



Mining human-scale insights from
log data with machine learning

David Andrzejewski - @davidandrzej
Data Sciences Engineering, Sumo Logic
OC Big Data Meetup, September 17, 2014

OC Big Data Meetup



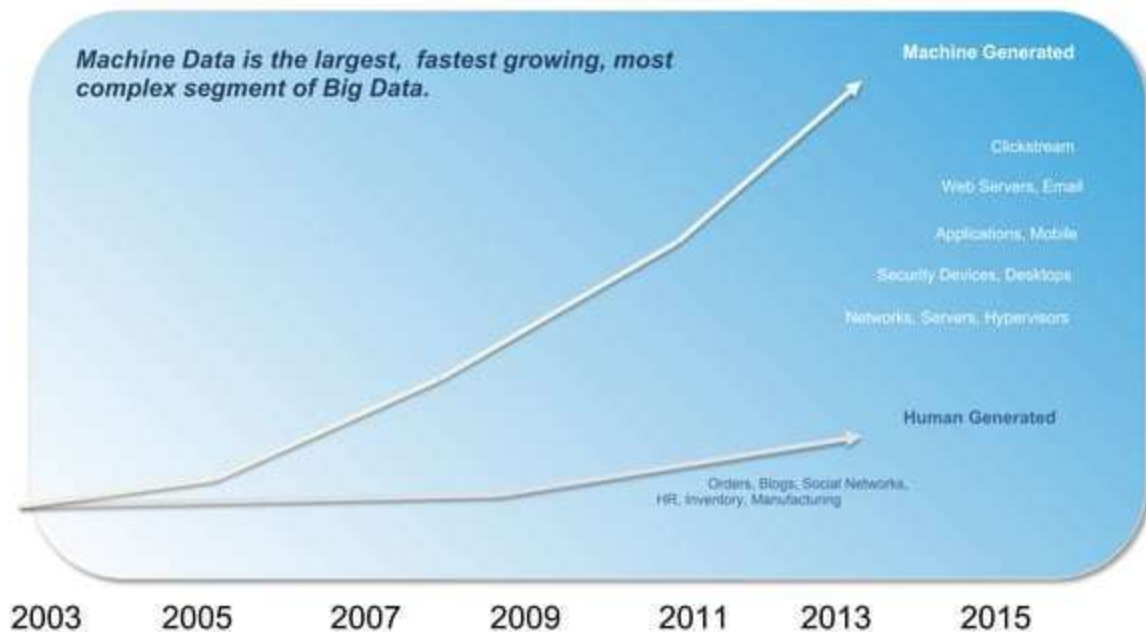
DVARVM FIGV-

REVM QVAE MOVEN
 mudi schizogenitas. Capulas
 conuulsas insinuat, ultra, que
 detestantur proprio corpore
 volens ac cerebello, et dicitur in
 figura dissecata in duos partes,
 dicitur interius, hanc omnia
 insinuat in mente, et in
 appropinquat, dicitur propter
 quod natus fuerit in dicitur
 autem per se, formam dicitur in
 non latere conuulsas, quod
 quod et dicitur in dicitur
 non quodammodo dicitur in
 natus latere hanc dicitur in
 figura hanc propter in dicitur
 pectus in dicitur in dicitur
 propter in dicitur, in dicitur
 dicitur propter, figura hanc
 dicitur, et dicitur in dicitur
 dicitur in dicitur propter
 dicitur, hanc in dicitur in
 natus hanc in dicitur propter
 hanc in dicitur in dicitur
 dicitur.

CVAKA.

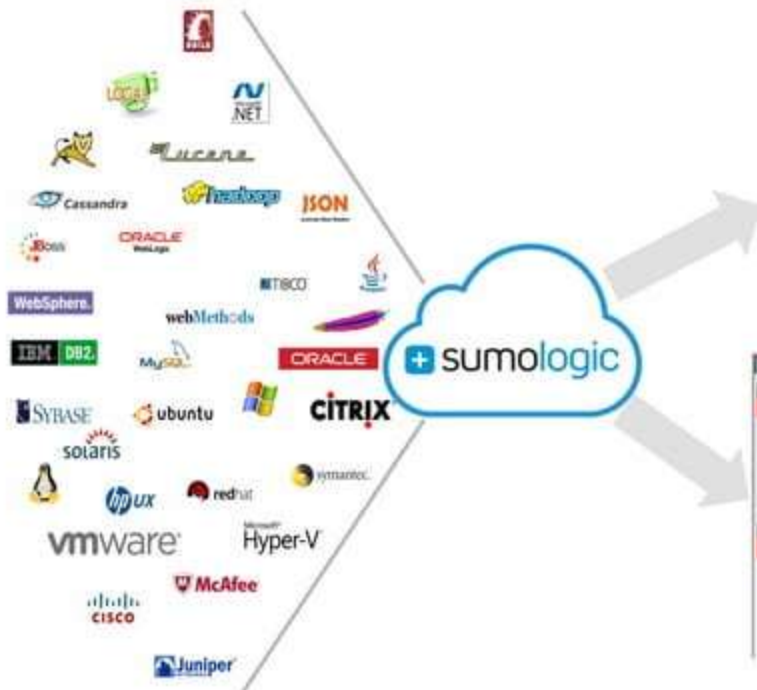
The Problem We Solve

"More Logs Are Created In A Single Day Now Than in All of FY 2003," Gartner



Sumo Logic

"Turning Machine Data Into IT and Business Insights"



Search, monitor, visualize



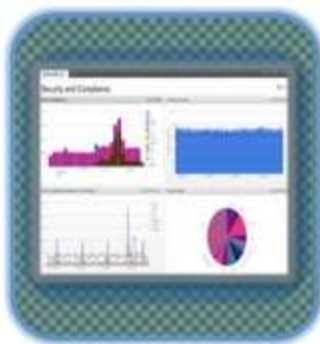
Learn, classify, predict



Use Cases



Availability &
Performance



Security and
Compliance

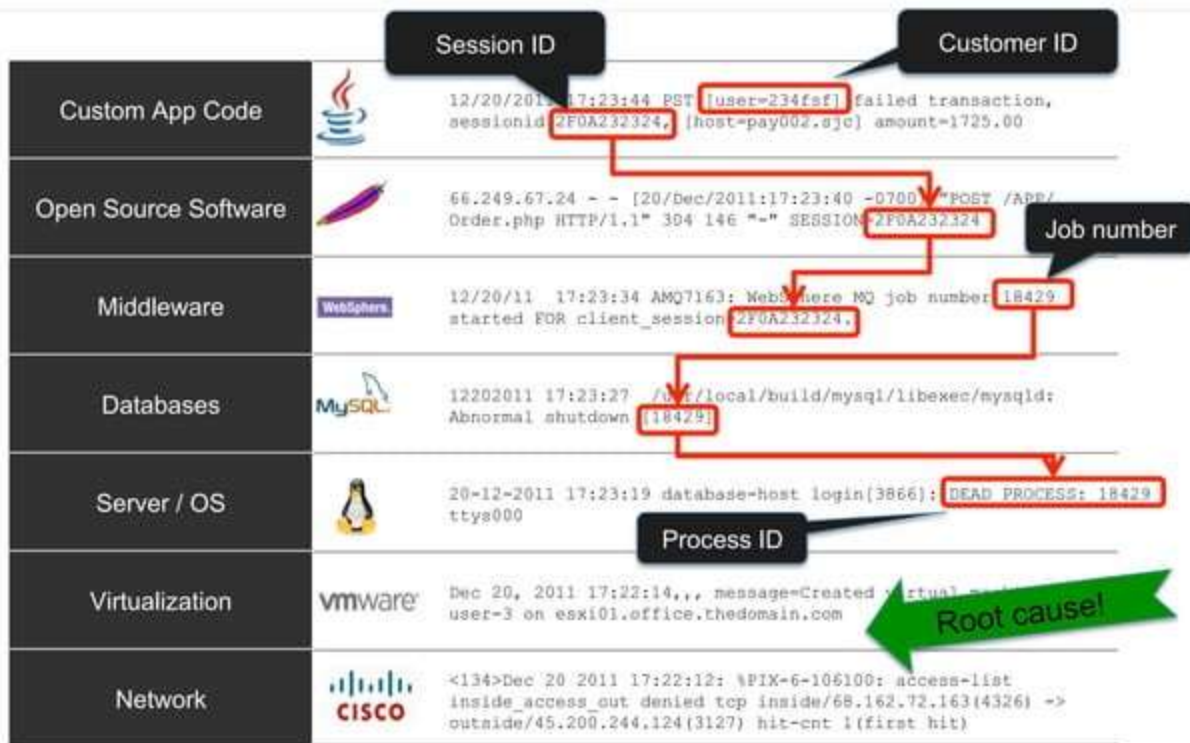


Customer
Insights

Monitoring and reporting



Troubleshooting and root cause analysis

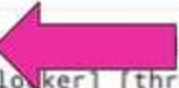


Anatomy of a log message: Five W's

```
2012-05-22 18:47:26,807 -0700 INFO [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:0000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'000000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000000483D'
```

Anatomy of a log message: Five W's

2012-05-22 18:47:26,807 -0700 I [redacted] =long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '000000000000483D'



- **When?** Timestamp with time zone

Anatomy of a log message: Five W's


```
2012-05-22 18:47:26,807 -0700 INFO [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:0000000000000000483D:00000000000000005:false]
[remote_ip=184.73.74.54] [web_session=Me...CS...] Pile for customer:
'0000000000000000005', ID: '8000000064076370', track: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000000000483D'
```




- **When?** Timestamp with time zone
- **Where?** Host, module, code location

Anatomy of a log message: Five W's

```
2012-05-22 18:47:26,807 -0700 INFO [frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePipePipeline-3]
[auth=Collector:prod-cass-raw-8:0000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'000000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000000483D'
```



- **When?** Timestamp with time zone
 - **Where?** Host, module, code location
 - **Who?** Authentication context
 - **What?** Log level and key-value pairs
- 

Inhuman scale

- Logs: like “computer tweets”
- Twitter 2013*
 - Peak @ ~144k TPS
 - Avg ~6k tweets / second
- Log data
 - Example: 1 TB / day
 - Avg ~25k logs / second



* <https://blog.twitter.com/2013/new-tweets-per-second-record-and-how>

Inhuman complexity

“A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.” - Leslie Lamport



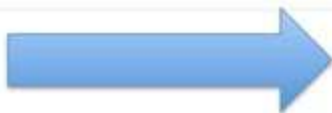
All-too-human messiness and variety

- (wildly) varying formats
 - printf, JSON, XML, Windows, X-delimited, ...
- Specialized knowledge



```
[2008-05-07 09:50:08.450 'App' 3560 verbose]  
[VpxdHeartbeat] Invalid heartbeat from  
10.17.218.46
```

Q: how to get **human-scale** insights from log data?



Q: how to get **human-scale** insights from log data?



A: machine learning (and friends)

- Unsupervised pattern discovery
- Anomaly / outlier detection
- Supervised classification
- Time-series data modeling
- Graph analysis
- Probabilistic data structures

Too many logs! "data disorientation"



~60k results: 30 minutes, one component



20.1-2846



Search

Anomalies

Dashboards

Settings

Innamed Search

Unnamed Search



7:30 PM

STATUS: Done gathering results ELAPSED TIME: 00:00:06 RESULTS: 59,063



Messages

Page: 1 of 3938

LogReduce

Time

Message

02/05/2014
19:59:54.333

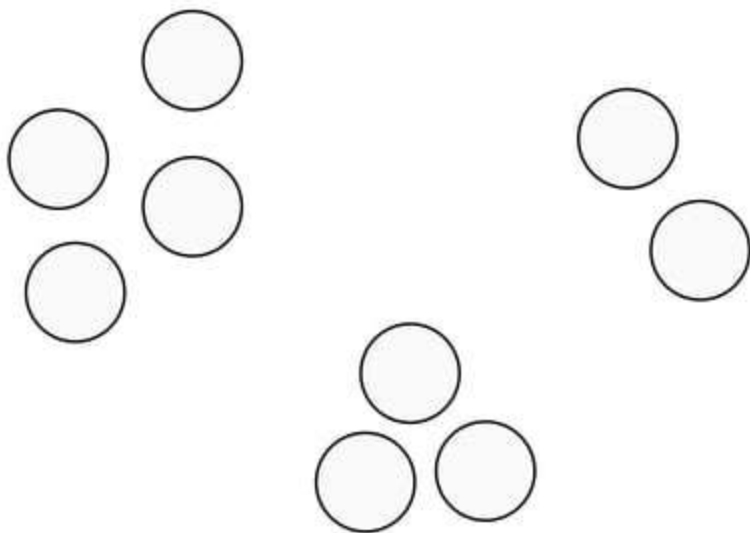
2014-02-05 19:59:54,333 -0800 INFO [hostId=nite-katta-1] [module=KATTA]
[logger=katta_sumo.node.ShardDiskCache] [thread=160184096@qtp0-14] Shard
waiting, session ID: FFFFFFFFFFFFFFFF

Host: nite-katta-1 Name: /usr/sumo/katta-sumo-20.1-1821/logs/katta.log Category: katta

Unsupervised clustering

$\hat{f}(x)$

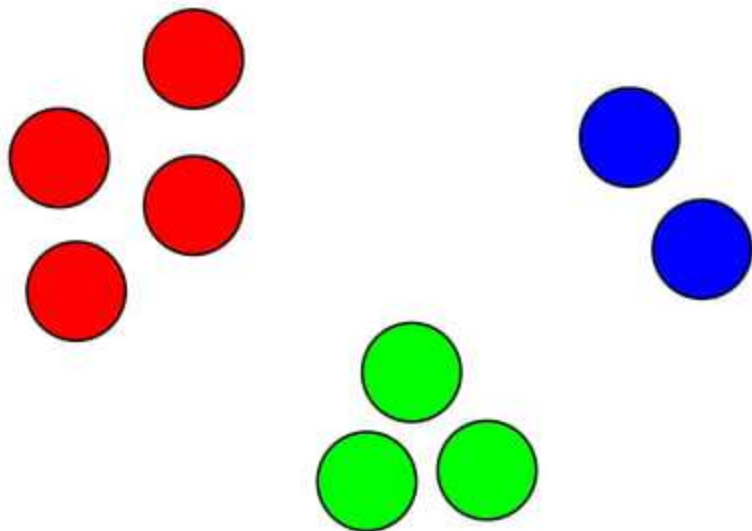
- **Given:** set of items
- **Do:** group similar items



Unsupervised clustering

$$\hat{f}(x)$$

- **Given:** set of items
- **Do:** group similar items



Distill logs down to **underlying structure**

```
$DATE INFO [hostId=stag-katta-*) [module=KATTA]  
[localUserName=katta] [logger=katta_sumo.node.FetchQueue]  
[thread=ShardDiskCache-*) Queue wait time for object:  
'*****#shard', ms: '*' with queue depth: '*',  
immediate: '****', fetch time: '*', total time: '**'
```

Results "compressed" ~1000x



 20.1.2046 Search Anomalies Ma

Unnamed Search +

7:30 PM STATUS: Done gathering results

Messages

Summarize

Page: 1 of 2

#	Select	Count ▼	Relevance	Actions	Signature
1	<input type="checkbox"/>	25,856	4.60	   	\$DATE INFO [hostId=nite-katta-*] [modu count with sessionId=*, shard count=*
2	<input type="checkbox"/>	23,824	0.49	   	\$DATE INFO [hostId=nite-katta-1] [modu with count with sessionId=*, * hits,
3	<input type="checkbox"/>	2,479	7.42	   	\$DATE INFO [hostId=nite-katta-2] [modu ***** fetched, ***** , session ID: *

In the beginning, there was the printf()

```
printf("Health status check: %s is %s",  
      hostid, hoststatus)
```



Log generation

```
Health status check: zim-5 is OK  
Health status check: gir-3 is OK  
Health status check: gir-2 is TIMED OUT  
Health status check: dib-1 is OK
```

Reverse engineering printf()

```
printf("Health status check: %s is %s",  
      hostid, hoststatus)
```



Log generation

```
Health status check: zim-5 is OK  
Health status check: gir-3 is OK  
Health status check: gir-2 is TIMED OUT  
Health status check: dib-1 is OK
```



"magic"

```
Health status check: *** is ***
```

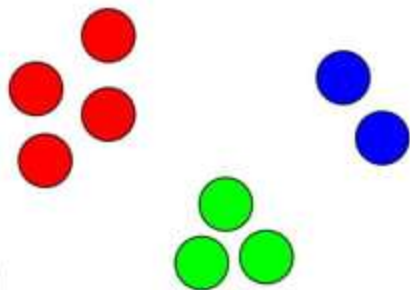
$\hat{f}(x)$ Unsupervised clustering

- **Given:** log messages
- **Do:** group by “signature”

1. Define string **distance function**

A B C ~~D~~ E
↓
A Z C E

$$d(\ell_1, \ell_2) = 2$$



2. Do **distance-based clustering**

Drill-down into the original raw logs



Messages		Summarize	
Page: 1		of 2	
#	Select	Count	Actions
1	<input type="checkbox"/>	8,497	
		Signature	
		\$DATE INFO [hostId [thread=IPC Server shards=[000000000000C	

2013-04-24 09:20:53,997 -0700 INFO [hostId=nite-katta-1] [module=KATT
[thread=IPC Server handler 8 on 20000] STARTED Calling getDetailsBatch
shards=[0000000000000131-99F822FECEBFE19C#shard], docIds.length=2500
Host: nite-katta-1 ▼ Name: /usr/sumo/katta-sumo-20.1-658/logs/katta.log ▼ Category

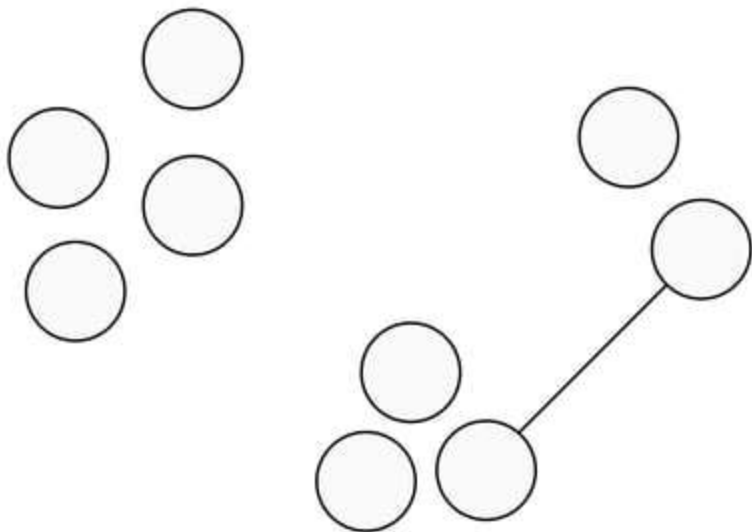
2013-04-24 09:20:53,674 -0700 INFO [hostId=nite-katta-1] [module=KATT
[thread=IPC Server handler 6 on 20000] FINISHED Calling getDetailsBatch
shards=[0000000000000131-99F822FECEBFE19C#shard], docIds.length=2497 at
Host: nite-katta-1 ▼ Name: /usr/sumo/katta-sumo-20.1-658/logs/katta.log ▼ Category

2013-04-24 09:20:53,462 -0700 INFO [hostId=nite-katta-1] [module=KATT

Partially supervised clustering

$\hat{f}(x)$

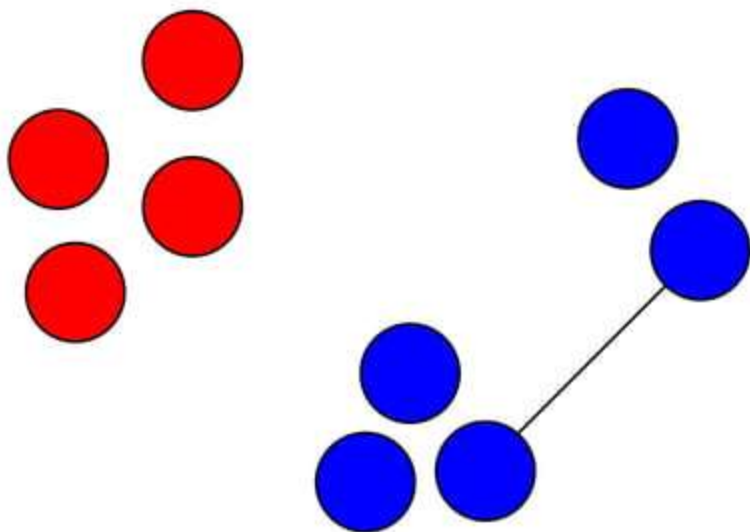
- **Given:** set of items + side info
- **Do:** group similar items



Partially supervised clustering

$\hat{f}(x)$

- **Given:** set of items + side info
- **Do:** group similar items



Too many wildcards!



```
$DATE INFO [hostId=long-katta-*] [module=KATTA  
[thread=IPC Server handler * on 20000] ***** ED  
shards=[000000000000000005-*****]
```



Not enough wildcards!



```
$DATE INFO [hostId=long-frontend-1]  
[logger=scala.config.protocol.handler  
[auth=User:scott@sumologic.com:000000  
[remote_ip=173.228.89.151] [web_sessi
```



"Hint" from human user



```
$DATE INFO [hostId=long-frontend-1]  
[logger=scala.config.protocol.handler  
[auth=User:scott@sumologic.com:000000  
[remote_ip=173.228.89.151] [web_sessi
```



```
$DATE INFO [hostId=long-frontend-1]  
[logger=scala.config.protocol.handler  
[auth=User:*****:false:DefaultSumoSy  
getDashboard(*****) after ns
```



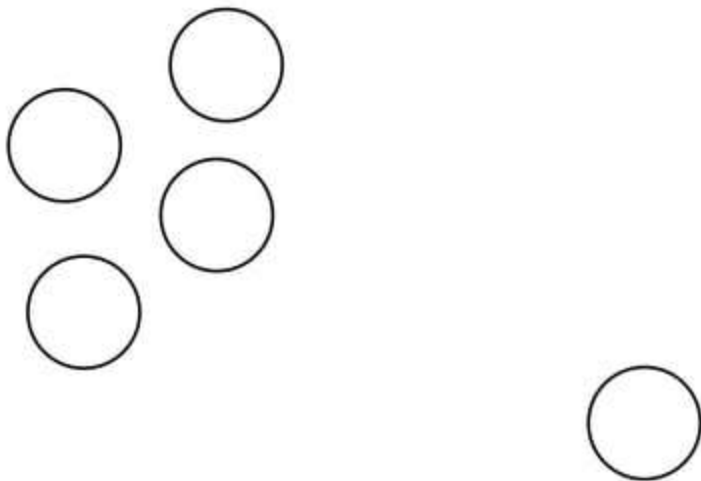
unknown unknowns



$$\hat{f}(x)$$

Outlier detection

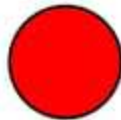
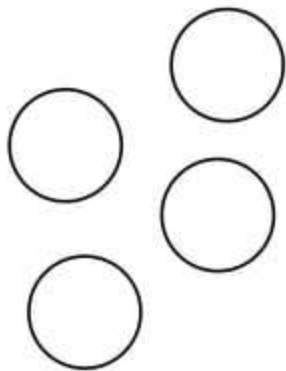
- **Given:** data points
- **Do:** identify outliers



$$\hat{f}(x)$$

Outlier detection

- **Given:** data points
- **Do:** identify outliers



$\hat{f}(x)$ Anomaly detection

- **Given:** log data
- **Do:** flag anomalies

Health check OK	$\begin{bmatrix} 33 \\ 29 \\ 3 \end{bmatrix}$	$\begin{bmatrix} 30 \\ 26 \\ 6 \end{bmatrix}$	$\begin{bmatrix} 31 \\ 27 \\ 732 \end{bmatrix}$
Request processed			
Txn timeout, retry			
	t_1	t_2	t_3

Stock Trader - Web App

0

0

0

1

$\hat{f}(x)$

Anomaly detection

- **Given:** log data
- **Do:** flag anomalies

9:00 AM

10:00 AM

11:00 AM

12:00 PM

1:00 PM

Unlabeled Event 7894

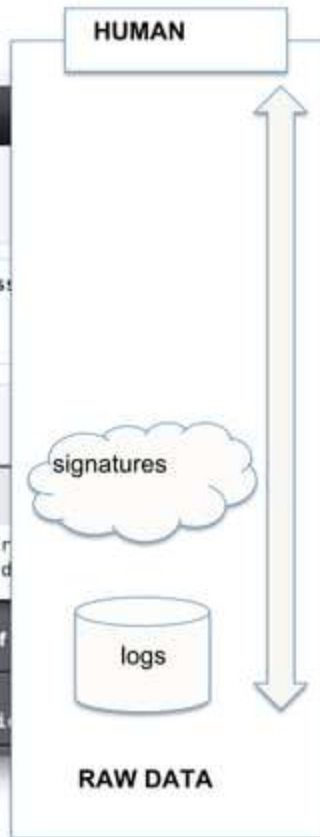
Health check OK	$\begin{bmatrix} 33 \\ 29 \\ 3 \end{bmatrix}$	t_1	,	$\begin{bmatrix} 30 \\ 26 \\ 6 \end{bmatrix}$	t_2	,	$\begin{bmatrix} 31 \\ 27 \\ 732 \end{bmatrix}$	t_3
-----------------	---	-------	---	---	-------	---	---	-------

Investigate and annotate events

The screenshot shows the Sumologic web interface. At the top is a navigation bar with the Sumologic logo, a date '19.74-02', and links for 'Search', 'Anomalies', and 'Dashboards'. Below this is a breadcrumb trail 'Stock Trader - App Dev' and the main title 'Database Timeout and User Issues'. The event details section shows the 'Event Name' as 'Database Timeout and User Issues' and the 'Description' as 'Database timeout is activity'. The 'Severity' is set to 'High' with a red flag icon. Below this is a 'Signatures' section with a table of results.

#	Score	Change	Signature
1	Blue circle	Up arrow	\$DATE StockTraderWebApplicationServiceClient.sell Er System.Exception: Database timeout creating sell ord
2	Blue circle	Down arrow	
3	Blue circle	Down arrow	

Below the table, there is a section 'Messages with this Signature' with pagination controls. The first message is dated '02/05/2014' at '13:19:59.000' and contains the text '2014-02-05 21:19:59 StockTraderWebApplicationServiceCli'.



Investigate and annotate events

The screenshot shows the Sumologic web interface. At the top, there's a navigation bar with the Sumologic logo, a date '19.74-02', and links for 'Search', 'Anomalies', and 'Dashboards'. Below this, the breadcrumb 'Stock Trader - App Dev' leads to the event title 'Database Timeout and User Issues'. The event details show 'Event Name' as 'Database Timeout and User Issues' and 'Description' as 'Database timeout is activity'. The 'Severity' is set to 'High' with a red square icon. A red starburst graphic highlights the 'Signatures' section. This section contains a table with columns '#', 'Score', and 'Change'. The first row shows a score of 1 with a blue circle and an upward arrow. To the right of this table, the 'Signature' is displayed as '\$DATE StockTraderWebApplicationServiceClient.sell Er System.Exception: Database timeout creating sell ord'. Below the signature, there's a section 'Messages with this Signature' with pagination controls (Page: 1 of 1). The first message entry shows a timestamp '02/05/2014 13:19:59.000' and the same signature text.

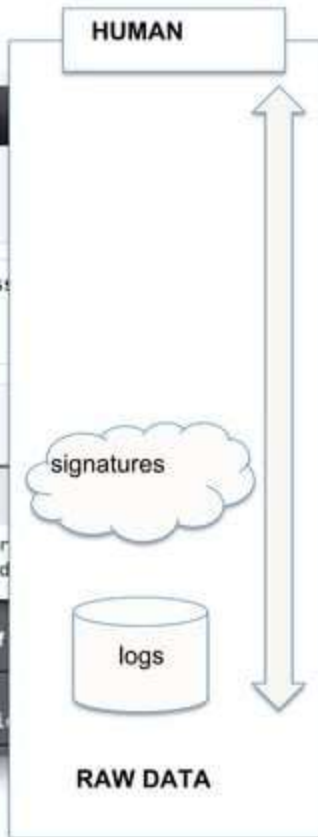
#	Score	Change
1	1	↑
2		↔
3		↓

Signature

\$DATE StockTraderWebApplicationServiceClient.sell Er System.Exception: Database timeout creating sell ord

Messages with this Signature Page: 1 of 1

1	02/05/2014 13:19:59.000	2014-02-05 21:19:59 StockTraderWebApplicationServiceCli
---	-------------------------	---



Investigate and annotate events

The screenshot shows the Sumologic interface for investigating an event. The event name is "Database Timeout and User Issues" with a severity of "High". The description is "Database timeout is activity". The event is highlighted with a red box. Below the event details, there is a section for "Signatures" with a table of signatures and a list of messages with this signature.

Event Details:

Event Name	Description
Database Timeout and User Issues	Database timeout is activity

Severity: High

Signatures:

#	Score	Change	Signature
1	Blue circle	Up arrow	\$DATE StockTraderWebApplicationServiceClient.sell Er System.Exception: Database timeout creating sell ord
2	Blue circle	Down arrow	
3	Blue circle	Down arrow	

Messages with this Signature:

	02/05/2014 13:19:59.000	2014-02-05 21:19:59 StockTraderWebApplicationServiceCli
1		

Diagram:

The diagram on the right illustrates the data flow from RAW DATA to HUMAN interaction. It shows a vertical flow from RAW DATA (logs) through signatures and events to HUMAN interaction. The event and signatures are highlighted with red boxes.

Investigate and annotate events

The screenshot shows the Sumologic web interface. At the top, there's a navigation bar with the Sumologic logo, a date '19.74-82', and links for 'Search', 'Anomalies', and 'Dashboards'. Below this, the page title is 'Stock Trader - App Dev' followed by 'Database Timeout and User Issues'. The main content area has a form for 'Event Name' (filled with 'Database Timeout and User Issues') and 'Severity' (set to 'High' with a red icon). To the right, a 'Description' field contains 'Database timeout iss activity'. Below the form, there's a 'Signatures' section with a table. The table has columns for '#', 'Score', 'Change', and 'Signature'. It lists three signatures, with the first one expanded to show 'Messages with this Signature'. The messages table shows a single entry with a timestamp '02/05/2014 13:19:59.000' and a message '2014-02-05 21:19:59 StockTraderWebApplicationServiceCli'.

sumologic 19.74-82 Search Anomalies Dashboards

Stock Trader - App Dev
Database Timeout and User Issues

Event Name Database Timeout and User Issues Description Database timeout iss activity

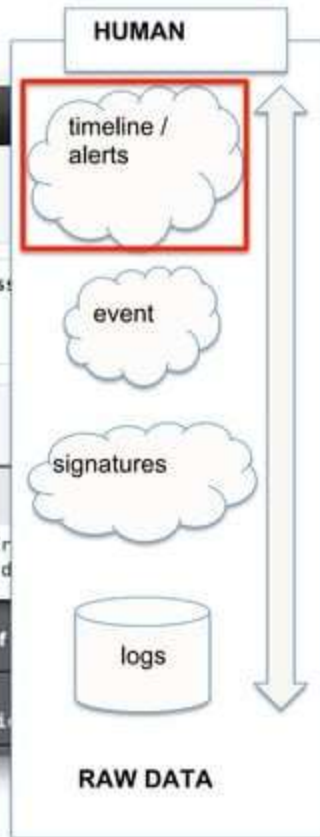
Severity High

Signatures

#	Score	Change	Signature
1			\$DATE StockTraderWebApplicationServiceClient.sell Er System.Exception: Database timeout creating sell ord
2			
3			

Messages with this Signature Page: 1 of

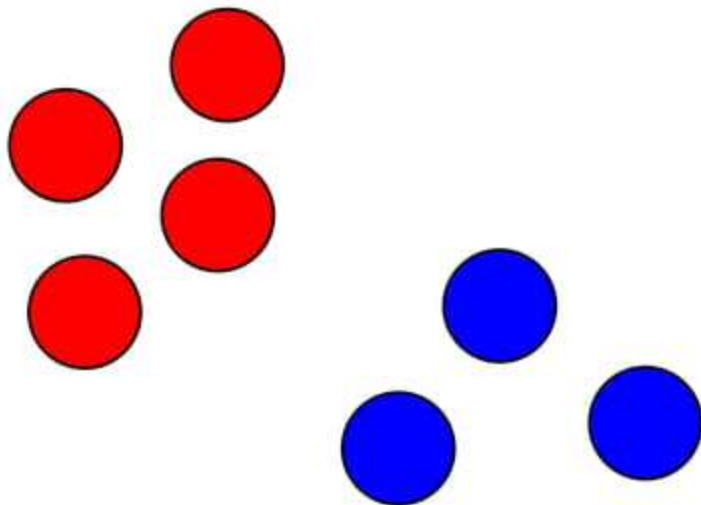
	02/05/2014 13:19:59.000	2014-02-05 21:19:59	StockTraderWebApplicationServiceCli
1			



$$\hat{f}(x)$$

Supervised classification

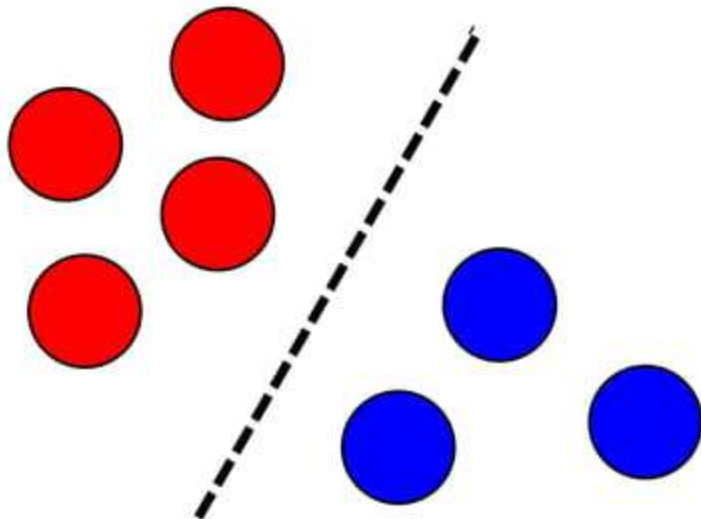
- **Given:** labeled data points
- **Do:** predict future labels



$$\hat{f}(x)$$

Supervised classification

- **Given:** labeled data points
- **Do:** predict future labels



Supervised classification

- $\hat{f}(x)$
- **Given:** log data, annotated events
 - **Do:** classify new occurrences



Database Timeouts



Event Name	Database Timeouts
Time Range	10:25 AM - 10:30 AM
Description	Database Timeouts



User action webID=7F92

Connected components

- $\hat{f}(x)$
- **Given:** nodes/edges
 - **Do:** identify component

PROCESSING FAILED: webID=79F92



webID
7F92



User action webID=7F92

webID
7F92



User action webID=7F92

Initiating requestID=082A for webID=7F92 ...



User action webID=7F92

Initiating requestID=082A for webID=7F92 ...

... orderID=34C8 received for requestID=082A ...



User action webID=7F92

Initiating requestID=082A for webID=7F92 ...

... orderID=34C8 received for requestID=082A ...

Retrieving userID=11D2 for requestID=082A ...



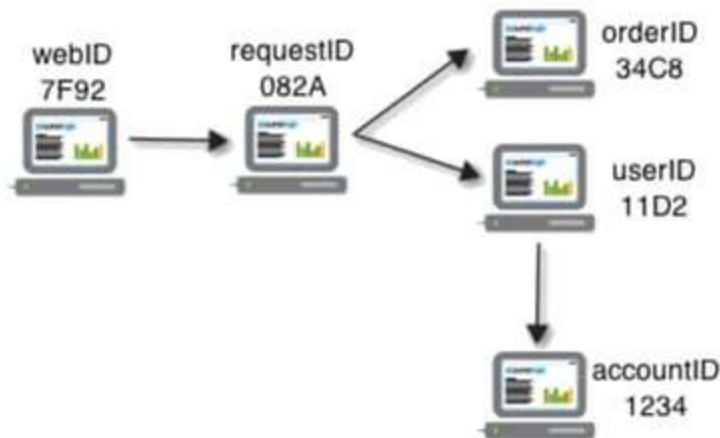
User action webID=7F92

Initiating requestID=082A for webID=7F92 ...

... orderID=34C8 received for requestID=082A ...

Retrieving userID=11D2 for requestID=082A ...

... accountID=1234 access, userID=11D2 ...



User action webID=7F92

Initiating requestID=082A for webID=7F92 ...

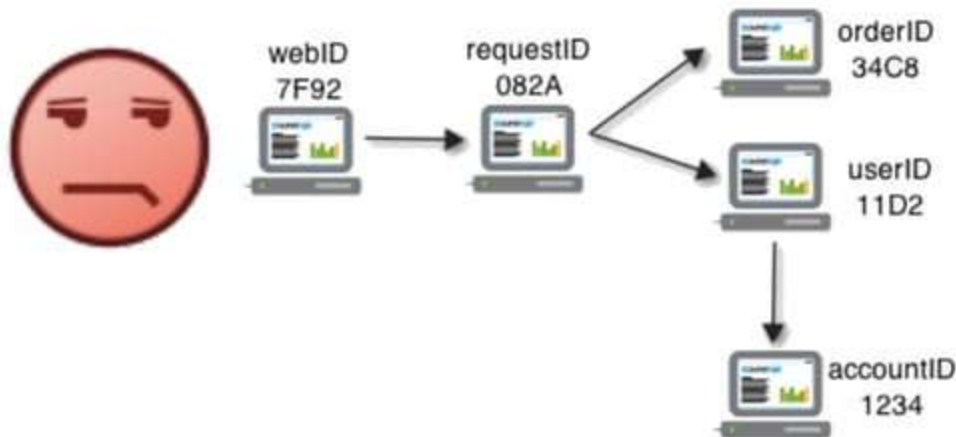
... orderID=34C8 received for requestID=082A ...

Retrieving userID=11D2 for requestID=082A ...

... accountID=1234 access, userID=11D2 ...

ERROR accountID=1234 not found!

PROCESSING FAILED: webID=79F92



Time-series detection

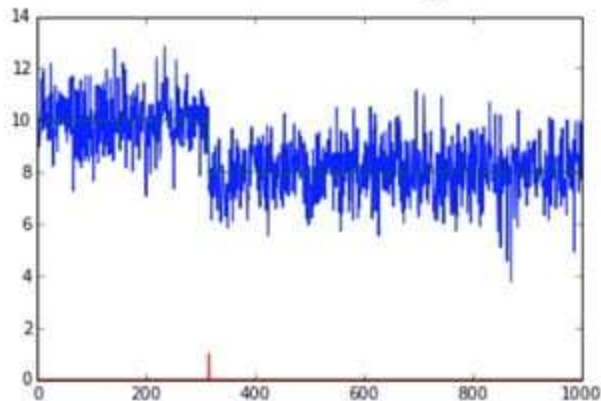
$\hat{f}(x)$

- **Given:** time-series metric data
- **Do:** identify unusual data pts

Time-series detection

- $\hat{f}(x)$
- **Given:** time-series metric data
 - **Do:** identify unusual data pts

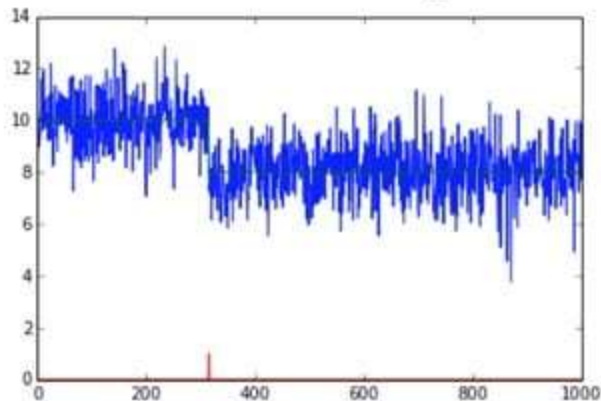
Level change



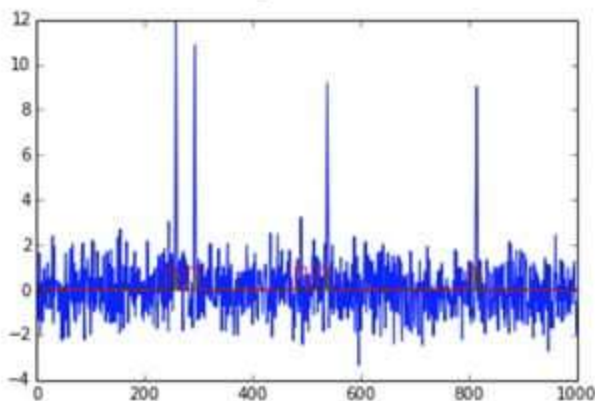
Time-series detection

- $\hat{f}(x)$
- **Given:** time-series metric data
 - **Do:** identify unusual data pts

Level change

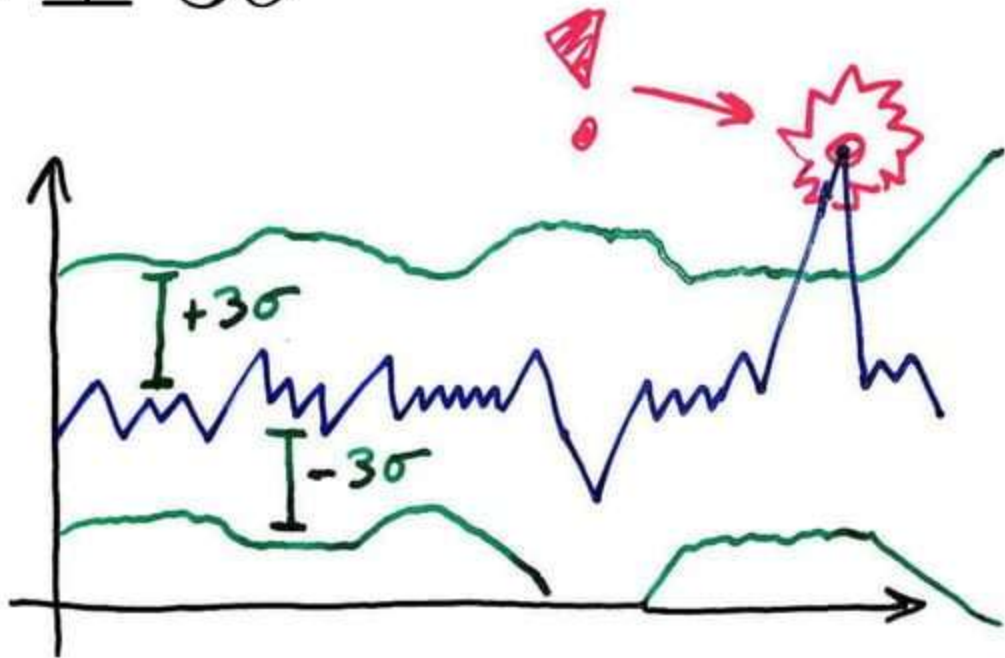


Spikes



“Bollinger bands” – rolling window approach

$$\mu \pm 3\sigma$$



$\hat{f}(x)$

Top-K identification

- **Given:** stream of observations
- **Do:** identify k most frequent (WITH FIXED MEMORY!)

Top-K identification

$\hat{f}(x)$

- **Given:** stream of observations
- **Do:** identify k most frequent (WITH FIXED MEMORY!)



Top-K identification

 $\hat{f}(x)$

- **Given:** stream of observations
- **Do:** identify k most frequent (WITH FIXED MEMORY!)



	4
	3
	2
	2

Top-K identification

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify k most frequent (WITH FIXED MEMORY!)



	4
	3
	2
	2

TOP 2

Top-K identification

$\hat{f}(x)$

- **Given:** stream of observations
- **Do:** identify k most frequent (WITH FIXED MEMORY!)

Count-Min Sketch

(Cormode & Muthukrishnan, 2003)

$$\hat{c} - c \geq 0$$

Top-K identification

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify k most frequent (WITH FIXED MEMORY!)

Count-Min Sketch

(Cormode & Muthukrishnan, 2003)

$$\hat{c} - c \geq 0$$

$$\hat{c} - c \leq \epsilon N$$

$$\text{w.p.} \geq 1 - \delta$$

Top-K identification

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify k most frequent (WITH FIXED MEMORY!)

Count-Min Sketch

(Cormode & Muthukrishnan, 2003)

$$\hat{c} - c \geq 0$$

$$\hat{c} - c \leq \boxed{\epsilon} N$$

$$\text{w.p.} \geq 1 - \boxed{\delta}$$

Cardinality estimation

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify number of distinct items (WITH FIXED MEMORY!)

Cardinality estimation

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify number of distinct items (WITH FIXED MEMORY!)



Cardinality estimation

- $\hat{f}(x)$
- **Given:** stream of observations
 - **Do:** identify number of distinct items (WITH FIXED MEMORY!)



$$|\{\text{coffee cup}, \text{skull}, \text{monkey}, \text{cocktail glass}\}| = 4$$

Cardinality estimation

$\hat{f}(x)$

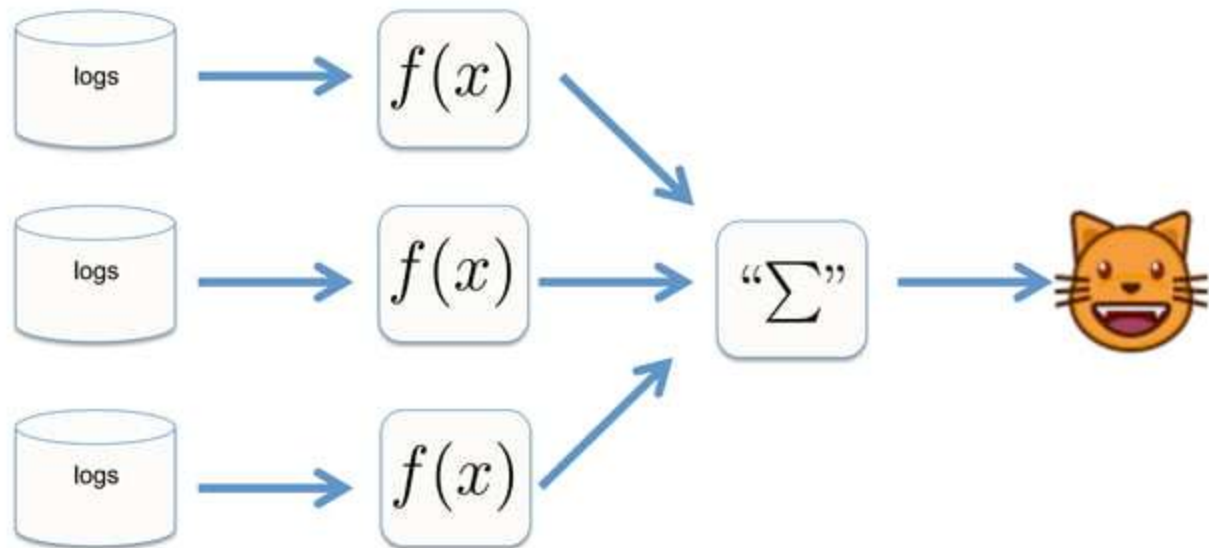
- **Given:** stream of observations
- **Do:** identify number of distinct items (WITH FIXED MEMORY!)

HyperLogLog

(Flajolet et al, 2007)

Hooray! Monoid homomorphism!

$$f(s_1 + s_2) = f(s_1) \oplus f(s_2)$$



< FINAL OBLIGATORY PLUG >

freesumo.com

Get Sumo Logic Free

Get a fully functioning version of our enterprise cloud-based log management and analytics service **FREE**



Sumo Logic Free delivers real-time troubleshooting, proactive application management and powerful IT and business insights. Our free version allows for up to three users and 500 MB per day with seven days of data retention.

SUMO LOGIC FREE

First name *

Last name *

Email *

Login credentials will be sent to this address

Company *

Phone *

[Sign Up Now](#)

By clicking on the "Sign Up Now" button, you agree to accept our [Terms of Service](#)