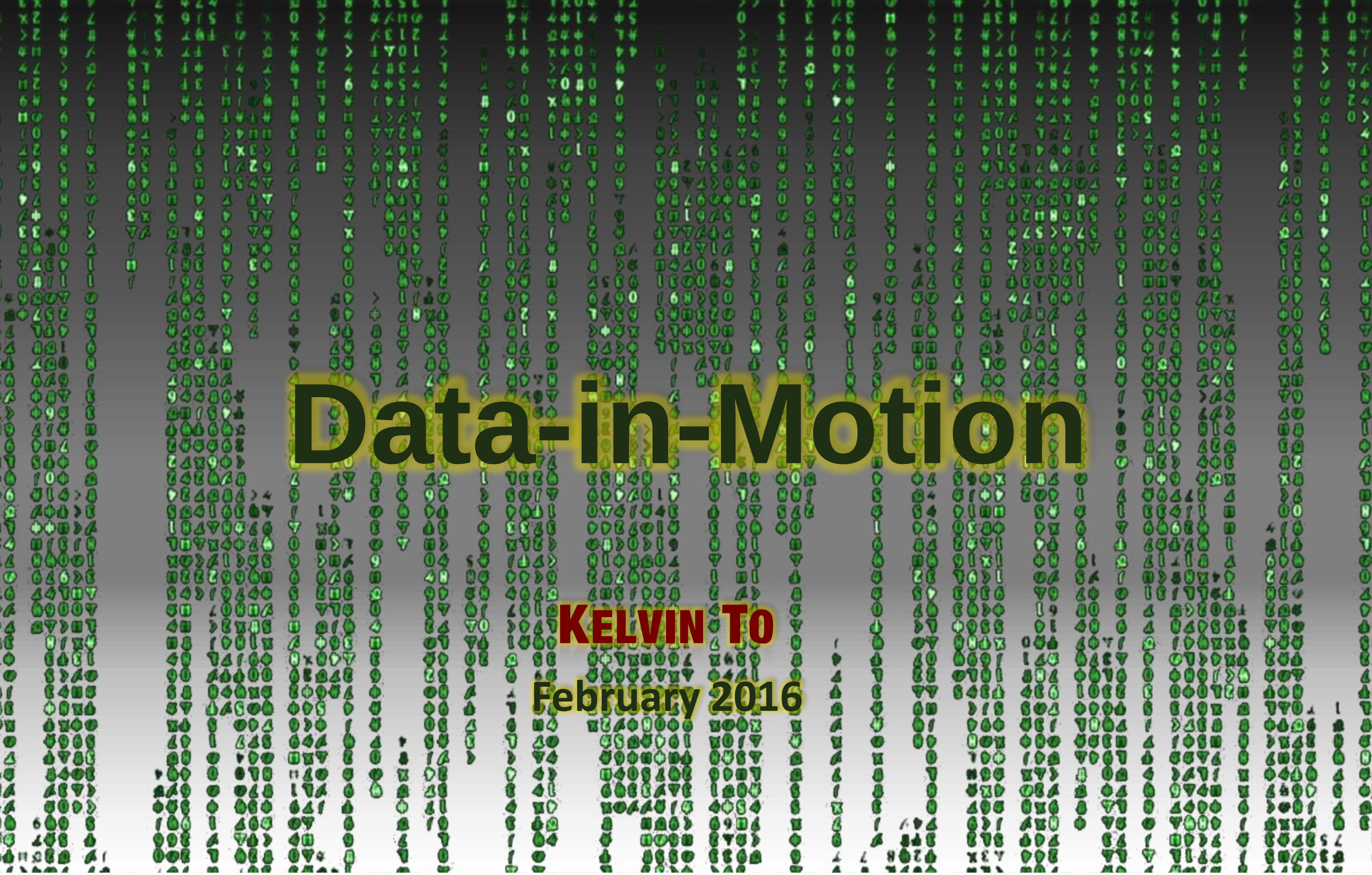


Data-in-Motion

Kelvin To

February 2016

Bio and Introduction

http://www.databoiler.com

Kelvin To

Founder and President

Data Boiler Technologies, LLC

Big Data | Big Picture | Big Opportunities

^ **Simulation / Data Science** to support your enduring business strategies	✓ **WorkPlace Organization** to ease the related reporting & administrative burden
+ **Datafication / Text Mining** to pluck diamonds from semi-structured or unstructured data	− **Value Chain Management** to align vendors / distributors to work cohesively together
× **Infomediaries / Aggregation** to cross tabulate data for higher value purposes	÷ **I.T. Agility Optimization** to enable big data success with a lean & scalable architecture

FSR BITS • • •   212° DATA BOILER TECHNOLOGIES, LLC

Table of Contents

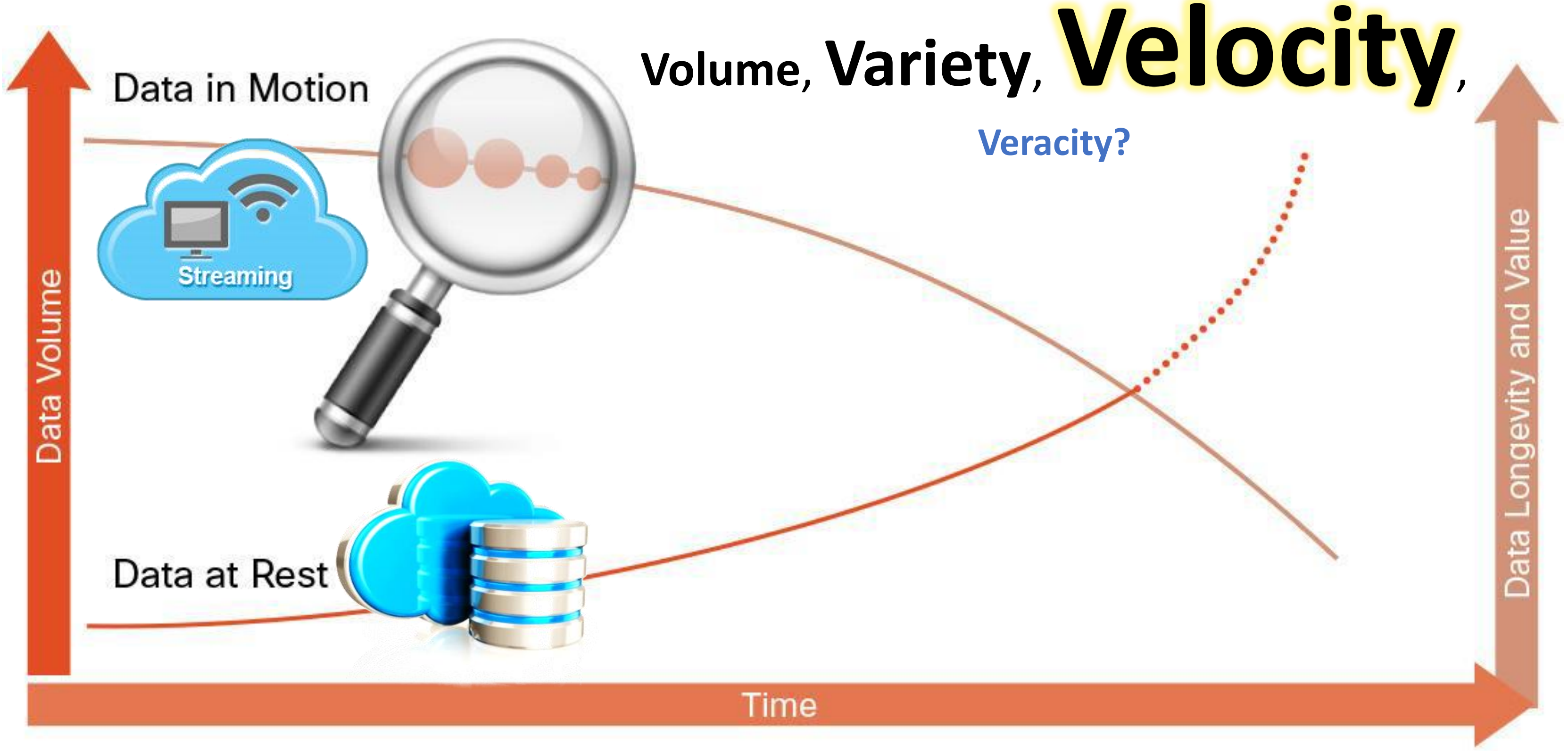
- What it is? DIM Characteristics and 4Vs
- Why you should care? Values and Relevance
- What are the challenges? Extraction, Speed, Balance
- Where to start? Roadmap
- How to harness it? Locate DIM
- In-memory analytics (Memory Forensics)
- Distributed environment (CAP theorem / ACID)
- DLP infrastructure (In Motion, At Rest, In Use, Disposed)
- Encryption isn’t everything (Eradication, Obfuscation)
- Proprietary techniques versus sharing (Security Compliance)
- The best is yet to come: (generating BIs vs privacy hazard)

Data-in-Motion: *Characteristics*

Easier to capture than Data-at-Rest, but its value diminishes with time

Volume, **Variety**, **Velocity**,

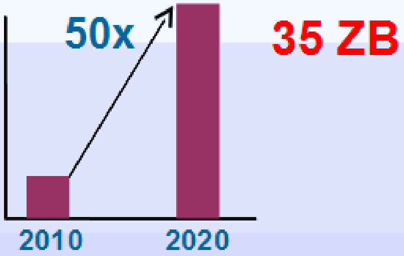
Veracity?



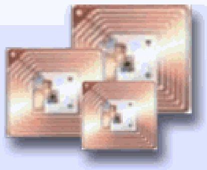
Source: CISCO http://unleashingit.com/docs/B13/IoE%20Data%20Motion/increase_the_value_relevance_of_data_in_motion.pdf

Data-in-Motion: *4Vs*

Cost efficiently processing the growing **Volume**

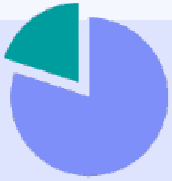


Responding to the increasing **Velocity**



30 Billion RFID sensors and counting

Collectively analyzing the broadening **Variety**



80% of the worlds data is unstructured

Establishing the **Veracity** of big data sources

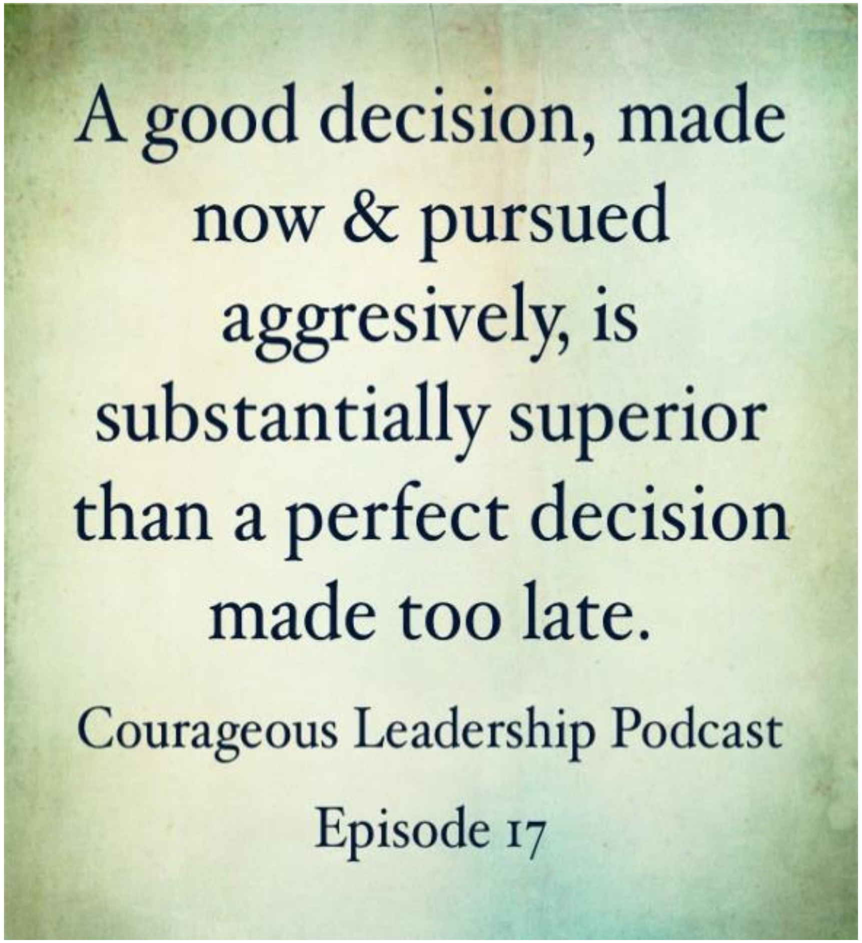
1 in 3 business leaders don't trust the information they use to make decisions

Source: https://www.ibm.com/developerworks/community/blogs/5things/entry/5_things_to_know_about_big_data_in_motion?lang=en

Data-in-Motion: *Values & Relevance*

- Mass Customization/ Deeper Insights
- Self Selection / Price Discrimination
- Interactive, Relevant Experiences
- Risk and Compliance

⇒ **Faster and Better Decision**

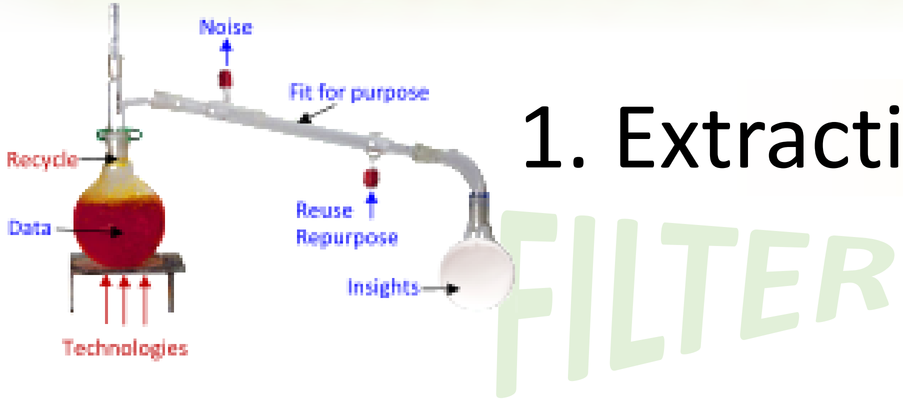


A good decision, made now & pursued aggresively, is substantially superior than a perfect decision made too late.

Courageous Leadership Podcast

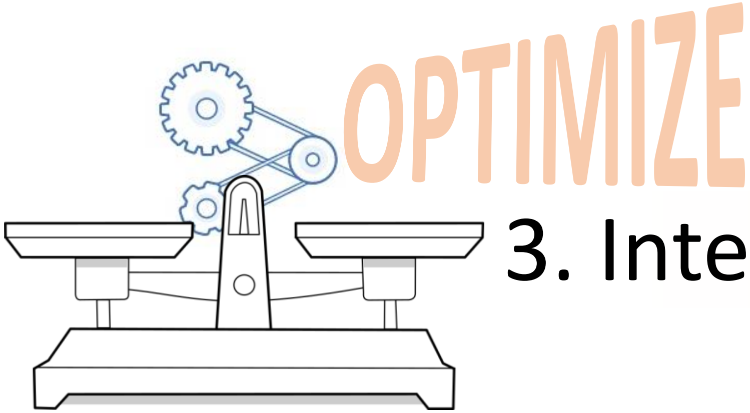
Episode 17

Data-in-Motion: *Challenges*



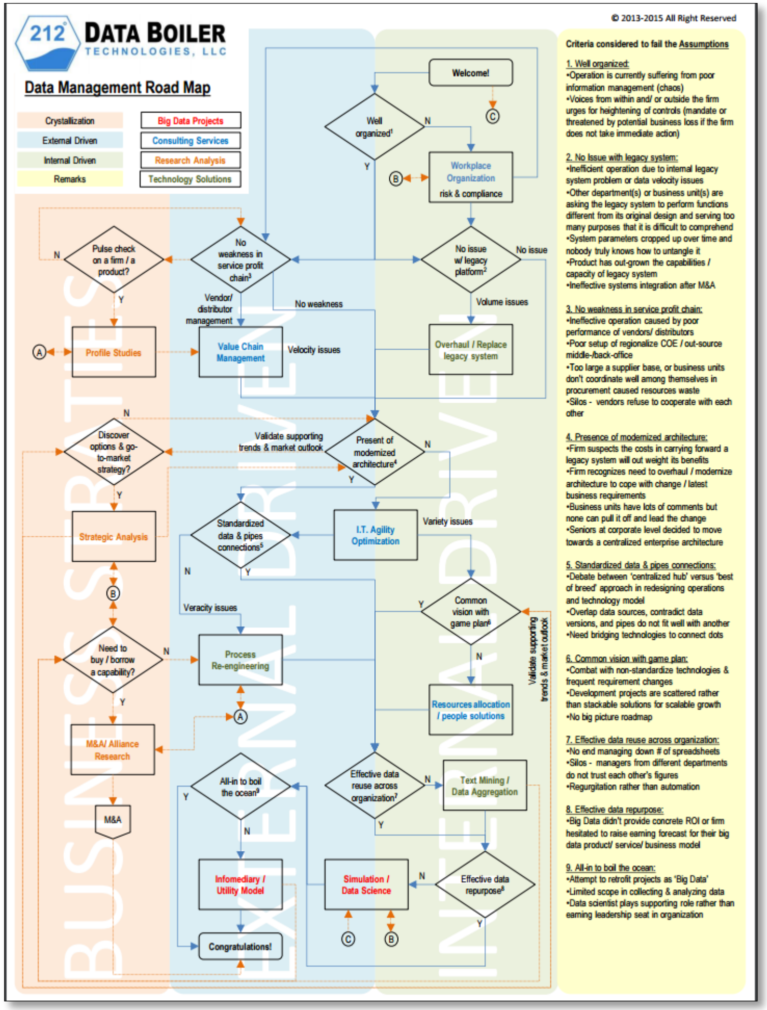
1. Extraction (Mining) / Discovery

2. OLTP (Process)/ RTAP (Analyze)



3. Integrate / Privacy & Security

Data-in-Motion: *Getting Started*


212° DATA BOILER TECHNOLOGIES, LLC
Data Management Road Map
Crystallization
Big Data Projects
External Driven
Consulting Services
Internal Driven
Research Analysis
Remarks
Technology Solutions
BUSINESS STRATEGIES
EXTERNAL DRIVEN
INTERNAL DRIVEN
Welcome!
C
Well organized¹
N
Y
Workplace Organization
risk & compliance
B
No issue w/ legacy platform²
No issue
Volume issues
Overhaul / Replace legacy system
Pulse check on a firm / a product?
N
Y
No weakness in service profit chain³
Vendor/ distributor management
No weakness
A
Profile Studies
Value Chain Management
Velocity issues
N
Discover options & go-to-market strategy?
Y
Validate supporting trends & market outlook
Present of modernized architecture⁴
N
Y
Strategic Analysis
Standardized data & pipes connections⁵
I.T. Agility Optimization
Variety issues
N
Y
B
Veracity issues
Common vision with game plan⁶
Y
N
Validate supporting trends & market outlook
Need to buy / borrow a capability?
N
Y
Process Re-engineering
A
Resources allocation / people solutions
M&A/ Alliance Research
All-in to boil the ocean⁹
Y
N
Effective data reuse across organization⁷
N
Y
Text Mining / Data Aggregation
M&A
Infomediary / Utility Model
Simulation / Data Science
N
Effective data repurpose⁸
Y
C
B
Congratulations!

© 2013-2015 All Right Reserved

Criteria considered to fail the Assumptions

1. Well organized:
•Operation is currently suffering from poor information management (chaos)
•Voices from within and/ or outside the firm urges for heightening of controls (mandate or threatened by potential business loss if the firm does not take immediate action)

2. No Issue with legacy system:
•Inefficient operation due to internal legacy system problem or data velocity issues
•Other department(s) or business unit(s) are asking the legacy system to perform functions different from its original design and serving too many purposes that it is difficult to comprehend
•System parameters cropped up over time and nobody truly knows how to untangle it
•Product has out-grown the capabilities / capacity of legacy system
•Ineffective systems integration after M&A

3. No weakness in service profit chain:
•Ineffective operation caused by poor performance of vendors/ distributors
•Poor setup of regionalize COE / out-source middle-/back-office
•Too large a supplier base, or business units don't coordinate well among themselves in procurement caused resources waste
•Silos - vendors refuse to cooperate with each other

4. Presence of modernized architecture:
•Firm suspects the costs in carrying forward a legacy system will out weight its benefits
•Firm recognizes need to overhaul / modernize architecture to cope with change / latest business requirements
•Business units have lots of comments but none can pull it off and lead the change
•Seniors at corporate level decided to move towards a centralized enterprise architecture

5. Standardized data & pipes connections:
•Debate between 'centralized hub' versus 'best of breed' approach in redesigning operations and technology model
•Overlap data sources, contradict data versions, and pipes do not fit well with another
•Need bridging technologies to connect dots

6. Common vision with game plan:
•Combat with non-standardize technologies & frequent requirement changes
•Development projects are scattered rather than stackable solutions for scalable growth
•No big picture roadmap

7. Effective data reuse across organization:
•No end managing down # of spreadsheets
•Silos - managers from different departments do not trust each other's figures
•Regurgitation rather than automation

8. Effective data repurpose:
•Big Data didn't provide concrete ROI or firm hesitated to raise earning forecast for their big data product/ service/ business model

9. All-in to boil the ocean:
•Attempt to retrofit projects as 'Big Data'
•Limited scope in collecting & analyzing data
•Data scientist plays supporting role rather than earning leadership seat in organization

1. How well is your workplace organized?
2. Any issue with legacy systems, any weakness in the service profit chain?
3. Presence of modernized architecture?
4. Nature of data and pipes connections?
5. Common vision with game plan?
6. Data re-use / repurpose?
7. "All-in" to boil the ocean?

Download:
http://www.databoiler.com/index_htm_files/DataBoiler%20RoadMap.pdf

Data-in-Motion: *Harnessing*

"Locate DIM

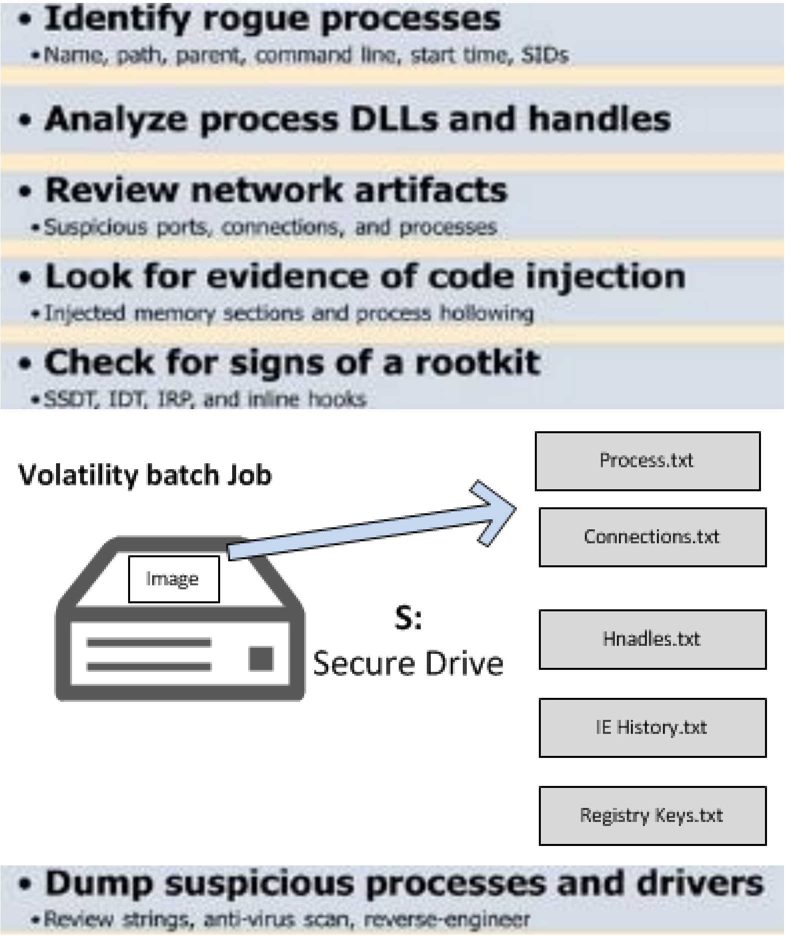
- Assessing the data trajectory
- Gaining visibility into the network traffic itself
- Determining whether certain network devices are storing sensitive data or related information
- Inspecting specific gateway devices such as mail servers and proxies"

Source: SANS Institute
https://www.sans.org/reading-room/whitepapers/analyst/regulations-standards-encryption-applies-34675

In-Memory Analytics: *Forensic*

- Querying data stream directly in RAM instead of storage media to allow BI to support faster decisions.
- Disk I/O bottlenecks and related CPU-intensive activities are either eliminated or moved to DRAM, graphics memory.
- Tuning of data structures to boost performance between analytics vs write access speed.

Forensic analysis of Malware

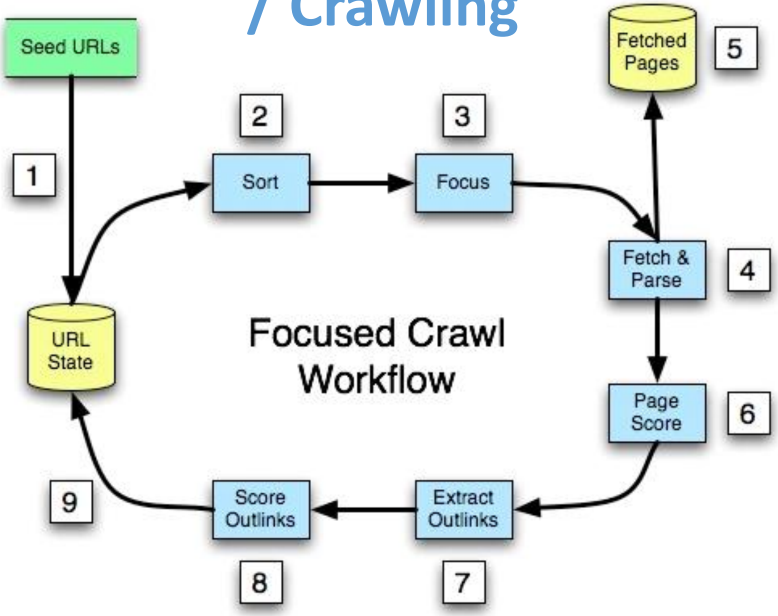


- Identify rogue processes
  - Name, path, parent, command line, start time, SIDs
- Analyze process DLLs and handles
- Review network artifacts
  - Suspicious ports, connections, and processes
- Look for evidence of code injection
  - Injected memory sections and process hollowing
- Check for signs of a rootkit
  - SSDT, IDT, IRP, and inline hooks

- Dump suspicious processes and drivers
  - Review strings, anti-virus scan, reverse-engineer

DIM: *Distributed Environment*

(A) Internet Data Extraction / Crawling



(B) Integrate/ Aggregate:

- Enables users to develop and repurpose applications
- Rapidly process and analyze information that is in-coming from thousands of real-time sources.
- Address the issue of data only being able to 'add' to HDFS files but 'not change' in a Hadoop batch processing environment.

(C) Managed File Transfer:

- Incorporates higher level of security, scalability, integration, reporting to bring order, predictability and security to file movement.

CAP theorem / ACID compliance

- Consistency
- Availability
- Partition tolerance

A distributed system cannot satisfy all three of these guarantees at the same time

- Atomicity
- Consistency
- Isolation
- Durability

Set of properties that guarantee that database transaction are processed reliably

DLP Infrastructure

Data	Typical protection methods, but not all
In Motion	Use SSL/TLS or IPsec and HAIPE encryption of the data transmitted over the network , Layer 2 “100Gbps” & “ESS” requirements
At Rest	Incorporate encryption by the Storage Area Network (SAN) Protect data at rest from access by a rogue host or from physical *Note: virtual disk appears unencrypted to the server, and administrator with full access rights could see the unencrypted data even if s/he is not authorized to do so. Hence, n/a for DIU*
In Use	Because DIU has to be decrypted before it can be used, therefore tokenization is needed to encrypt more narrowly, or use database encryption to mitigate full exposure of data (i.e. restrict usage to specific range of data fields and/ or records)
Disposed	Use a degausser to thoroughly wiped device before it is discarded

DIM: *Beyond Encryption*

Eradication: (e.g. automated key zeroing techniques)

- removal of data as soon as it has been transmitted/ used

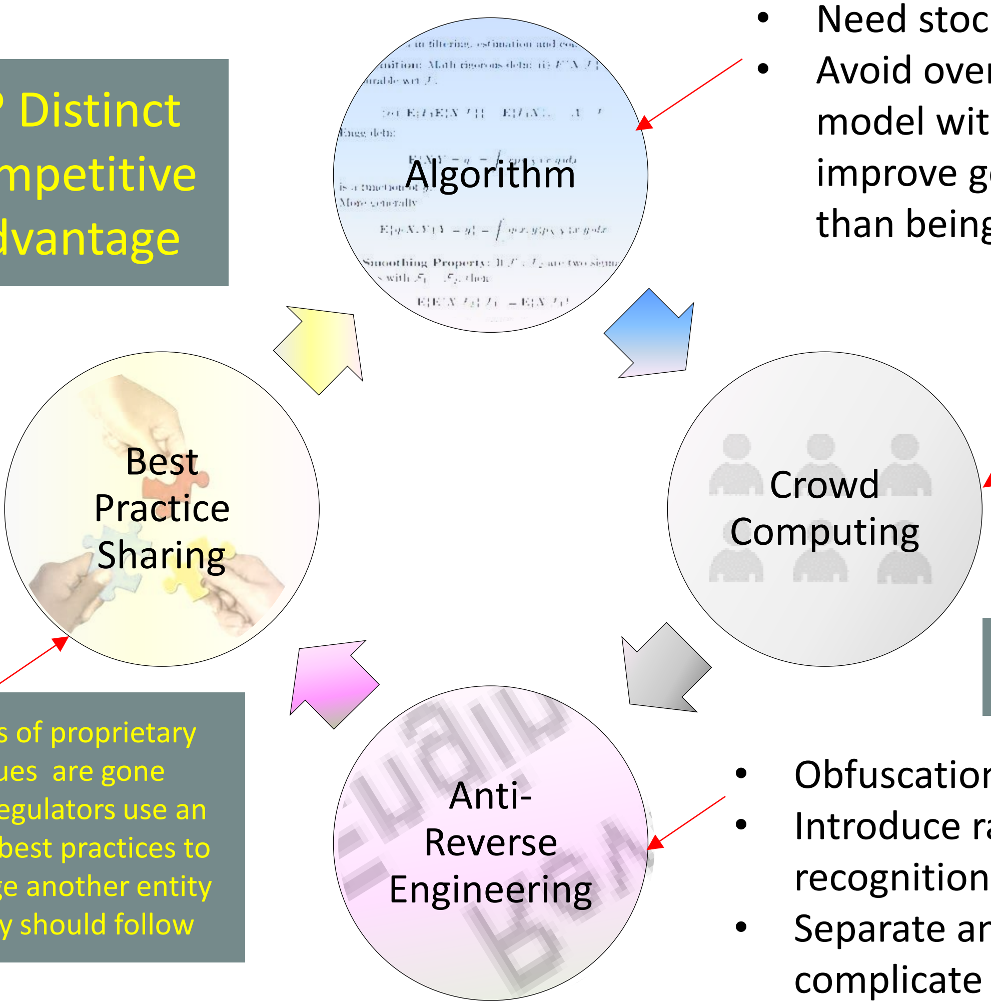
Obfuscation: (e.g. traffic flow security to mask patterns, de-identification)

- "This involves modifying the stored format of data so that it is not easily readable or accessible. This option is often employed for transaction-related information such as credit card numbers. *For example, full payment card tracking data may be sent to a processor at the time of use, but the merchant may only store the last four digits of the card number and obfuscate the rest.*" sans.org

Bare-metal cloud, size of counter for scalability, secure in-band / out-band management, certificate generation, hardware tempering controls, …

Proprietary Techniques versus Sharing

?? Distinct competitive advantage



Algorithm

Crowd Computing

Anti-Reverse Engineering

Best Practice Sharing

- Need stochastic estimation control
- Avoid over-fitting / under-fitting the model with data (take out outliers to improve general predictive power rather than being distorted)

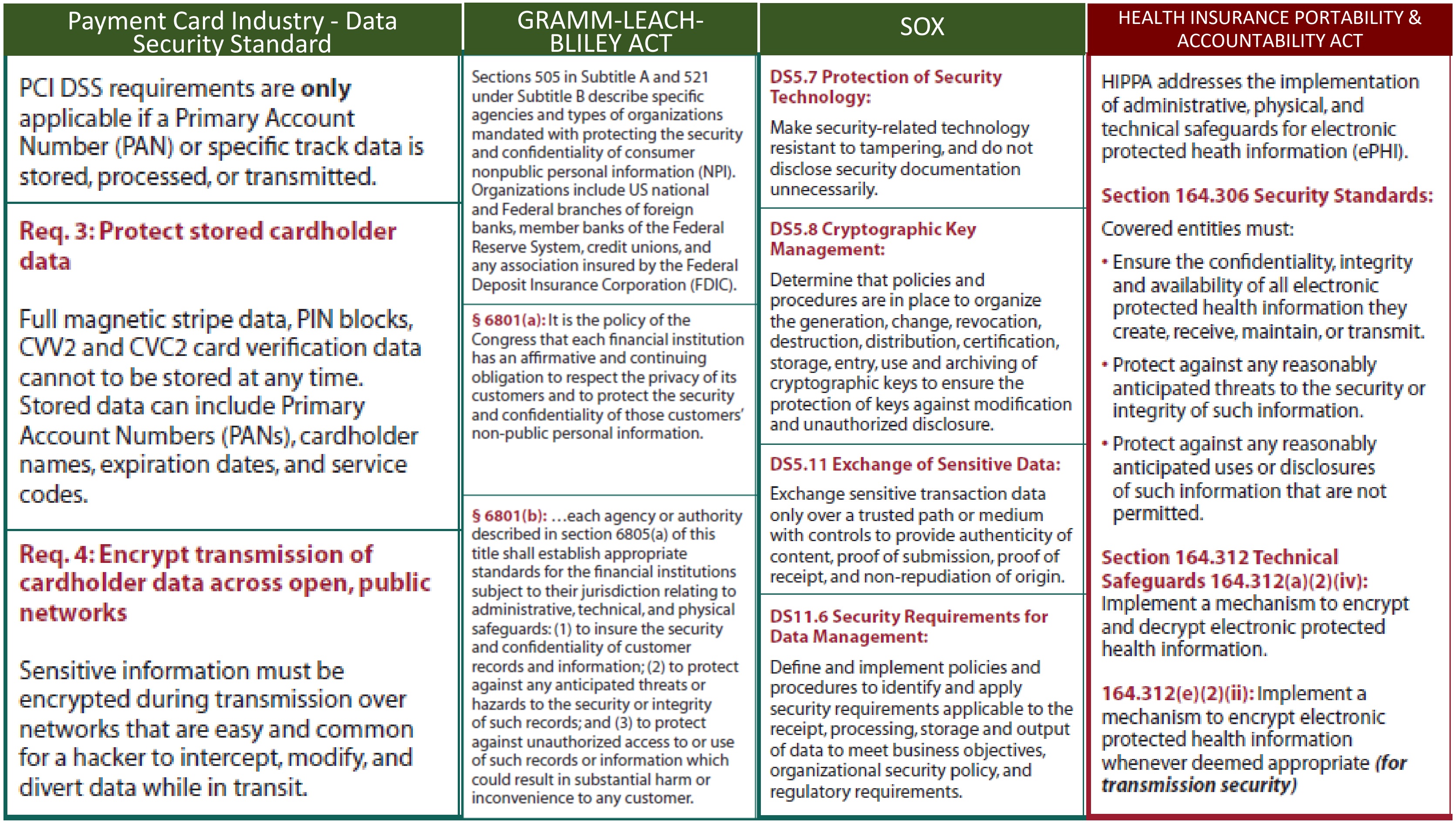
- Crowd collective intelligence addresses limitations with any standalone algorithm
- Dynamic upgrades to make system smarter than ever

Benefit from the crowd versus making it meaningless for others to avoid front-running of trade strategies

- Obfuscation to make it incompatible
- Introduce randomness to resist pattern recognition techniques
- Separate and scramble to make it more complicate or aggregate rollup to simplify it

The days of proprietary techniques are gone where regulators use an entity's best practices to challenge another entity that they should follow

Security Compliance



Payment Card Industry - Data Security Standard	GRAMM-LEACH-BLILEY ACT	SOX	HEALTH INSURANCE PORTABILITY & ACCOUNTABILITY ACT
PCI DSS requirements are **only** applicable if a Primary Account Number (PAN) or specific track data is stored, processed, or transmitted. **Req. 3: Protect stored cardholder data** Full magnetic stripe data, PIN blocks, CVV2 and CVC2 card verification data cannot to be stored at any time. Stored data can include Primary Account Numbers (PANs), cardholder names, expiration dates, and service codes. **Req. 4: Encrypt transmission of cardholder data across open, public networks** Sensitive information must be encrypted during transmission over networks that are easy and common for a hacker to intercept, modify, and divert data while in transit.	Sections 505 in Subtitle A and 521 under Subtitle B describe specific agencies and types of organizations mandated with protecting the security and confidentiality of consumer nonpublic personal information (NPI). Organizations include US national and Federal branches of foreign banks, member banks of the Federal Reserve System, credit unions, and any association insured by the Federal Deposit Insurance Corporation (FDIC). **§ 6801(a):** It is the policy of the Congress that each financial institution has an affirmative and continuing obligation to respect the privacy of its customers and to protect the security and confidentiality of those customers' non-public personal information. **§ 6801(b):** …each agency or authority described in section 6805(a) of this title shall establish appropriate standards for the financial institutions subject to their jurisdiction relating to administrative, technical, and physical safeguards: (1) to insure the security and confidentiality of customer records and information; (2) to protect against any anticipated threats or hazards to the security or integrity of such records; and (3) to protect against unauthorized access to or use of such records or information which could result in substantial harm or inconvenience to any customer.	**DS5.7 Protection of Security Technology:** Make security-related technology resistant to tampering, and do not disclose security documentation unnecessarily. **DS5.8 Cryptographic Key Management:** Determine that policies and procedures are in place to organize the generation, change, revocation, destruction, distribution, certification, storage, entry, use and archiving of cryptographic keys to ensure the protection of keys against modification and unauthorized disclosure. **DS5.11 Exchange of Sensitive Data:** Exchange sensitive transaction data only over a trusted path or medium with controls to provide authenticity of content, proof of submission, proof of receipt, and non-repudiation of origin. **DS11.6 Security Requirements for Data Management:** Define and implement policies and procedures to identify and apply security requirements applicable to the receipt, processing, storage and output of data to meet business objectives, organizational security policy, and regulatory requirements.	HIPPA addresses the implementation of administrative, physical, and technical safeguards for electronic protected heath information (ePHI). **Section 164.306 Security Standards:** Covered entities must: • Ensure the confidentiality, integrity and availability of all electronic protected health information they create, receive, maintain, or transmit. • Protect against any reasonably anticipated threats to the security or integrity of such information. • Protect against any reasonably anticipated uses or disclosures of such information that are not permitted. **Section 164.312 Technical Safeguards 164.312(a)(2)(iv):** Implement a mechanism to encrypt and decrypt electronic protected health information. **164.312(e)(2)(ii):** Implement a mechanism to encrypt electronic protected health information whenever deemed appropriate ***(for transmission security)***

Source: https://www.sans.org/reading-room/whitepapers/analyst/regulations-standards-encryption-applies-34675

Cybersecurity Information Sharing Act

- Agencies must continue to educate stakeholders to improve preparedness, while creating a blueprint for sharing defensive measures and cyber threat indicators between the government and other entities and establish consensus-based voluntary best practices between the agencies to improve security and reduce cyber threats.
- CISA protects the liability of private sector entities when sharing and receiving cyber threat information. It also establishes the personal data that needs to be removed before data sharing can occur and how quickly individuals must be notified their information was shared.

Source: https://www.congress.gov/bill/114th-congress/senate-bill/754
https://www.sec.gov/investment/im-guidance-2015-02.pdf

Generating BIs versus Privacy Hazard

- https://www.fdic.gov/regulations/examinations/financialprivacy/handbook/

- "**Privacy** is an important challenge facing the growth of the connected Web and the propagation of various transaction models supported by it. Decentralized distributed models of computing are used to mitigate privacy breaches by eliminating a single point of failure. However, end-users can still be attacked in order to discover their private information."

An empirical research

- *The financial industry wants to retain accuracy in generating Business Intelligence from DIM/ big data analytics, while they need to keep relatively large parts of the consumer profile obfuscated.*
- *How to enable accurate collaborative filtering recommendations, while reducing privacy hazard?*

Stay tuned for future session …