Assessing Geocoding Solutions

Carrie Muenks & Chris Lawrence

September 9, 2014

HS SEDI
Homeland Security Systems Engineering and Development Institute

Homeland Security Systems Engineering and Development Institute

The Homeland Security Systems Engineering and Development Institute (hereafter “HS SEDI” or “SEDI”) is a federally funded research and development center (FFRDC) established by the Secretary of Homeland Security under Section 305 of the Homeland Security Act of 2002. The MITRE Corporation operates SEDI under the Department of Homeland Security (DHS) contract number HSHQDC-09-D-00001.

SEDI’s mission is to assist the Secretary of Homeland Security, the Under Secretary for Science and Technology, and the DHS operating elements in addressing national homeland security system development issues where enterprise, lifecycle, and/or acquisition systems engineering expertise is required. SEDI also consults with other government agencies, nongovernmental organizations, institutions of higher education, and nonprofit organizations. SEDI delivers independent and objective analyses and advice to support systems development, decision making, alternative approaches, and new insight into significant acquisition issues. SEDI’s research is undertaken by mutual consent with DHS and is organized by Tasks in the annual SEDI Research Plan.

This briefing does not necessarily reflect official DHS opinion or policy.

This briefing was prepared for public release.

Approved for Public Release; Distribution Unlimited. Case Number 14-2511

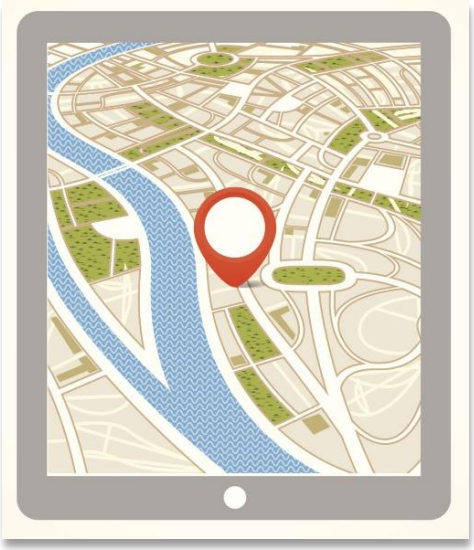
Outline

- **Definitions**
- **Geocoding Requirements**
- **Methodology**
- **Quantitative Test Procedure**
  - Test Datasets
  - User Simulation
  - Accuracy Analysis
- **Qualitative Test Procedure**
  - Software Requirements
- **Scalability of Assessment**

HS SEDI
Homeland Security Systems Engineering and Development Institute

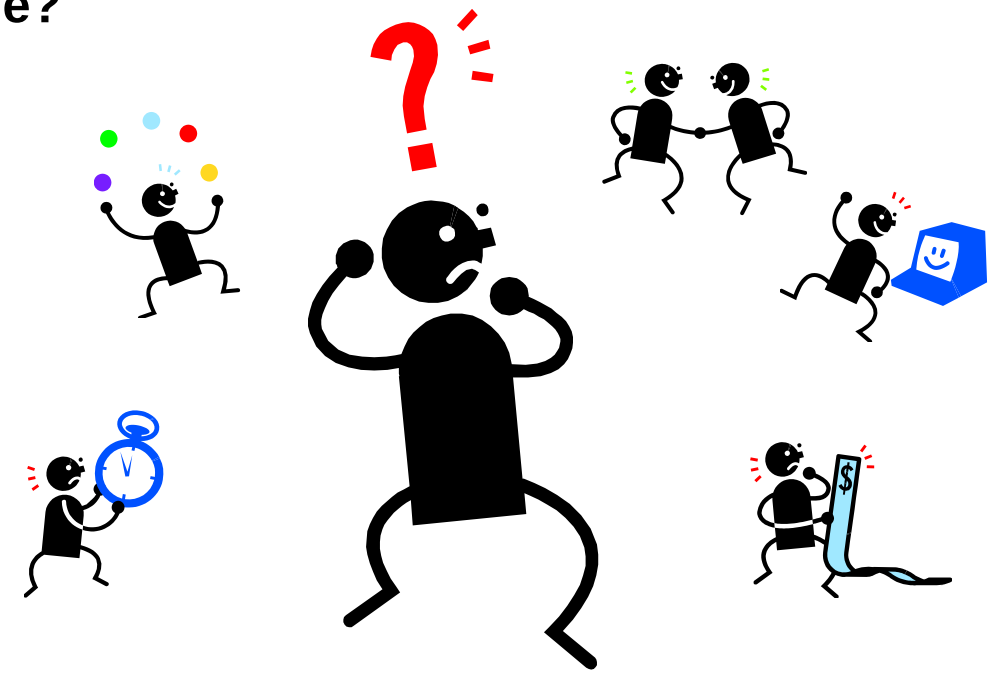
Definitions

- **Geocoding – process for converting street addresses into spatial data that can be displayed as features on a map**
- **Geocoding solution – comprised of a geocoding engine and geocoding reference data**
  - Geocoding engine - entity in the geocoding framework that drives the geocoding process.
    - The engine maps to the reference data source, based on the geographic places (e.g., country code) listed in the non-spatial data file
    - Then, the engine determines the appropriate algorithms for standardizing the addresses and matching them to the reference data
    - Finally, the engine defines parameters for reading address data, matching address data to the reference data, and creating output
  - Geocoding reference data - data that a geocoding service uses to determine the geometric representations for locations



I need a geocoding solution…

- What are the requirements for a geocoding solution and/or output?
- Do you just need information from the vendor or do you want to independently test the geocoding solution?
- How soon do you need to make a decision?
- Is cost an issue?



Geocoding Requirements

- **Accuracy**
  - Latitude and longitude coordinate in relation to what is true on the ground
- **Precision**
  - The level of precision (i.e., decimal places) needed within the latitude and longitude coordinates
- **Positional accuracy**
  - Acceptable latitude and longitude coordinates are dependent on the use case, especially for international locations

- **Reference data coverage**
  - The reference data can affect the accuracy, precision, and positional accuracy of the output
- **Geodetic aspects**
  - Knowledge of the coordinate system and projection of output data is needed

Geocoding Requirements

- **Processing environment**
  - Currency and reference data can differ between disconnected and web-based environments
- **Data structure**
  - Structured vs unstructured data formats can affect the ability of the geocoder to assign latitude and longitude coordinates
- **Output information**
  - Output should include: latitude and longitude coordinates, positional accuracy, address associated with the coordinates, and a confidence score at minimum for users to understand the output properly.
- **Cost**
  - Cost can be a limiting factor and would likely influence any decision for a geocoding solution.

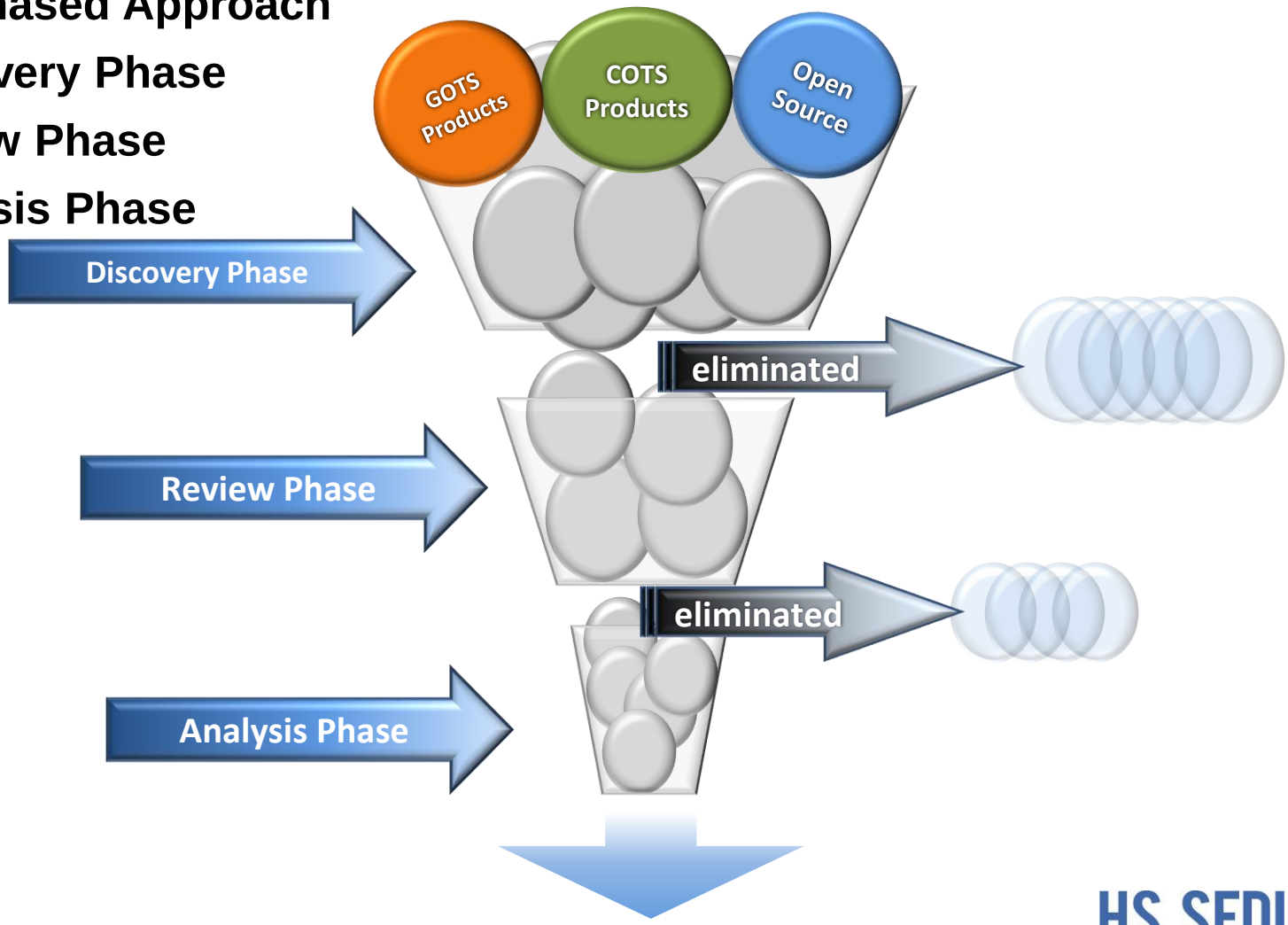
Geocoding Requirements

Larger enterprise considerations

- **Performance**
  - Amount of addresses processed within a specified time and the number of concurrent users
- **Customization**
  - The ability to customize the user experience or processing options. Up-front customizations might reduce the processing time

Methodology

- Three-Phased Approach
  - Discovery Phase
  - Review Phase
  - Analysis Phase

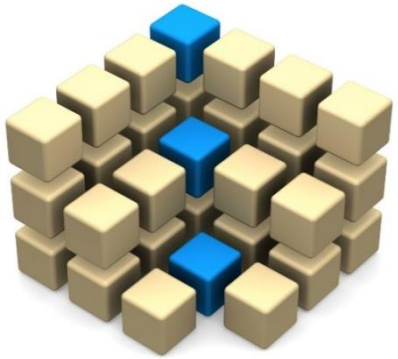


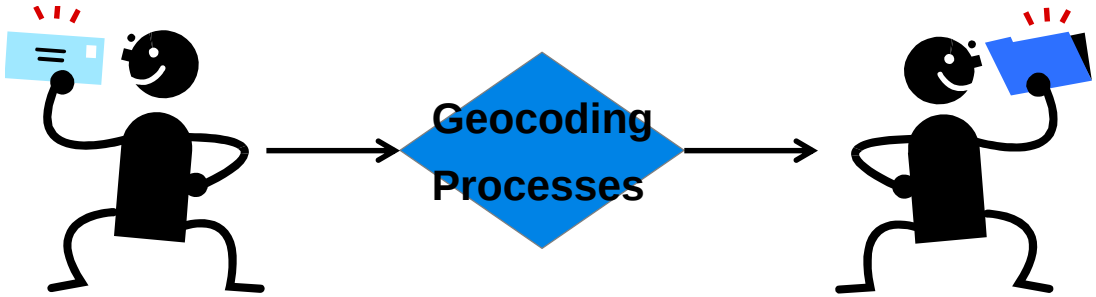
Methodology

- **Discovery Phase**
  - Exploration (i.e., industry survey) of the product space to identify geocoding solutions.
- **Review Phase**
  - Qualitative study of product capabilities according to vendor-provided and publicly available resources against a set of requirements.
- **Analysis Phase**
  - Quantitative and qualitative study of product capabilities and performance based on hands-on use.
    - Cost Proposals would be solicited during the Analysis Phase.

Quantitative Test Procedure

- **Test procedure (Analysis Phase) comprised of 2 approaches**
  - **User Simulation**
    - 3-tiered method to simulate increasing degrees of the user's perceived understanding in using each of the geocoders.
  - **Accuracy Assessment**
    - Generate the latitude and longitude coordinates for each grouping of test data and assess the additional output fields.





Test Datasets

- **Truth Data**
  - Records obtained from authoritative sources where the latitude, longitude, and positional accuracy of the addresses were considered trusted
- **Synthetic Data**
  - Truth data records where elements within the address were intentionally and systematically altered to simulate “dirty data”

Nonexistent state	Incorrect street type
Incorrect state abbreviation	Incorrect ordering
Nonexistent country code	Incomplete address
Incorrect postal code	Misspelling
Spacing	Multiple issues

User Simulation

- **Designed to simulate the various skills of users and how they typically approach interacting with new software**

1. Tester launched the geocoder software and attempted to geocode the test dataset without any external guidance and/or documentation
   - Simulate a user who tries to figure out the software on his/her own

2. Tester read the geocoder's documentation and then attempted to geocode
   - Simulate a user who wanted to be more informed and was driven to be so, prior to working with a new software product
   - Simulate a user who had failed attempts with the first non-methodical approach

3. Tester contacted the vendor to confirm the recommended approach to processing the test dataset via the geocoder
   - Simulate a trained or less novice user
   - Simulate a user who experienced failed attempts on the previous methods and who was now seeking help desk support

Accuracy Analysis

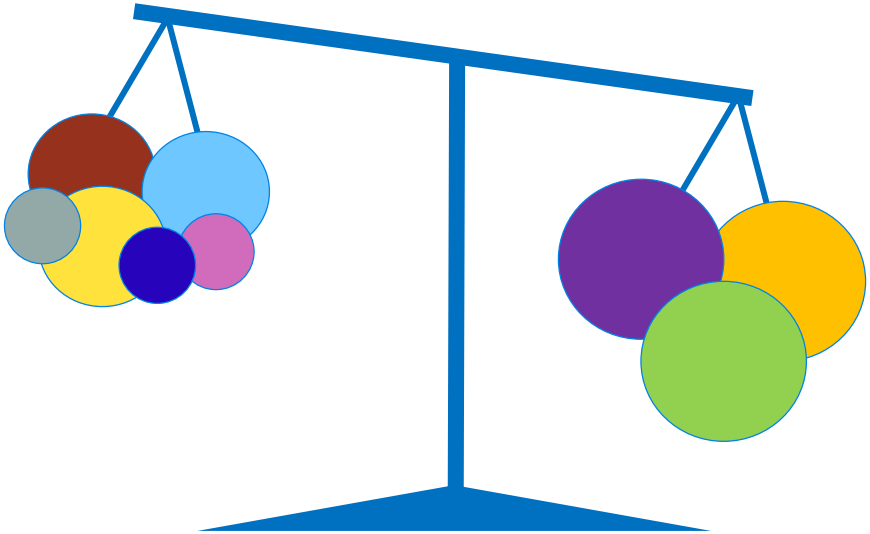
- **Results were binned according to the test data's positional accuracy for each record and the positional accuracy of the geocoder's output for that same record.**
  - The bins were directional, which means if the geocoder's output was at a less coarse level (e.g., parcel) than the truth data (e.g., street range), the output was categorized as such and different from where the truth data was less coarse than the output

Bin	Truth Positional Accuracy	Geocoder Positional Accuracy
1	Parcel	Parcel
2	Parcel	Street Centerpoint
3	Street Range	Street Centerpoint
4	Street Range	Parcel
5	…	

- **Accuracy was determined by calculating the distance between the truth latitude and longitude coordinates and the geocoder's output latitude and longitude coordinates for the same address**

Qualitative Assessment

- **The qualitative analysis (Analysis Phase) focused on areas that are not easily quantifiable but were important to this assessment**
  - Total scores for each geocoding solution were calculated based on answers to the qualitative questions
  - The importance of a factor was handled through weighting



Qualitative Requirements

- **Availability**
  - Amount of time the system must be operational and available for use
- **Reliability**
  - The probability of failure on demand
- **Data Retention**
  - Amount of time data must be stored and archived
- **Robustness**
  - The degree to which system is able to handle error conditions gracefully, without failure
- **Scalability**
  - The ability to handle a wide variety of system configuration sizes

Qualitative Requirements

- **Interoperability**
  - The ability of two or more diverse systems or components to exchange information and use the information that has been exchanged
- **Maintainability**
  - The ability of a system to be maintained through updates, upgrades, and failure
- **Portability**
  - A property of software that enables it to be transferred from one environment to another

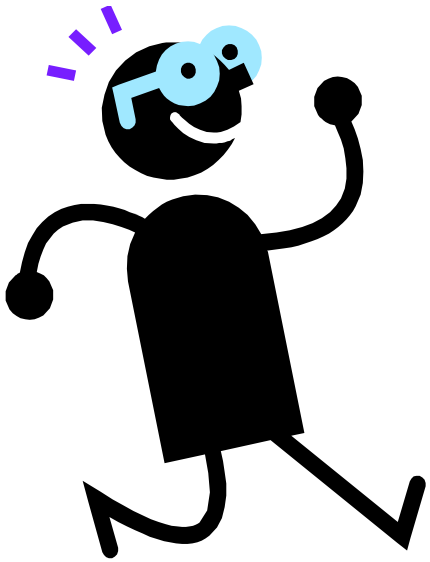
- **Security**
  - The ability to protect systems, information, and services from unintended or unauthorized access, change, or destruction
- **Auditability**
  - The ability to log, review, and analyze events, transactions, and effectiveness

Qualitative Requirements

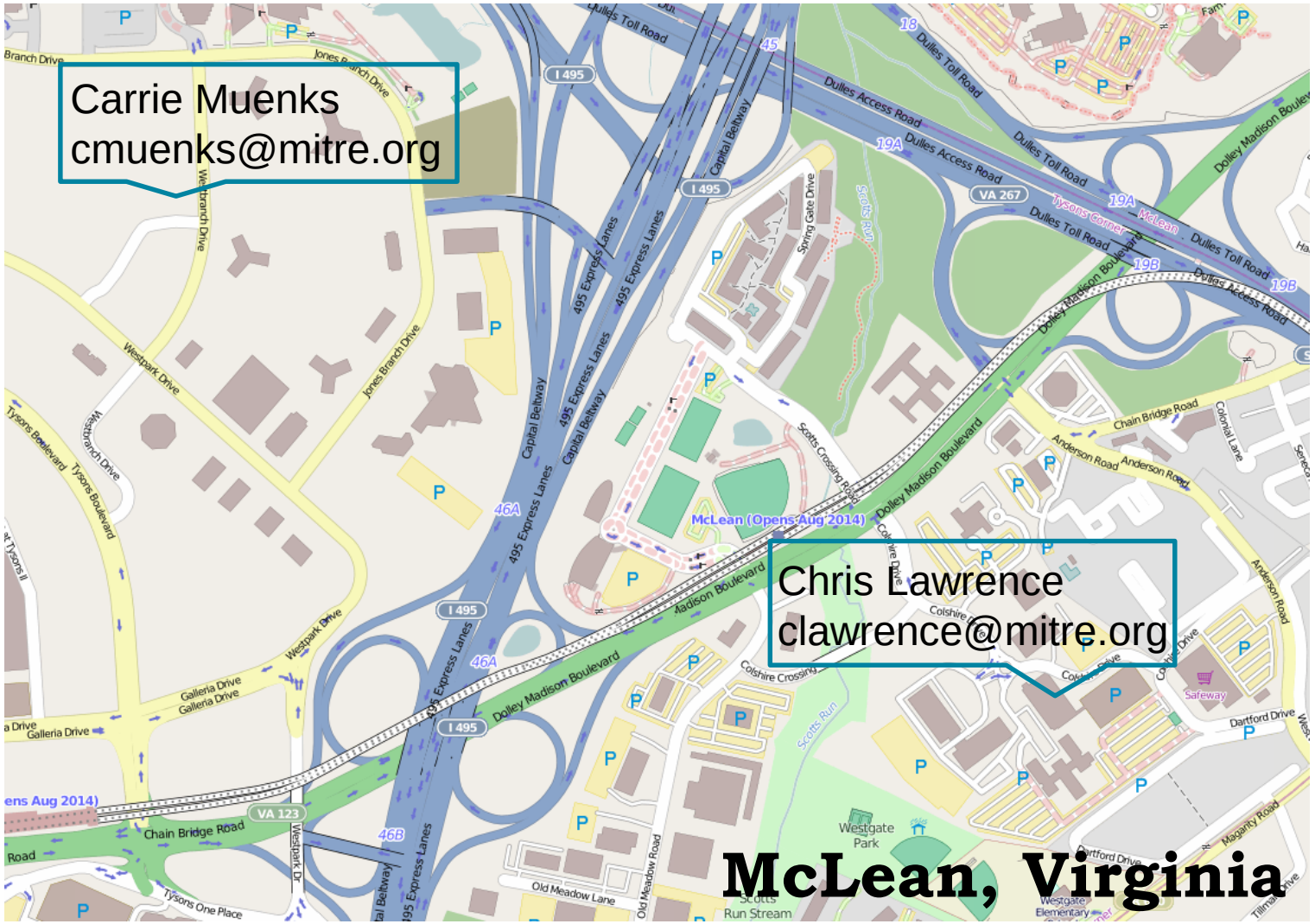
- **Transition**
  - The ability to load required data from various sources into the system for operations with data changed as necessary for system use
- **Usability/Human Factors/User Interface/Aesthetics**
  - The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction
- **Documentation**
  - Conditions for user-focused and/or technical materials that describe the use and the operation of the system
- **Resources/Resource Management**
  - These conditions address responsibilities for acquisition or monitoring of personnel and equipment for the development, operation, and support of the product
- **Regulatory/Programmatic**
  - Specific laws and regulations that constrain COTS solution selection

Scalability of Assessment

- Number of primary requirements (functional and non-functional)
- Consideration of secondary requirements
- User's characteristics
- Extent of testing needed
- Time frame and manpower
- Availability of a test environment and trial licenses
- Lifespan of the solution
- Needed scalability of the solution within your organization



Questions?



HS SEDI
Homeland Security Systems Engineering and Development Institute