The Big Data Paradigm Shift

Insight Through Automation

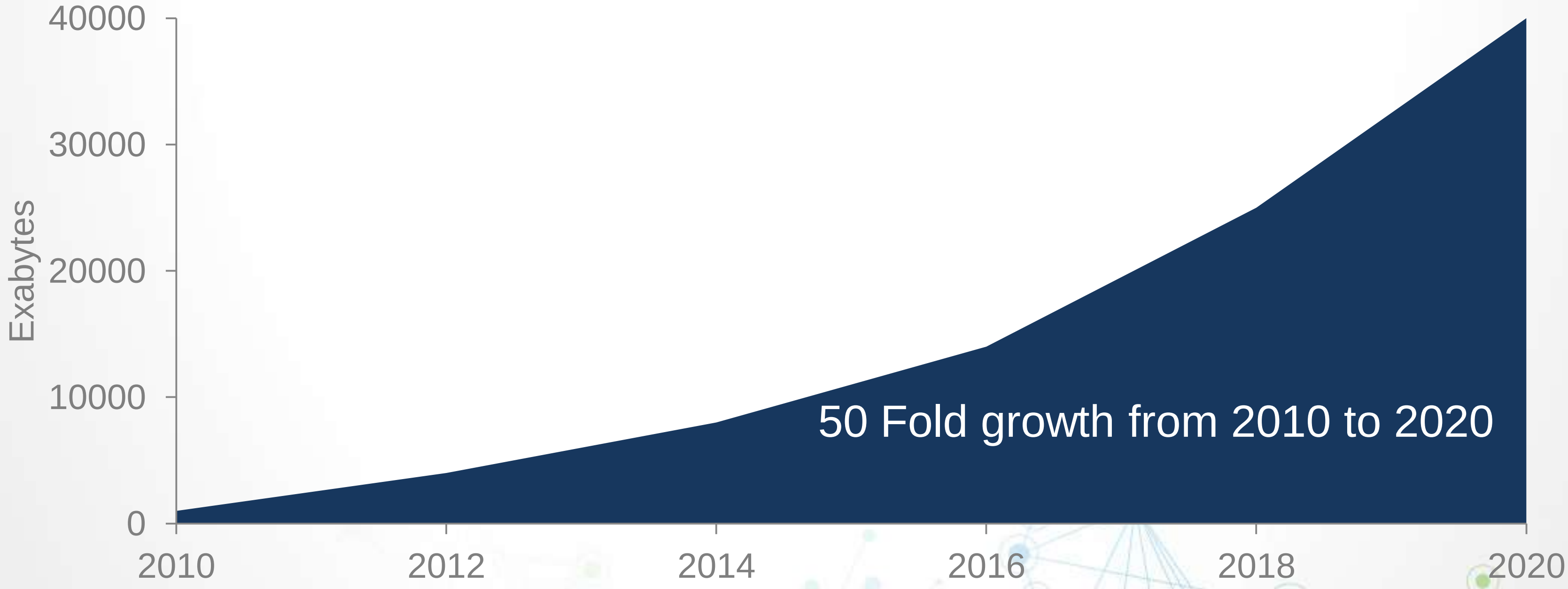
emcien™

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
- **<u>Less than 1% </u>** of data is analyzed



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

emcien™

Old Paradigm: Manually Intensive Analysis

- Unpredictable
- Slow
- Expensive

Collect → Analyze → Report

emcien™

New Paradigm: Automation of Analysis

PREDICTABLE

FAST

ECONOMICAL

Collect → Solve → Review & Act

emcien™

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 <u>solve business problems</u>

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

Machine Data
Network log files

Social, Blogs, Newsfeeds

Marketing Data **Sales Data** **Corporate Data**

Email Data

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

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

emcien™

Another View of the Big Data Stack

Our Focus

Value Layer

Use Cases
(Industry specific or cross industry)

Sectors

Banking
Insurance
Manufacturing
Retail
Internet
Telco
Intel/Security
Healthcare

Analysis Layer

Algorithms and Analytics

Infrastructure

Processor

Storage

Data

Click stream
Machine Data
Corporate Data

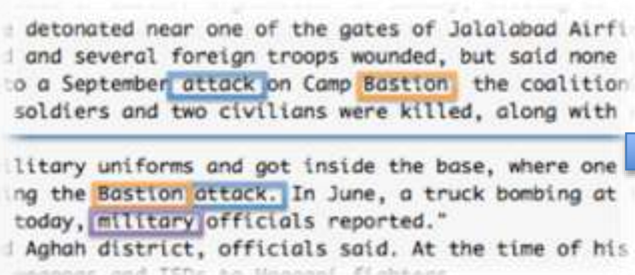
emcien™

How Does Our Solution Work?

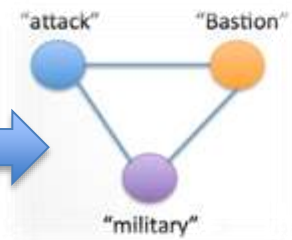
- Big Data problems need graph analysis
- Framework for analyzing relationships
- Highly scalable representation

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

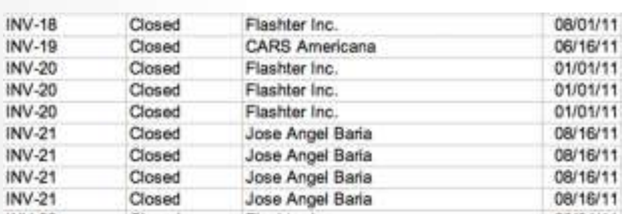
Unstructured


detonated near one of the gates of Jalalabad Airfi
and several foreign troops wounded, but said none
o a September attack on Camp Bastion the coalition
soldiers and two civilians were killed, along with

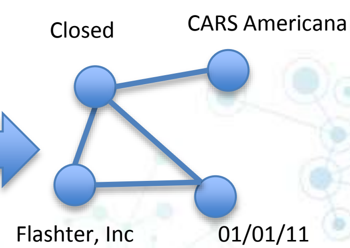
litary uniforms and got inside the base, where one
ng the Bastion attack. In June, a truck bombing at
today, military officials reported."
Aghah district, officials said. At the time of his



Structured



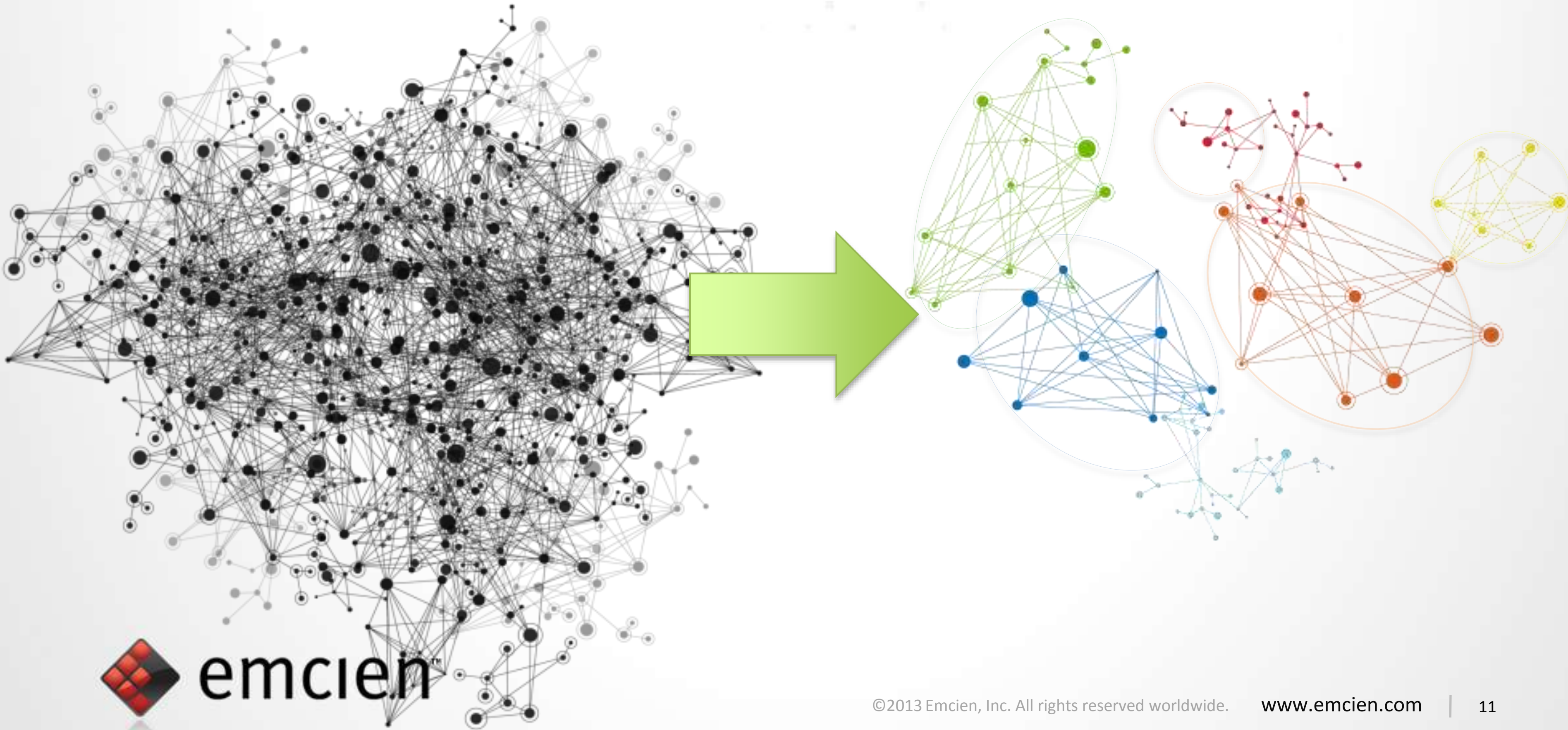
INV-18	Closed	Flashter Inc.	08/01/11
INV-19	Closed	CARS Americana	06/16/11
INV-20	Closed	Flashter Inc.	01/01/11
INV-20	Closed	Flashter Inc.	01/01/11
INV-20	Closed	Flashter Inc.	01/01/11
INV-21	Closed	Jose Angel Baria	08/16/11
INV-21	Closed	Jose Angel Baria	08/16/11
INV-21	Closed	Jose Angel Baria	08/16/11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



emcien™

Algorithms Solve to Extract Patterns

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



emcien

Data Patterns That Reveal "The Insight"

Algorithms designed to reveal graph constructs that solve a business problem

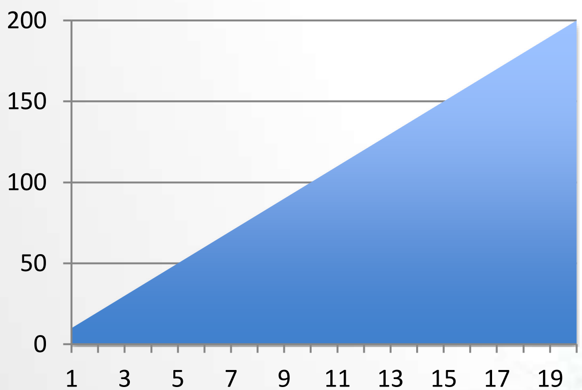
Solving a Graph Problem	Results in Solving a Business Problem
Loosely Federated Communities	- Reveals groups that behave similarly - Reveals dimensions that bind the group - Impossible to detect in a typical querying system
Cliques	- Highly correlated elements - Optimal query that would lead to insight
People Network	- Reveals influence network of individual - Highly predictive for adoption behaviors
Substitute Nodes	- Nodes that behave very similarly - ID theft or product substitutes

emcien™

Algorithms Are Highly Scalable For Big Data

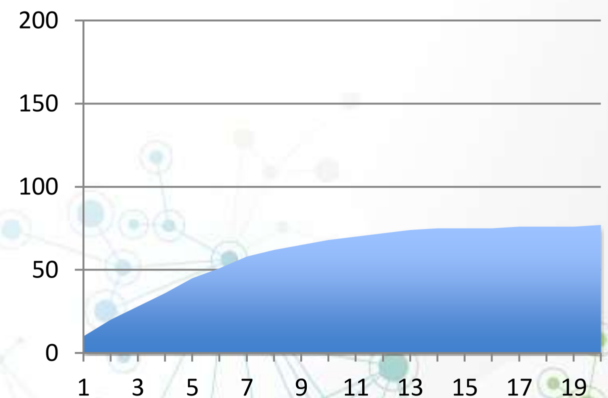
Traditional Data Storage:

- 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



emcien™

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

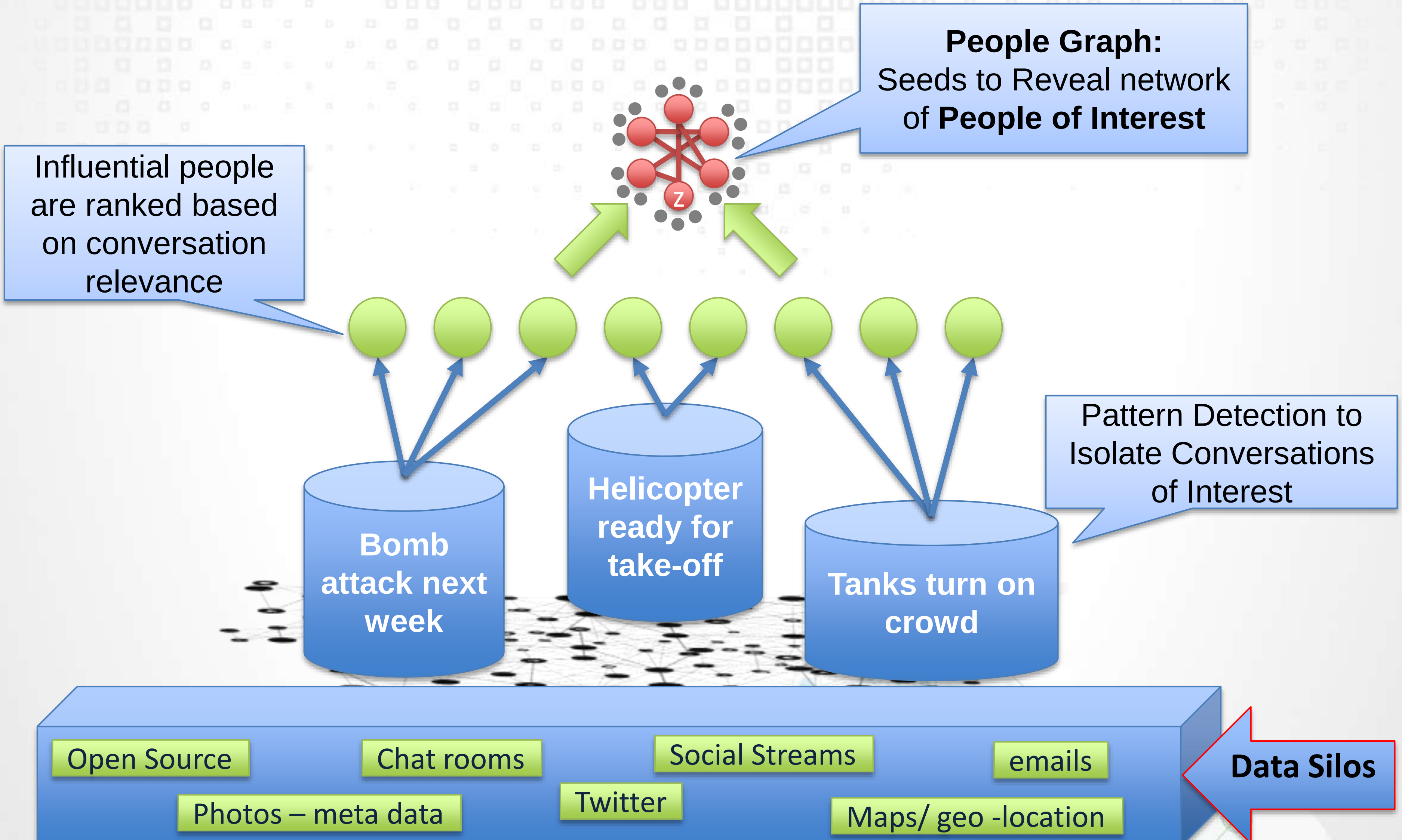
Customer Overview And Current Situation

- 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

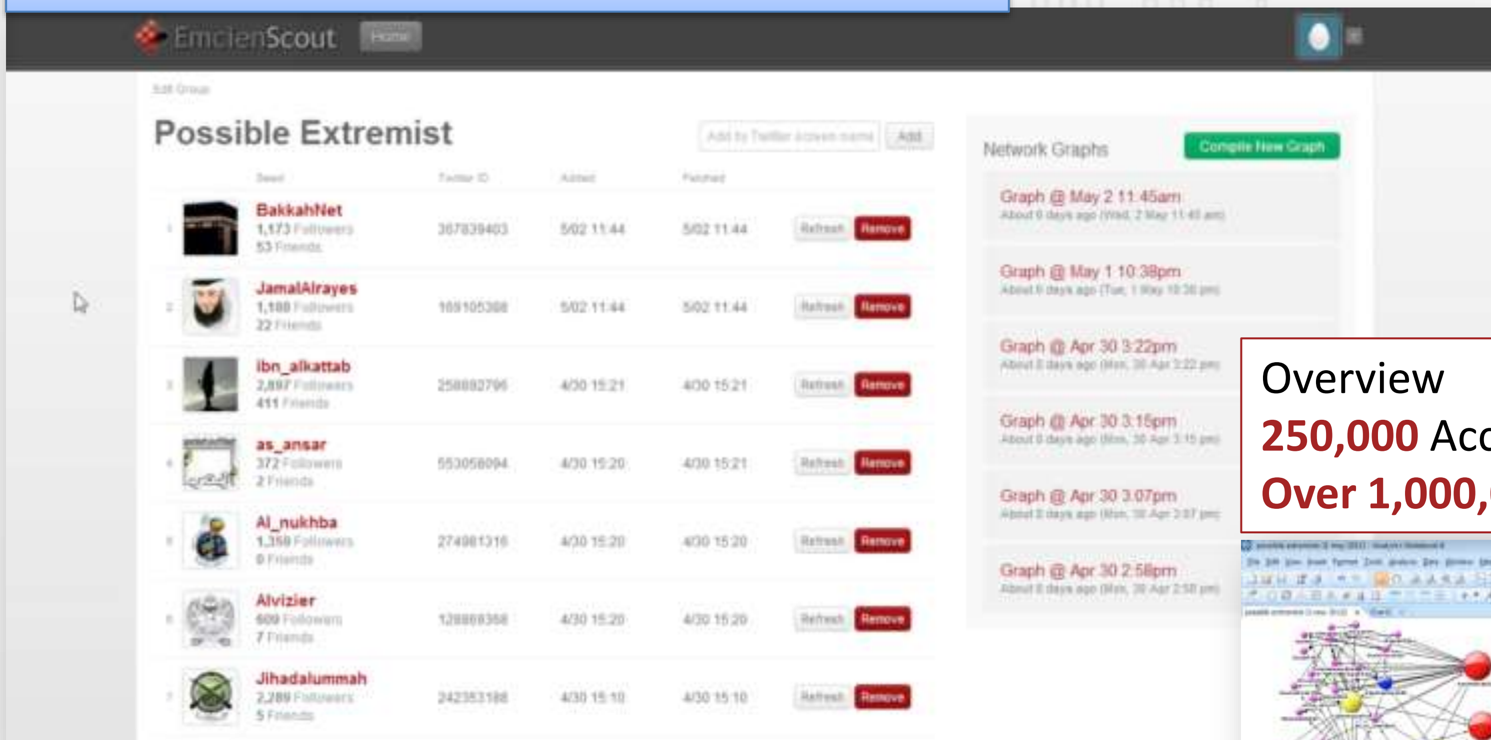
- Federal agency requires fast methods to process high volume open source data
- Need automated methods to highlight **conversations of interest**
- Need automated link analysis to focus on **people of interest**
- Fast and continuous data processing to **keep up with the speed of crime**

Intel Case Study (2/4)
Layers of Analysis for Cyber Threat Monitoring

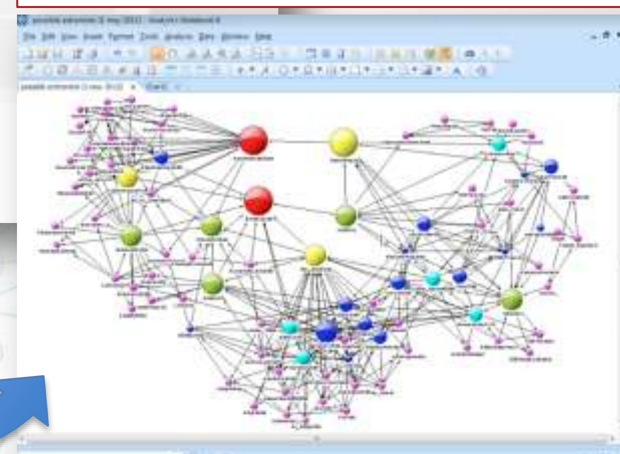


Intel Case Study (3/4)
Cyber Threat Monitoring with Open Source Data

Influential people ranked based on conversations



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



Highly relevant" People of Interest" network

emcien

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

The silent signal – Automatically detecting a sleeper cell

Overview

250,000 Accounts Analyzed

Over 1,000,000 Connections

1 Account of Interest

emcien

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.)

emcien

Network Traffic Log Files (2/6)
Example Use Cases

Example Use Cases

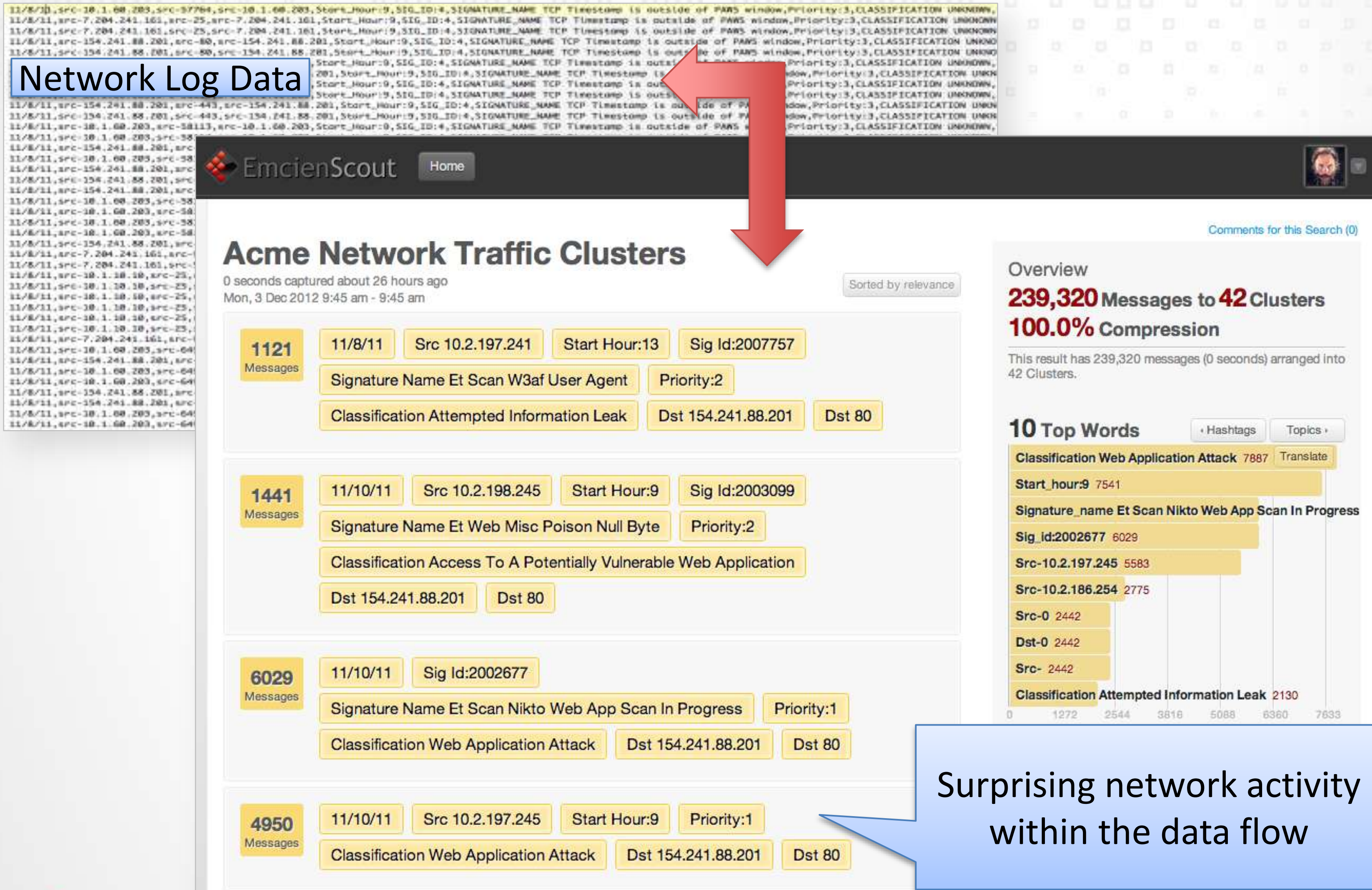
1. Summarize and Rank Log File data based on "Surprising flow patterns"

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"

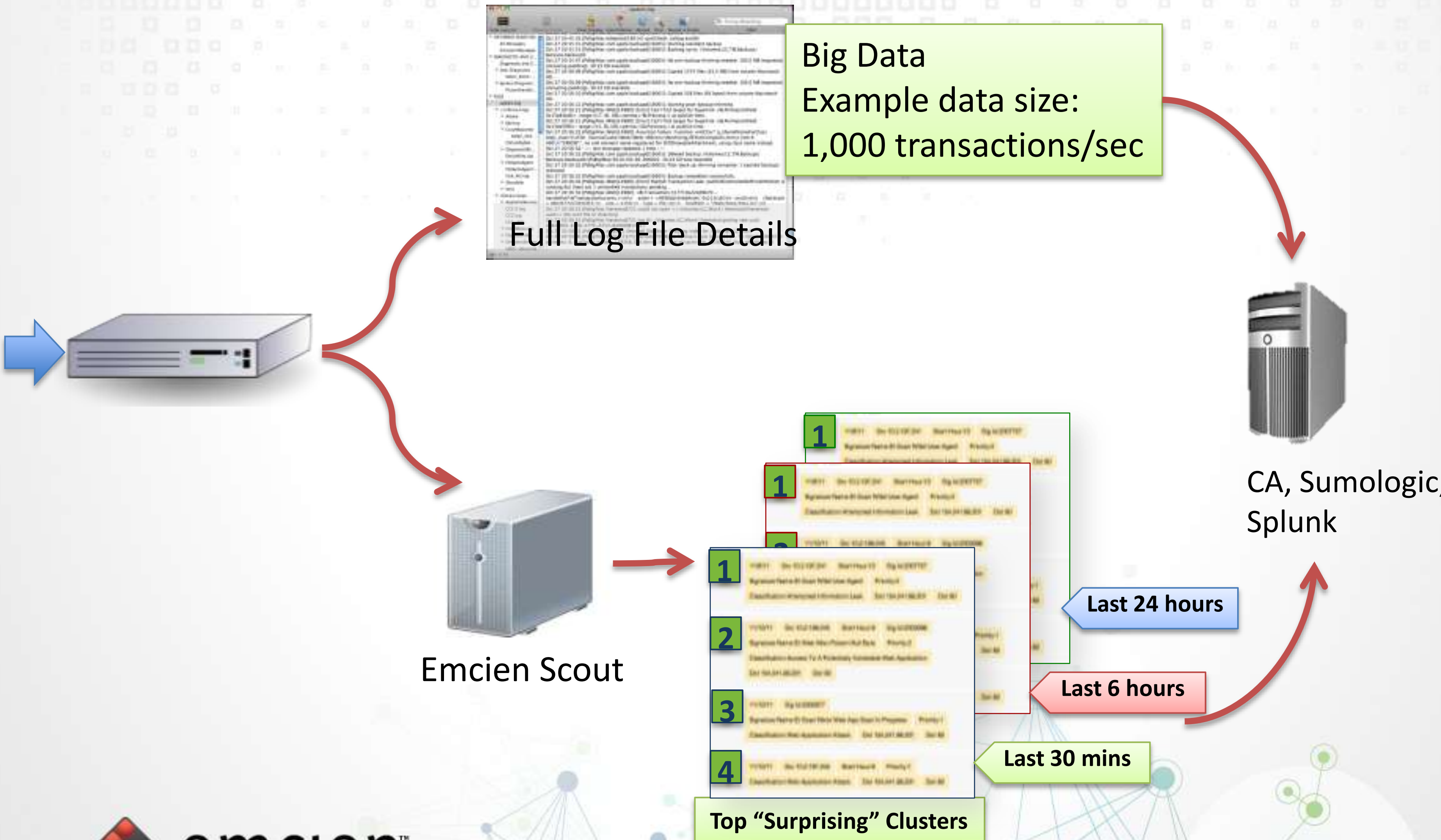
emcien™

Network Traffic Log Files (3/6)
Reveal Surprising Patterns In Network flow Data



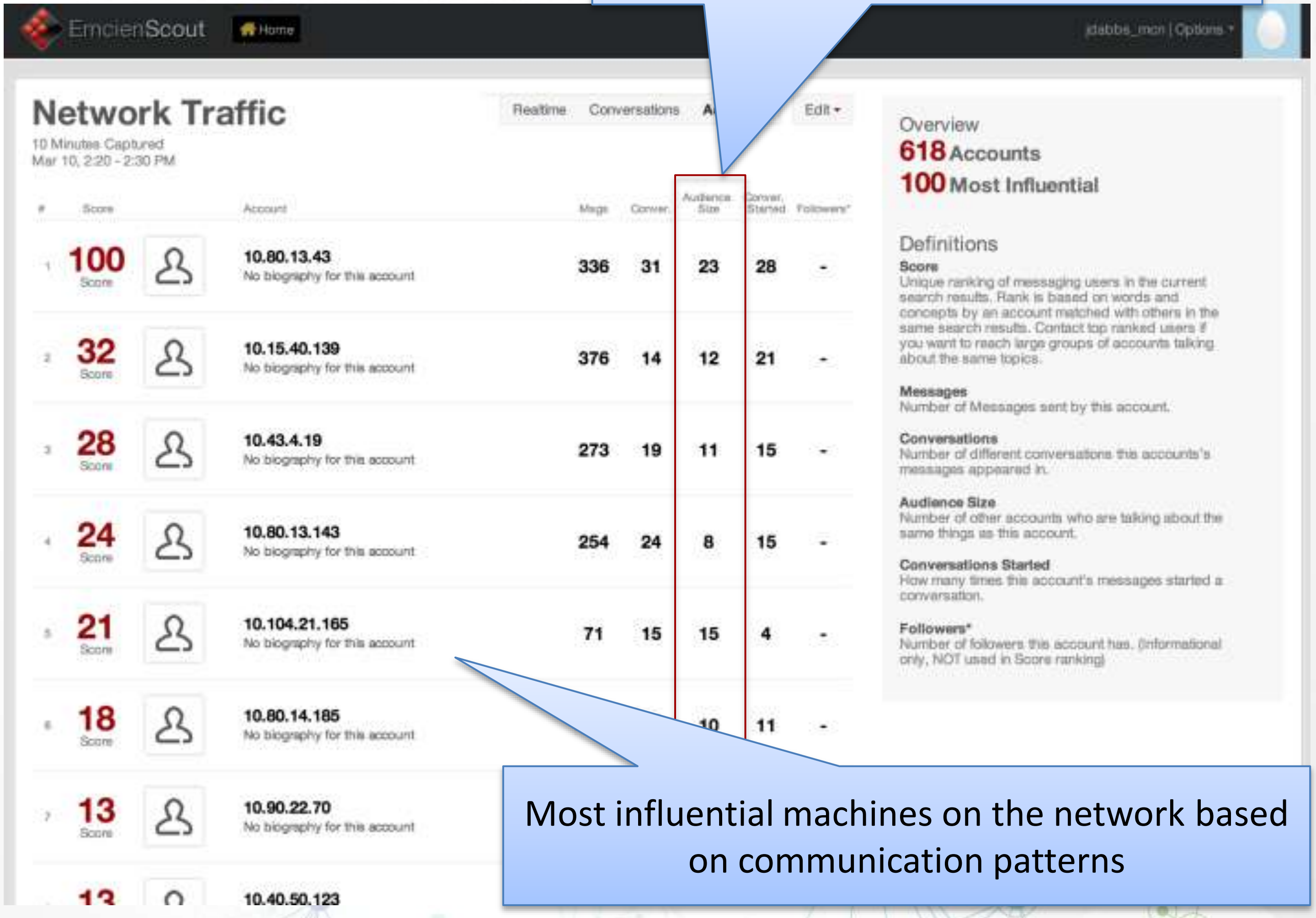
Network Traffic Log Files (4/6)

Ranked Summary of “Surprising” Events



Network Traffic Log Files (5/6)
Most "Influential" Nodes on network

Emcien Scout



Audience size for each machine

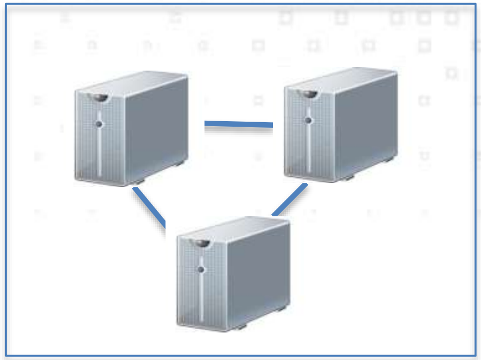
Most influential machines on the network based on communication patterns

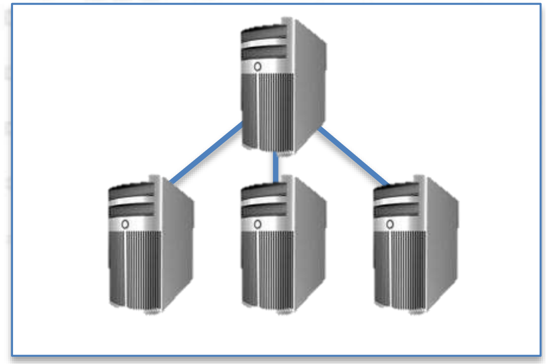
Network Traffic Log Files (6/6)
Machine Communities based on “how they talk”

Lab A

Lab B

Physical Connections

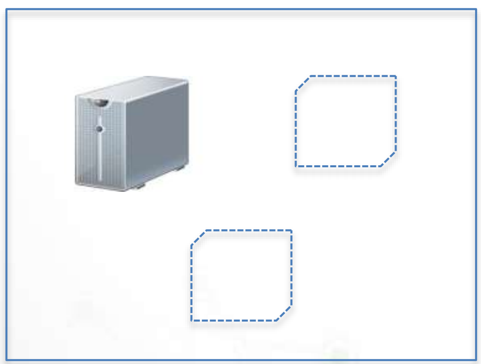


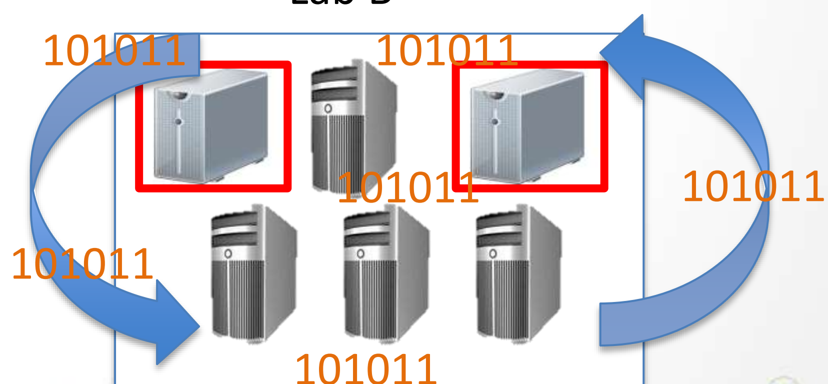


Lab A

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

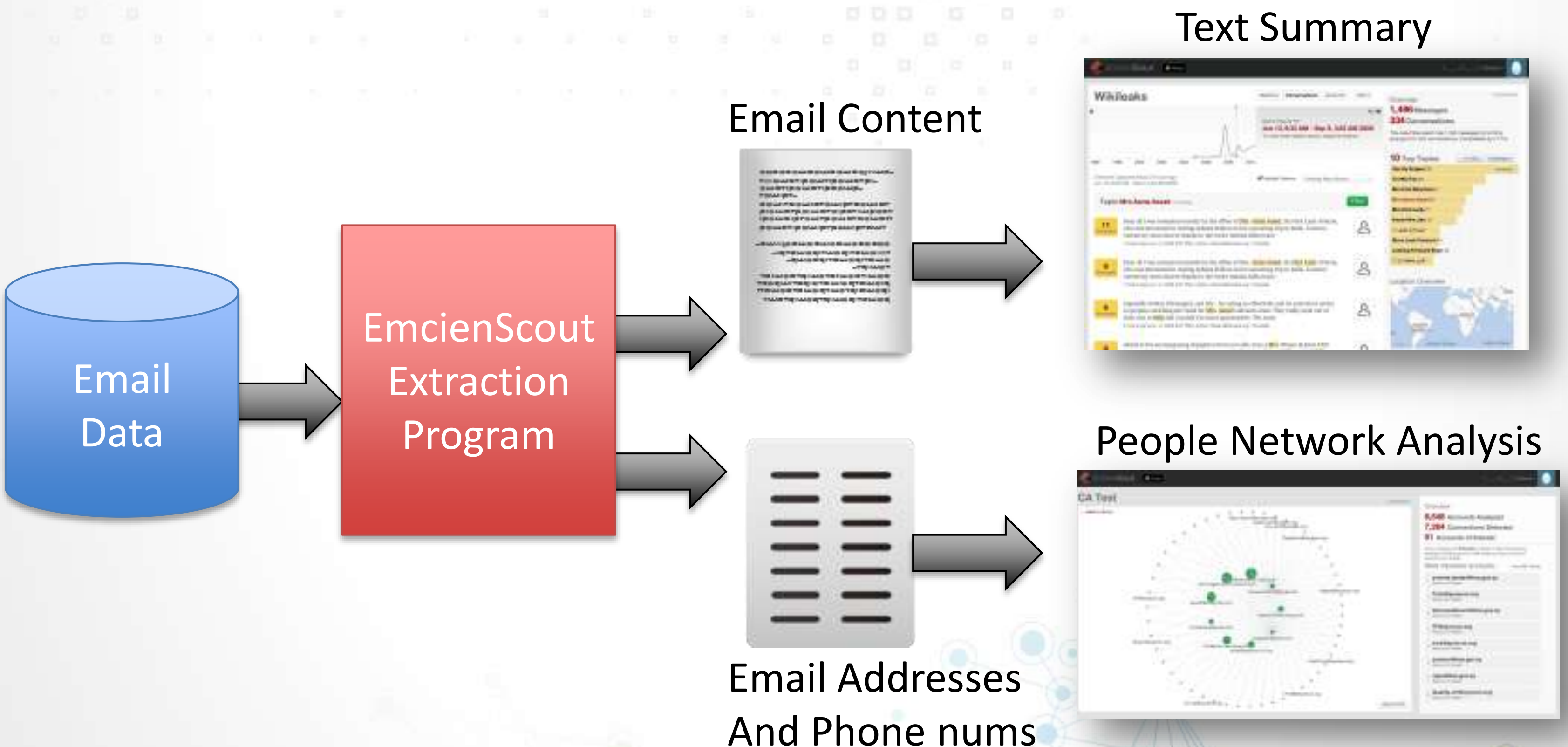
Customer Objective

- 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.leff@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.leff@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….

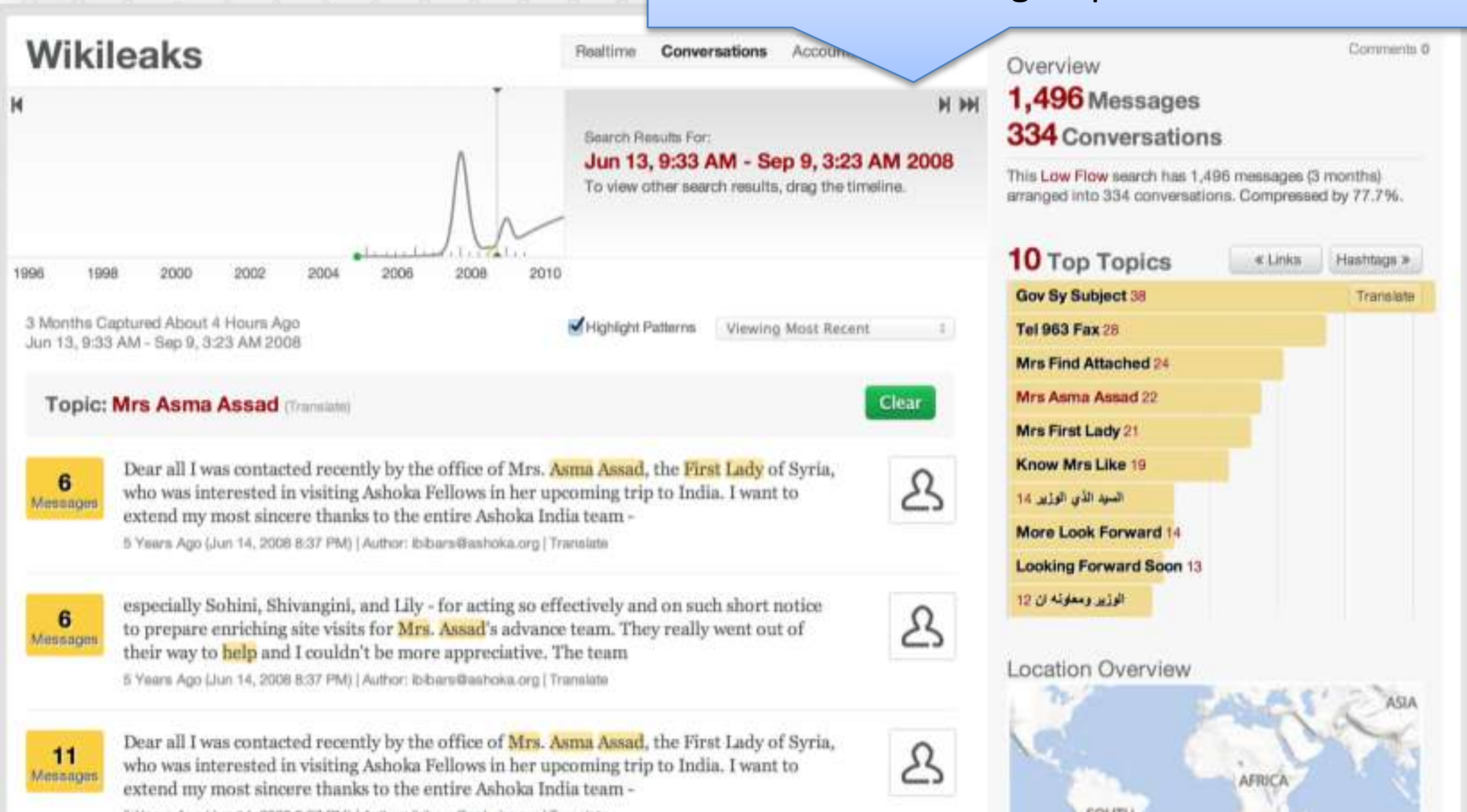
We | are | working | to | gather

emcien™

Intel Case Study (4/6)

Automatic Email Summarization

Summarize content from emails to better understand group conversations



emcien™

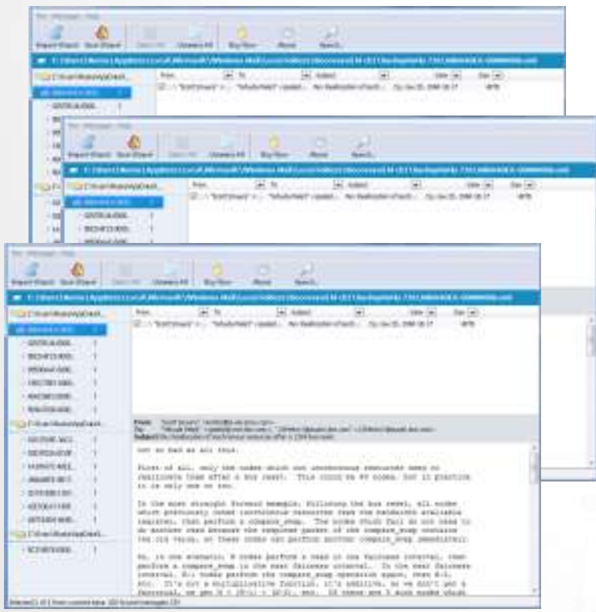
Intel Case Study (5/6)

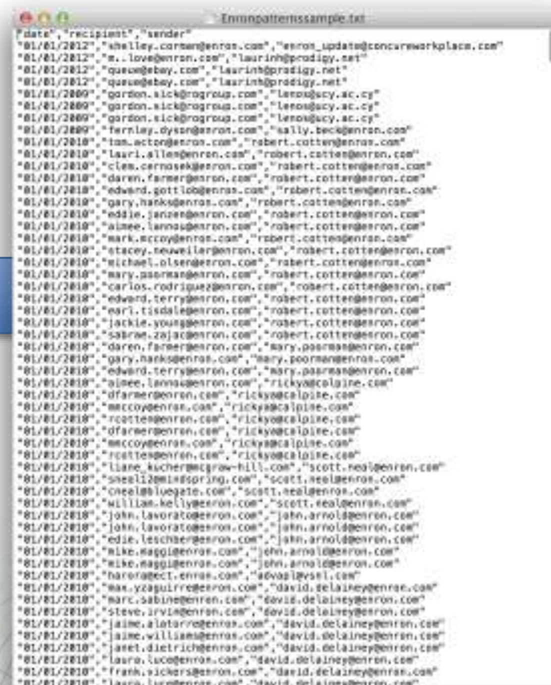
People Graph (1/2)

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

Newly created contacts file is loaded into Scout People Graph

Large complex graph is created using emails and phone number connections



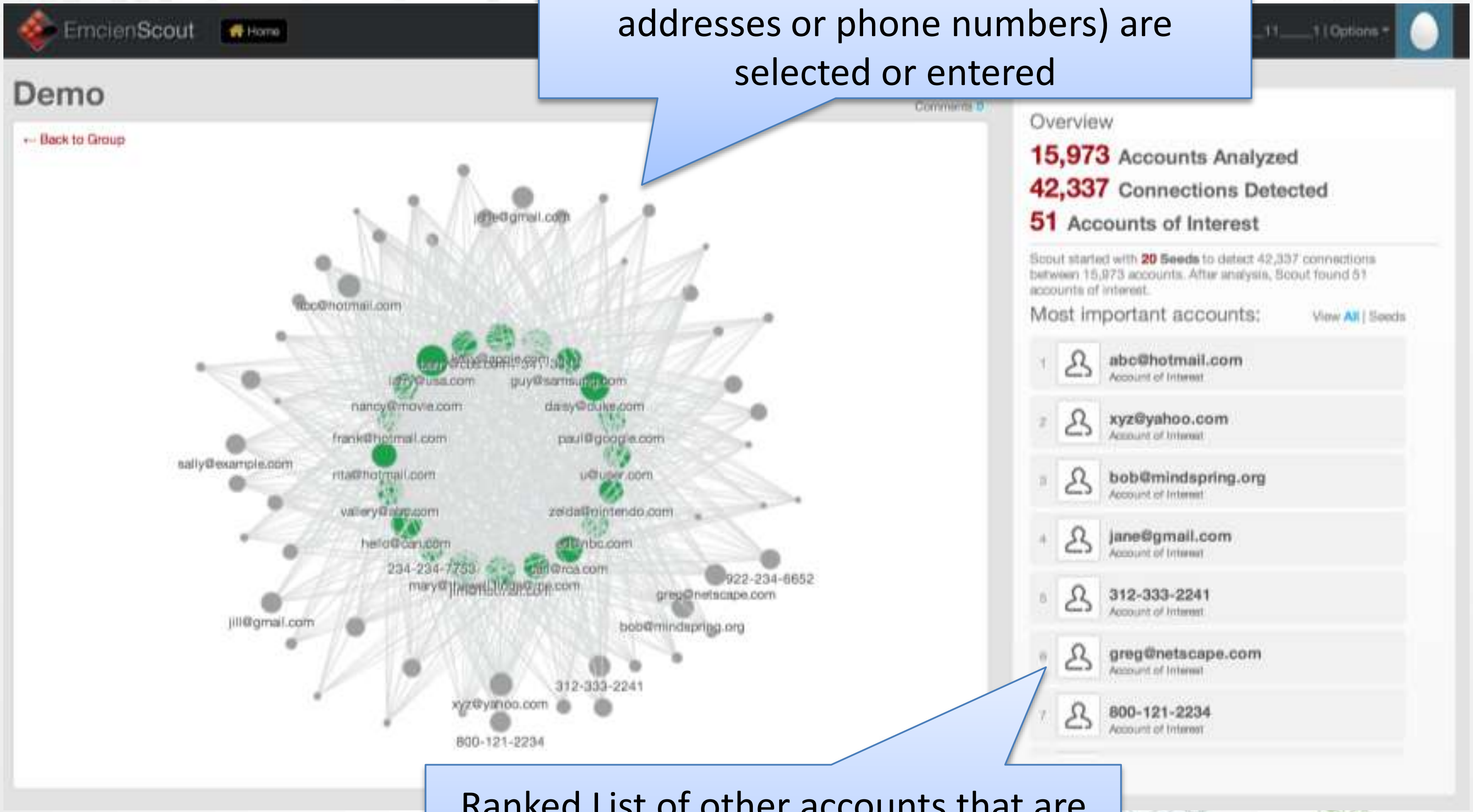


emcien

Intel Case Study (6/6)

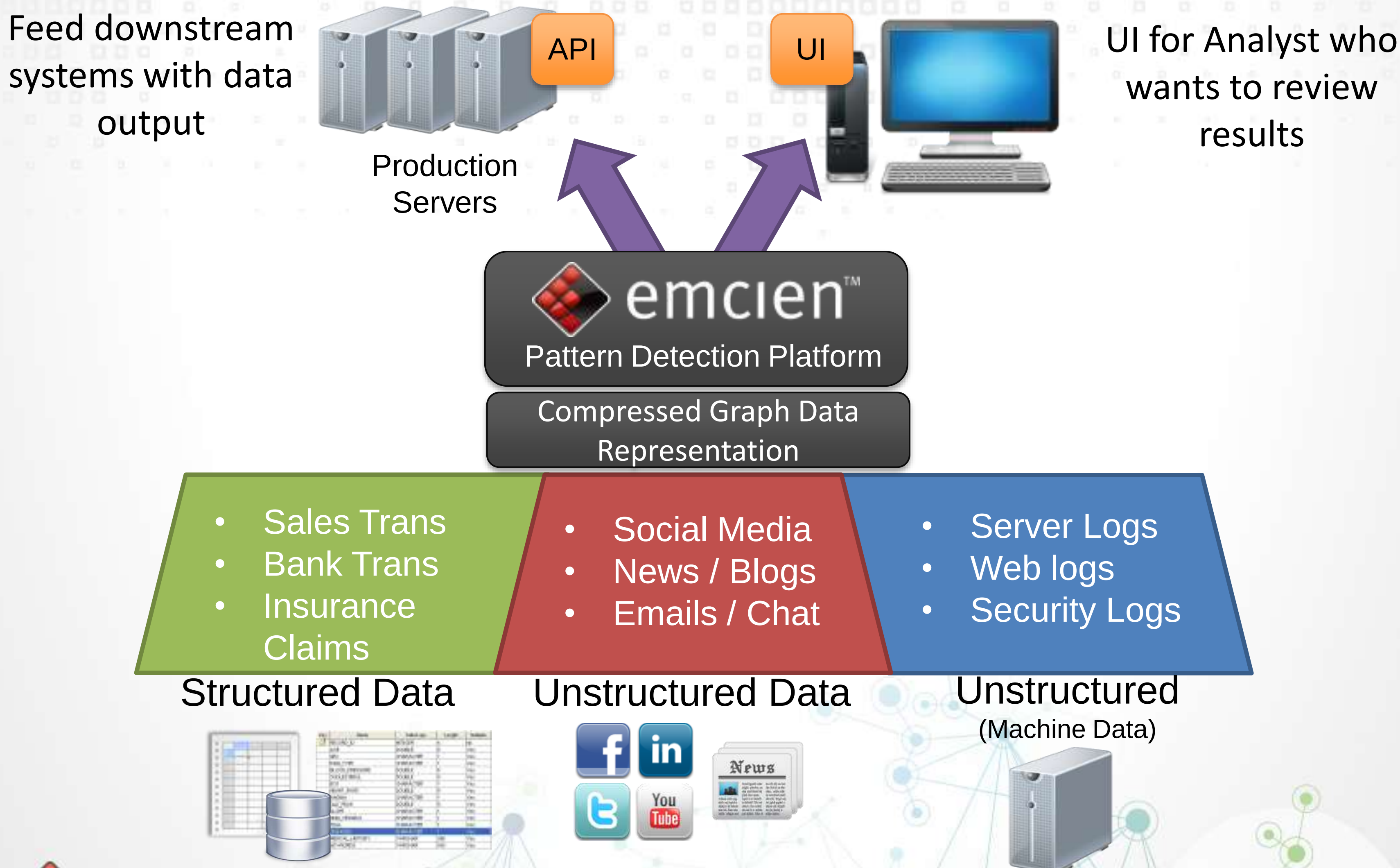
Algorithm Computes People Graph (2/2)

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



Questions?

Types of Data

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

emcien

Limitations of Current Solutions

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

emcien™