RSAConference2016

San Francisco | February 29 – March 4 | Moscone Center

Connect to Protect

SESSION ID: PDAC-W02

Applying Auto-Data Classification Techniques for Large Data Sets

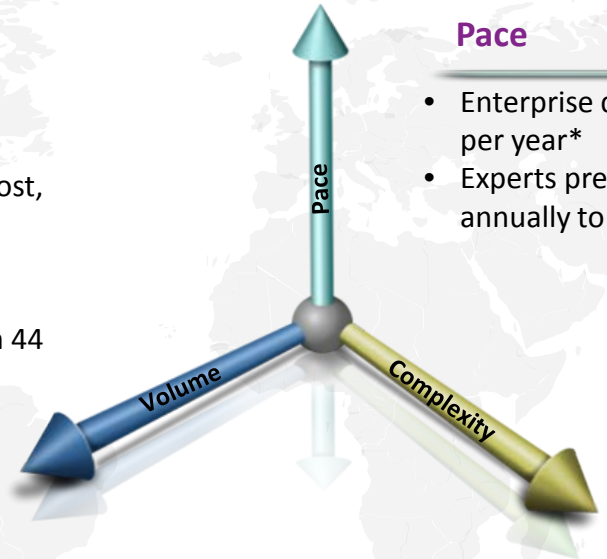
Anchit Arora
Program Manager
InfoSec, Cisco

#RSAC

The proliferation of data and increase in complexity

Volume

- Big data architectures, low storage cost, Increase of data retention
- 80% of data generated today is unstructured
- Data generated worldwide will reach 44 zettabytes by 2020*



Pace

- Enterprise data collection to increase 40 to 60 % per year*
- Experts predict the amount of data generated annually to increase 4300% by 2020 *

Complexity

- Complex work models: always accessible, remote & mobile workers
- Definition of perimeter: Cloud, Customer & partners
- Users choose devices (BYOD)

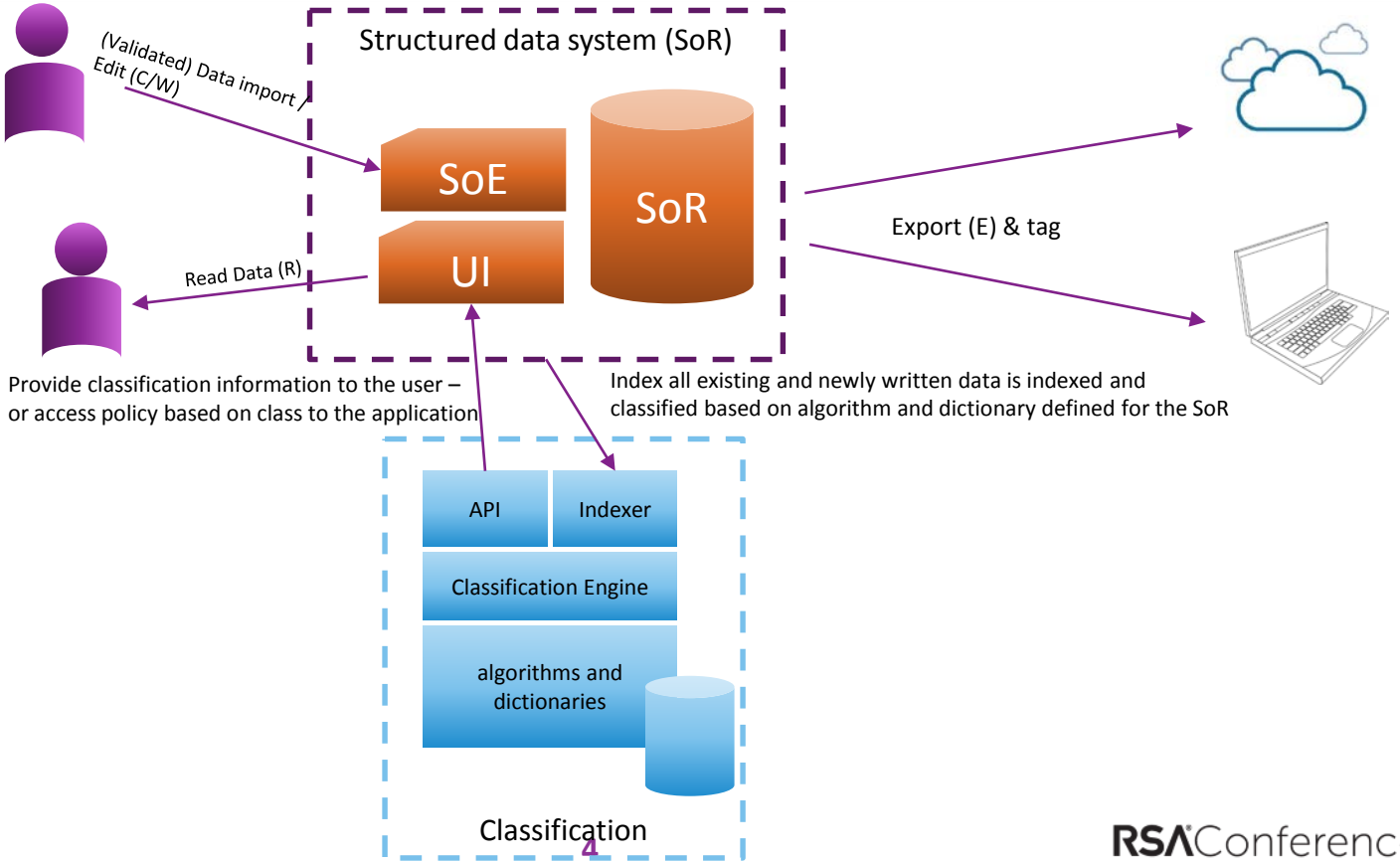
cisco

* Numbers and statistics from Gartner, Gigaom Research, CSC, Seagate

Auto-classification: The why and what

- Desired business outcome: At Cisco we want to provide additional sensitivity context to structured and unstructured data, to be able to apply controls more effectively
- Scope: Our aim is to have an automated classification capability for all structured data systems, and provide capability to better govern/control generation of unstructured data which is created as a result of export from structured data systems using label/field association to each record set

Use-case: From structured to unstructured



An unstructured data use-case: box.com

- Box.com is an external cloud platform used by Cisco for collaboration and storage of data
- Security questions to ask:
  - What is this data?
  - What's the source of the data?
  - Who owns this data?
  - What's the sensitivity of the data?
  - Is all data equally sensitive (this is the essence for optimal security)?
  - What's the level of security required?

Should we ask the user to govern security?

- Can we expect the user to make the right security decision with all this complexity involved in decision making?
  - The user needs to be very knowledgable to make the right decision
- The answer is No: But however many systems are designed to have users govern security -
  - Recognize data categories in systems with unstructured data
  - Classify data in any data system
  - Set data securitypolicy
  - Securely export data out of the system
- Making the shift from user governed to data owner governed

How to make the shift to a data owner model?

Governance of Data by Data Owner

Data Management

Data Taxonomy	Recognize Data Type	Classify Sensitivity	Tag	Policy Enforcement

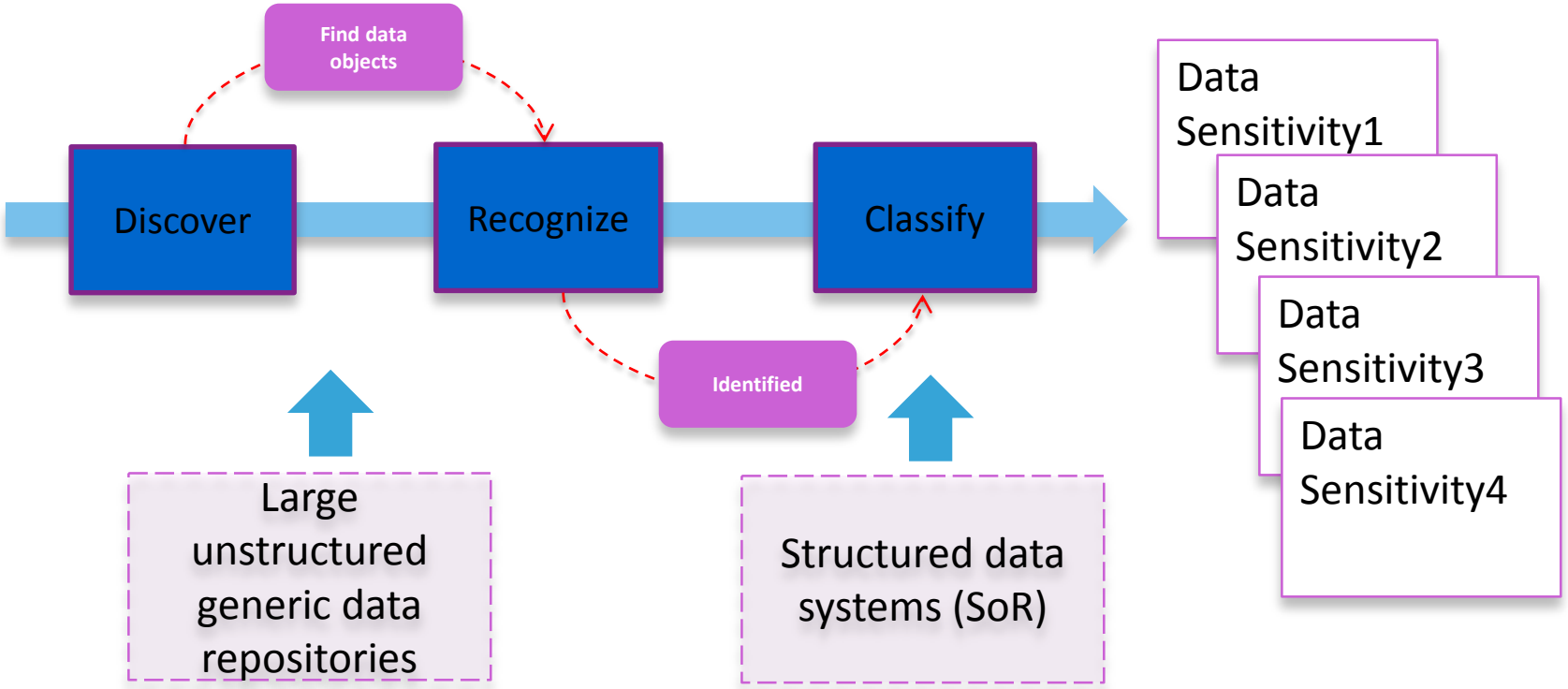
Governance of Data by End User

Data Intelligence & monitoring capabilities

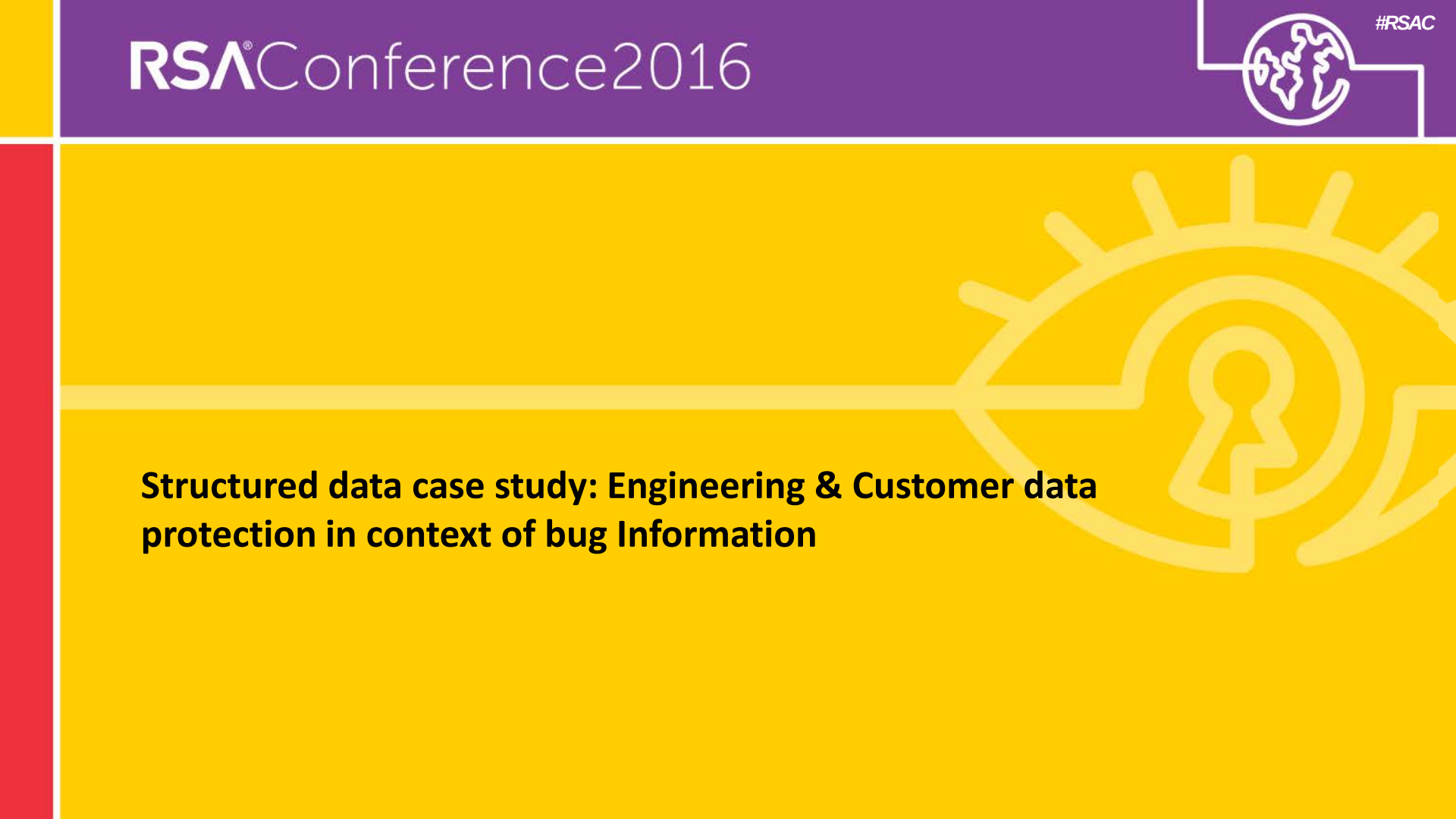
Data Protection capabilities

Across various data types: Engineering, Customer, Finance, HR

Conceptual approach



Classification mostly unknown



Structured data case study: Engineering & Customer data protection in context of bug Information

A case study: Bug information

Millions of bugs + product bugs, 3 approaches available to protect:

1. Treat all bugs equally, and apply ‘very strict’ controls on all bugs
   - In heterogenic data models , most data is ‘Over’-protected
   - Limits business ability and User experience
2. Treat all bugs equally, and apply ‘loose’ controls on all bugs
   - Results in ‘Under’-protected data
3. Apply the right amount of protection on a bug, based on sensitivity
   - Balanced security and cost applied – just the right amount of security!

Setting the foundation for auto-class

Inventory Process

Identify

- Identify the most sensitive IP and IP's appropriate owner(s)

Define

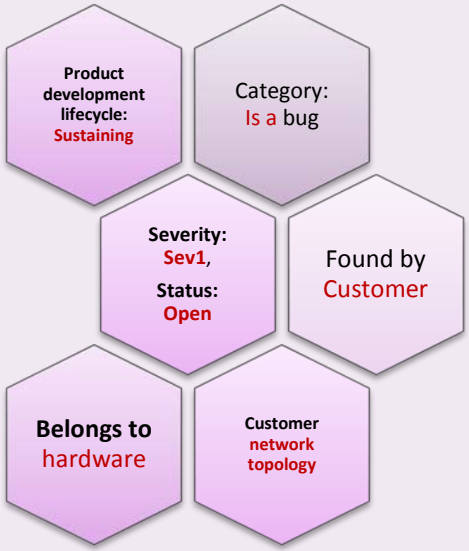
- Define data use and access rules for the most sensitive IP

Translate

- Translate rules into IT enforceable policies

The inventory process engages the business to build out the data taxonomy and a model of the sensitivity

A Sensitive software bug in CDETS



The proof is in the numbers!!

Manual approach

Parameter	Value
Average time to classify a single bug	5 minutes
Total number of bugs	7 Million
Time to classify	**35 Million minutes**
Cost/min of SME analyst	$ 0.83/Min
Cost to classify	**$ 29 Million**

Additional costs to consider for manual:

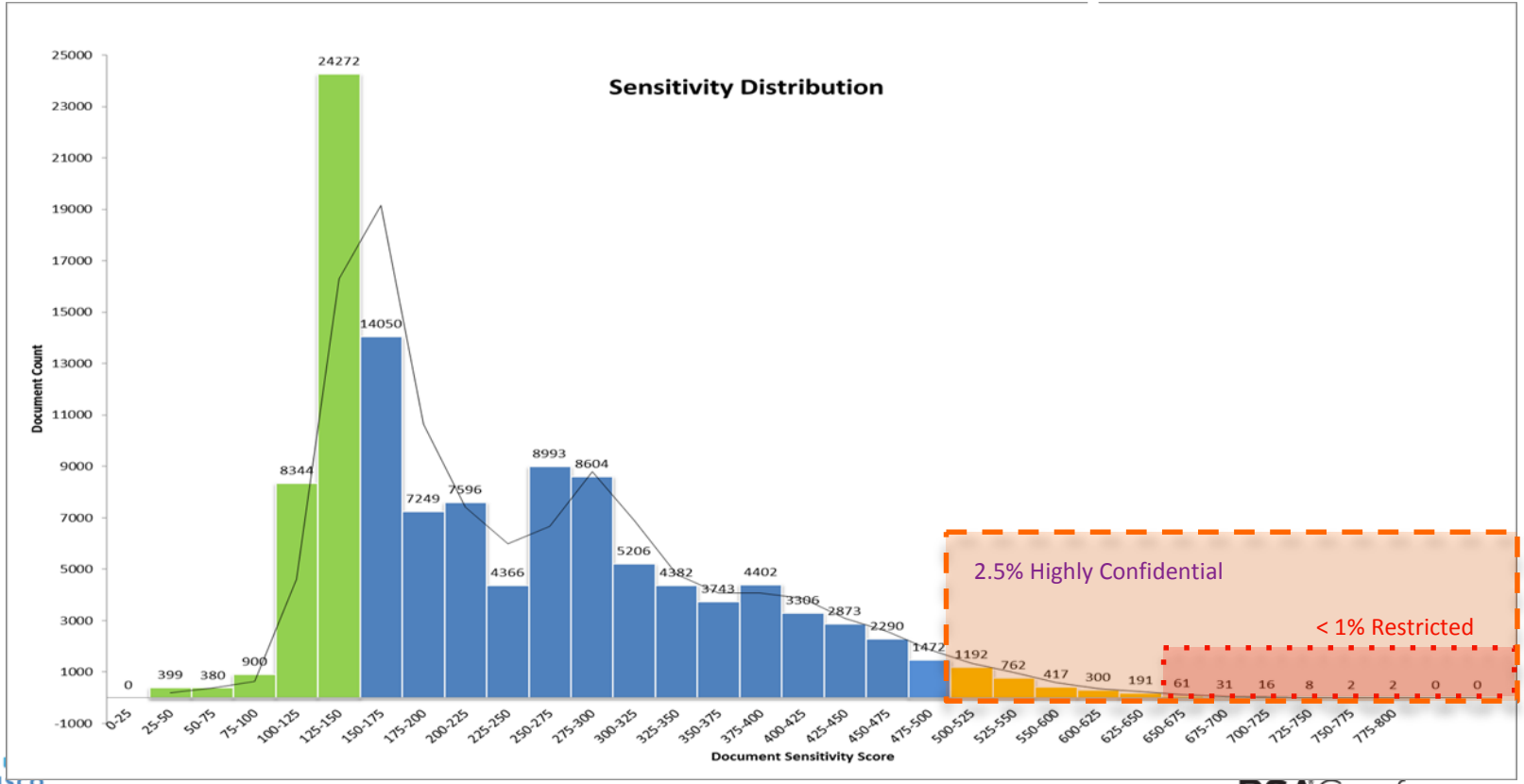
Training: For consistent user behavior
Change to business: Cleaning legacy
Change to applications and Infrastructure

Auto-Classification approach

Parameter	Value
Average time to classify a single bug*	0.002 minutes
Total number of bugs	7 Million
Time to classify	**14,000 Minutes**
Estimated cost for Infrastructure and resources required to classify	**$ 0.25 Million**

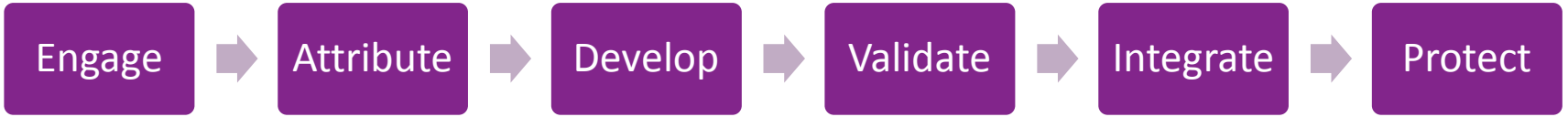
Accuracy Results
83%

The most sensitive data is just a small portion



How did we execute the methodology?

A 6 step workflow, for structured data (SoR)



#	Phase	Scope
1	Engage	Identify SoR and engage stakeholders to communicate expectations, R&R, Identify data workflow (user stories) and data categories. Plan and establish scope and planning of the SoR integration
2	Attribute	Analysis of data, database fields, record and build a data sensitivity model / algorithm to be able to classify the data
3	Develop	Development of attribution and scoring algorithm into the classification engine and perform indexing of datasets
4	Validate	Validation and tuning of classification results of the classification engine to ensure accuracy of the output
5	Integrate	Integration of classification data with the source system
6	Protect	Planning and implementation of protective measures in the source system for sensitive data classes

Building an attribution model

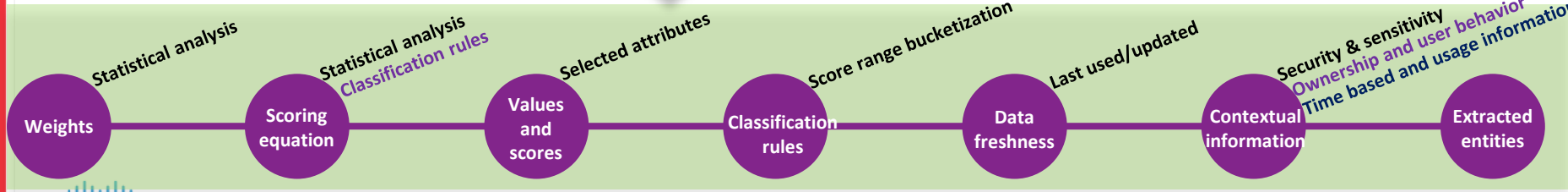
Attribute A, Attribute B, Attribute C ……
Attribute L, Attribute M, Attribute N……
Attribute X, Attribute Y, Attribute Z……

All available source system built-in attributes

Selected attributes and values

Extracted entities from free-text fields and attachments:

Attribution model

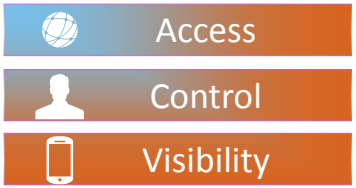


How to create a similar solution for your organization?

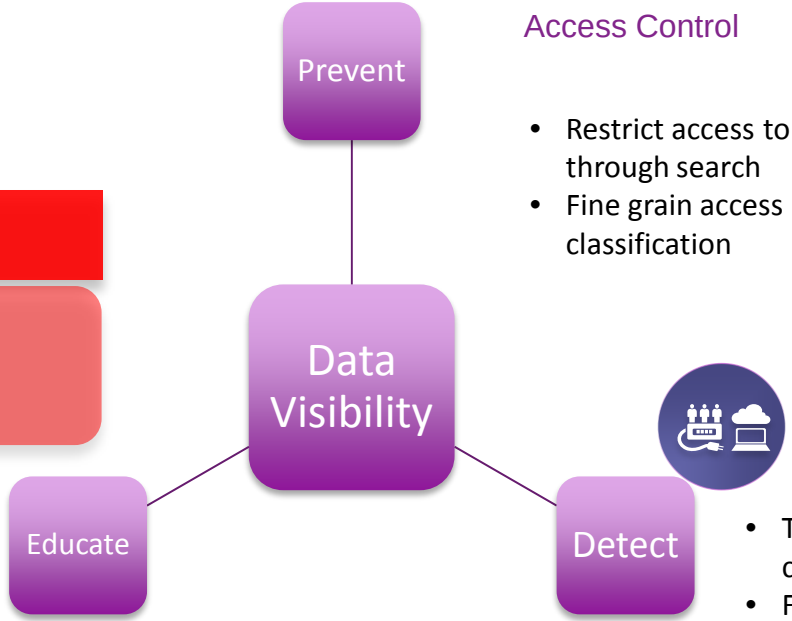
Engage	Attribute	Develop	Validate	Integrate	Protect
•System Identification •Stakeholder identification •Source system data fields •Field analysis •Field type analysis •Data record analysis •Define Dictionary •Candidate fields •Feasibility •Socialization	•Field value assignment •Field correlation •Weight scoring •Sensitivity scoring	•Classification engine Infrastructure Setup •Classification engine configuration •Coding of classification algorithm	•Sample size scoping •Sample size indexing •Validation of sample set •Statistical validation of sample set •Tune •Result socialization	•Design •User stories •Source system tagging (application tagging) •Stakeholder Socialization	•Access control •Behavior monitoring •Source System Secure design •Source System compliance •Export control •Import control •Data Loss

Now what? - Prevent, Detect and Educate

Policy Driven, Context-Based Access Control



- Restrict access to the application and through search
- Fine grain access based on data classification



Restricted

Why

- **Bug Status**: Open
- **Bug Severity**: Critical
- **Keywords**: **Customer**:

- Tag source systems and docs w/ classification metadata
- Focus on most sensitive data
- Integration with DLP solutions
- Data science

Q&A

Anchit Arora
Program Manager
InfoSec, Data Security Analytics Team
ancarora@cisco.com