

COLLECTIVE INTELLIGENCE

SECURITY INTELLIGENCE IS LIVING, SOCIAL DATA.

COLLABORATIVE DATA-DRIVEN SECURITY FOR
HIGH PERFORMANCE NETWORKS 2010
CLAIMID.COM/WESYOUNG

SOME CONTEXT, REN-ISAC

- 300+ Participating Institutions
- 700+ Active Members
- Two overlapping federations:
 - ‘General’ -- appointed by your CIO for “meeting work related requirements”
 - ‘XSec’ -- double vouched by your peers
- every single shop is a snowflake!
- so how do you scale information sharing?

WHERE WE LEFT OFF...

- Our Security Event System, “SES” has 10+ Sites Sharing between 50 and 20,000 data-points per day per site.
- (SSH | Telnet | FTP | VNC | Pushdo | Darknet) Scanners.
- Near realtime in most cases, from live sensors as well as honeypots
- Leveraging Snort, Nepenthes, syslogs, Custom Darknet scripts via the current SES API (libprelude)
- We create a “correlated scanners” (multi-location) into a mitigation feed for sites to pull down.
- We also have a web page users can manually enter malicious domain-names, malware drop sites, botnet C&C into which produce various other mitigation feeds (stuff they’ve manually investigated).
- Solve that problem, uncover 10 more problems...

DATA... DATA... AND MORE DATA... (IT'S ALL ABOUT OPS)

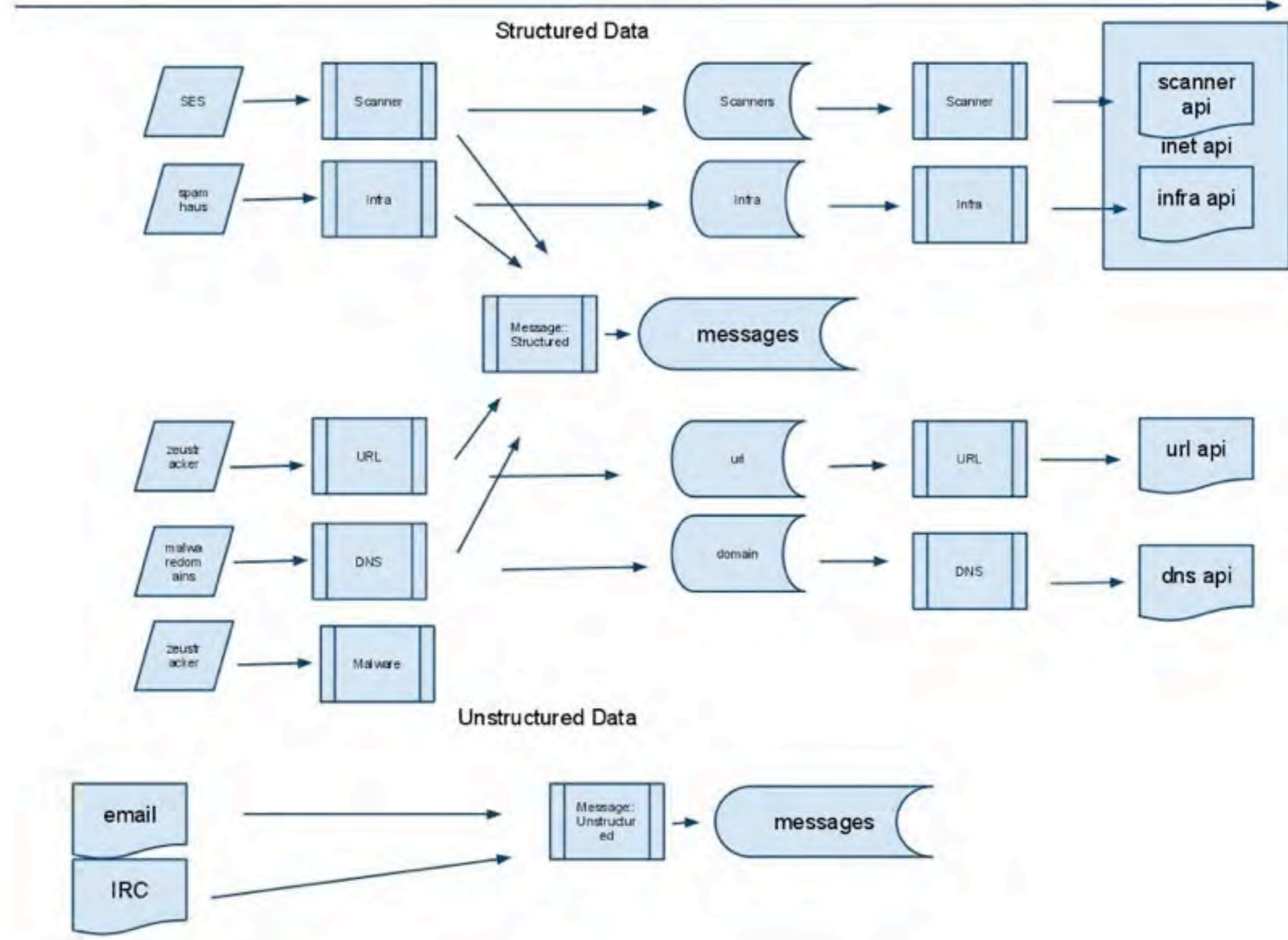
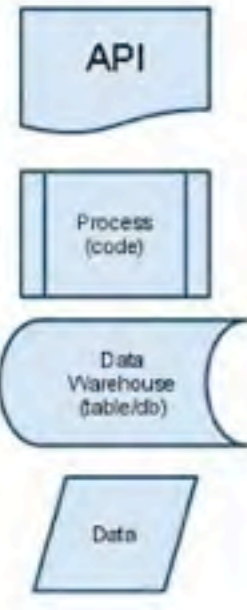
- Locally correlated Events (typically malicious ip-infrastructure)
- Spamhaus DROP list (hijacked networks)
- Malwaredomains.com feed (malware hashes, malware domains, malware ip-infrastructure)
- Malwaredomainlist.com feed (malware urls, malware domains)
- DShield List(s) (scanning ip-infrastructure)
- Phishtank Data (phishing urls, phishing ip-infrastructure)
- Zeustracker data (binary urls, config urls, domains, ip-infrastructure)
- From each domain, you have massive potential intelligence from the name-servers involved with each domain.
- Whitelists (domains, ip-infrastructure... dnswhl.org)
- Passive domain lookup data (not necessarily malicious addresses, but a good reference to have along side your intelligence).
- Locally discovered intel (potentially all of the above)

OUR PROBLEM(S)

- Data Normalization (format, confidence, severity, etc).
- Largely diverse (and usually large) data-sets
- data is “living”, it’s only as fresh as your last record or trend. (as the insert() completes, it’s already become stale, regardless if you’ve updated the “lastUpdated” column).
- Even within similar data-sets, some intel may become stale more quickly than others (scanners vs botnet C&C’s)
- ultimately data is from PEOPLE (eg: human beings). Whether it’s a sensor that was programmed by someone with a bias towards something, or a forensics investigation. We must interpret that data from their context to our application EVERY TIME before we can make use of it.
- search vs feeds and distinguishing the difference (presentation)
- i can has API? (application integration, reaching an intelligence driven infrastructure)

FRAMEWORK

- Takes data from public and private sources, pre-processes it, normalizes it down to your favorite standard (eg: IDMEF, IODEF, ICSG, json keypairs, etc...) and stores in along side it's counter part data points.
- Malware metadata is stored along side suspicious networks data.
- Malicious Domains data is stored along side phishing url data.
- The main intelligence stream warehouses everything in blob's and uses 'cookie cutter' style index partitions (eg: regular tables) to be derived from the specific parts of the data worth using in analytics / mitigation's.

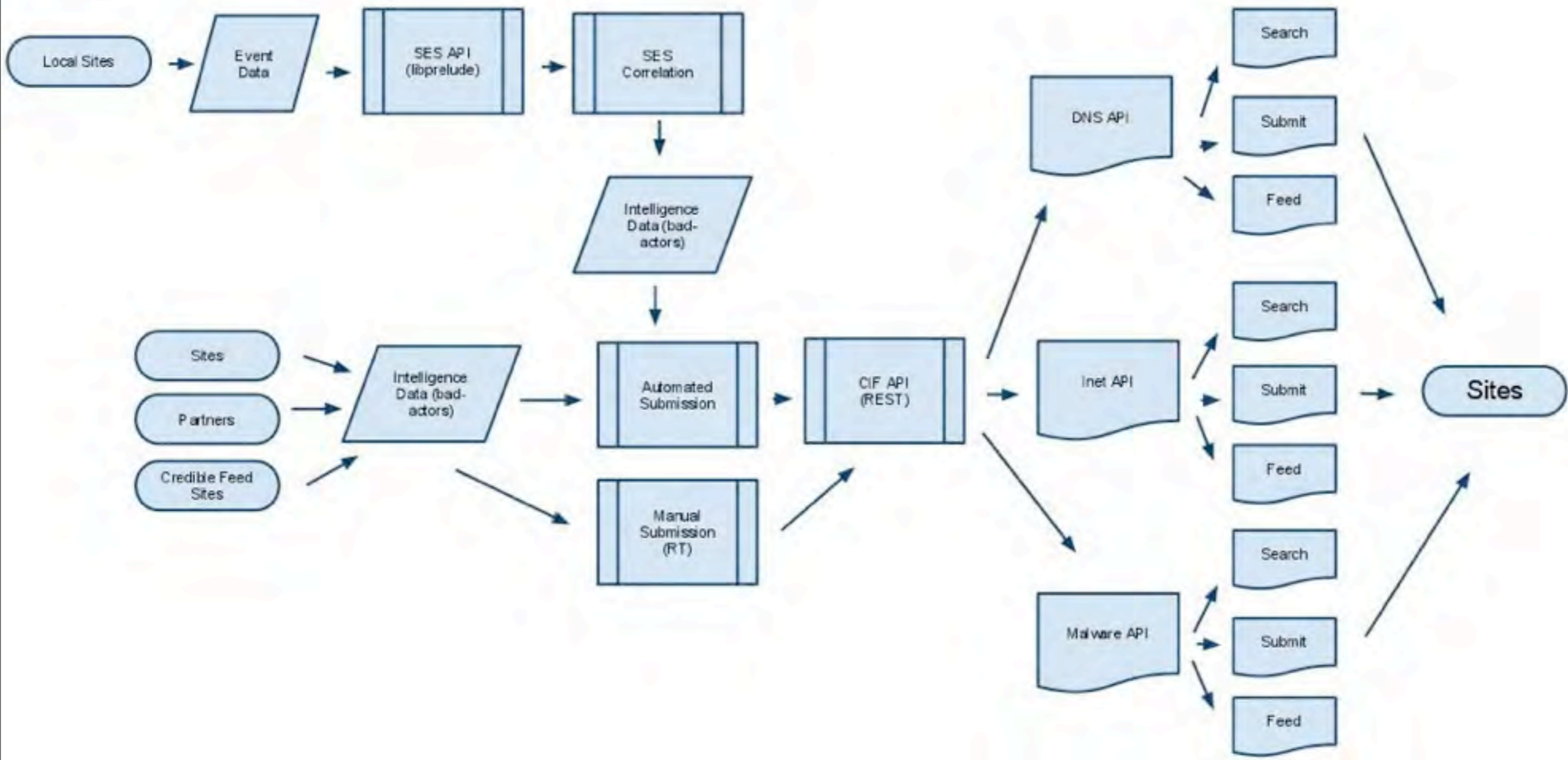


WAREHOUSING

- schema-less data
- store anything and everything (xml, plain-text, binary blobs, etc).
- If you wanna add / remove something, just alter the table (no index locking issues).
- structure what you can (eg: xpath searches as a last resort), even if it's a simple key-pair. (hint: standards help document the data, but isn't required). If you **must** do a raw data search, you at-least know the xpath query.
- creating mitigation lists "on the fly" (time is expensive, do it the first time and in a distributed manner to spread out the processing costs across nodes). Leave aggregation at the end when you're ready to generate the feeds.
- unstructured message integration (sometimes good intel is in e-mail form)
- partitioning, long-term storage.

OPS

- There are lots of things Security Event Management got right.
- lots of diverse input's
- distributed normalization processing
- correlation (pre-reputation)
- centralized, peer-able repositories
- A SEM accelerates when it is highly specialized at handling security events
- Intelligence is a type of event. It's a streaming event. A known measurement at a specific point in time
- Take the "event specialization" and make it a component, not the main driver.
- Instead of correlating purely on event data, correlate on rolling reputation data.



PROJECT INFO

- [code.google.com / p / collective-intelligence-framework /](http://code.google.com/p/collective-intelligence-framework/)
- [www.ren-isac.net / ses](http://www.ren-isac.net/ses)