SOFT SALES
Exquisite
REPLICAS
YOUR PASS TO THE RICH WORLD
#1 Online Pharmacy
Generic & Brand medications
ULTIMATE REPLICA
LUXURY AT AN AFFORDABLE PRICE!
GLAV MED
VIRILITY
ViaGrow
THE WORLD'S BEST MALE ENHANCEMENT
OEM Soft Store
Easy and Fast Download
AUTHORIZED SOFTWARE RESELLER
PREMIER ONLINE STORE FOR POPULAR SOFTWARE
DIAMOND REPLICAS
Impeccable luxury at affordable prices
Dr. MaXman
Max Penis Enlarger Pills

# Our work

- **Goal**: classify storefronts by affiliate program;
disrupt their operation to undermine spam business model

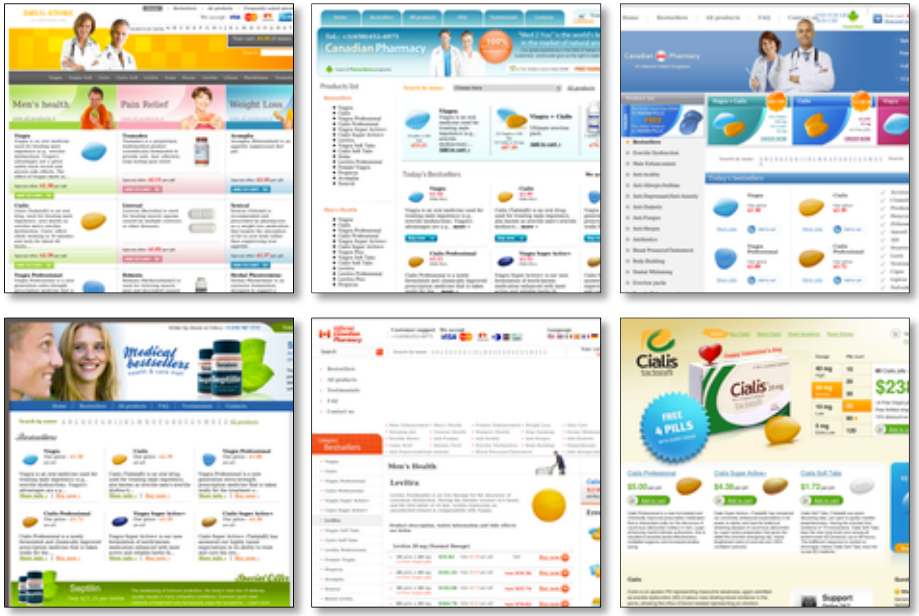


- **Approach**: HTML bag-of-words, nearest neighbor classification (**automated** system)
- **Takeaway**: highly accurate — even with simple classifier & limited labeled examples

# Challenges

# Challenges

1. Web pages that render very differently are often linked to the same affiliate program



Men's health

Pain Relief

Weight Loss

Canadian Pharmacy

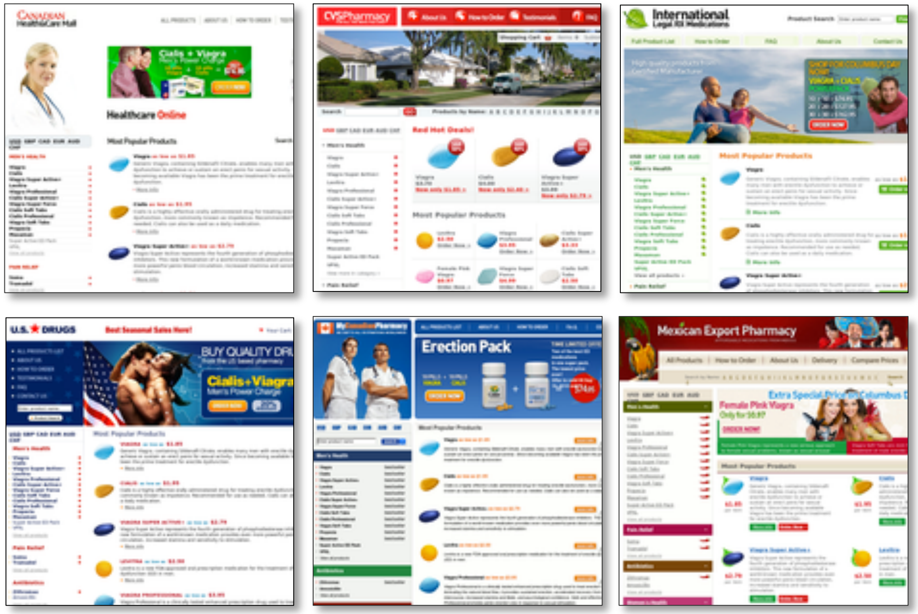
Medical bestsellers

Levitra

Cialis

FREE 4 PILLS

GlavMed



Healthcare Online

Cialis + Viagra

CVSPharmacy

International Legal RX Medications

U.S. DRUGS

BUY QUALITY DRUGS

Erection Pack

Mexican Export Pharmacy

Female Pink Viagra

EvaPharmacy

# Challenges

2. Difficulty in acquiring training data

# Challenges

2. Difficulty in acquiring training data

# Challenges

2. Difficulty in acquiring training data

U.S. ★ DRUGS

Best Seasonal Sales Here!

BUY QUALITY DRU

Cialis+Viagra

Men's Power Charge

Most Popular Products

# Challenges

2. Difficulty in acquiring training data

Eva

# Challenges

2. Difficulty in acquiring training data

Eva

U.S. ★ DRUGS

Best Seasonal Sales Here!

BUY QUALITY DRU

Cialis+Viagra

Men's Power Charge

Most Popular Products

Prestige

PRESTIGE

LUXURY FOOTWEAR AND BAGS

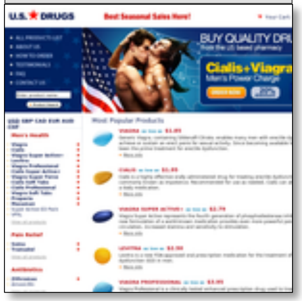
PRESTIGE Luxury Footwear

PRESTIGE 2008 Best Selling Luxury Shoes

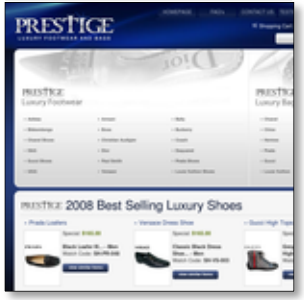
# Challenges

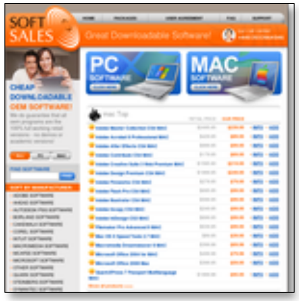
2. Difficulty in acquiring training data

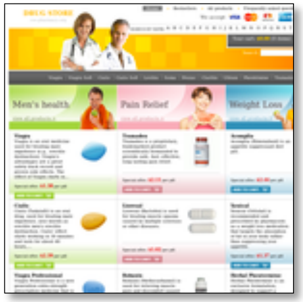
Eva

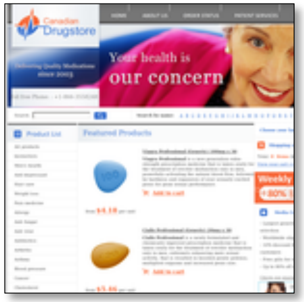


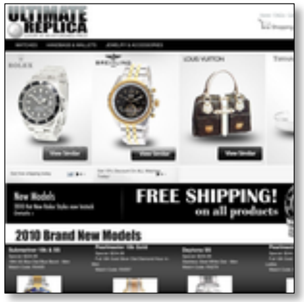
Prestige




SOFT SALES
Great Downloadable Software!
PC SOFTWARE
MAC SOFTWARE
CHEAP DOWNLOADABLE OEM SOFTWARE!







# Challenges

## 2. Difficulty in acquiring training data

expert labeling is slow & tedious!

Eva

Prestige

SoftSales

SOFT SALES
Great Downloadable Software!
PC SOFTWARE
MAC SOFTWARE
CHEAP DOWNLOADABLE OEM SOFTWARE!

GlavMed

Greenline

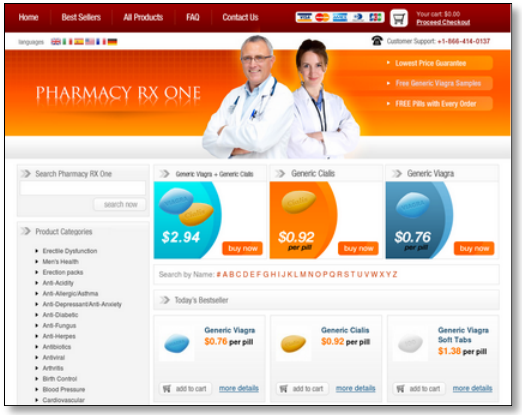
Ultimate

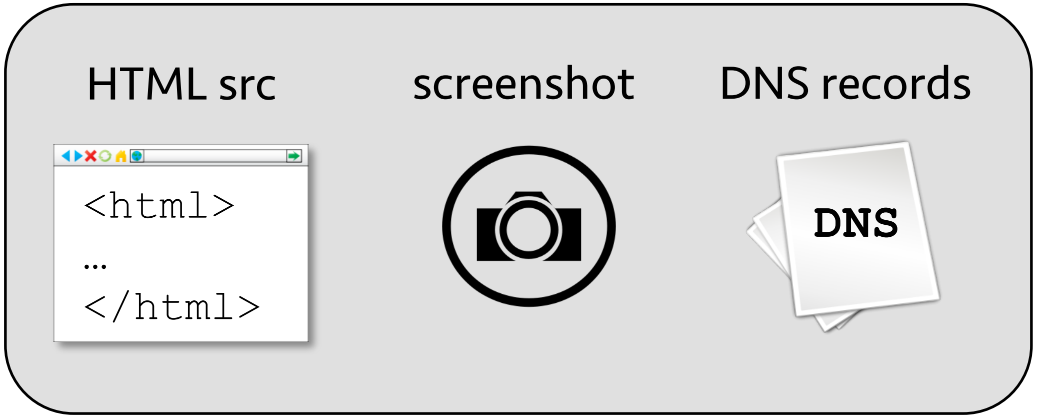
# A role for machine learning

- Security experts labeled 178k storefronts
  - Estimated **~200 person-hours**
  - Painstaking manual process
    - Inspect HTML source for signals, encode with regular expressions

- **NOT** once-and-for-all effort
  - Storefronts change over time, regexs go stale
- Ripe opportunity for machine learning — a more automated approach to aid security practitioners

# Feature extraction



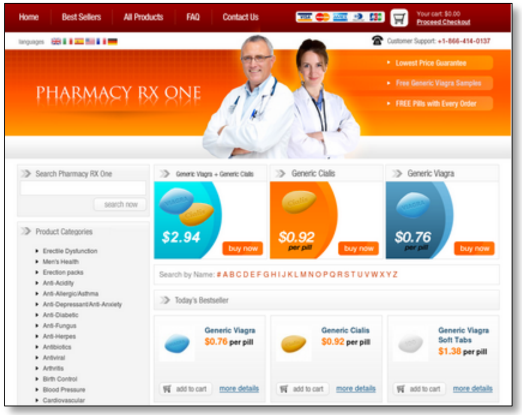


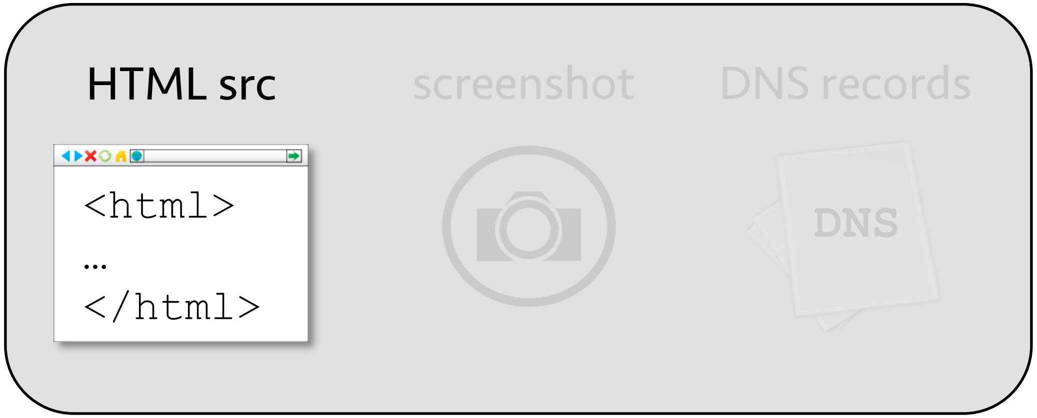
HTML src

screenshot

DNS records

# Feature extraction

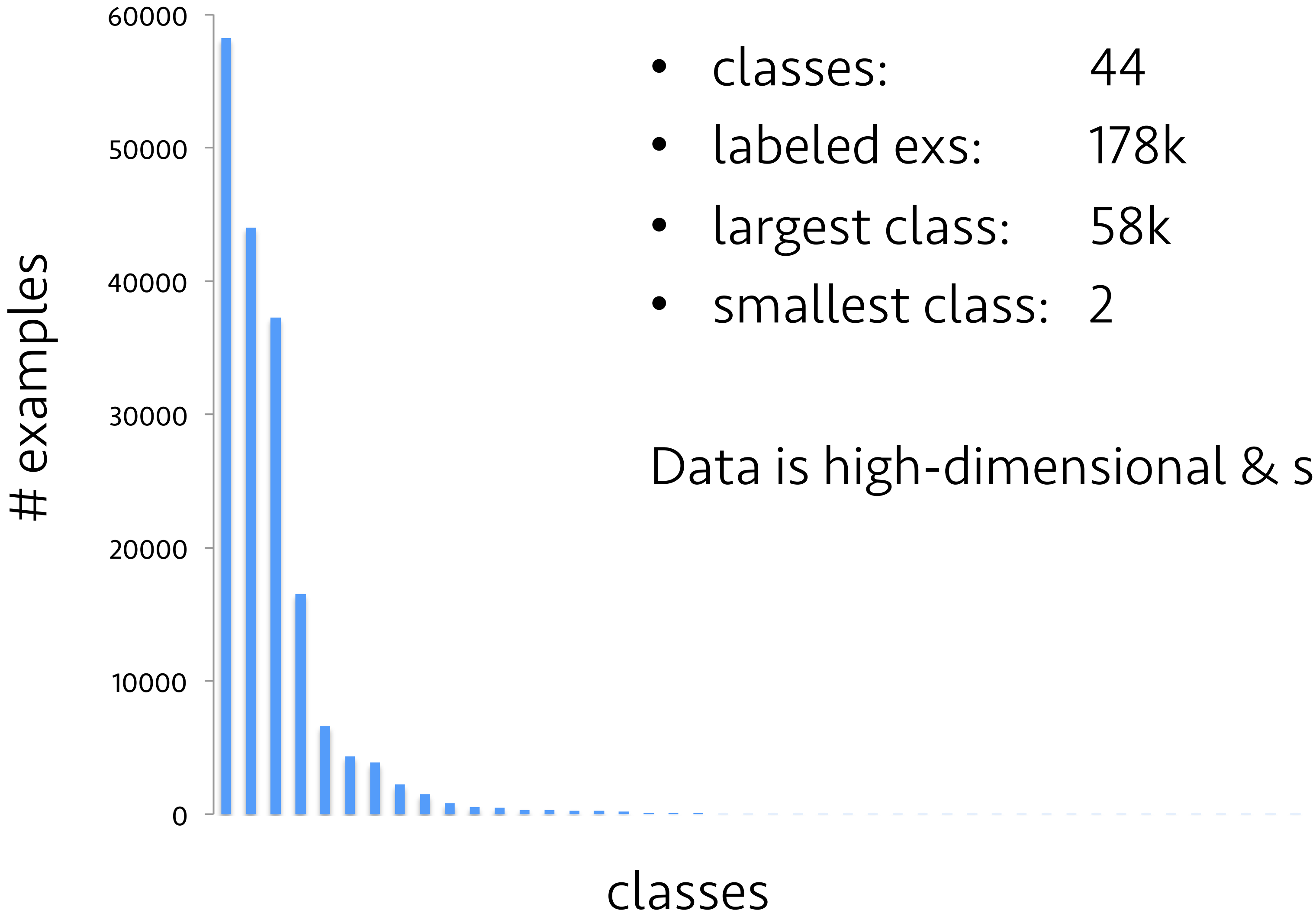




```
<html>
…
</html>
```

- Affiliate programs use in-house software engines to generate storefront templates
- HTML contains distinctive signatures
- Bag-of-words on HTML – automated!

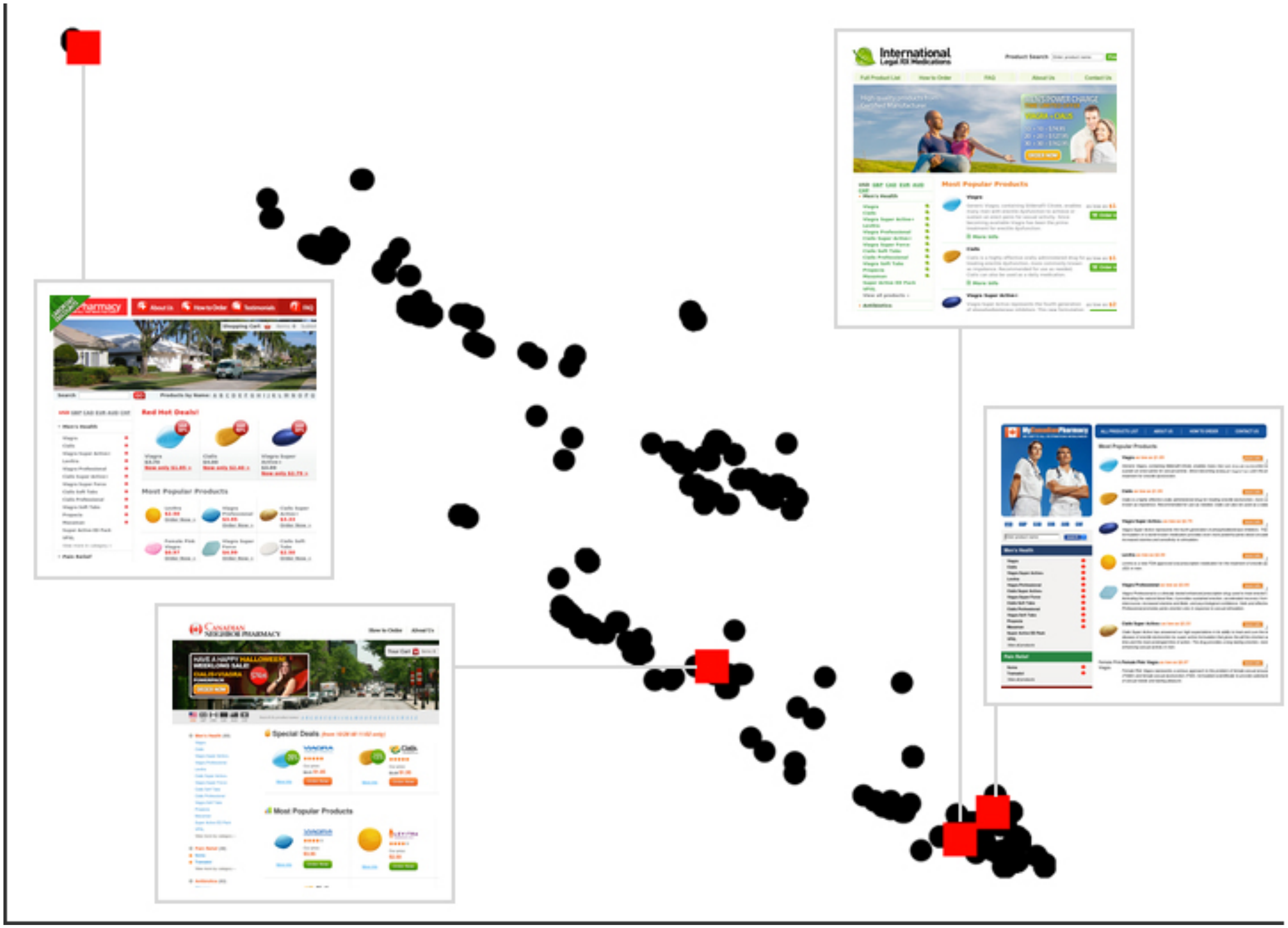
# Data set



- classes: 44
- labeled exs: 178k
- largest class: 58k
- smallest class: 2

Data is high-dimensional & sparse

# Visualization of EvaPharmacy



# Proof-of-concept experiment

- **Question**: are these HTML features enough to distinguish affiliate programs:
  - From one another?
  - From the noise in the Web crawl?
    - 43k unlabeled Web pages → "other" class
- Favorable setting: plenty of labeled data
- 45-way 1-nearest neighbor classification, 10 random 70/30 train/test splits

# Proof-of-concept experiment

- **Question**: are these HTML features enough to distinguish affiliate programs:
  - From one another?
  - From the noise in the Web crawl?
    - 43k unlabeled Web pages → "other" class
- Favorable setting: plenty of labeled data
- 45-way 1-nearest neighbor classification, 10 random 70/30 train/test splits

Avg accuracy = 99.95%

# Proof-of-concept experiment

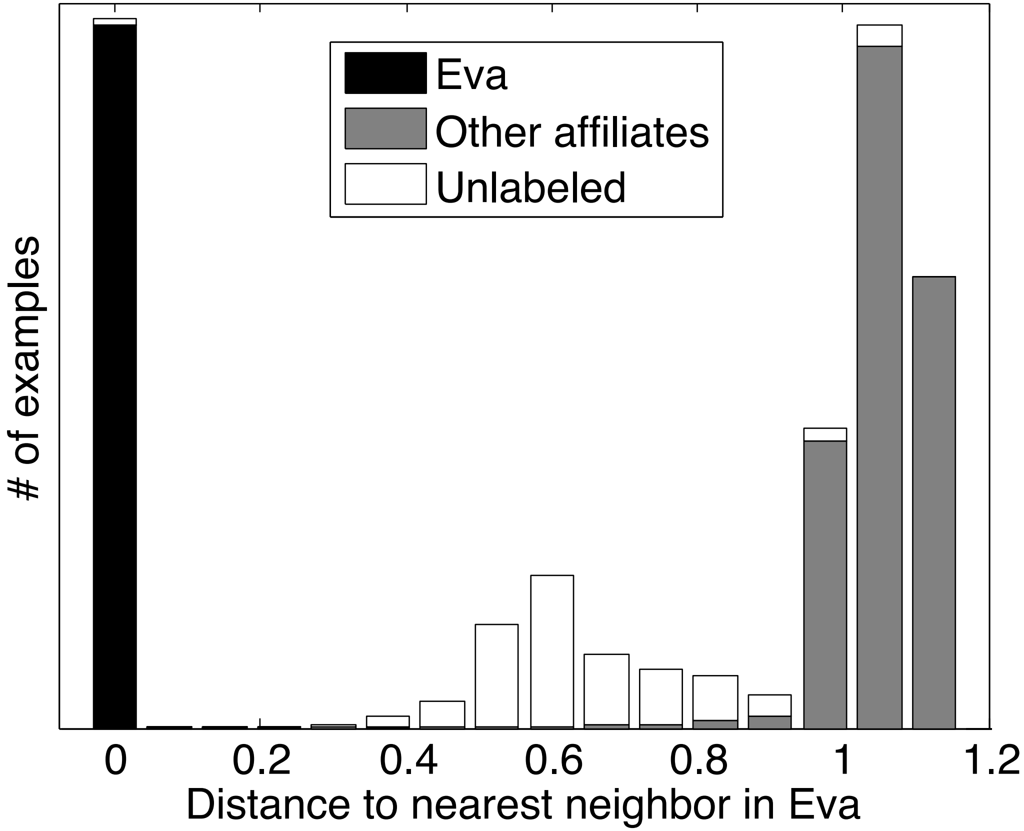
- **Question**: are these HTML features enough to distinguish affiliate programs:
  - From one another?
  - From the noise in the Web crawl?
    - 43k unlabeled Web pages → "other" class
- Favorable setting: plenty of labeled data
- 45-way 1-nearest neighbor classification, 10 random 70/30 train/test splits

Avg accuracy = 99.95%

How so good?!

# HTML distances are highly predictive

Distances from *every* point to nearest neighbor in EvaPharmacy



# Mimicking an operational deployment

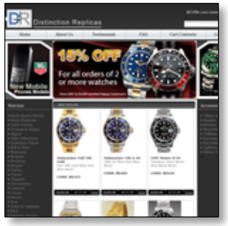
- Experts must label **some** storefronts, but how many?
- Learning from scratch:  only small initial seed of labeled storefronts

# Mimicking an operational deployment

- Experts must label **some** storefronts, but how many?
- Learning from scratch: only small initial seed of labeled storefronts

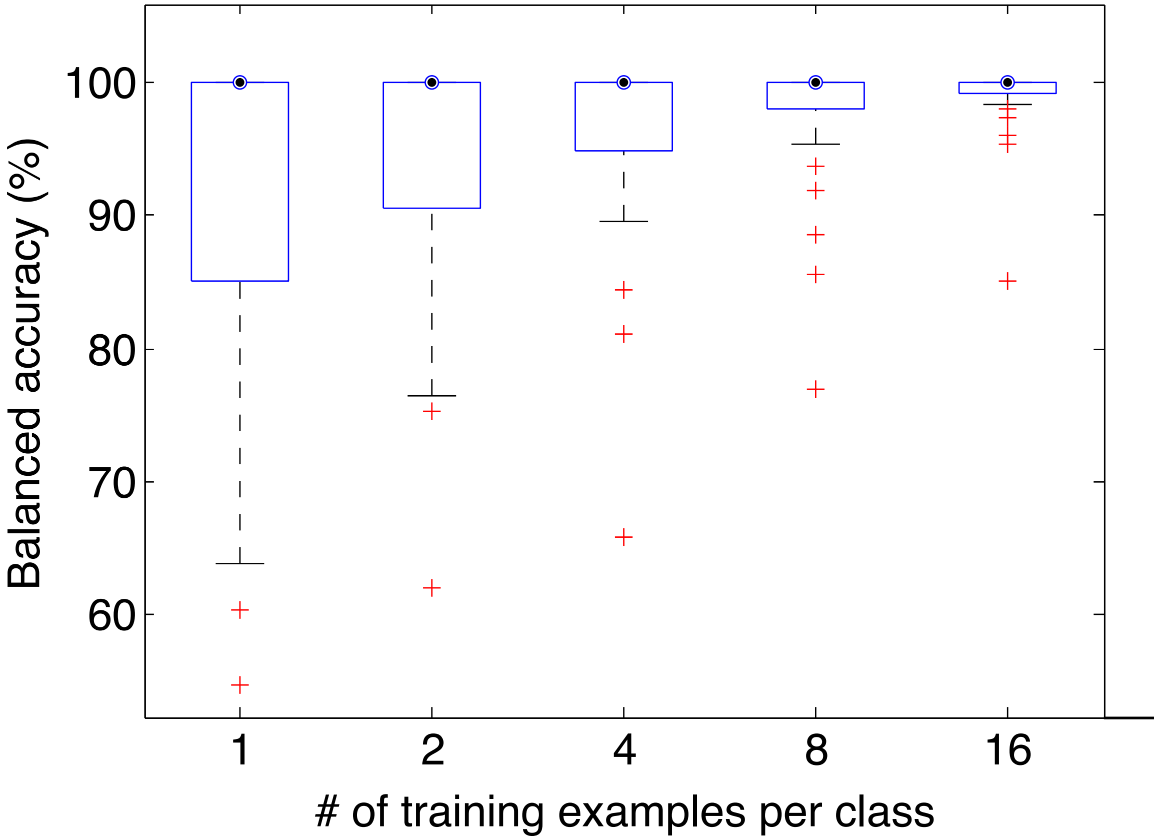
# Mimicking an operational deployment

- Experts must label **some** storefronts, but how many?
- Learning from scratch:  only small initial seed of labeled storefronts



# Classification in operational setting

Avg accuracy 75% 85% 93% 97% 98% 99.95%



• • •

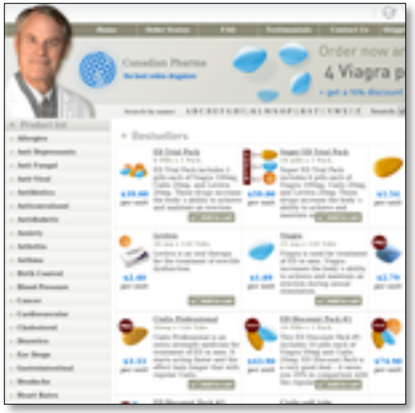
ALL
(200 hours)

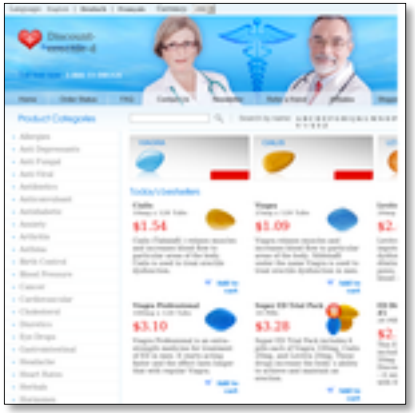
# One-shot learning

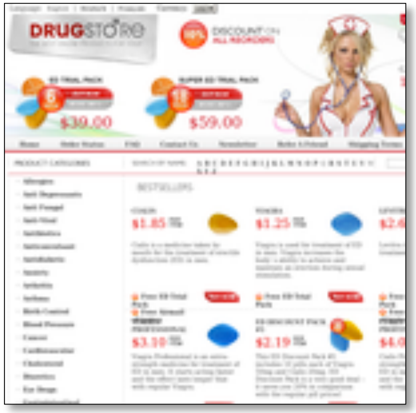
Singly labeled storefront

Correctly classified storefronts

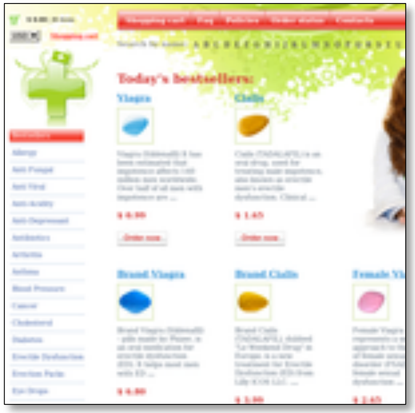
33drugs

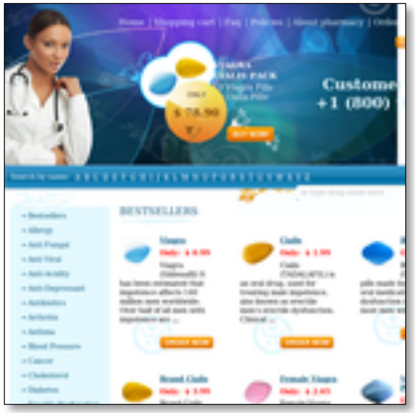


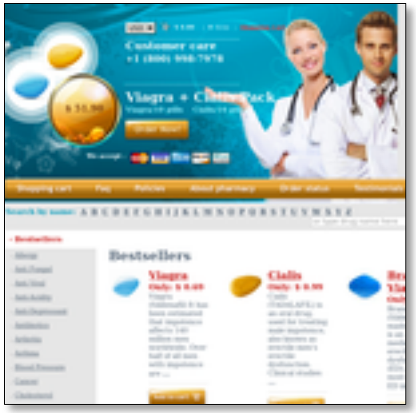




RX-Promotions

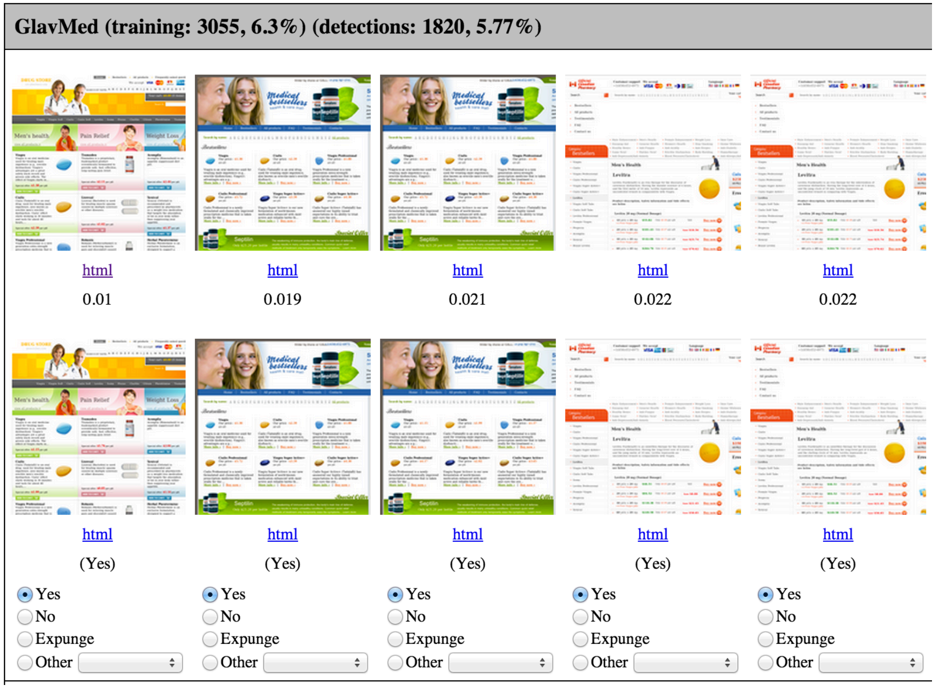






# Further results

- Found & labeled 3,785 additional storefront pages



GlavMed (training: 3055, 6.3%) (detections: 1820, 5.77%)

| | | | | |
|---|---|---|---|---|
| html | html | html | html | html |
| 0.01 | 0.019 | 0.021 | 0.022 | 0.022 |
| html | html | html | html | html |
| (Yes) | (Yes) | (Yes) | (Yes) | (Yes) |
| ● Yes<br>○ No<br>○ Expunge<br>○ Other | ● Yes<br>○ No<br>○ Expunge<br>○ Other | ● Yes<br>○ No<br>○ Expunge<br>○ Other | ● Yes<br>○ No<br>○ Expunge<br>○ Other | ● Yes<br>○ No<br>○ Expunge<br>○ Other |

- Clustered unlabeled Web pages to identify new affiliate programs

# Conclusion

- Automated system for identifying affiliate programs behind illegal online storefronts
- Simple model is highly accurate
  - Templatized storefronts, many near-duplicates
  - Affiliate programs' efforts to operate at scale make automated defense possible
- Big win for security practitioners
  - Modest labeling effort is enough to bootstrap the system
- Security a Big Data field; ML an invaluable tool for large-scale Web crawls

# Thank you!

Questions?

UCSDCSE
Computer Science and Engineering