Building a Web Intelligence Machine

Staffan Truvé, PhD
CTO, Recorded Future
truve@recordedfuture.com
Who are we?

• Founded in 2009
• US-Swedish startup
  • 40 persons, 50/50 US/SE
  • Boston (HQ), DC, Göteborg, Stockholm
  • and a few more places around the globe
• Focus only on web intelligence, primarily for governments and Fortune 500
• Backed by Google Ventures, Atlas Venture, Balderton, IA Ventures, and I-Q-T
Avoid surprise — grasp opportunity
Silicon Valley executives head to Vail, Colo. next week for the annual Pacific Crest Technology Leadership Forum.

The carrier may select partners to set up a new carrier as early as next month.

Unilever will hold their UK launch event early next week in Manchester.

“2010 is the year when Iran will kick out Islam. Ya Ahura we will.”

“...Dr Sarkar says the new facility will be operational by March 2014.”

“...opposition organizers plan to meet on Thursday to protest...”

Estimated study completion date: November 2014.

...the transaction is expected close in late 2013...

...new facility will be operational by March 2013.

“Excited to see Morsi speak this weekend.”

Strange new Russian worm set to unleash botnet on 4/1/2013.

“According to TechCrunch China’s new 4G network will be deployed by mid-2010”
Organizing the Web for Analysis
Kuo expects that Apple will introduce an iPhone 5S around June or July of this year.

Drought and malnutrition hinder next spring’s expansion plans in Kabul...

“...opposition organizers plan to meet on Thursday to protest...”
Time

**Facebook reported about**

**Cyber attack by India on Jul 21, 2013**

"Three days ago, the official website of congress mouthpiece, monthly Hindi journal, ‘Congress Yuva Sandesh’, was hacked by "Anonymous Kashmir" and it displays many "messages" to India."

Jul 24, 2013  7:23  Facebook  

**Twitter and 1 more reported about**

**Cyber attack against India on Aug 14, 2013**

"@hamiltonatdell Pakistani hacktivists plan on celebrating their Independence Day, 8/14, by launching cyberattacks against India http://t.co/p341GTZN4D."

Jul 17, 2013  17:42  Twitter  

"Pakistani Hackers Threaten to Attack India on August 14."

Jul 17, 2013  12:08  Team Cymru Internet Security News

“NOW”

Jul 24

July 17  July 21  July 24  August 12  August 14
Structure – entities, events, ontologies

Associated Press
At least 10000 dead in Haiti as earthquake hit late afternoon local time.
2010-01-13 08:00

steve23
Terrible earthquake in Haiti – many dead
2010-01-12 22:00

Port au Prince
NaturalDisaster
2010-01-12 22:00-24:00
Fatalities: 10000

NaturalDisaster
2010-01-12 22:05
Fatalities: ?

Country

City

ON THE WEB

IN THE WORLD

ONTOLOGY

Country-Capital

NaturalDisaster
Kind:Earthquake
Location:

Haiti: Country

Port-au-Prince: City

NaturalDisaster
Kind:Earthquake
Location:

NaturalDisaster
Kind:Earthquake
Location:
A year of protests

Top 14 protest countries, relative, 7 day rolling average

[Graph showing protest activity over time for various countries]

- Bangladesh
- Israel
- Libya
- Syria
- Canada
- Spain
- China
- Bahrain
- Brazil
- Pakistan
- United Kingdom
- India
- Turkey
- Egypt
Protests in Libya
Predicting Egyptian Protests

![Graph showing predicted dates for Egyptian protests]

The graph illustrates the predicted dates for Egyptian protests, with the dates of the prediction on the vertical axis and the dates in question on the horizontal axis. The colors indicate the likelihood of protests, with darker colors representing higher likelihoods.
Threat Analysis Applications

THREAT INTELLIGENCE TEAM

LOCATIONS SECURITY

EXECUTIVE SECURITY

PRODUCT SECURITY

C Y B E R S E C U R I T Y

Threat Signals

Analysis Insights

INTERNAL CUSTOMERS

Regional Security
International Travel
Executive Protection
Product Security
Supplier Risk
CERT
Cyber Forensics
Online Fraud
Monitor

- Track discussion about your business locations: offices, data centers, etc.
- Future events in supplier locations, travel sites, and event venues.
- Coded according to your security profile and current threat level.
LOCATIONS

Nigeria

 Clamp Down On Boko Haram Commended
 "Governor Rochas Aksed To Declare State Of Emergency On Imo State University Law School And Others EFCC Chairman advocates for regular anti-corruption training for the military Fayemi, Fashola are worse than military dictators SSS Parades Suspected Bombers Of Jaji Military Baracks Nigeria Police Force Launches New Code Of Conduct."  
 Jun 26, 2013, 20:54 · spyghanacom · [ GRjZWOA5DHb ]

Nigeria: Nigeria Food Security Update June 2013
 "Source: Famine Early Warning System Network Country: Nigeria Crisis levels of food insecurity persist in the northeast due to the effects of conflict KEY MESSAGES □ Boko Haram-related conflict, as well as a declared state of emergency, is still ongoing in the northeast."  
 Jun 26, 2013, 19:55 · Relief Web Int · [ GRjYeoA46eG ]

You have received this email because you are subscribing to Recorded Future corporate security application alerts. If you no longer wish to receive these alerts please send an email to support@recordedfuture.com and ask to be removed from the alert subscriber list.
EVALUATE

> Threats scored for sentiment, volume of coverage, and uniqueness.

> One click to source material for quick evaluation of threat credibility.

> Data can be filtered by date, source location, and other parameters.
Visualization tools for deep dive, multi-lingual open web analysis.

Email alerts, personal watchlists, and workbooks.

Data export and reporting tools for dissemination.
Web Intelligence – at Web Scale

• Processing 100s of millions of documents
• Sources from all over the world, in 8 languages – English, Arabic, Chinese x 2, Russian, Spanish, French, Farsi
• From government sites and big media to blogs and social media
• 10B “facts” - growing fast
• 25+ entity types
• 100+ event types
• Metrics, signals, alerts
Challenges – running a web intelligence machine is a process industry!

- Real time process – from web to analytic insights
- 24/7 – no down time acceptable
- Scalability
  - Volumes growing over time – sources, data, amount of analysis
  - Repeated, massive re-analysis
  - Interactive UX with very large data quantities
- Continuous development and deployment
- Coordinating a very distributed team
Some numeric facts

- 130 servers
- 50 TB of data
- About 50 deployments per week
- 500,000 different metrics collected
  - 5000 metric values collected per second
- Developers at 6 locations around the globe, 4 time zones
- Deployed to 3 different data centers
- 500,000 lines of code
- 100,000 lines of infrastructure code
- 4000 RabbitMQ messages/second
How to run a software startup in 2013

• Do less
• Own less
• Achieve more
On the theme of “do less”:

Next 5 slides borrowed from Andreas Ehn (ex-Stardoll, ex-Spotify, Wrapp, etc.)

Thank you!
Overarching trends that vastly simplify starting a company

- The central theme is that commonalities are identified and factored out
- \(ab + ac = a(b + c)\)
- Same thing with code
- And services, and infrastructure
- Commoditization
In the future...

- Platforms getting more sophisticated, rising through the abstraction layers
- What other pieces of infrastructure that we build ourselves now will move to the platforms?
  - Login and user management
  - CRM, retention, notifications
  - Social channels
  - Other growth mechanics
  - Deployment and operations
How we do it – under the hood
NLP / TEXT ANALYSIS

UX / API

STORAGE / ANALYTICS BACKEND
Länspoliser ska skydda Obama

- Ett tjugo tal poliser från Västerbotten har kallats till Stockholm för att bistå i säkerhetsarbetet för världens mäktigaste man.

Grafik: Kolla in Air Force One
- Obamas flygplan är utrustad med operationsbord.

TV: Brunnslock svetsas
- Trafikkaos och svetsade brunnslock inför Obamas besök.

Barack brann i Sorsele
- En fristående barrack ska vara övertänd i Sorsele.
Text document

Calais

Language detection

POS tagging

Named entity extraction

Entity resolution & disambiguation

Event extraction

Dependency Parsing

Sentiment & attitude analyses

Culture categorization

Topic categorization

Temporal extraction

Entity/Event-Time matching

Post processing: Attribute merging and sanity controls

Analysis document

Analysis document

Post processing:
Attribute merging and sanity controls
Client side: Javascript, jQuery, D3.js, SVG, Google Maps+translate
Backend model: pipelined, message based, persistent queues, horizontal scaling, sharded nosql database, scalable search index

RSS
Special sources
Social media

Actors written in
Java
Scala
Python

Text analysis

Data analysis

mongoDB

elasticsearch
Under the hood

- Chef
- ASGARD
- Graphite
- Ruby
- sensu
- ubuntu
- logstash
- Amazon Web Services
Testing and Monitoring at Web Scale
**Current Performance and Availability Status**

<table>
<thead>
<tr>
<th>Service / Website</th>
<th>Performance and Availability Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>api.recordedfuture.com/query</td>
<td>Service is operating normally</td>
</tr>
<tr>
<td><a href="http://www.recordedfuture.com">www.recordedfuture.com</a></td>
<td>Service is operating normally</td>
</tr>
</tbody>
</table>

**Performance and Availability History**

<table>
<thead>
<tr>
<th>Service / Website</th>
<th>Sep 23</th>
<th>Sep 22</th>
<th>Sep 21</th>
<th>Sep 20</th>
<th>Sep 19</th>
<th>Sep 16</th>
<th>Sep 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>api.recordedfuture.com/query</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><a href="http://www.recordedfuture.com">www.recordedfuture.com</a></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

- ✔️: Service is operating normally
- 🚨: Performance issues
- ⚠️: Service disruption
- 🌟: Informational message

Recorded Future Inc - Public health page real-time websites availability and performance status. This site shows if Recorded Future Inc - Public health page sites are down or have performance issues right now, and provides Recorded Future Inc - Public health page uptime and performance history.

Public Website Health Status for Recorded Future Inc - Public health page

![Graph showing performance and availability history over time](chart.png)
Monitor, monitor, monitor...
Language - EventTypes Occurrences
Distributed team – synchronized with Skype!

20 augusti 2013
Lars Olsson
Upgraded chef components basis, rf_metadata, and ha_document_analyzer to svn rev 43813.
No explicit actions taken for the ranker
10:41

21 augusti 2013
Filip Reesalu
Posted 910 documents for re-analysis. (Entity type: CVEIdentifier)
10:08

23 augusti 2013
Ulf Måansson
Will deploy rev 43719 of UI, API and admin
17:32

Ulf Måansson
Reverted back to 43432
18:09
Conclusions

- Web Intelligences requires Web Scale all the way
- Do less, own less, achieve more!
- Standing on the shoulders of a lot of giants – but of course you still have to solve many things yourself
- No lack of fun challenges
  - Linguistics – humans are way too creative!
  - Scalability and reliability
  - Creating analytic insight algorithmically