

The Big Data Paradigm Shift

Insight Through Automation



emcien™

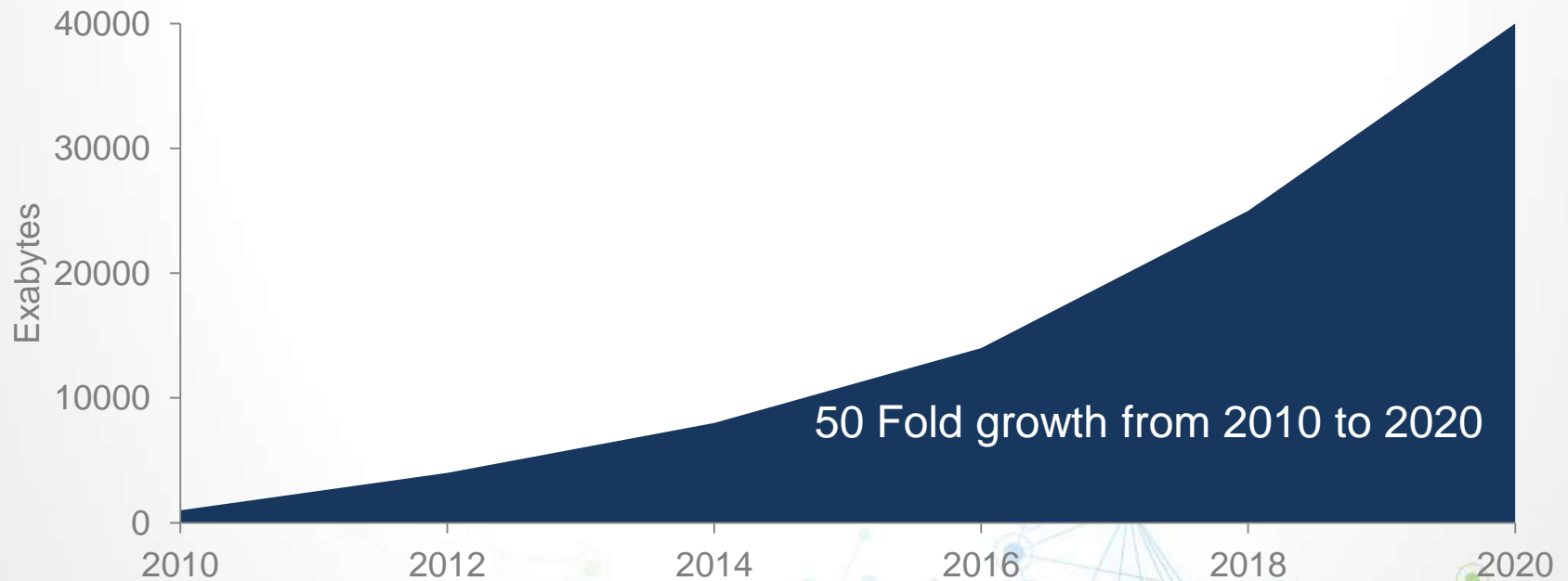


Agenda

- The Problem
- Emcien's Solution:
 - Algorithms solve data related business problems
- How Does the Technology Work?
- Case Studies

The Problem

- Data is growing at an unprecedented rate
- **Less than 1%** of data is analyzed



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Old Paradigm: Manually Intensive Analysis



Unpredictable

Slow

Expensive

Collect

Analyze

Report

New Paradigm: Automation of Analysis



PREDICTABLE

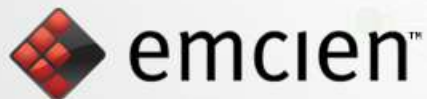
FAST

ECONOMICAL

Collect

Solve

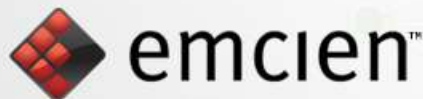
Review & Act



Emcien's Unique Value Proposition

Emcien's **automatic pattern-detection platform** delivers timely mission critical insights from data

- Automated analysis for fast, predictable, accurate insight
- Applicable across all data types:
 - Structured & Unstructured data, Text or Numeric
- Algorithms designed to solve business problems



Types of Data: Structured, Unstructured, Static, Streaming...



**Machine Data
Network log
files**

Social, Blogs, Newsfeeds



Email Data



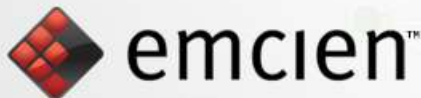
Click stream data



**Marketing
Data**

**Sales
Data**

**Corporate
Data**



Limitations of Current Solutions

Manually Intensive

- Very slow and unreliable
- Search or query based
- Visualization as a means for discovery → High error

Only certain data types

- Numerical analysis only
- Text only, NLP methods, very high set up cost

Data staging

- Streaming data and recent analysis
- At-rest data and historic analysis

Lack of Scalability

Current approaches focus too much on storage methods

Another View of the Big Data Stack

Our Focus

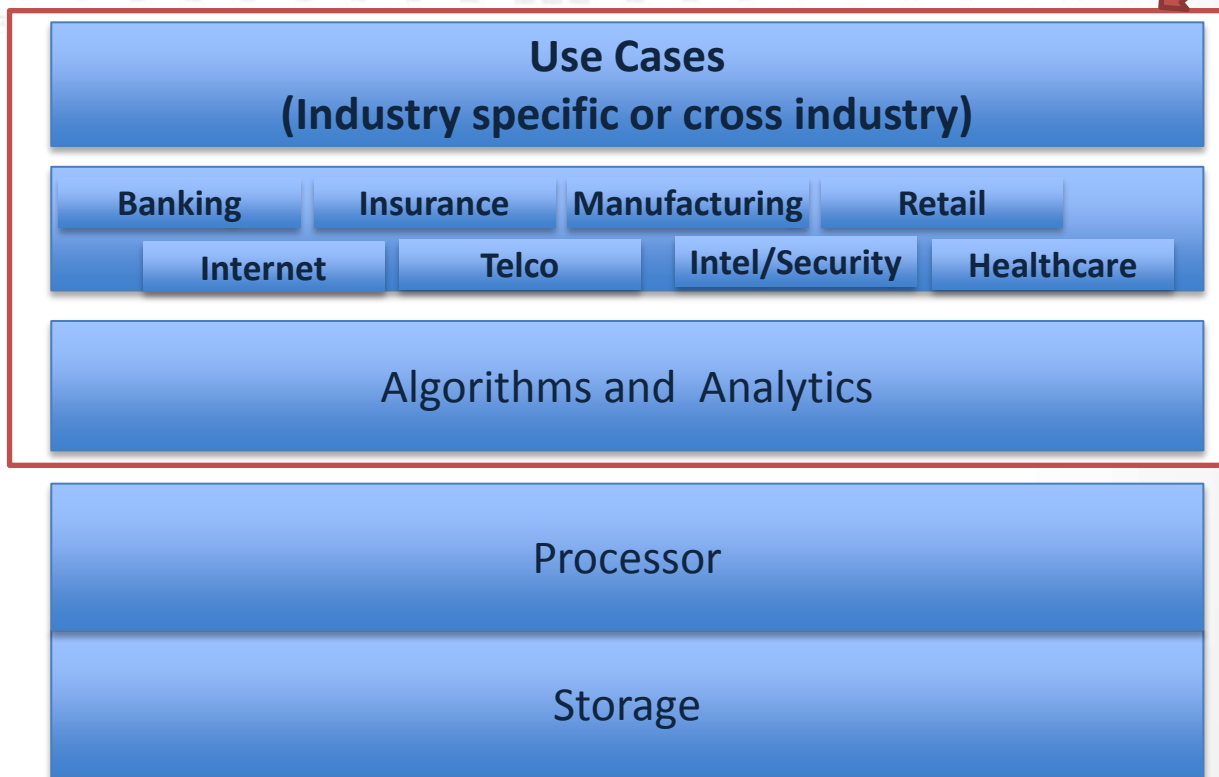
Value Layer

Sectors

Analysis Layer

Infrastructure

Data



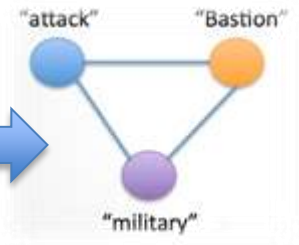
How Does Our Solution Work?

- Big Data problems need graph analysis
- Framework for analyzing relationships
- Highly scalable representation

Data values → Tokens → Graph
Tokens are linked if they occur together

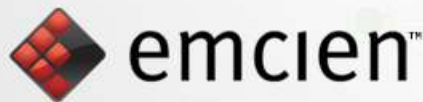
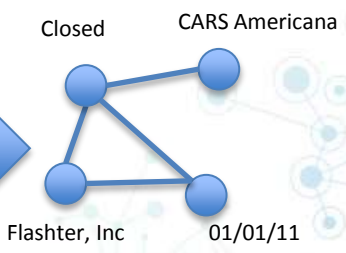
Unstructured

detonated near one of the gates of Jalalabad Airfield and several foreign troops wounded, but said none to a September attack on Camp Bastion the coalition soldiers and two civilians were killed, along with military uniforms and got inside the base, where one of the Bastion attack. In June, a truck bombing at today, military officials reported. Aghah district, officials said. At the time of his wedding and TEDs to Hassan Flahter



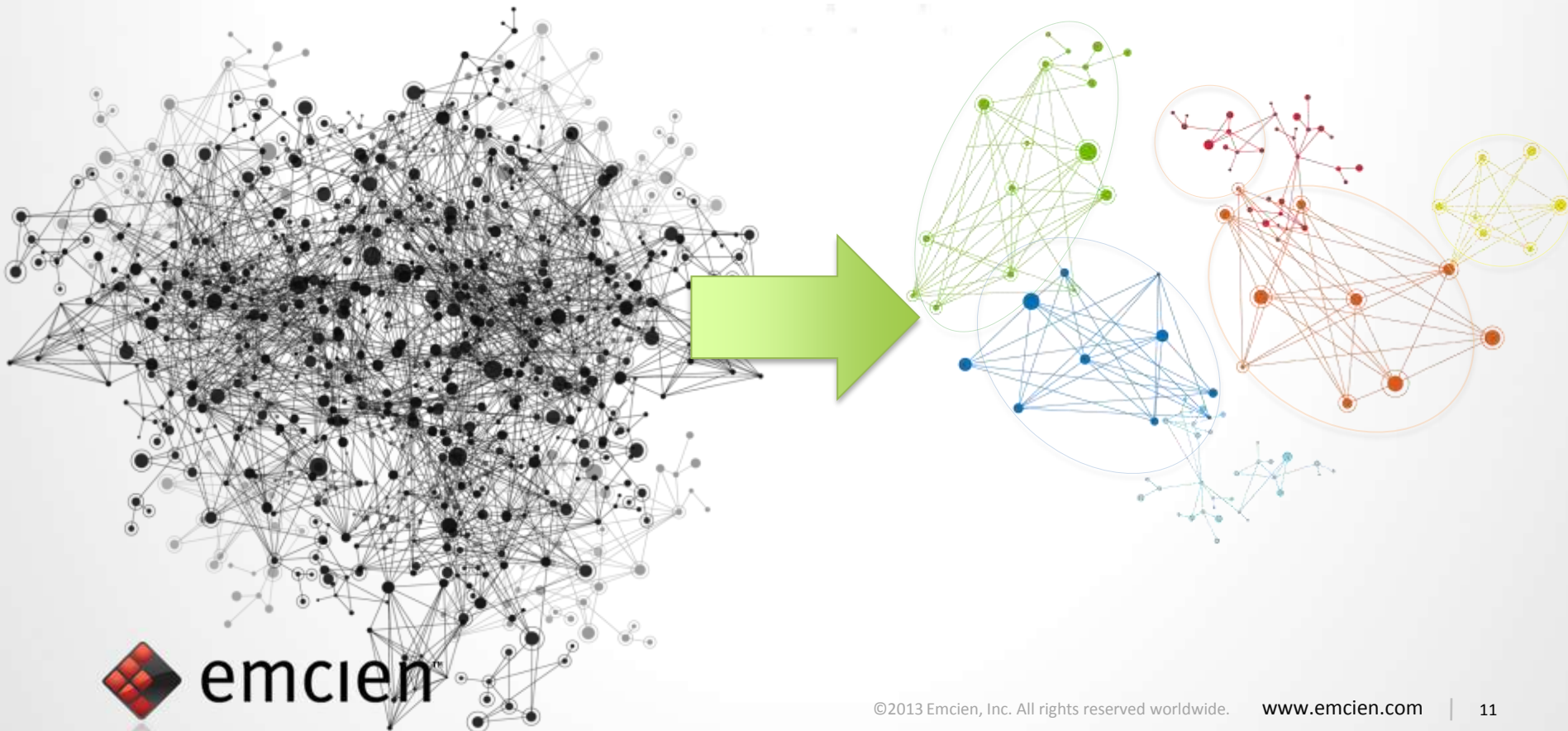
Structured

INV-18	Closed	Flashter Inc.	08/01/11
INV-19	Closed	CARS Americana	06/16/11
INV-20	Closed	Flashter Inc.	01/01/11
INV-20	Closed	Flashter Inc.	01/01/11
INV-21	Closed	Jose Angel Baria	08/16/11
INV-21	Closed	Jose Angel Baria	08/16/11
INV-21	Closed	Jose Angel Baria	08/16/11
INV-21	Closed	Jose Angel Baria	08/16/11



Algorithms Solve to Extract Patterns

- Algorithms surface the highly relevant dependencies
 - Defocus the redundant/noise to surface the signal



Data Patterns That Reveal “The Insight”

Algorithms designed to reveal graph constructs that solve a business problem

Solving a Graph Problem



Results in Solving a Business Problem

Loosely Federated Communities

- Reveals groups that behave similarly
- Reveals dimensions that bind the group
- Impossible to detect in a typical querying system

Cliques

- Highly correlated elements
- Optimal query that would lead to insight

People Network

- Reveals influence network of individual
- Highly predictive for adoption behaviors

Substitute Nodes

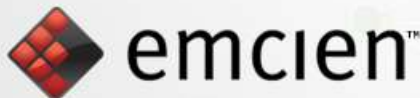
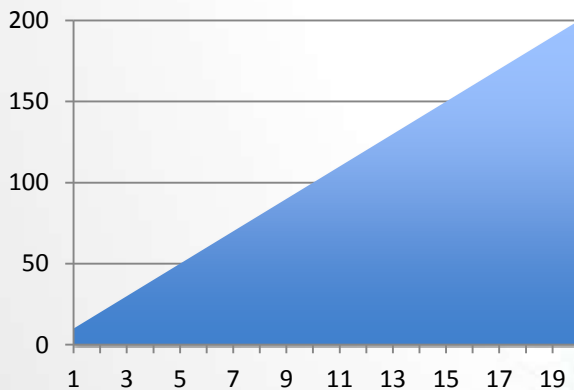
- Nodes that behave very similarly
- ID theft or product substitutes

Algorithms Are Highly Scalable For Big Data



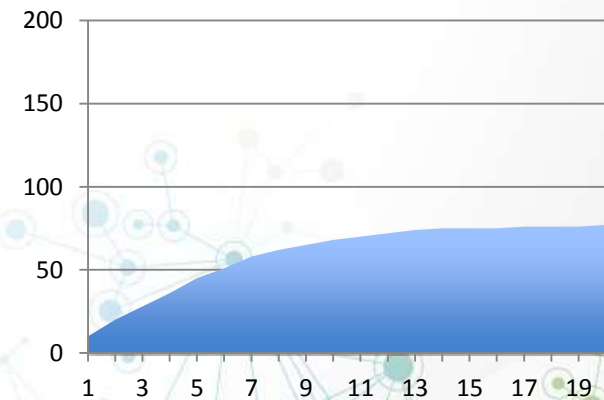
Traditional Data Storage:

- Linear growth with transactions
- Very large storage requirements are
- Increases response time



Graph Data Storage:

- Size of total number of entities
- E.g. Store has 500,000 items → graph has 500,000 nodes
- Weights updated with transactions
- Delivers a global view of the data



Speed of Data → Answers

Access Time vs. Processing Speed



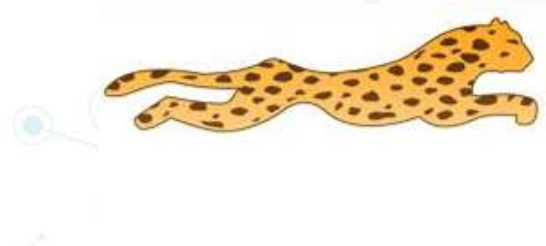
Traditional Data Storage:

- Limit is query speed
- In-memory, hadoop cluster approaches
- Highly dimensional data is a problem
- Unstructured content is a problem



Graph Data Storage:

- All results are Pre-computed (like Google)
- Pre-computing speed: 50K trans/sec compute on 1-core 8GB RAM system
- Speed of response is “access time”



The Business Problems We Solve Across Different Types of Data

Automatically Extract Dependencies

- Web click-stream - Reveal click patterns & market segments
- Sales data - Reveal consumption patterns and propensity
- Clinical trials - extract hypothesis for testing

Social Patterns & People Network

- Marketing - Reveal conversation patterns, people communities
- For Intel - Reveal bad actors based on conversation patterns

Surprising Streaming Content

- Machine Network traffic – Reveal network intrusion
- Sales transactions – Reveal fraud based on unusual patterns

Entity Resolution / Cleansing

- Patterns automatically clusters similar entities.
- Example – credit card transactions, insurance claims with varying merchant names

Intel Case Study (1/4)

Cyber Threat Monitoring with Open Source Data

Customer Overview And Current Situation

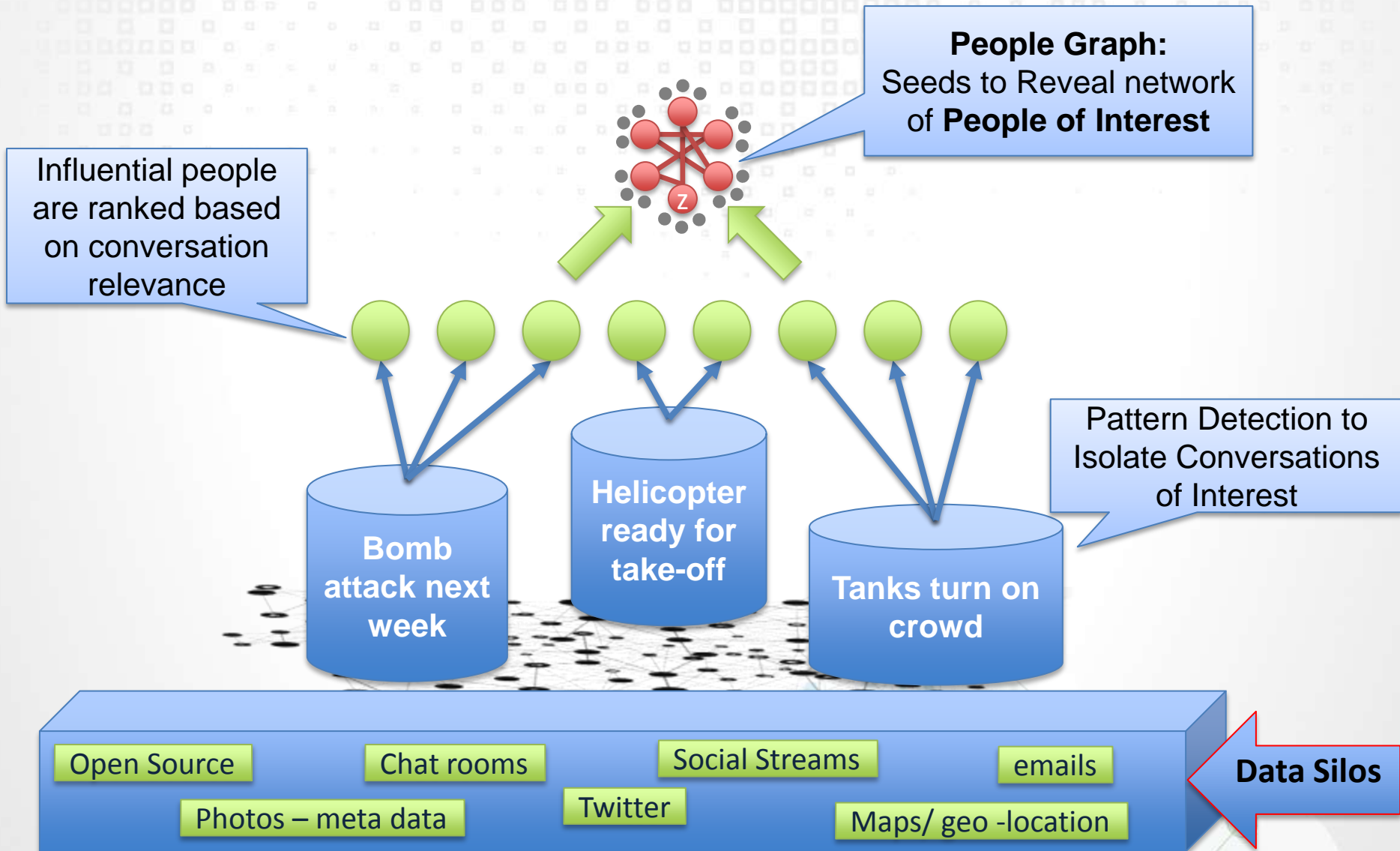
- Federal agency is failing to keep up with the activity and data in open source
- Open Source (social, IRC, blogs, etc.) are a key source of communication for underworld
- Link analysis leads to **people of interest network** – which is key for intelligence

Customer Objective

- Federal agency requires fast methods to process high volume open source data
- Need automated methods to highlight **conversations of interest**
- Need automated link analysis to focus on **people of interest**
- Fast and continuous data processing to **keep up with the speed of crime**

Intel Case Study (2/4)

Layers of Analysis for Cyber Threat Monitoring



Intel Case Study (3/4) Cyber Threat Monitoring with Open Source Data

Influential people ranked based on conversations

The screenshot shows the EmcienScout interface. On the left, a table lists 'Possible Extremist' accounts. On the right, a 'Network Graphs' section shows a list of generated graphs with their respective dates and times.

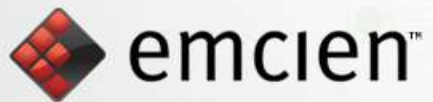
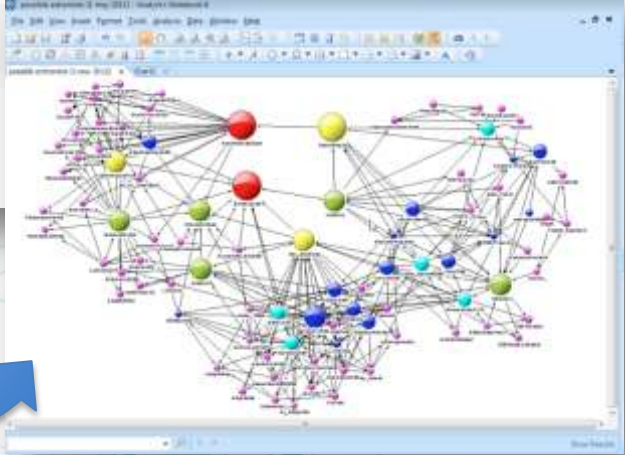
Rank	Avatar	Account Name	Followers	Friends	Twitter ID	Added	Featured	Refresh	Remove
1		BakkahNet	1,173 Followers	53 Friends	367839403	5/2 11:44	5/2 11:44	Refresh	Remove
2		JamalAlrayes	1,188 Followers	22 Friends	189105388	5/2 11:44	5/2 11:44	Refresh	Remove
3		ibn_alkattab	2,897 Followers	411 Friends	258882796	4/30 15:21	4/30 15:21	Refresh	Remove
4		as_ansar	372 Followers	2 Friends	553058094	4/30 15:20	4/30 15:21	Refresh	Remove
5		Al_nukhba	1,358 Followers	8 Friends	274981316	4/30 15:20	4/30 15:20	Refresh	Remove
6		Alvizier	609 Followers	7 Friends	128888358	4/30 15:20	4/30 15:20	Refresh	Remove
7		Jihadalummah	2,289 Followers	5 Friends	242352188	4/30 15:10	4/30 15:10	Refresh	Remove

Network Graphs Complete New Graph

- Graph @ May 2 11:45am: About 9 days ago (Wed, 2 May 11:45 am)
- Graph @ May 1 10:38pm: About 9 days ago (Tue, 1 May 10:38 pm)
- Graph @ Apr 30 3:22pm: About 9 days ago (Mon, 30 Apr 3:22 pm)
- Graph @ Apr 30 3:15pm: About 9 days ago (Mon, 30 Apr 3:15 pm)
- Graph @ Apr 30 3:07pm: About 9 days ago (Mon, 30 Apr 3:07 pm)
- Graph @ Apr 30 2:58pm: About 9 days ago (Mon, 30 Apr 2:58 pm)

Overview
250,000 Accounts
Over 1,000,000 Connections

Highly relevant” People of Interest” network



Intel Case Study (4/4)

Cyber Threat Monitoring with Open Source Data

The silent signal – Automatically detecting a sleeper cell



The screenshot shows a Twitter profile for 'الجهاد Cihad' (@ruhulcihad). The profile has 96 followers. The 'Followers' list includes accounts such as @thory2012, @apo_marwan, @albo5o5o, @alassaf69, @salafeymojahid, and @amineghabri. A red box highlights the following statistics:

- Overview
- 250,000 Accounts Analyzed
- Over 1,000,000 Connections
- 1 Account of Interest

Network Traffic Log Files (1/6)

Revealing Patterns In Machine-to-Machine Data

Customer Overview

- Research Institute has thousands of users on their network
- Must provide controlled safe access for the internal working labs and the outside network
- Control illegal intrusions, malicious malware and illegal data transmissions

Customer Objective – Automate Process of Intrusion Detection

- Scan streaming machine-to-machine log file output
- Detect surprising/interesting anomalies/beacons
- Automatically send short list of top ranked “questions” to ask of the data into existing tools (such as CA, Sumologic, Splunk, etc.)



Network Traffic Log Files (2/6)

Example Use Cases

Example Use Cases

1. Summarize and Rank Log File data based on “Surprising flow patterns”
2. Determine Machine network based on flow patterns.
 - Rank Machines based on their “influence” in the network
3. Detect “communities of machines” based on how they “talk to each other”

Network Traffic Log Files (3/6)

Reveal Surprising Patterns In Network flow Data

Network Log Data



EmcienScout Home

Acme Network Traffic Clusters

0 seconds captured about 26 hours ago
Mon, 3 Dec 2012 9:45 am - 9:45 am

Sorted by relevance

- 1121 Messages**
11/8/11 Src 10.2.197.241 Start Hour:13 Sig Id:2007757
Signature Name Et Scan W3af User Agent Priority:2
Classification Attempted Information Leak Dst 154.241.88.201 Dst 80
- 1441 Messages**
11/10/11 Src 10.2.198.245 Start Hour:9 Sig Id:2003099
Signature Name Et Web Misc Poison Null Byte Priority:2
Classification Access To A Potentially Vulnerable Web Application
Dst 154.241.88.201 Dst 80
- 6029 Messages**
11/10/11 Sig Id:2002677
Signature Name Et Scan Nikto Web App Scan In Progress Priority:1
Classification Web Application Attack Dst 154.241.88.201 Dst 80
- 4950 Messages**
11/10/11 Src 10.2.197.245 Start Hour:9 Priority:1
Classification Web Application Attack Dst 154.241.88.201 Dst 80

Overview

239,320 Messages to **42** Clusters
100.0% Compression

This result has 239,320 messages (0 seconds) arranged into 42 Clusters.

10 Top Words

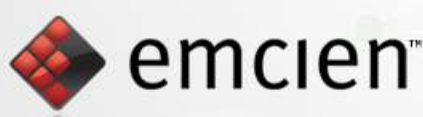
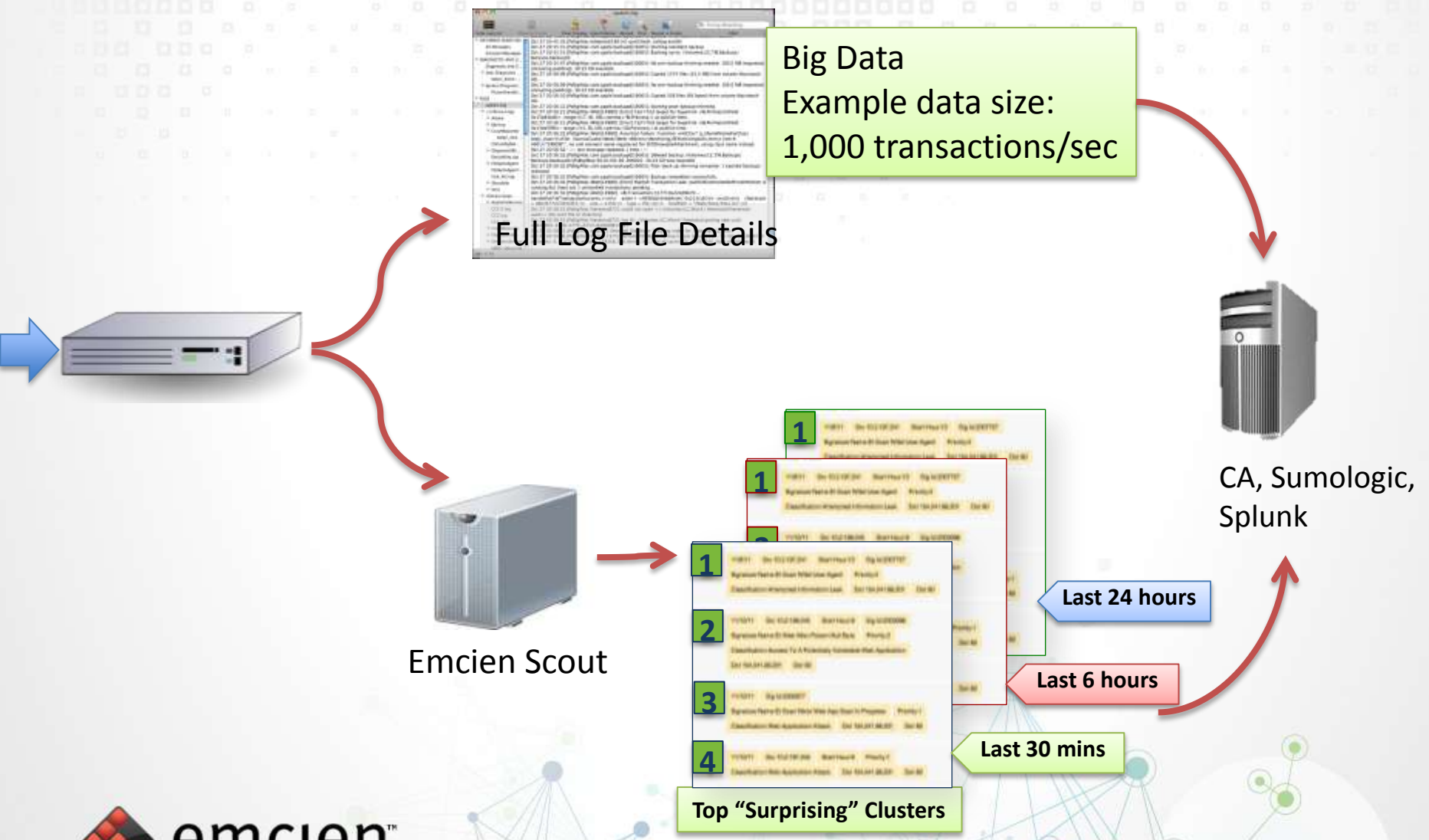


Surprising network activity within the data flow



Network Traffic Log Files (4/6)

Ranked Summary of "Surprising" Events



Network Traffic Log Files (5/6)

Most "Influential" Nodes on network

Audience size for each machine



Emcien Scout

EmcienScout Home jtabbs_mon | Options

Network Traffic

10 Minutes Captured
Mar 10, 2:20 - 2:30 PM

#	Score	Account	Mgs	Conver.	Audience Size	Conver. Started	Followers*
1	100 Score	10.80.13.43 No biography for this account	336	31	23	28	-
2	32 Score	10.15.40.139 No biography for this account	376	14	12	21	-
3	28 Score	10.43.4.19 No biography for this account	273	19	11	15	-
4	24 Score	10.80.13.143 No biography for this account	254	24	8	15	-
5	21 Score	10.104.21.165 No biography for this account	71	15	15	4	-
6	18 Score	10.80.14.185 No biography for this account			10	11	-
7	13 Score	10.90.22.70 No biography for this account					
8	13 Score	10.40.50.123 No biography for this account					

Overview
618Accounts
100Most Influential

Definitions

Score
Unique ranking of messaging users in the current search results. Rank is based on words and concepts by an account matched with others in the same search results. Contact top ranked users if you want to reach large groups of accounts talking about the same topics.

Messages
Number of Messages sent by this account.

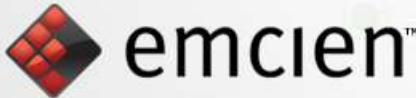
Conversations
Number of different conversations this account's messages appeared in.

Audience Size
Number of other accounts who are talking about the same things as this account.

Conversations Started
How many times this account's messages started a conversation.

Followers*
Number of followers this account has. (Informational only, NOT used in Score ranking)

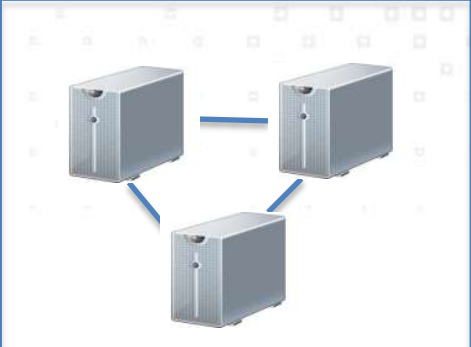
Most influential machines on the network based on communication patterns



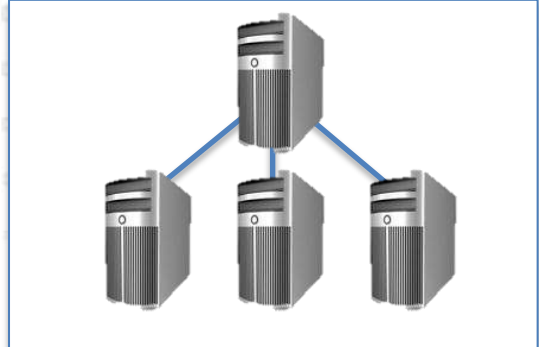
Network Traffic Log Files (6/6)

Machine Communities based on “how they talk”

Lab A

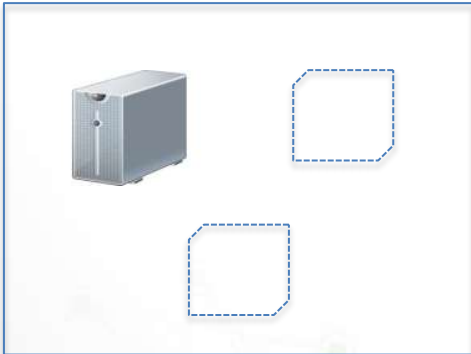


Lab B

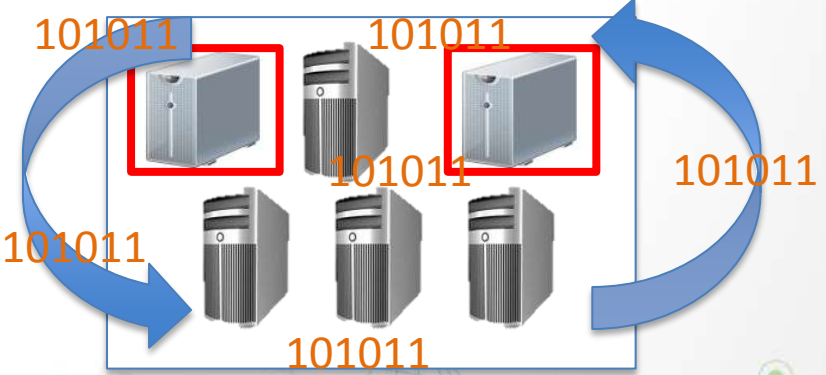


Physical Connections

Lab A



Lab B



Machine Communication Communities

Intel Case Study (1/6)

Reveal Conversation Patterns & Network of Actors in Email Data

Customer Overview And Current Situation

- Federal agency is failing to keep up with the activity and data in email
- Too much data and current tools are manually intensive

Customer Objective

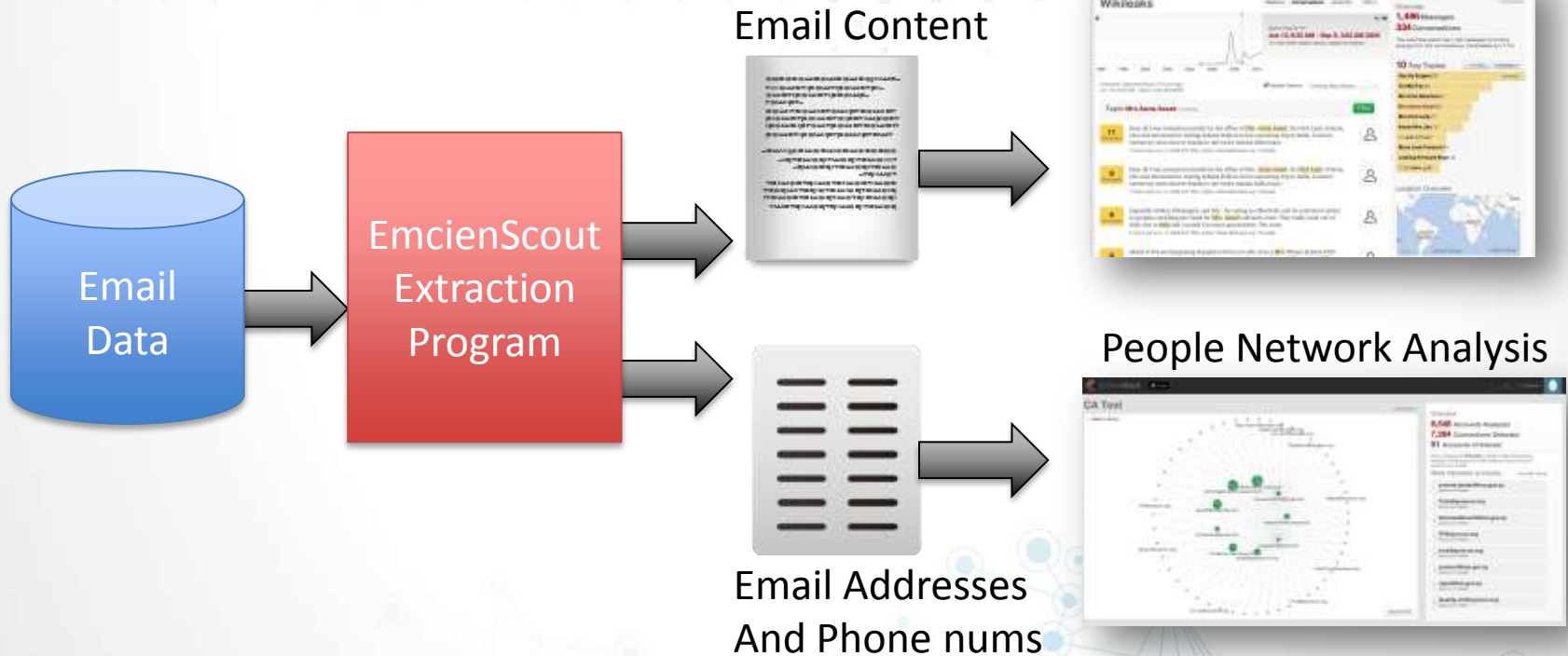
- Federal agency requires fast methods to process high volume of email data
- Need automated methods to highlight **conversations of interest**
- Need automated link analysis to focus on **people of interest based on emails**
- Fast and continuous data processing to **keep up with the speed of crime**



Intel Case Study (2/6)

Automatic Data Collectors

- Content extracted from emails
- Addresses extracted and linked



Intel Case Study (3/6)

Automatic Email Extraction

From: daniel.brown@enron.com
To: dan.jeff@enron.com, david.delaney@enron.com
Subject: FW: EES Employee Issues
Cc: kalen.pieper@enron.com, judy.gray@enron.com
Bcc: kalen.pieper@enron.com, judy.gray@enron.com
Date: Wed, 12 Dec 2001 09:28:51 -0800 (PST)

Dan/Dave,

We are working to gather as much information as possible on our exposure to relocated former and current domestic and international employees impacted by Enron's bankruptcy filing. Lloyd has outlined our position on the urgent issues below. Please keep in mind that regardless of our obligation, the courts have only approved \$15K per employee for all expenses less the \$4500 payment if applicable.

We will continue to work on getting a comprehensive listing over the next couple of days.

Daniel

Extracts all Addresses in Header AND Body

From: daniel.brown@enron.com
To: dan.jeff@enron.com, david.delaney@enron.com
Subject: FW: EES Employee Issues
Cc: kalen.pieper@enron.com, judy.gray@enron.com
Bcc: kalen.pieper@enron.com, judy.gray@enron.com

Messages extracted, each word tokenized and connected into graph.

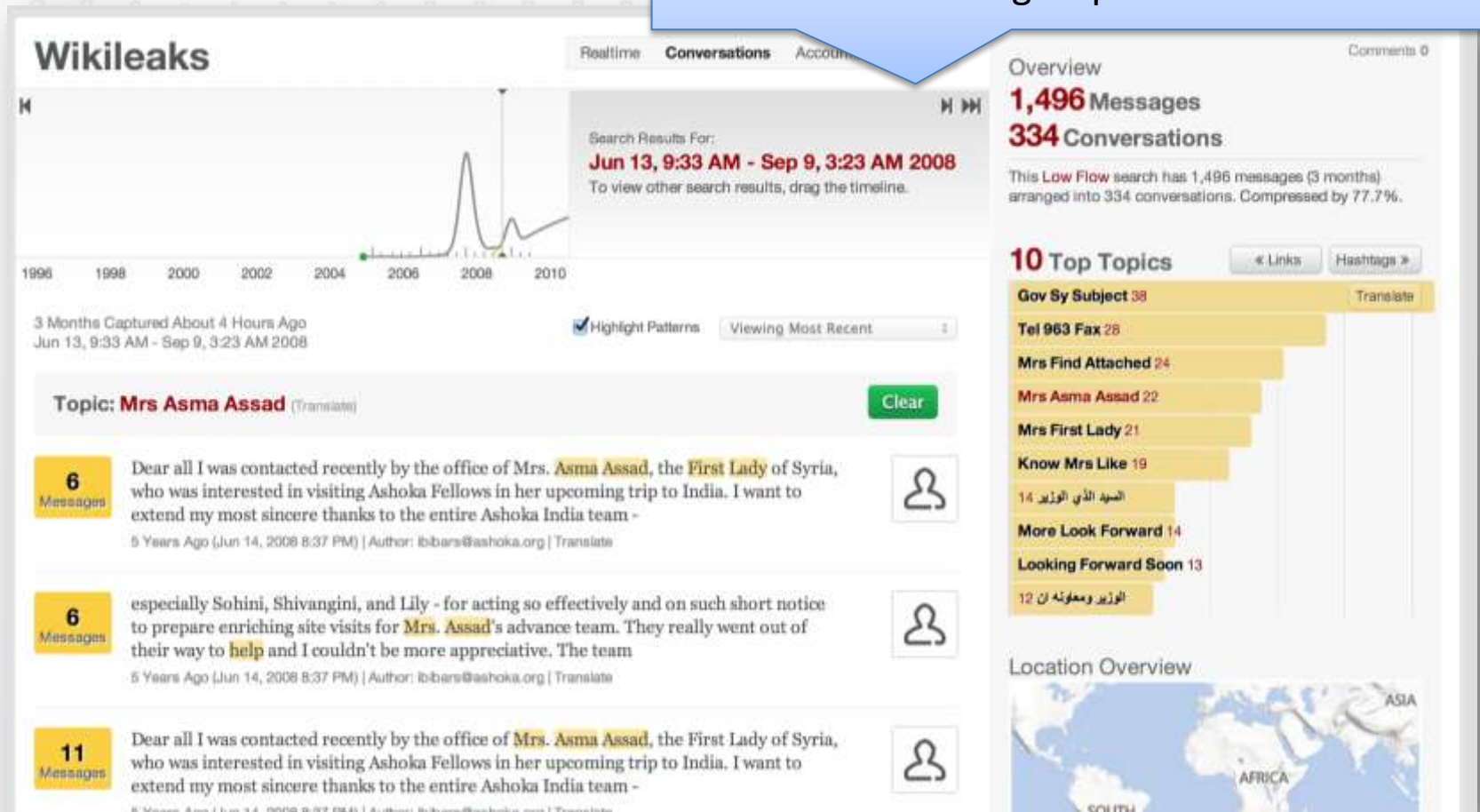
We are working to gather as much information as possible....



Intel Case Study (4/6)

Automatic Email Summarization

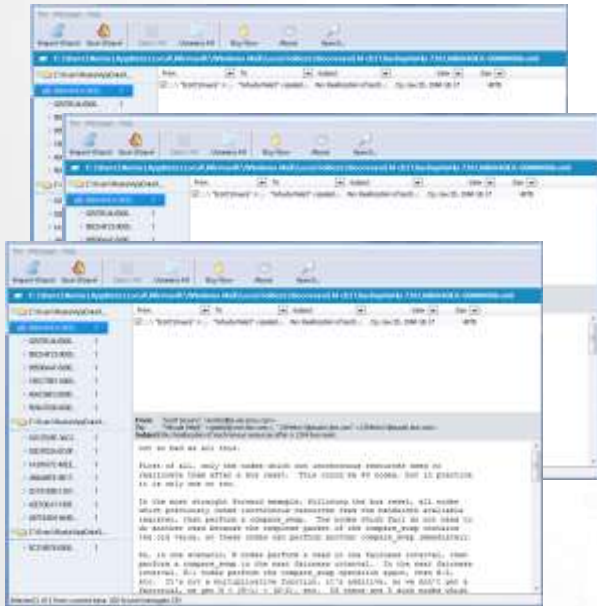
Summarize content from emails to better understand group conversations



Intel Case Study (5/6)

People Graph (1/2)

Program extracts
To/From email addresses
and phone numbers
from suspects email
account

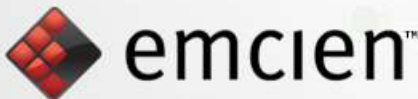


Newly created contacts
file is loaded into Scout
People Graph

```

data: "recipient", "sender"
"01/01/2012", "shelia.corn@genron.com", "lennis_upd@atconcurworkplacs.com"
"01/01/2012", "m_love@genron.com", "laurinh@prodigy.net"
"01/01/2012", "quame@ebay.com", "laurinh@prodigy.net"
"01/01/2009", "gordon.sick@progroup.com", "lensodacy.ac.cy"
"01/01/2009", "gordon.sick@progroup.com", "lensodacy.ac.cy"
"01/01/2009", "fernlay.dyas@genron.com", "sally.becc@genron.com"
"01/01/2010", "tam.otto@genron.com", "robert.cott@genron.com"
"01/01/2010", "Lauri.silve@genron.com", "robert.cott@genron.com"
"01/01/2010", "clara.cernosek@genron.com", "robert.cott@genron.com"
"01/01/2010", "daren.farmer@genron.com", "robert.cott@genron.com"
"01/01/2010", "dawn.gott@genron.com", "robert.cott@genron.com"
"01/01/2010", "gary.harks@genron.com", "robert.cott@genron.com"
"01/01/2010", "eddie.janper@genron.com", "robert.cott@genron.com"
"01/01/2010", "aimee.lomas@genron.com", "robert.cott@genron.com"
"01/01/2010", "mark.mccoy@genron.com", "robert.cott@genron.com"
"01/01/2010", "stacey.nouvel@genron.com", "robert.cott@genron.com"
"01/01/2010", "michael.olson@genron.com", "robert.cott@genron.com"
"01/01/2010", "mary.pasman@genron.com", "robert.cott@genron.com"
"01/01/2010", "carlos.rodriguez@genron.com", "robert.cott@genron.com"
"01/01/2010", "edward.terry@genron.com", "robert.cott@genron.com"
"01/01/2010", "earl.tisdale@genron.com", "robert.cott@genron.com"
"01/01/2010", "jackie.young@genron.com", "robert.cott@genron.com"
"01/01/2010", "sabrae.zajac@genron.com", "robert.cott@genron.com"
"01/01/2010", "daren.farmer@genron.com", "robert.cott@genron.com"
"01/01/2010", "gary.harks@genron.com", "mary.pasman@genron.com"
"01/01/2010", "edward.terry@genron.com", "mary.pasman@genron.com"
"01/01/2010", "diane.lomas@genron.com", "ricky@calipre.com"
"01/01/2010", "dfrance@genron.com", "ricky@calipre.com"
"01/01/2010", "mccoy@genron.com", "ricky@calipre.com"
"01/01/2010", "scott@genron.com", "ricky@calipre.com"
"01/01/2010", "laine.scher@genron.com", "scott.heal@genron.com"
"01/01/2010", "shelli2@msgring.com", "scott.heal@genron.com"
"01/01/2010", "cristina.lucy@genron.com", "scott.heal@genron.com"
"01/01/2010", "william.kelly@genron.com", "scott.heal@genron.com"
"01/01/2010", "john.lavorat@genron.com", "john.arnold@genron.com"
"01/01/2010", "john.lavorat@genron.com", "john.arnold@genron.com"
"01/01/2010", "edie.lescher@genron.com", "john.arnold@genron.com"
"01/01/2010", "mike.magg@genron.com", "john.arnold@genron.com"
"01/01/2010", "mike.magg@genron.com", "john.arnold@genron.com"
"01/01/2010", "harolabelt@genron.com", "david.de@genron.com"
"01/01/2010", "mike.yaguir@genron.com", "david.de@genron.com"
"01/01/2010", "marc.sabine@genron.com", "david.de@genron.com"
"01/01/2010", "steve.arvin@genron.com", "david.de@genron.com"
"01/01/2010", "laine.lavorat@genron.com", "david.de@genron.com"
"01/01/2010", "aizick.williams@genron.com", "david.de@genron.com"
"01/01/2010", "anet.dietrich@genron.com", "david.de@genron.com"
"01/01/2010", "laura.lucot@genron.com", "david.de@genron.com"
"01/01/2010", "frank.wickers@genron.com", "david.de@genron.com"
"01/01/2010", "laura.lucot@genron.com", "david.de@genron.com"
  
```

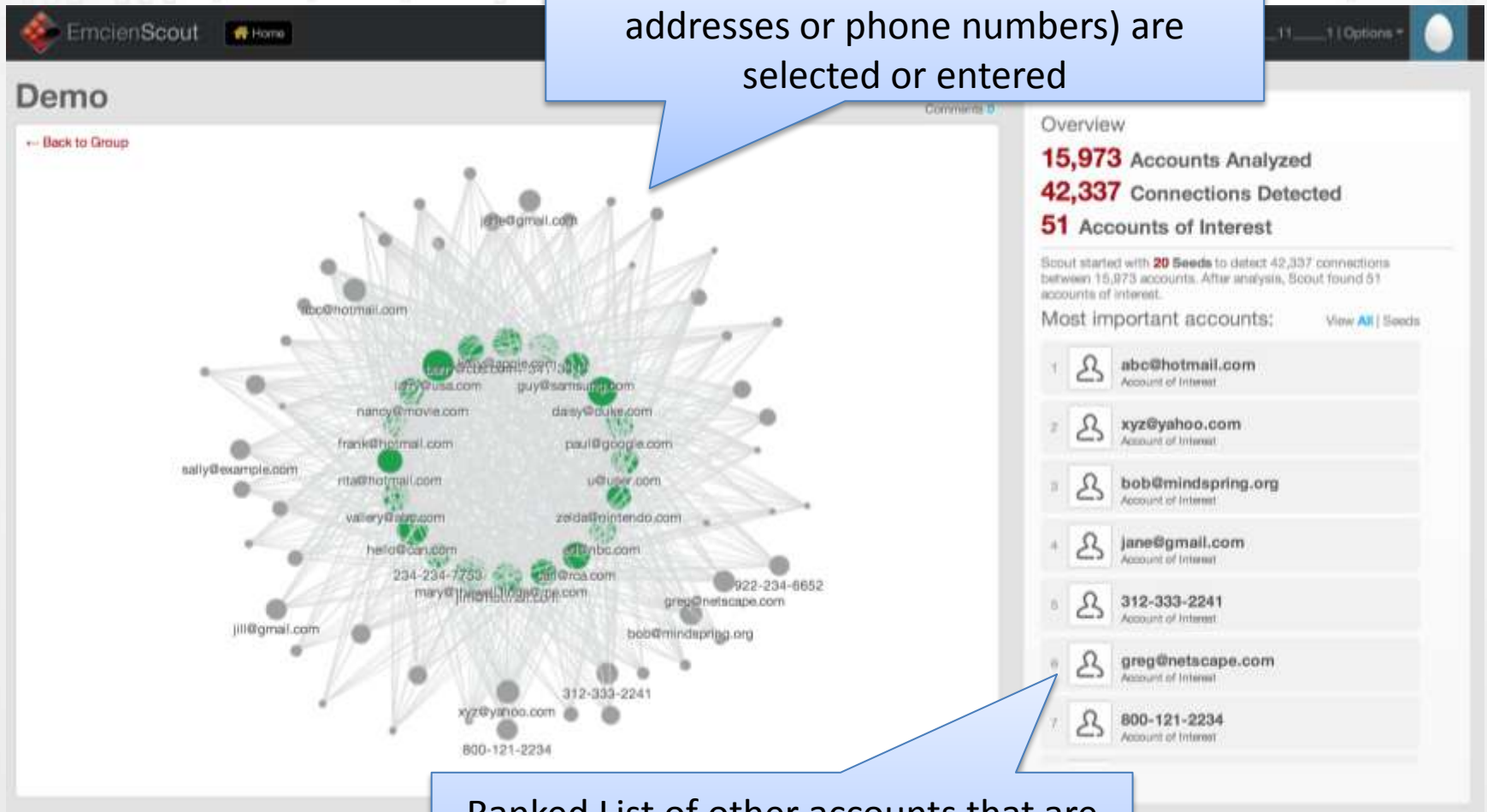
Large complex graph is
created using emails and
phone number
connections



Intel Case Study (6/6)

Algorithm Computes People Graph (2/2)

Initial “bad actor” seed accounts (emails addresses or phone numbers) are selected or entered



Ranked List of other accounts that are likely involved with seed accounts based on their connections.

How Emcien Fits Into Your Ecosystem

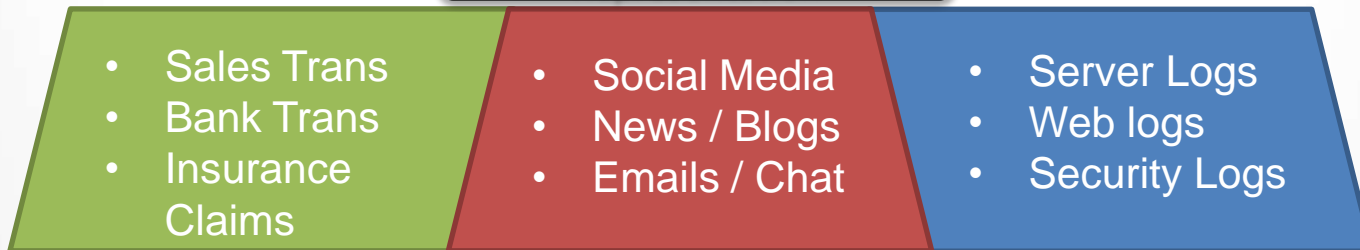
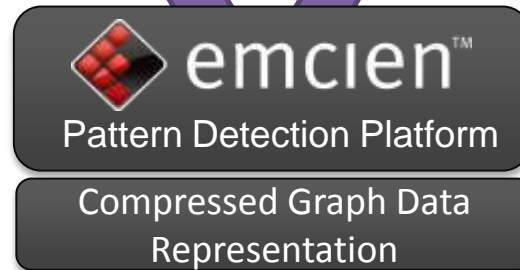
Feed downstream systems with data output



Production Servers



UI for Analyst who wants to review results



Unstructured (Machine Data)



Questions?



emcien™



Types of Data

- Many types of Data
 - Structured, Unstructured
 - Text, Numeric, Machine
- In many states
 - Static (slow batch)
 - Streaming or fast batch



Social, Blogs,
Newsfeeds



Email Data



Machine Data
Network log files



Click stream
data

Limitations of Current Solutions

- Manually Intensive: Very slow and unreliable
 - Search or query based
 - Visualization as a means for discovery → High error
- Limitation based on data types
 - Numerical analysis only
 - Text only, NLP methods, very high set up cost
- Limitation based on data staging
 - Streaming data and recent analysis
 - At-rest data and Historic analysis
- Scalability
 - Current approaches focus too much on storage methods