INSTITUTE FOR DEFENSE ANALYSES

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO) (Presentation)

Shelley M. Cazares
Jacob B. Bartel

October 2020

Approved for public release;
distribution is unlimited.

IDA Document NS D-18431

Log: H 20-000445

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Project AM-2-1528, “Assessment of Traditional and Emerging Approaches to the Detection and Classification of Surface and Buried Unexploded Ordnance (UXO),” for the Director, Environmental Security Technology Certification Program (ESTCP) and Strategic Environmental Research and Development Program (SERDP), under the Deputy Assistant Secretary of Defense (Environment). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

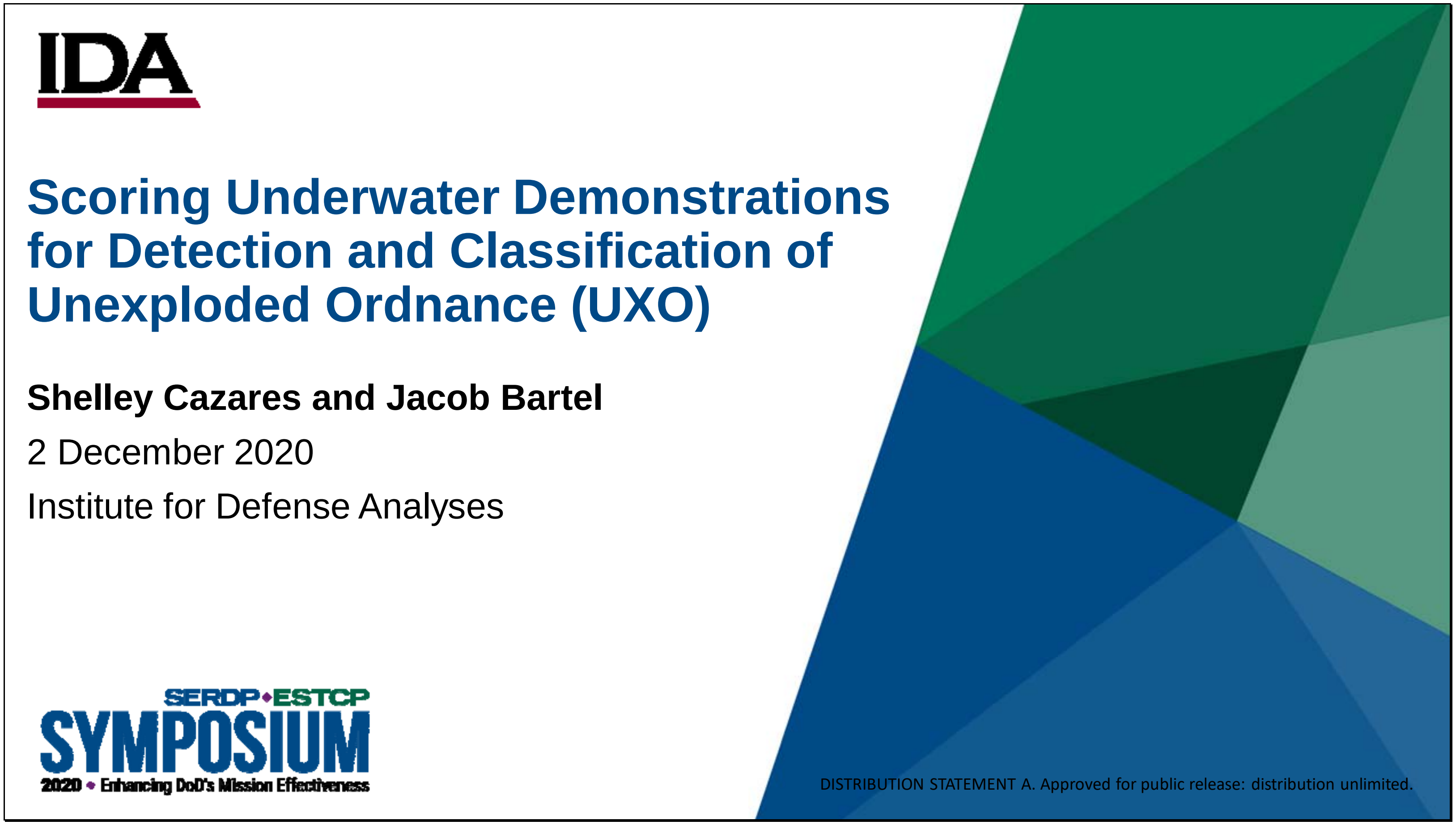
Shelley M. Cazares, Project Leader
scazares@ida.org, 703-845-6792

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Copyright Notice

Executive Summary

SERDP/ESTCP is sponsoring the development of novel systems for the detection and classification of Unexploded Ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems. Scoring these demonstrations is a complicated process. The Institute for Defense Analyses (IDA) designed and implemented the scoring process for ESTCP’s previous terrestrial demonstrations from 2007 – 2017. In some cases, the lessons learned from the terrestrial demonstrations can be leveraged in the underwater demonstrations. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment. This presentation will provide an overview of the main considerations for scoring underwater demonstrations for UXO detection and classification.



IDA

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO)

Shelley Cazares and Jacob Bartel

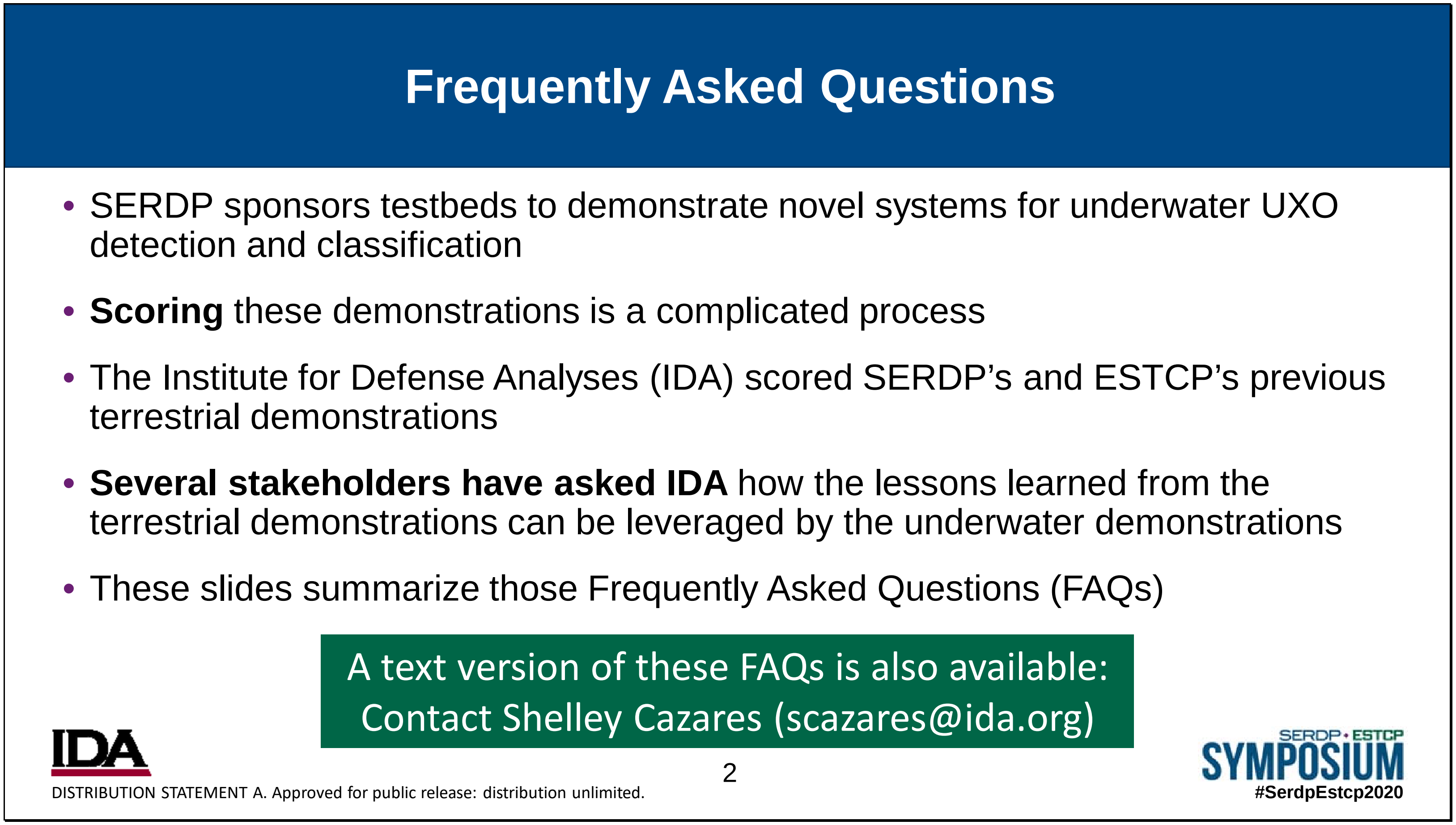
2 December 2020

Institute for Defense Analyses

SERDP•ESTCP SYMPOSIUM 2020 • Enhancing DoD's Mission Effectiveness

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

The Institute for Defense Analyses (IDA) is a non-profit corporation operating three Federally Funded Research and Development Centers (FFRDCs). IDA's mission is to provide objective analyses of national security issues, particularly those requiring scientific, technical, and analytic expertise.



Frequently Asked Questions

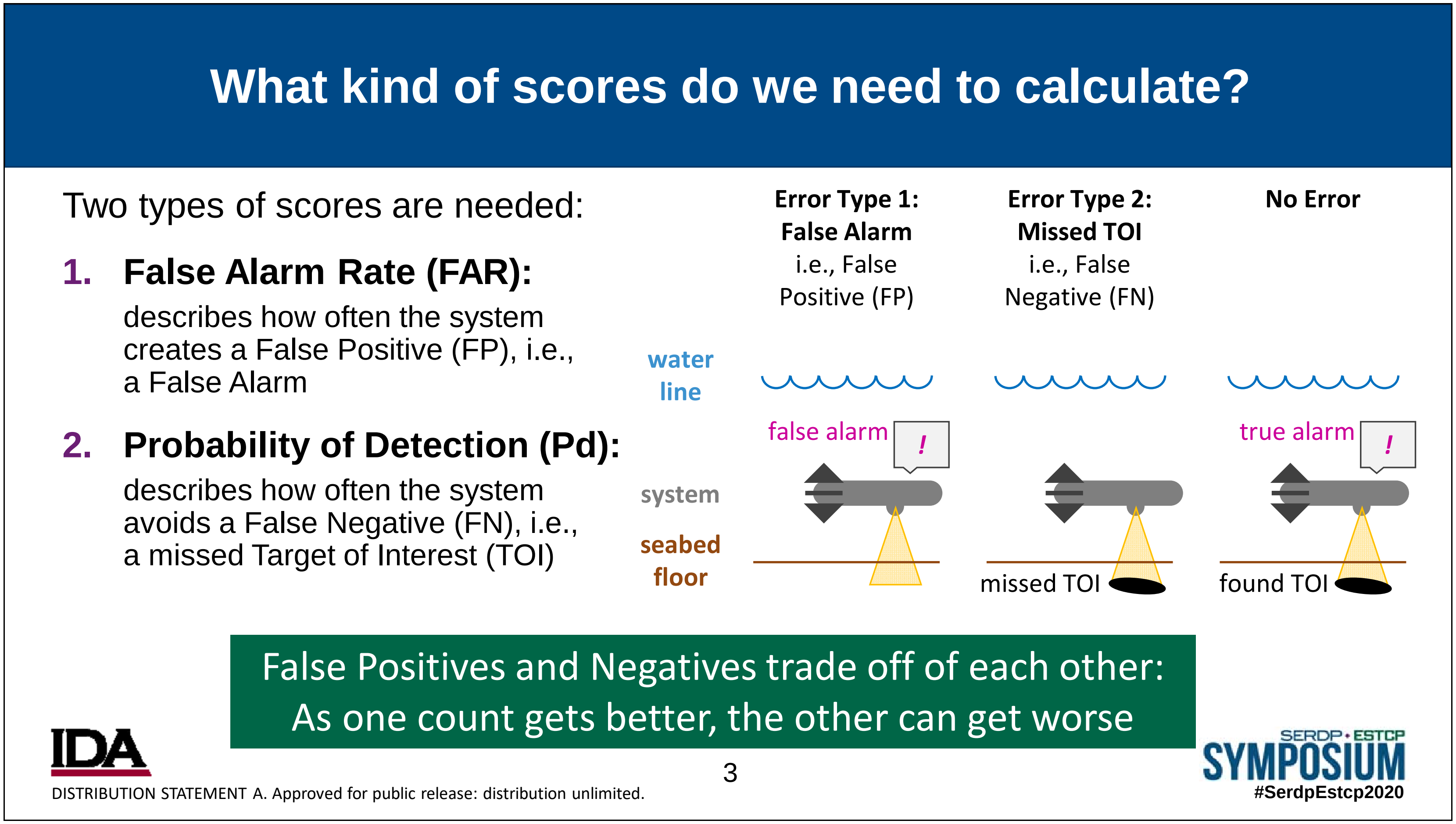
- SERDP sponsors testbeds to demonstrate novel systems for underwater UXO detection and classification
- **Scoring** these demonstrations is a complicated process
- The Institute for Defense Analyses (IDA) scored SERDP's and ESTCP's previous terrestrial demonstrations
- **Several stakeholders have asked IDA** how the lessons learned from the terrestrial demonstrations can be leveraged by the underwater demonstrations
- These slides summarize those Frequently Asked Questions (FAQs)

A text version of these FAQs is also available:
Contact Shelley Cazares (scazares@ida.org)

The Strategic Environmental Research and Development Program (SERDP) sponsors underwater testbeds for demonstrating Unexploded Ordnance (UXO) detection and classification systems. Scoring these underwater demonstrations is a complicated process. IDA designed and implemented the scoring process for previous *terrestrial* demonstrations sponsored by SERDP and its sister organization, the Environmental Security Technology Certification Program (ESTCP). Several stakeholders in the *underwater* UXO remediation community have asked IDA if the lessons learned in the terrestrial demonstrations can be leveraged underwater. The answer is: Sometimes. In some cases, the lessons learned on land *can* be leveraged underwater. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment.

IDA has compiled the resultant conversations into a set of Frequently Asked Questions (FAQs). The audience is intended to be site managers of underwater testbeds. However, other stakeholders could also gain valuable information from these FAQs, including system developers, environmental regulators, and the site managers who wish to use these novel systems to remediate their sites of UXO.

These slides provide a very high-level summary of the FAQs, with an accompanying text document that is available upon request.



What kind of scores do we need to calculate?

Two types of scores are needed:

1. **False Alarm Rate (FAR):**
   describes how often the system creates a False Positive (FP), i.e., a False Alarm

2. **Probability of Detection (Pd):**
   describes how often the system avoids a False Negative (FN), i.e., a missed Target of Interest (TOI)

Error Type 1:
False Alarm
i.e., False
Positive (FP)

Error Type 2:
Missed TOI
i.e., False
Negative (FN)

No Error

False Positives and Negatives trade off of each other:
As one count gets better, the other can get worse

IDA

3

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

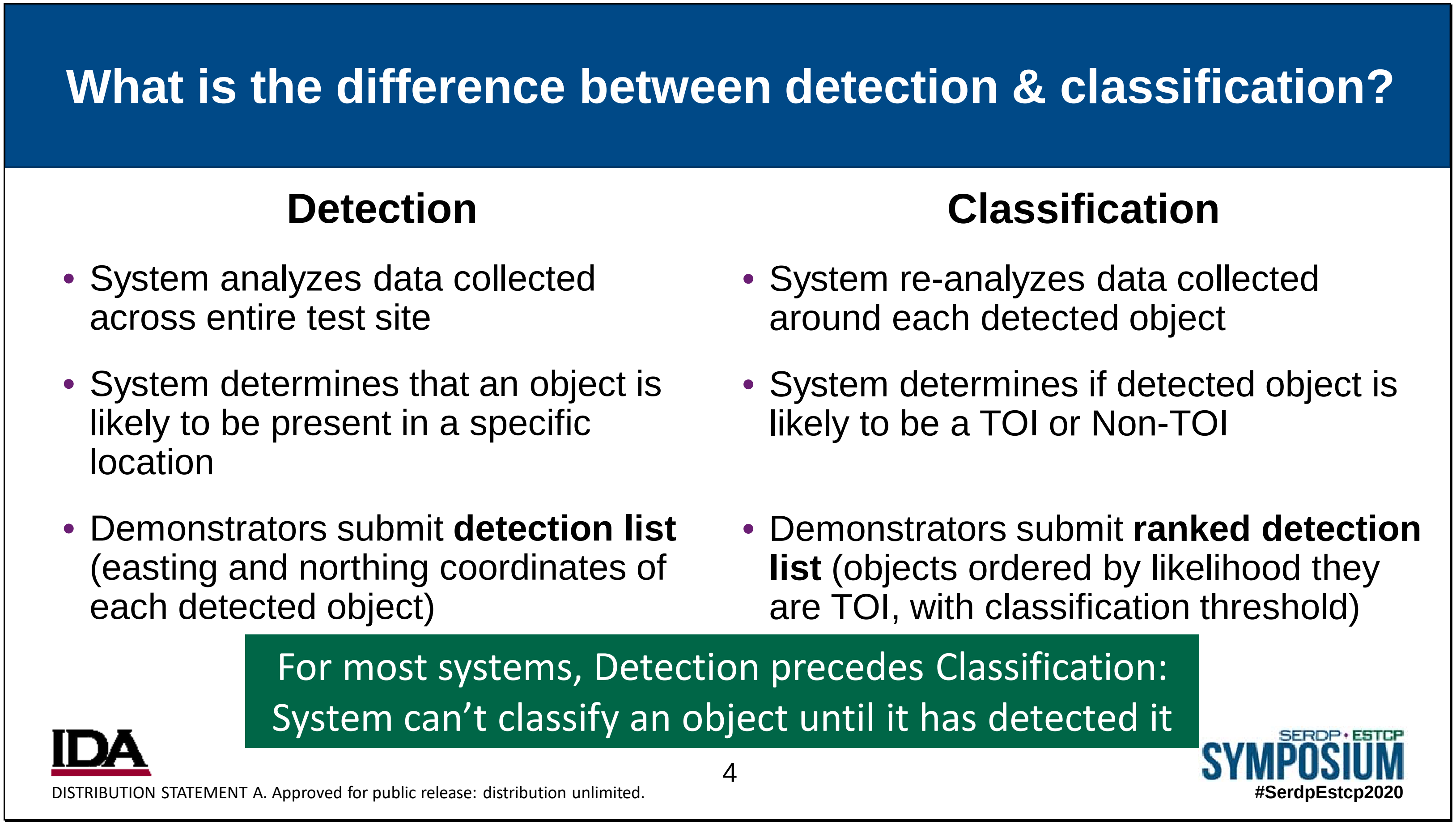
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2020

One question that is often asked is: What kind of scores do we need to calculate?

Scores must quantify how well the system can detect and classify Targets of Interest (TOIs) like UXO from Non-TOIs like clutter. Two types of scores are needed to describe the two types of errors in this figure:

- False Alarm Rate (FAR) describes how often the system creates a False Positive (FP), otherwise known as a False Alarm. A False Alarm is depicted on the left of the figure.
- Probability of Detection (Pd) describes how often the system avoids a False Negative (FN), otherwise known as a Missed TOI. A Missed TOI is shown in the middle of the figure.

Generally, False Positives and False Negatives– otherwise known as False Alarms and Missed TOIs– trade off of each other. As one count goes down, the other can go up. That is why *both* types of errors must be tallied, separately.



What is the difference between detection & classification?

Detection

- System analyzes data collected across entire test site
- System determines that an object is likely to be present in a specific location
- Demonstrators submit **detection list** (easting and northing coordinates of each detected object)

Classification

- System re-analyzes data collected around each detected object
- System determines if detected object is likely to be a TOI or Non-TOI
- Demonstrators submit **ranked detection list** (objects ordered by likelihood they are TOI, with classification threshold)

For most systems, Detection precedes Classification: System can’t classify an object until it has detected it

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

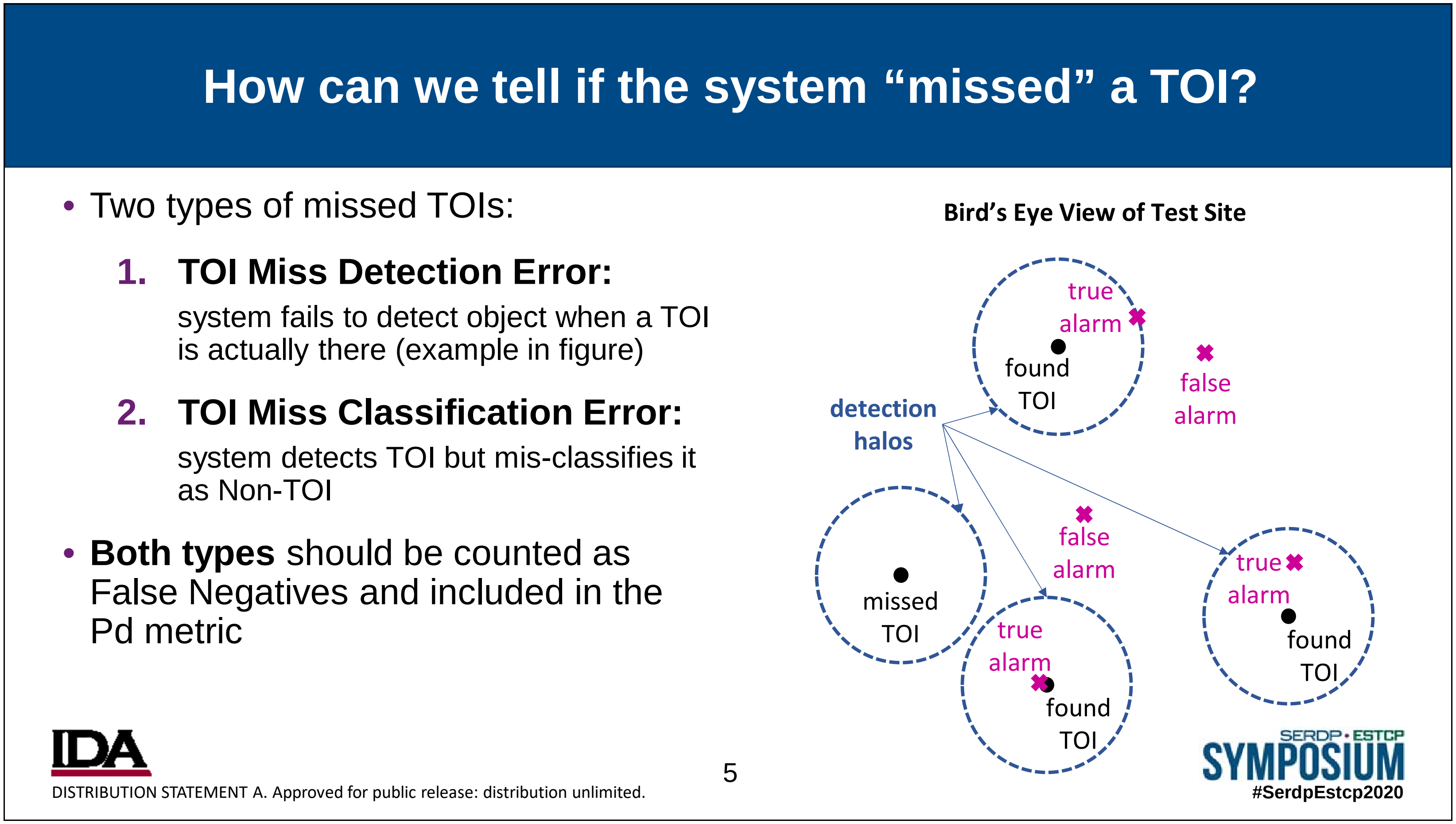
#SerdpEstcp2020

So far, we have been loose with our language. We have referred to the system “finding” a TOI. To be more precise, we should have referred to the system “detecting” and “classifying” a TOI. This leads to the next question: What is the difference between Detection and Classification?

Detection and Classification are the two main steps of *data processing* in a UXO remediation project.

- During *Detection*, the system analyzes the data collected across the entire test site and determines if one or more objects are likely to be present, and, if so, at what specific location(s). At the end of the Detection step, the demonstrator submits his or her *detection list*, which is a list of the estimated Easting and Northing coordinates of each detected object. *To ease scoring down the line, it’s best to submit coordinates in Universal Transverse Mercator (UTM) units, rather that latitude and longitude.*
- Classification is the next step. During *Classification*, the system re-analyzes data collected around each detected object, one by one, and determines if each detected object is likely to be a TOI or a Non-TOI. At the end of the Classification step, a demonstrator submits his or her *ranked* detection list. This list consists of *all* of the same detected objects as the original detection list. Now, however, the detected objects must be ordered according to their *likelihood of being a TOI*. The *first* detected object is the *most* likely to be a TOI, while the *last* detected object is the most likely to be a *Non*-TOI. The demonstrator also indicates the *classification threshold,* such that all detected objects above the threshold are classified as TOI and all below are Non-TOI.

For most systems, the Detection step must come before the Classification step. A system cannot classify an object until it has detected it.



How can we tell if the system "missed" a TOI?

- Two types of missed TOIs:

  1. **TOI Miss Detection Error:**
     system fails to detect object when a TOI is actually there (example in figure)

  2. **TOI Miss Classification Error:**
     system detects TOI but mis-classifies it as Non-TOI

- **Both types** should be counted as False Negatives and included in the Pd metric

Bird's Eye View of Test Site

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

5

#SerdpEstcp2020

The next two questions go together: How can we tell if the system missed a TOI, and how can we tell if the system had a False Alarm? This figure can help answer both questions. This figure is a notional bird's eye view of a test site. Black dots denote true TOIs, and pink Xs denote system alarms. Some of the black dot TOIs are scored as "found" while others are scored as "missed". Similarly, some of the pink X alarms are scored as "true" alarms, while others are scored as "false" alarms. In the next two slides, we will explain how to tell which is which.

First, we will discuss *Missed TOIs* on this slide. There are two types of Missed TOIs:

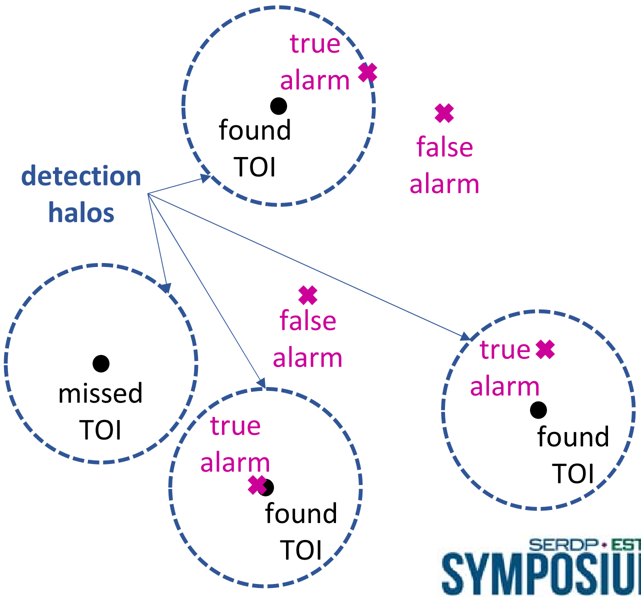
- The first type is a TOI Miss *Detection* Error. This kind of error occurs when the system fails to detect an object even though a TOI is actually, truly there. To determine if a TOI Miss *Detection* Error has occurred, the scoring team must know the *true* locations of all *TOI objects*– the Easting and Northing coordinates of the black dots, obtained from the ground truth. The scoring team must also know the *estimated* locations of all *detected objects*– the Easting and Northing coordinates of the pink Xs, obtained from the demonstrator's detection list. The scoring team must also decide how close an alarm must be to a TOI in order to declare that the system detected or "found" the TOI. This is done by drawing a circle or *detection halo* around each TOI– the blue dashed circles in the figure. Then:
  - If no alarms are in or on a detection halo– like the left-most TOI in the figure– then the scoring team should conclude that the system "missed" the TOI.
  - On the other hand, if at least one alarm is in or on a detection halo– like the other three TOIs in the figure– then the scoring team should conclude that the system "found" the TOI. Note, though, that at this point in the process, the system does not yet know if the detected object is a TOI or Non-TOI. Therefore, the system must pass the detected object onto the next step of the process, Classification.
- The second type of Missed TOI is a TOI Miss *Classification* Error. This kind of error occurs when the system detects a TOI but mis-classifies it as a *Non*-TOI. To determine if a TOI Miss *Classification* Error has occurred, the scoring team must know the true and estimated *types* of each detected object, obtained from the ground truth and the demonstrator's *ranked* detection list.

Both types of Missed TOIs should be counted as False Negatives and included in the Pd metric.

How can we tell if the system has a false alarm?

- Two types of false alarms:
  1. **False Alarm Detection Error:**
     system detects object even when no object is actually there (two examples in figure)
  2. **False Alarm Classification Error:**
     system mis-classifies Non-TOI as TOI
- **Only one type** (False Alarm Classification Error) should be counted as a False Positive and included in the FAR metric

Bird's Eye View of Test Site



DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

#SerdpEstcp2020

Approved for public release; distribution is unlimited.

There are also two different ways of having a *False Alarm*.

- The first type is a False Alarm *Detection* Error. Here, the system detects an object even when no object is actually there. This figure shows *two* False Alarm Detection Errors– the two lone pink Xs, one in the center and the other in the upper right. To determine if a False Alarm *Detection* Error has occurred, the scoring team must:
    - First, as discussed on the previous slide, determine if each TOI was “found” or “missed”, based on whether or not at least one alarm was in or on its detection halo.
    - Then, all remaining alarms must be flagged as *False* Alarms, like the two we already pointed out in the figure.
- The second type is a False Alarm *Classification* Error. Here, the system mis-classifies a detected object as TOI even though it is actually *Non*-TOI– or even it there is no real object at all. To determine if this is the case, the scoring team must know the detected object’s true type– either that it is truly a TOI on one hand, or truly a *non*-TOI– or *not present at all*– on the other hand.

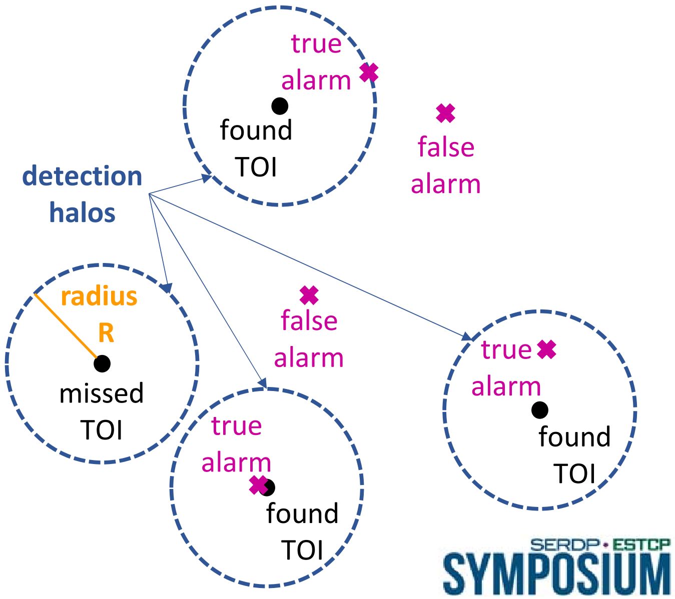
Only *one* of these types of errors should be counted as a False Alarm and included in the FAR metric– the False Alarm *Classification* Errors. That is due to the fact that, after the Detection step, *all* detected objects, even the incorrectly detected ones, are passed on to the *Classification* step, where the system gets a second chance to correctly classify them as *Non*-TOI. That is, during the Classification step, the system gets a second chance to correct the potential False Alarm mistakes it made in the Detection step. In fact, philosophically speaking, we shouldn’t even call them “mistakes” at all. They weren’t “*False* Alarms” because the system never claimed they were TOIs– instead, the system simply claimed they were worthy of further analysis– which is the whole point of the next step, Classification. That is why False Alarm *Detection* Errors should *not* be counted as False Alarms, and should *not* be included in the FAR metric. In contrast, any False Alarms that remain *after* the Classification step are False Alarm *Classification* Errors– and *these* should be included in the False Alarm count and the FAR metric.

How do we set the detection halo?

- The detection halo is a circle centered around each true TOI
- Its radius R depends on:
  - Requirements of subsequent steps of remediation process (e.g., retrieval)
  - Geolocation error
  - Sensor resolution

Proper selection of R is *the* trickiest part of underwater demonstration scoring

Bird's Eye View of Test Site



IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

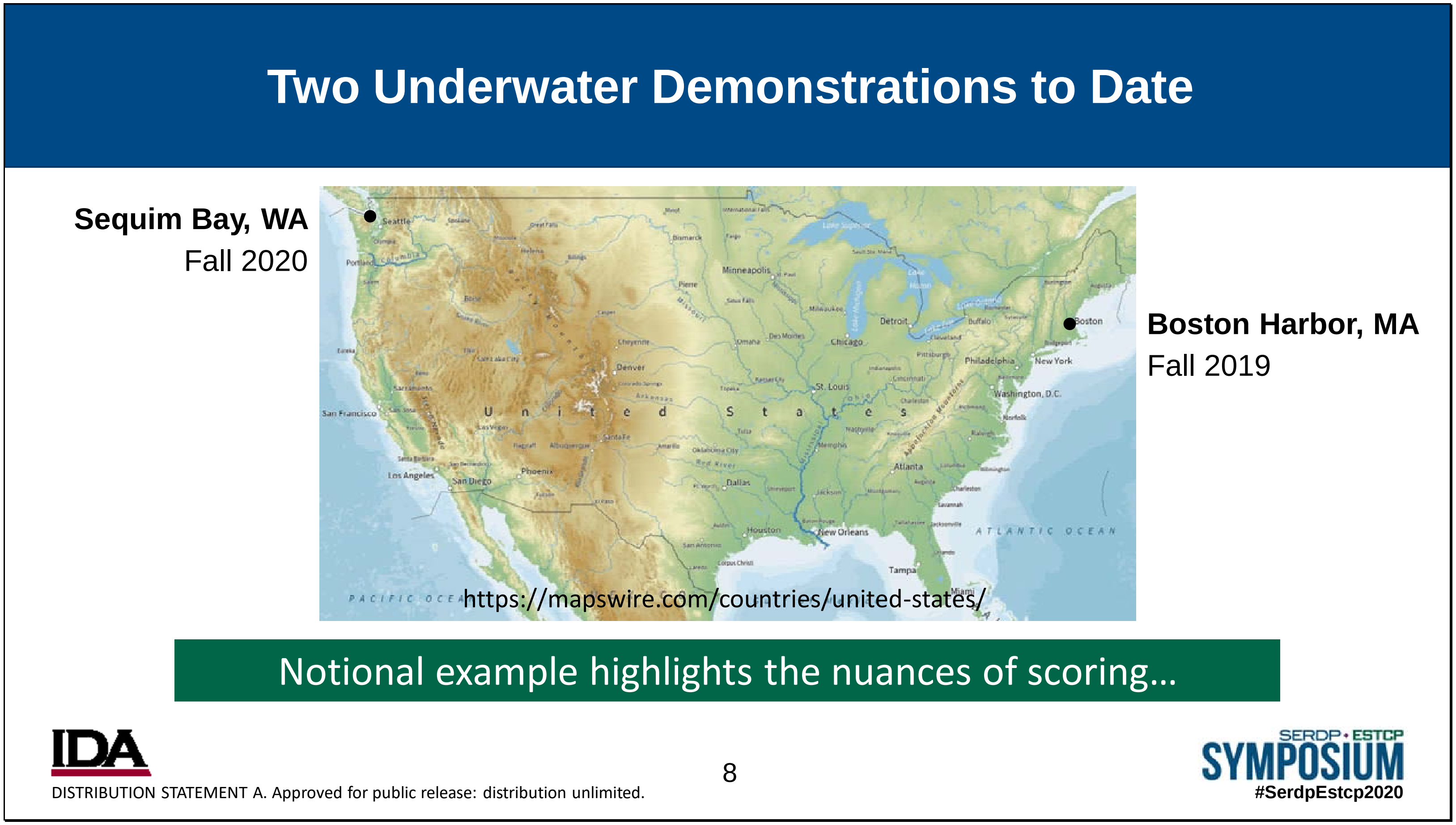
So far, we have discussed how to score Missed TOIs and False Alarms using the detection halo. The next question we always get is: How do we *set* the detection halo?

The detection halo is a circle centered around each true TOI. The scoring team must define the radius of the detection halo, labeled R in the figure. This radius R sets the maximum acceptable distance between a TOI's *true* versus *estimated* coordinates during Detection scoring.

- *In an ideal world*, the detection halo radius R would be based solely on the requirements of the next steps of the UXO remediation process– such as reacquisition for retrieval. In previous *terrestrial* demonstrations, R was usually set to be somewhere between 25 and 40 centimeters– roughly the width of a standard shovel used to dig up the UXO. However, in *underwater* projects, it is much more difficult to dig a single, clean hole. Therefore, R should be based on how far out from the estimated coordinates the diver should be expected to swim around to find each UXO for retrieval. Ideally, that would be around one meter in radius.
- *In the real world*, there are other factors that come into play, especially for *underwater* demonstrations:
  - The first is *geolocation error*, which is the uncertainty in the survey instruments used to measure location. *Terrestrial* demonstrations can make use of Real Time Kinematic (RTK) Global Positioning Systems (GPS) to provide *centimeter*-level accuracy in location estimates. Therefore geolocation error isn't a very large consideration in *terrestrial* demonstrations. In contrast, *underwater* demonstrations must use alternative surveying technology, for which geolocation error is much larger, likely on the order of *meters*. Thus geolocation error becomes a much larger consideration in *underwater* demonstrations.

  - The other additional consideration is what we are calling *sensor resolution* in these slides, a concept which refers to the ability of the system's sensors to tell multiple, closely-spaced objects apart. *Terrestrial* remediation projects make use of *Electromagnetic Induction (EMI)* sensors for which a data processing technique called "multi-source dipole inversions" can be used to resolve multiple, closely spaced objects– therefore sensor resolution is no longer a large consideration for terrestrial demonstrations. In contrast, some *underwater* remediation projects are likely to use *acoustic* sensors, for which sensor resolution may be a larger issue.

All of these issues must be considered when selecting the proper value for the detection halo radius. This is often the #1 trickiest step in underwater demonstration scoring. To date, IDA has been using an R of 2-3 meters. *System developers should begin communicating their expected geolocation error and sensor resolution to SERDP as early as possible, so that appropriate testbeds and scoring protocols can start being developed for their specific systems.*



Two Underwater Demonstrations to Date

Sequim Bay, WA
Fall 2020

Seattle
Portland
San Francisco
Los Angeles
San Diego
Phoenix
Denver
Dallas
Houston
New Orleans
Minneapolis
Chicago
Detroit
St. Louis
Atlanta
Tampa
Miami
Philadelphia
New York
Boston
Washington, D.C.
United States
ATLANTIC OCEAN
PACIFIC OCEAN

Boston Harbor, MA
Fall 2019

https://mapswire.com/countries/united-states/

Notional example highlights the nuances of scoring…

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

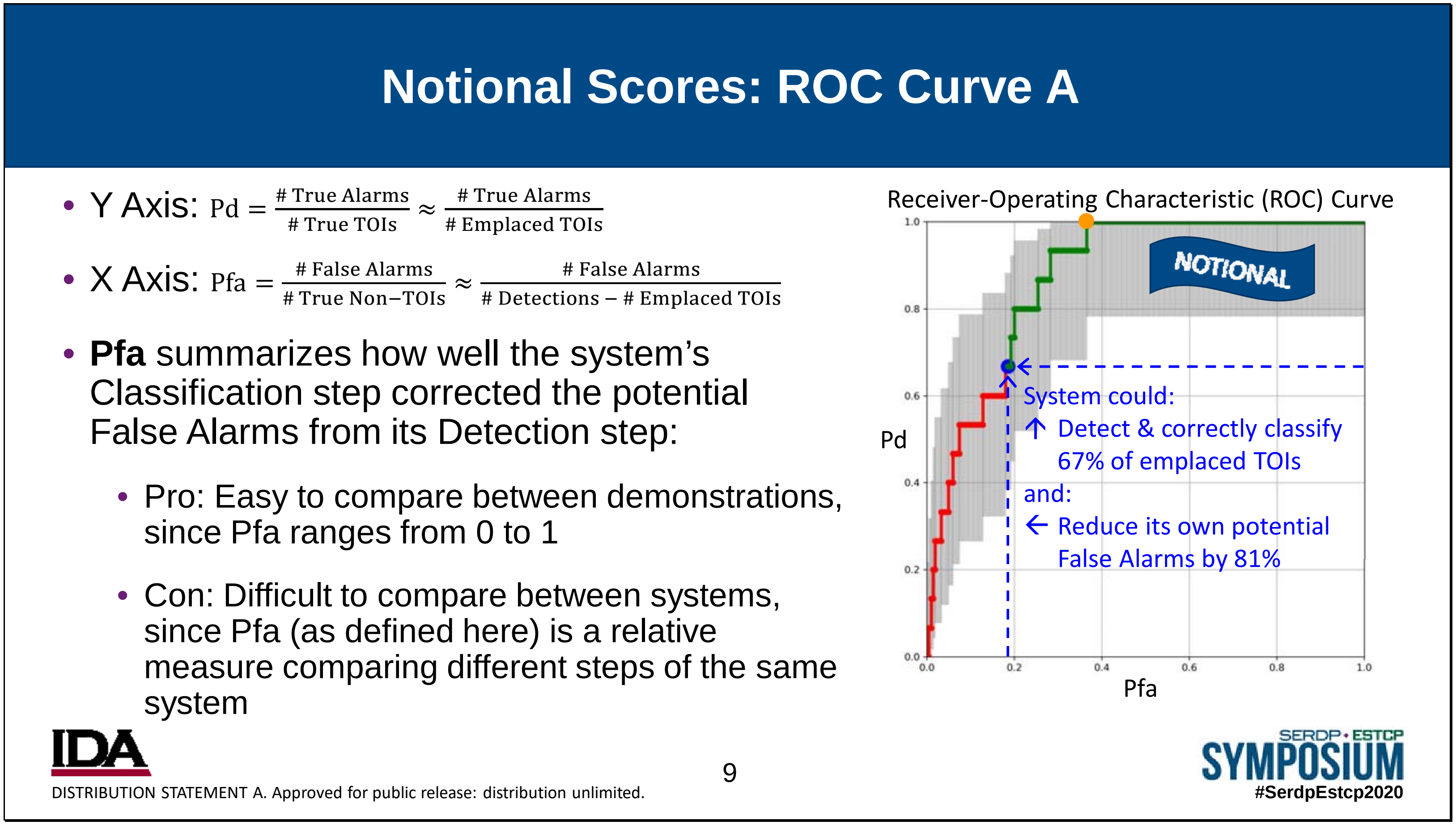
SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

Two underwater Blind Tests have occurred so far:

- Boston Harbor, MA in Fall 2019.
- Sequim Bay, WA in Fall 2020.

Scoring is still underway for both demonstrations.

On the next few slides, we will show some notional scores– *fake scores*– to highlight the nuances of scoring.



Notional Scores: ROC Curve A

- Y Axis: $\mathrm{Pd} = \frac{\#\,\text{True Alarms}}{\#\,\text{True TOIs}} \approx \frac{\#\,\text{True Alarms}}{\#\,\text{Emplaced TOIs}}$
- X Axis: $\mathrm{Pfa} = \frac{\#\,\text{False Alarms}}{\#\,\text{True Non−TOIs}} \approx \frac{\#\,\text{False Alarms}}{\#\,\text{Detections} - \#\,\text{Emplaced TOIs}}$
- **Pfa** summarizes how well the system's Classification step corrected the potential False Alarms from its Detection step:
  - Pro: Easy to compare between demonstrations, since Pfa ranges from 0 to 1
  - Con: Difficult to compare between systems, since Pfa (as defined here) is a relative measure comparing different steps of the same system

Receiver-Operating Characteristic (ROC) Curve

System could:

↑ Detect & correctly classify 67% of emplaced TOIs

and:

← Reduce its own potential False Alarms by 81%

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

This slide shows a Receiver-Operating Characteristic (ROC) curve for a *notional* system. We used a detection halo radius of 3 m for scoring.

ROC curves help us visualize the tradeoff between False Negatives and False Positives– between TOI Misses and False Alarms:

- Pd is plotted on the Y axis. Pd indicates how well the system can avoid False Negatives, otherwise known as Missed TOIs. Pd ranges from 0 to 1, with higher values considered better. Theoretically, Pd is supposed to be the number of *true alarms* divided by the number of *true TOIs*. Unfortunately, we don't know how many true TOIs were actually in this demonstration– it's possible that a munition was fired in the past but did not explode, creating a legacy UXO that has sat on the bottom of the seabed floor for decades, unknown. A full sweep of the seabed floor was not within the scope of this demonstration, and so we must simply assume that there were *no* legacy UXOs in the test area. Therefore we estimate the denominator of Pd as simply the number of *emplaced* TOIs. That is, we assume that the only TOIs in the test area are those that were emplaced for the purpose of the demonstration.
- Pfa is plotted on the X axis. Pfa an indication of False Alarms. Theoretically, Pfa is supposed to be the number of *false* alarms divided by the number of true *Non*-TOIs, where a "Non-TOI" can be either a Non-TOI object or the lack of any object at all. In practice, though, Pfa can really only be calculated for pure *classification* tests– it doesn't make sense to calculate Pfa for tests that involve a *detection* component. That's because there's actually an *infinite* number of location coordinates on the seabed floor where no object exists! However, we can still calculate a metric that is similar in *spirit* to Pfa. Here, we estimate the *denominator* of our Pfa-like metric as the total number of detected objects minus the number of emplaced TOIs. That is, we assume that the only Non-TOIs in the test area are those objects that were *detected* but were *not* truly TOIs.

This notional ROC curve plots our Pd versus Pfa-like metric. Each point on the ROC curve corresponds to a different rank level on the notional ranked detection list. Remember, the ranked detection list is the final deliverable that the demonstrator submits at the end of the Classification step– it consists of *all* detected objects, ordered by their likelihood of being a TOI. The blue dot represents the demonstrator's final TOI-vs-Non-TOI classification threshold. Other points on this curve represent other classification thresholds– other rank levels– that the demonstrator could have chosen instead.

A vertical grey line is drawn through each point, to indicate the 95% confidence interval around that point's Pd value on the Y axis. Each point's confidence interval was calculated with the beta distribution approximation to the binomial distribution, with no adjustments for multiple comparisons.

To interpret this ROC curve, we must consider what the X and Y axes actually mean:

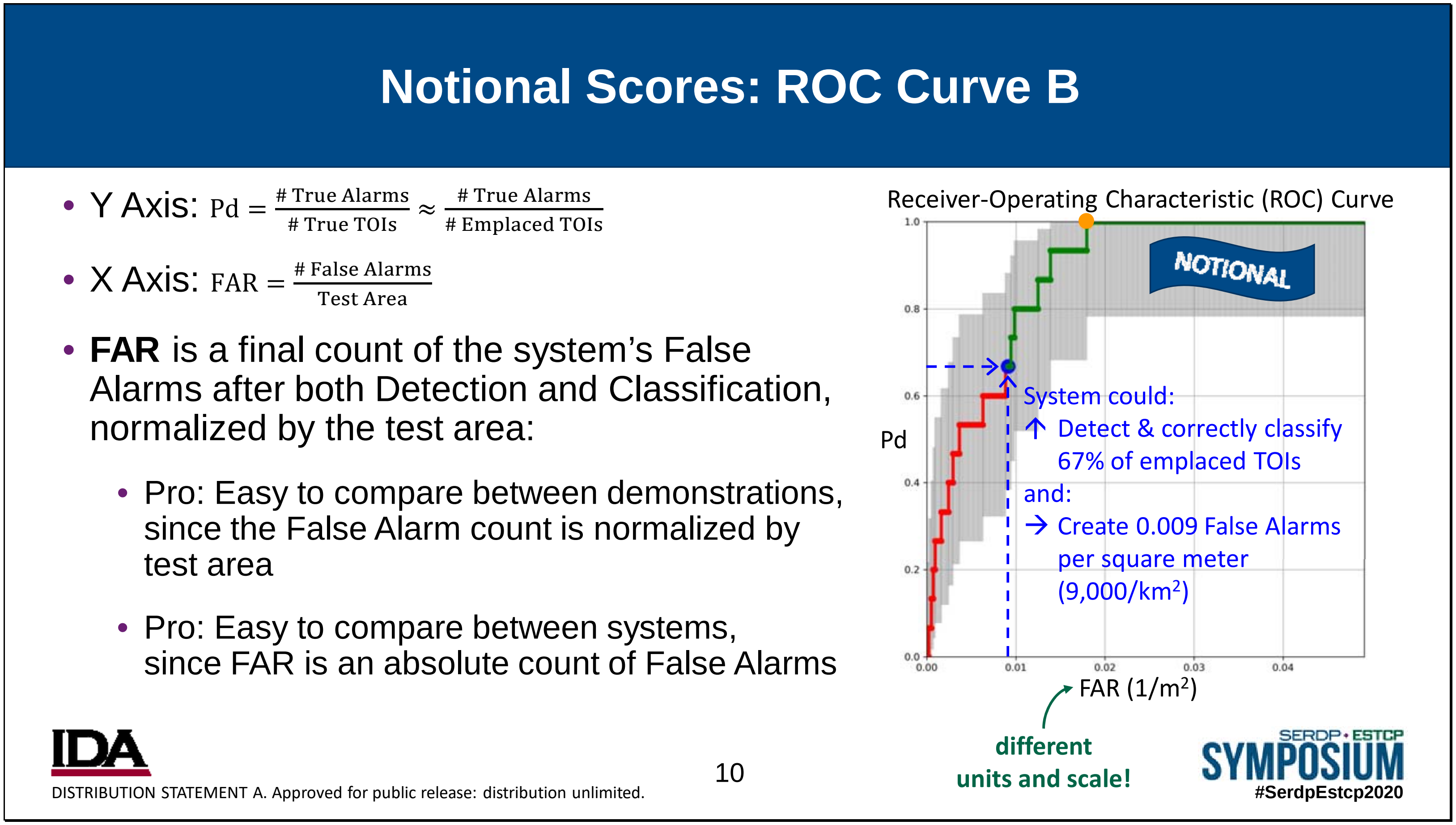
- Pd is on the Y axis. Its interpretation is fairly straightforward. It summarizes what *fraction* of TOIs the system was able to detect and correctly classify.
- Pfa is on the X axis. Its interpretation is more nuanced. It summarizes how well the system's Classification step corrected the potential False Alarms from its Detection Step.

The blue dot [0.19, 0.67] tells us that the system detected and correctly classified 67% of the emplaced TOIs (with a rather wide 95% confidence interval ranging from 0.38 to 0.88, due to the rather small number (15) of emplaced TOIs). At the same time, the system reduced its own potential False Alarms by 81%, from 1.0 on the right edge of the plot back to 0.19 at the blue dot. That is, during the Detection step, many true Non-TOI objects were detected. Had there been no subsequent Classification step, then *all* of those detections of true Non-TOIs would have remained False Alarms, leading to a Pfa of 1.0. Fortunately, there *was* a Classification step, in which 81% were correctly classified as Non-TOI, such that only 19% remained as False Alarms.

With 20/20 hindsight, the orange dot higher up and to the right [0.36, 1.0] tells us that the system could have detected and correctly classified *all* emplaced TOIs, while reducing its own potential False Alarms by 64%, from 1.0 on the right back to 0.36.

There are pros and cons to plotting Pfa on the X axis of the ROC curve:

- The main *advantage* is that ROC curves like this are easy to compare between *demonstrations*, since Pfa ranges from 0 to 1. Therefore a Pfa of 0.36 means the same thing from one demonstration to the next.
- The main *disadvantage* is that it is difficult to compare ROC curves between *systems*, since this Pfa-like metric, as defined here, is a *relative* measure comparing different steps of the same system– the system’s Detection step versus its Classification step.



Notional Scores: ROC Curve B

- Y Axis: $\text{Pd} = \frac{\#\,\text{True Alarms}}{\#\,\text{True TOIs}} \approx \frac{\#\,\text{True Alarms}}{\#\,\text{Emplaced TOIs}}$
- X Axis: $\text{FAR} = \frac{\#\,\text{False Alarms}}{\text{Test Area}}$
- **FAR** is a final count of the system's False Alarms after both Detection and Classification, normalized by the test area:
  - Pro: Easy to compare between demonstrations, since the False Alarm count is normalized by test area
  - Pro: Easy to compare between systems, since FAR is an absolute count of False Alarms

Receiver-Operating Characteristic (ROC) Curve

System could:
↑ Detect & correctly classify 67% of emplaced TOIs
and:
→ Create 0.009 False Alarms per square meter (9,000/km^2)

different units and scale!

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

10

SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

This slide shows the same notional scores from the same notional system in the same demonstration. The ROC curve has exactly the same shape as the previous slide. The numbers on the X axis are different, though– in both units and scale. That’s because this time, we plot a slightly different metric: FAR, the number of False Alarms divided by the total test area.

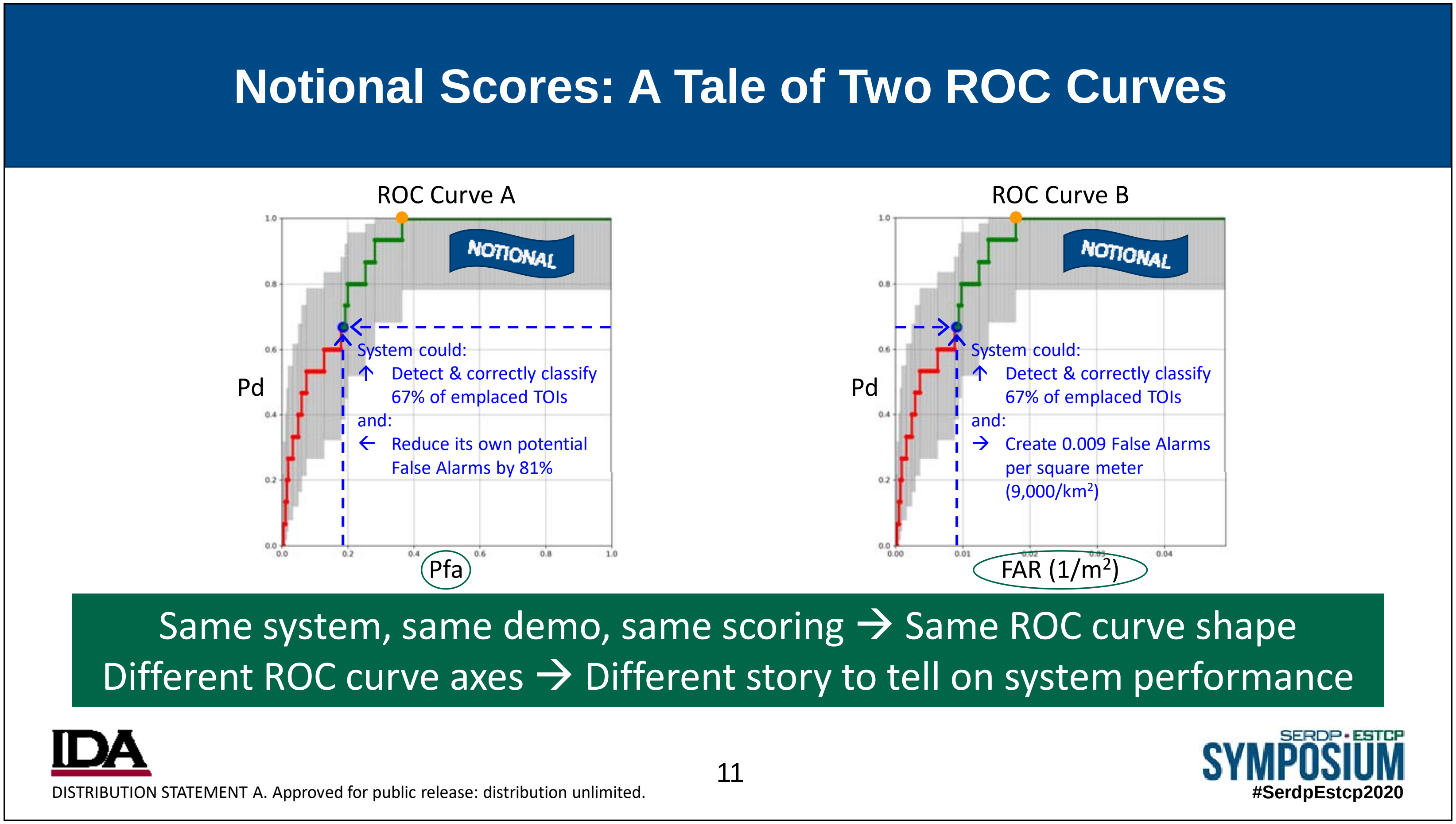
FAR is the final count of the system’s False Alarms after both the Detection *and* Classification steps– it’s an *absolute* count of False Alarms. Furthermore, this count is *normalized* by the test area, since some demonstrations are larger than others.

The blue dot [0.009, 0.67] tells us that the system detected and correctly classified 67% of the emplaced TOIs, while creating 0.009 False Alarms per square meter– roughly 9,000 False Alarms per square kilometer.

The orange dot [0.018, 1.0] tells us that with 20/20 hindsight, the system could have detected and correctly classified *all* emplaced TOIs, while creating about 0.018 False Alarms per square meter– roughly 18,000 False Alarms per square kilometer.

There are several benefits to plotting FAR on the X axis of the ROC curve:

- With FAR, it’s still easy to compare the ROC curves between *demonstrations*, since the False Alarm count has been normalized by the test area, and so the ROC curves can be compared between demonstrations of different sizes.
- Furthermore, with FAR, it’s now also easy to compare ROC curves between *systems*, since FAR is an *absolute* count of False Alarms, after both the Detection *and* Classification steps.



Notional Scores: A Tale of Two ROC Curves

ROC Curve A

System could:

↑ Detect & correctly classify 67% of emplaced TOIs

and:

← Reduce its own potential False Alarms by 81%

ROC Curve B

System could:

↑ Detect & correctly classify 67% of emplaced TOIs

and:

→ Create 0.009 False Alarms per square meter (9,000/km^2)

Same system, same demo, same scoring → Same ROC curve shape

Different ROC curve axes → Different story to tell on system performance

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

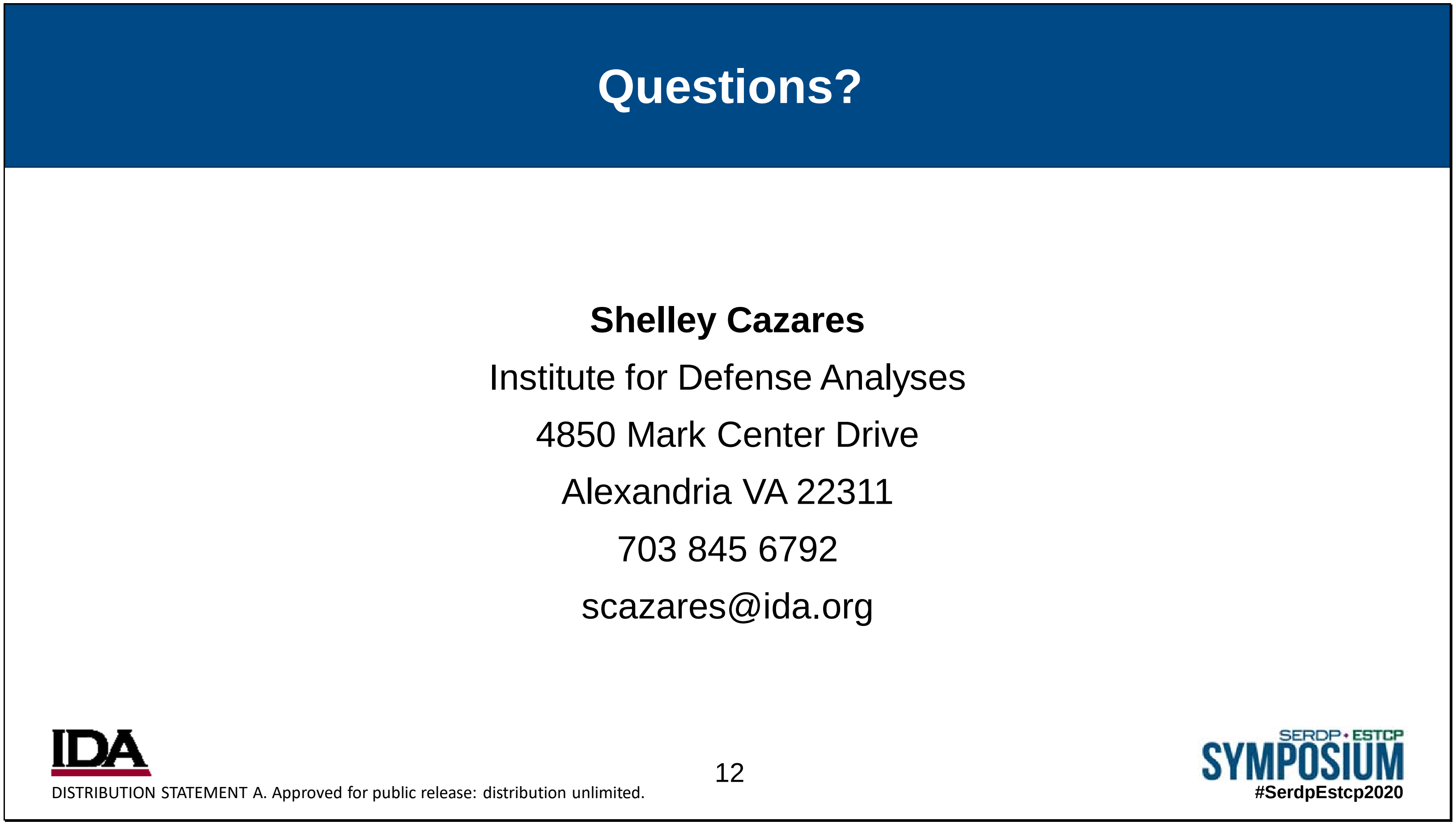
11

SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

The major point we'd like to leave you with is that there are many different scores that can be calculated from a demonstration, and each of those scores describes the system's performance from a slightly different perspective. The last two slides showed the same notional system from the same demonstration, scored with the same process, resulting in two ROC curves of the same shape. However, for each of those two ROC curves, we plotted a slightly different metric on the X axis, which allowed us to interpret the system's performance from two different perspectives. That is, we were able to tell two different stories about how well the system could find TOIs and avoid false alarms:

- ROC Curve A on the left described how well the system's Classification step cleaned up the potential False Alarms it made during the Detection step– a *relative* measure of performance, comparing the system to itself. ROC Curve A is helpful when we're focused *solely* on the system's Classification step, compared to a *standard* detection method.
- ROC Curve B on the right described how often the system as a whole created false alarms– an *absolute* measure of performance. ROC Curve B is helpful when we're focused on the system's Detection *and* Classification steps, *together*.

Both stories were legitimate, and both were fair assessments of the system. The question is: Which story does *this* community want to tell?



Questions?

Shelley Cazares
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria VA 22311
703 845 6792
scazares@ida.org

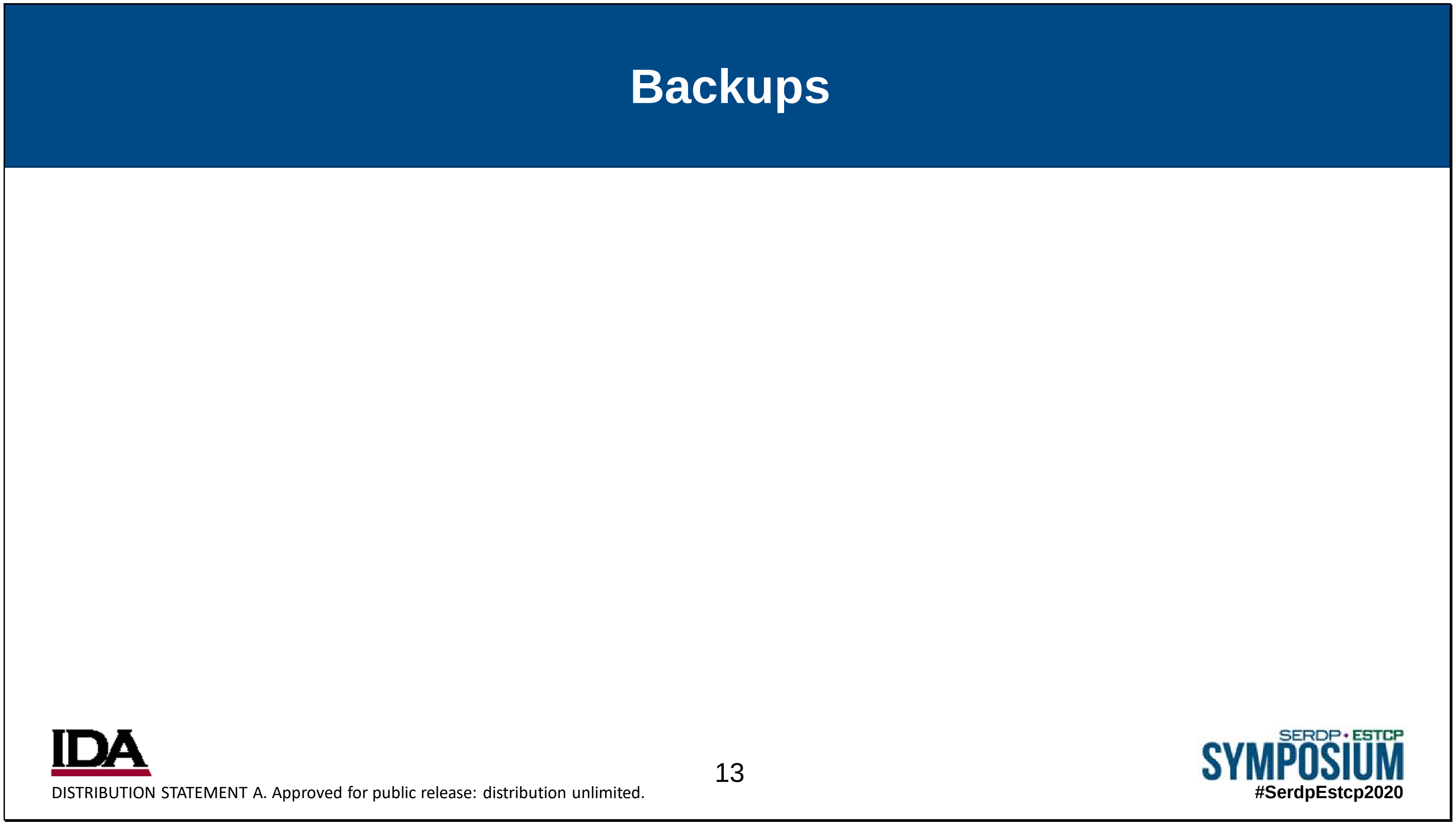
IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

12

SERDP • ESTCP
SYMPOSIUM
#SerdpEstcp2020

Please contact Shelley Cazares at IDA with any questions.



Backups

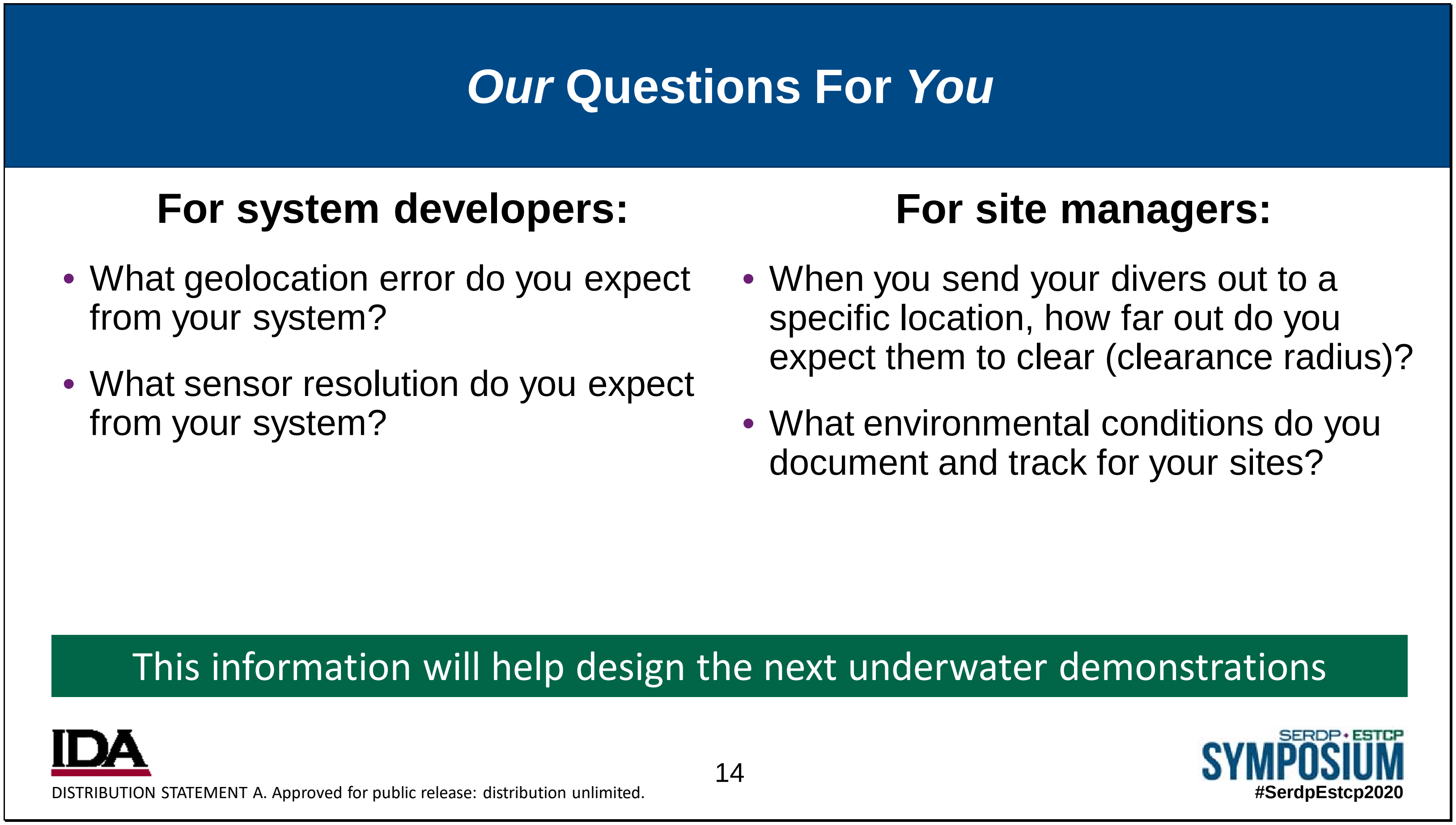
IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

SERDP • ESTCP SYMPOSIUM

#SerdpEstcp2020

The following slides may be useful in follow-on discussions.



Our Questions For *You*

For system developers:

- What geolocation error do you expect from your system?
- What sensor resolution do you expect from your system?

For site managers:

- When you send your divers out to a specific location, how far out do you expect them to clear (clearance radius)?
- What environmental conditions do you document and track for your sites?

This information will help design the next underwater demonstrations

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

14

SERDP • ESTCP SYMPOSIUM

#SerdpEstcp2020

This slide deck has summarized questions that the UXO remediation community has frequently asked IDA about scoring underwater demonstrations. As it turns out, *IDA* has questions for the *UXO remediation community*, as well.

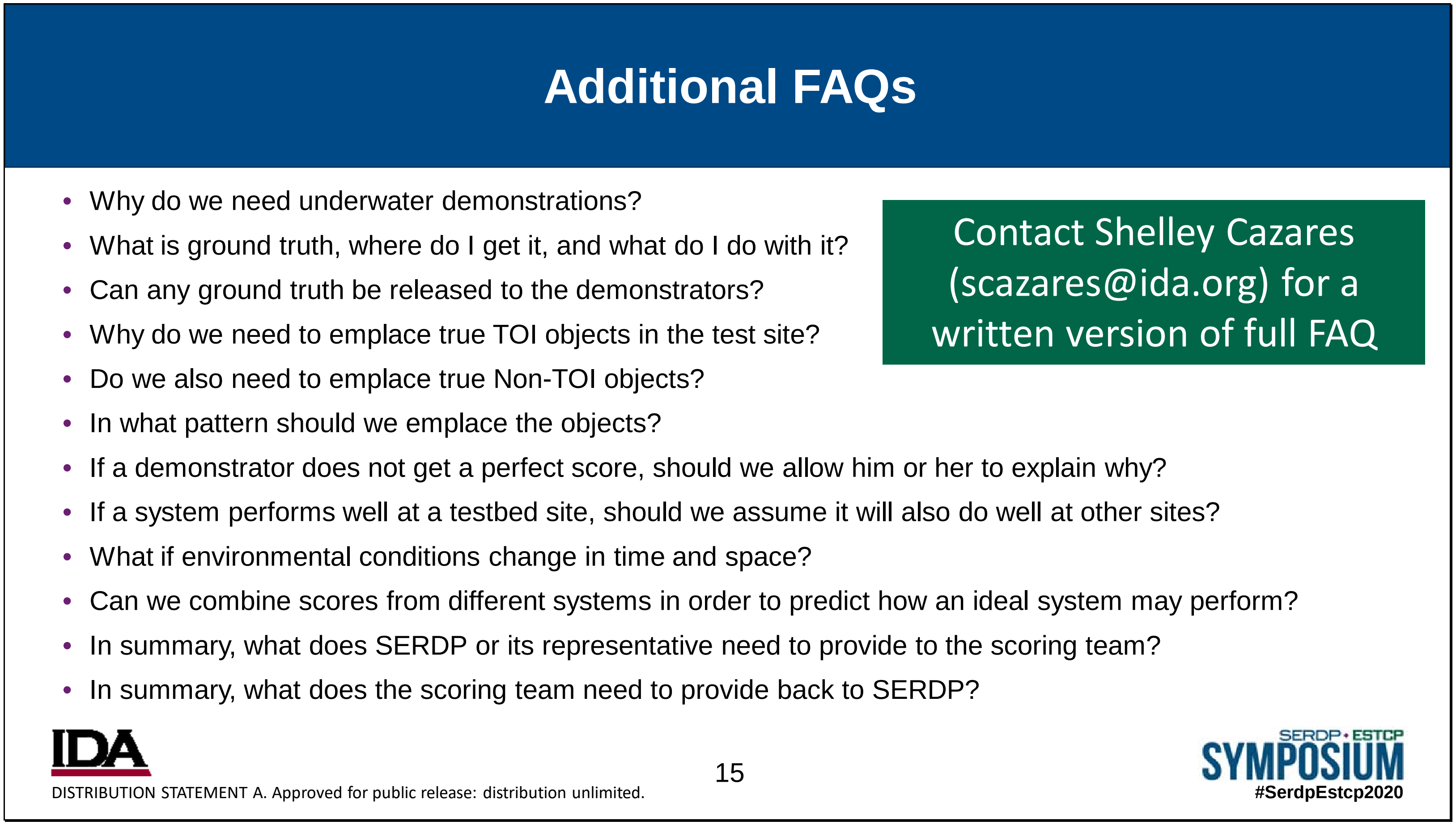
For system developers:

- What geolocation error and sensor resolution do you expect from your systems? (e.g., 2 m? 3 m?)

For site managers:

- When you send your divers out to a specific location coordinate to remediate a UXO, how far out do you expect them to clear? (e.g., 1 m radius? 2 m radius?)
- What environmental conditions do you document and track for your sites? (e.g., wind speed, sea state, etc)

This information will help SERDP design the next series of underwater demonstrations.



Additional FAQs

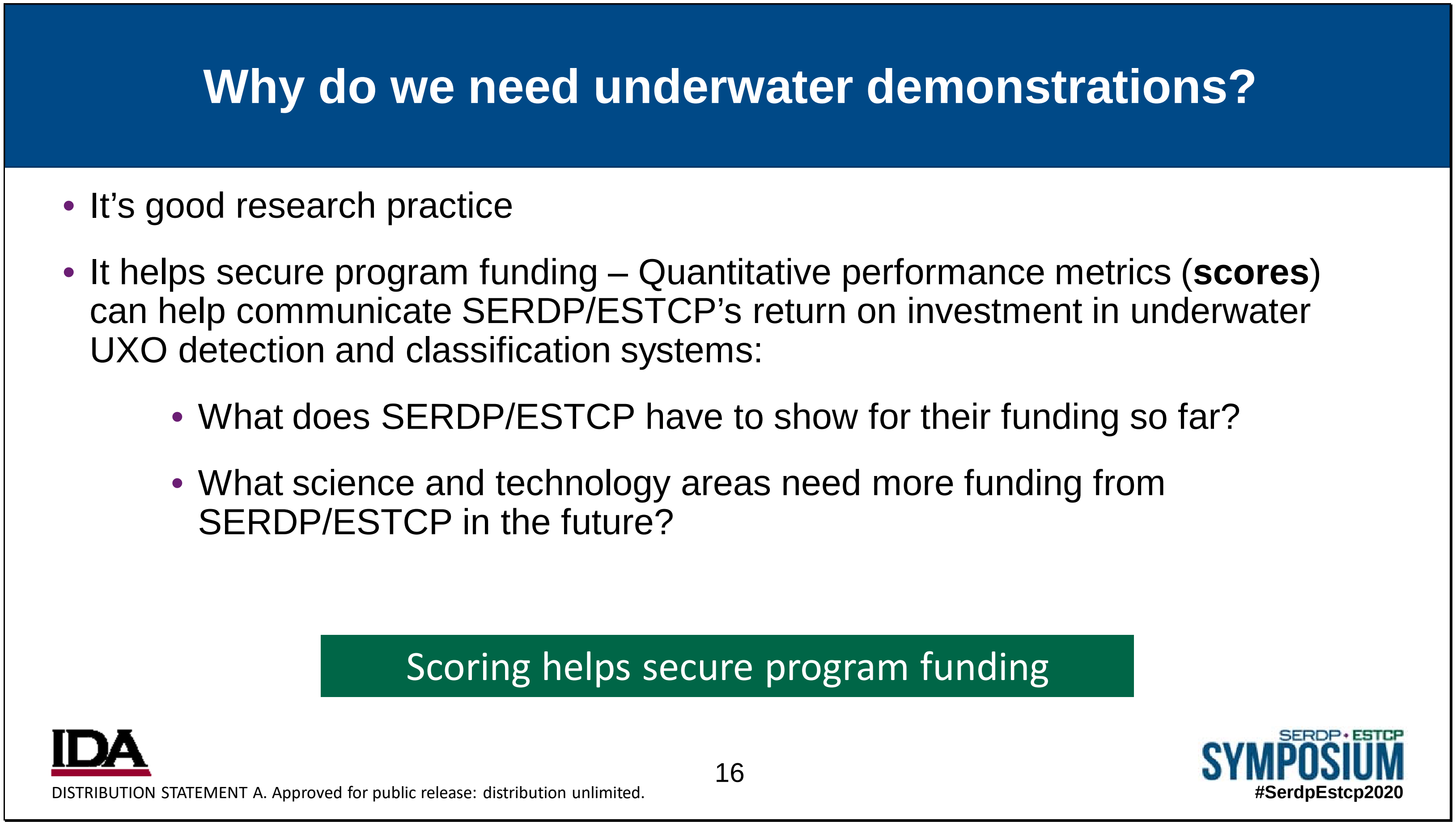
- Why do we need underwater demonstrations?
- What is ground truth, where do I get it, and what do I do with it?
- Can any ground truth be released to the demonstrators?
- Why do we need to emplace true TOI objects in the test site?
- Do we also need to emplace true Non-TOI objects?
- In what pattern should we emplace the objects?
- If a demonstrator does not get a perfect score, should we allow him or her to explain why?
- If a system performs well at a testbed site, should we assume it will also do well at other sites?
- What if environmental conditions change in time and space?
- Can we combine scores from different systems in order to predict how an ideal system may perform?
- In summary, what does SERDP or its representative need to provide to the scoring team?
- In summary, what does the scoring team need to provide back to SERDP?

Contact Shelley Cazares (scazares@ida.org) for a written version of full FAQ

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

#SerdpEstcp2020

We have received many other Frequently Asked Questions, which we briefly list here. We have written then all up in a text document. Please contact us for a copy of the written document. This is a living document that is not set in stone– we fully expect it to evolve as systems and demonstration techniques mature. Therefore we greatly welcome your feedback. Thank you to all who have already reviewed early versions of the document and provided such helpful suggestions.



Why do we need underwater demonstrations?

- It’s good research practice
- It helps secure program funding – Quantitative performance metrics (**scores**) can help communicate SERDP/ESTCP’s return on investment in underwater UXO detection and classification systems:
  - What does SERDP/ESTCP have to show for their funding so far?
  - What science and technology areas need more funding from SERDP/ESTCP in the future?

Scoring helps secure program funding

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

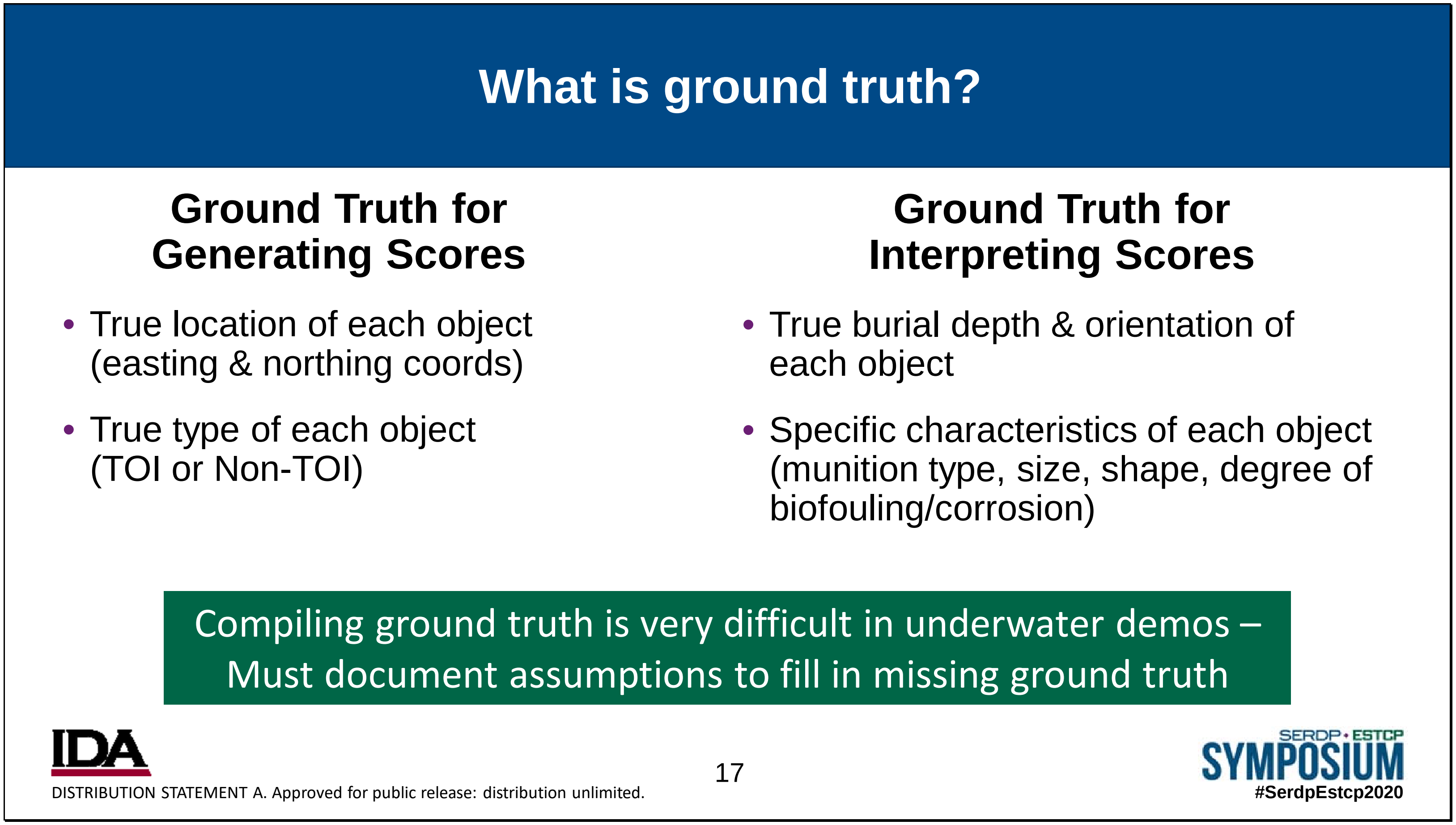
SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

One question often asked is: Why do we need underwater demonstrations in the first place? And why do we need to score them so rigorously?

The main reason is that it's just good research practice.

However, there's another reason: It helps secure funding of the program as a whole. Over the last several years, SERDP and ESTCP have spent tens of millions of dollars sponsoring the development of novel systems and processes for underwater UXO detection and classification. American taxpayers and lawmakers want to understand SERDP/ESTCP's return on investment (ROI). Demonstrations can communicate that ROI through quantitative performance metrics, i.e., scores. Scores can help tell two kinds of stories. First, scores can quickly summarize how far novel systems and processes have come. That is, scores can provide a clear indication of just what SERDP/ESTCP has to show for their funding so far. Second, scores can also clarify which science and technology areas require *additional* investment from SERDP/ESTCP in the future, so that the systems and processes can reach technical maturity.

In short, rigorously scored demonstrations can help secure program funding.



What is ground truth?

Ground Truth for Generating Scores

- True location of each object (easting & northing coords)
- True type of each object (TOI or Non-TOI)

Ground Truth for Interpreting Scores

- True burial depth & orientation of each object
- Specific characteristics of each object (munition type, size, shape, degree of biofouling/corrosion)

Compiling ground truth is very difficult in underwater demos – Must document assumptions to fill in missing ground truth

IDA

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

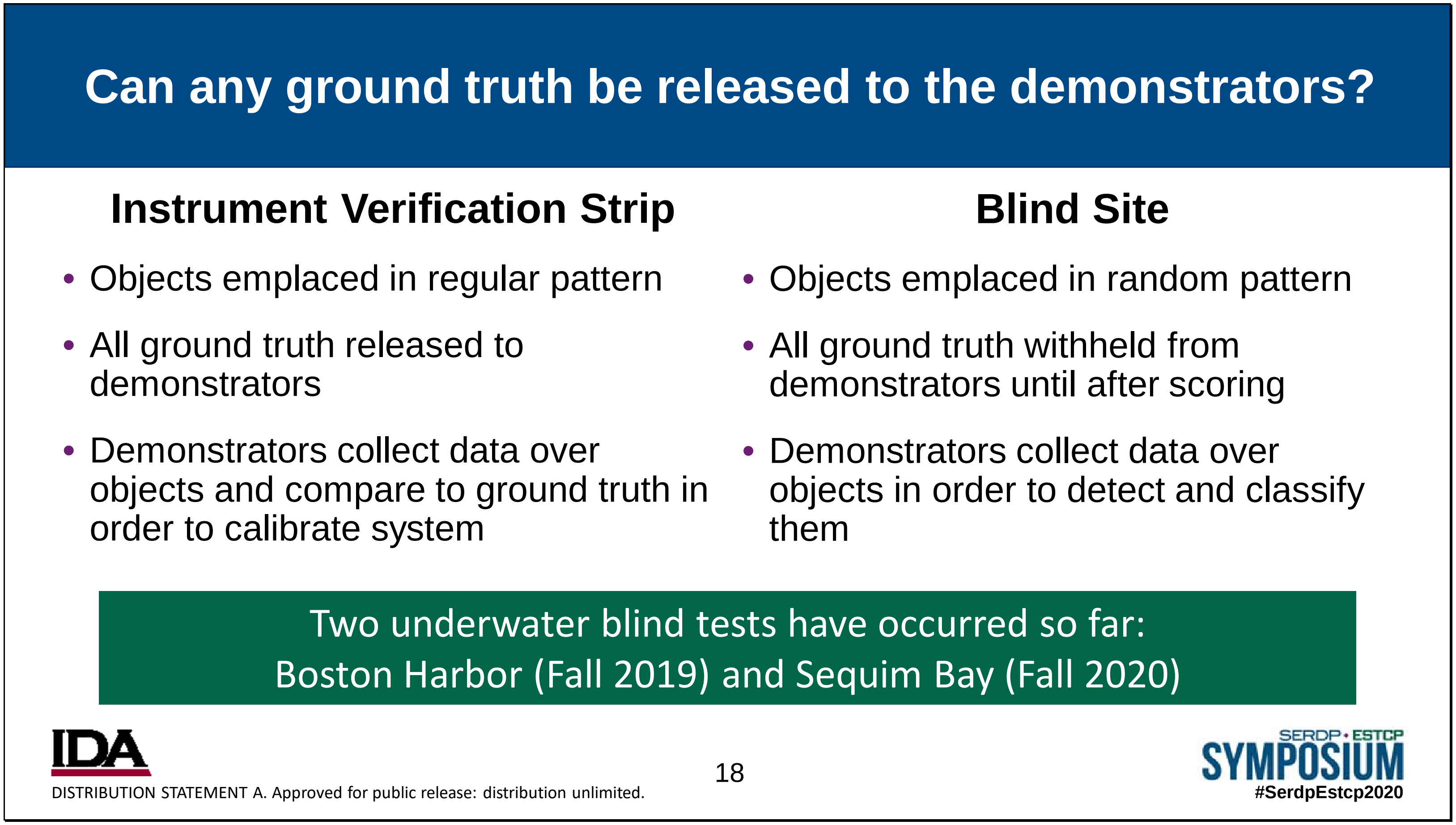
SERDP • ESTCP SYMPOSIUM #SerdpEstcp2020

Two additional questions involve ground truth. First, what is ground truth? There are two types of ground truth:

- The first type of ground truth is used to *generate* the scores in the first place. This ground truth consists of:
  - The true *location* of each object– the Easting and Northing coordinates of the black dots on that bird's eye view figure we showed on a previous slide. This ground truth is used to generate the *Detection* scores.
  - The true *type* of each object– the TOI or Non-TOI labels of each object. This ground truth is used to generate the *Classification* scores, which we summarize with ROC curves.
- There is additional ground truth that is used to help *interpret* the scores. Demonstrators can use this ground truth to help them perform their failure analyses and recommend corrective actions. This ground truth consists of:
  - The true burial *depth* and *orientation* (azimuth and inclination) of each object.
  - Specific *characteristics* of each object (munition type, size, shape, degree of biofouling or corrosion, etc).

Compiling all of this ground truth is tricky for *terrestrial* demonstrations. It's even more difficult– and sometimes impossible– for *underwater* demonstrations, due to the additional safety, logistical, and engineering constraints of the underwater environment. Therefore, some of this ground truth may be collected at a poor resolution or may simply be missing for some underwater demonstrations. For example, the true locations of the emplaced targets may be difficult to survey, due to the inability to use RTK GPS in an underwater environment. Similarly, the true burial depths of the emplaced targets may also be difficult to measure, due to the lack of a firm mud-to-water boundary for some sediments on the seabed floor. And finally, it may be cost prohibitive to even *attempt* to measure the ground truth for *all* detected objects, aside from those that correspond to targets purposefully emplaced for the demonstration.

Despite these challenges, the underwater demonstrations must still be scored. Therefore, the scoring team will have to make some *assumptions* about the missing ground truth, relying on theirs or others' subject matter expertise. *These ground truth assumptions must be clearly documented and included in the scoring report.*



Can any ground truth be released to the demonstrators?

Instrument Verification Strip

- Objects emplaced in regular pattern
- All ground truth released to demonstrators
- Demonstrators collect data over objects and compare to ground truth in order to calibrate system

Blind Site

- Objects emplaced in random pattern
- All ground truth withheld from demonstrators until after scoring
- Demonstrators collect data over objects in order to detect and classify them

Two underwater blind tests have occurred so far:
Boston Harbor (Fall 2019) and Sequim Bay (Fall 2020)

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

#SerdpEstcp2020

Another question is: Can *any* ground truth be released to the demonstrators?

It depends on whether the test makes use of an *Instrument Verification Strip (IVS)* and/or a *Blind Site*:

- An IVS consists of objects emplaced in a regular pattern—often at set intervals along a linear strip. All ground truth about the IVS objects should be released to the demonstrators even *before* the test begins. The demonstrators can collect data over the IVS objects at regular time intervals throughout the test– every few hours, if they'd like– and use this data to calibrate their system. IVSs are often called *calibration strips* or *calibration lanes*. If they'd like, demonstrators can detect and classify the objects in the IVS to produce detection lists and ranked detection lists. These lists can be scored, either formally or informally (self-scored). Tests like these are often called *Engineering Tests*.
- A Blind Site consists of objects emplaced in a much more random pattern. For a truly blind test, all ground truth about the Blind Site objects should be withheld from the demonstrators, at least until scoring is complete. Demonstrators collect data over the Blind Site and then process it without knowledge of ground truth, in order to form their detection lists and ranked detection lists, which are then formally scored (by a third-party organization like IDA). Tests like these are often called *Blind Tests*.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE October 2020	2. REPORT TYPE FINAL	3. DATES COVERED *(From–To)*

4. TITLE AND SUBTITLE

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO) (Presentation)

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cazares, Shelley M.
Bartel, Jacob B.

5d. PROJECT NUMBER
AM-2-1528

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document NS D-18431

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SERDP/ESTCP
4800 Mark Center Drive, Suite 16F16
Alexandria, VA 22350-3605

10. SPONSOR/MONITOR'S ACRONYM(S)

SERDP/ESTCP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (4 October 2020).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

SERDP/ESTCP is sponsoring the development of novel systems for the detection and classification of Unexploded Ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems. Scoring these demonstrations is a complicated process. The Institute for Defense Analyses (IDA) designed and implemented the scoring process for ESTCP's previous terrestrial demonstrations from 2007 – 2017. In some cases, the lessons learned from the terrestrial demonstrations can be leveraged in the underwater demonstrations. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment. This presentation will provide an overview of the main considerations for scoring underwater demonstrations for UXO detection and classification.

15. SUBJECT TERMS

acoustic color; demonstration test; electromagnetic induction (EMI); False Alarm Rate (FAR); Probability of Detection (Pd); Probability of False Alarm (Pfa); Receiver Operating Characteristic (ROC) curve; scoring; testbed; underwater; unexploded ordnance (UXO)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Bradley, David
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	39	19b. TELEPHONE NUMBER *(include area code)* 571-372-6388