Systematizing Confidence in Open Research and Evidence (SCORE)

Adam Russell
Defense Sciences Office

SCORE Proposers Day

June 8, 2018

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Proposal Tips

Read the BAA! (If the BAA differs from this presentation, be guided by the BAA)

- If in doubt, address the Heilmeier Catechism
- Don't overlook mandatory inclusions as highlighted by the BAA – a great idea can be sunk by ignoring the details
- Present a compelling, innovative approach that isn't addressed by current state of the art - describe how it will advance the science, provide new capabilities, and positively impact DoD
- Back up your ideas and technical approaches (e.g., theoretical arguments, models, past results, new data)
- Provide quantitative metrics and milestones to assist DARPA in evaluating feasibility and transparency of proposed work
- Where possible, go open-source. If you can't, provide strong justification.
- Don't forget to address risks! "Hope is not a management strategy."

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Objective

Automated tool to quantify the confidence DoD should have in social and behavioral science (SBS) research claims

Outcome:

- **Automated capabilities for assigning Confidence Scores** (CSs) for the Reproducibility and Replicability (R&R) of different SBS research claims
- **Automated mechanisms for updating Confidence Scores** based on new information (retractions, etc.) and/or new signals (social media, etc.)
- **Tailored, interpretable Confidence Scores** for different users and applications

Impact:

- Enhance DoD's capabilities to leverage SBS research
- Enable more effective SBS modeling and simulation
- Guide future SBS research towards higher CSs

The "R&R" of SCORE

Reproducibility: The extent to which results can be computationally reproduced by others

Replicability: The degree to which results can be replicated by others

DISTRIBUTION STATEMENT A. Approved for public release.

What is the problem SCORE is addressing?

Effective use of SBS research for national security is hampered by questions of its reproducibility and replicability

There is increasing evidence that there are widespread uncertainties in the confidence one should have in many SBS research claims (e.g., refs here).

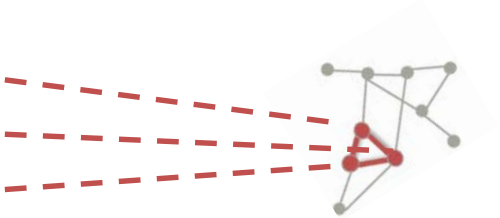
But this does not mean that all SBS research is untrustworthy.

How can an SBS consumer practically know the difference?

Search/evaluate SBS Literature

Create Models

Applications



SCORE Impact

Impact:

- Enhance DoD's capabilities to leverage SBS research
- Enable more confident SBS modeling and simulation
- Guide future SBS research towards higher Confidence Scores

Research Vetting:

Low Time Cost
Minutes

+

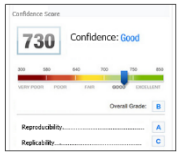
Low Financial Cost
Cents

+

Wide Coverage
Many signals

+

Quantified Results
Interpretable



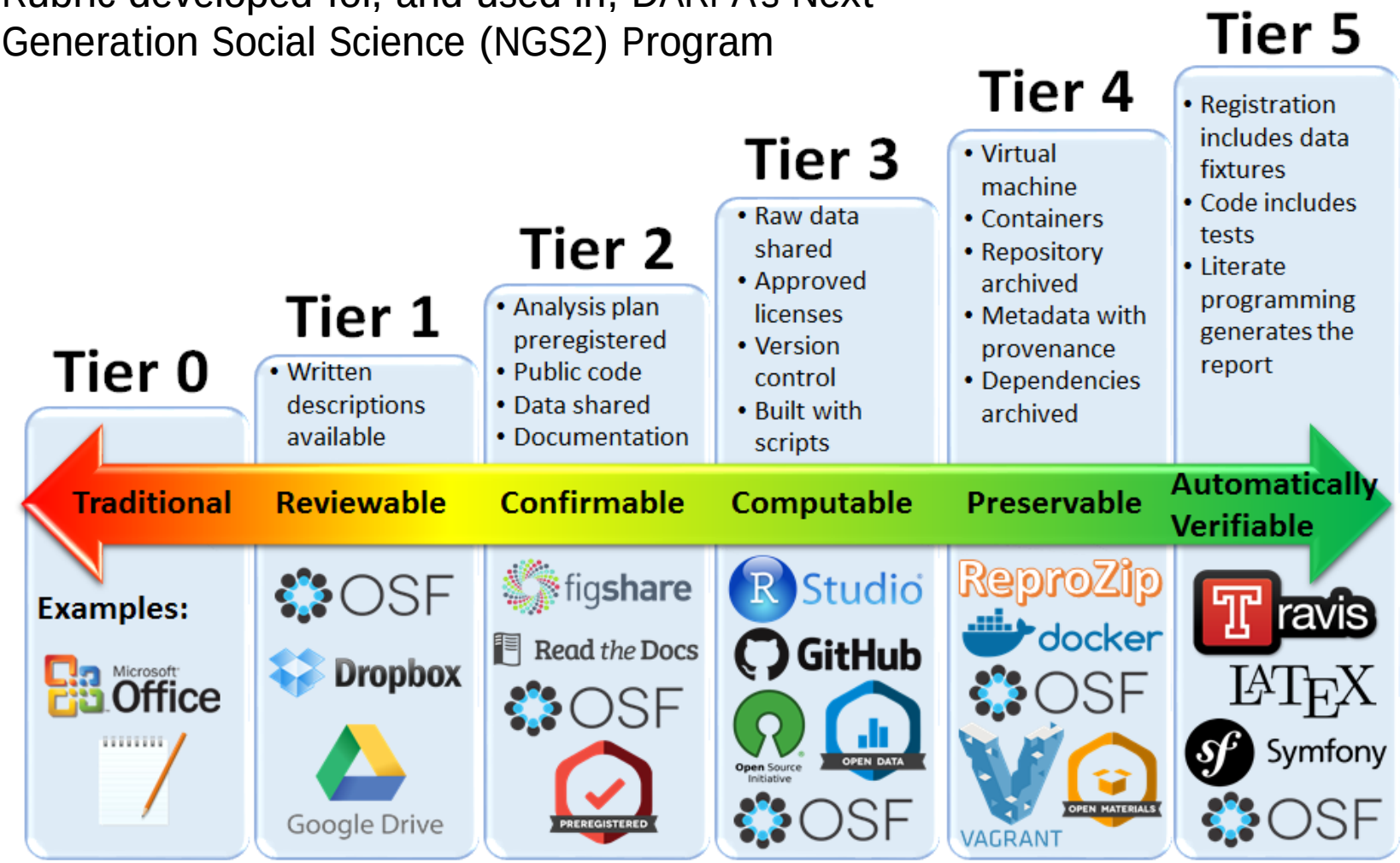
+

User Focused
Tailorable

SCORE will improve DoD's efficiency in evaluating SBS research, and increase confidence in how that research can be leveraged for the Human Domain

Computational Reproducibility Rubric

Rubric developed for, and used in, DARPA's Next Generation Social Science (NGS2) Program



https://goo.gl/ns1vDj

DISTRIBUTION STATEMENT A. Approved for public release.

Technical Challenges

Automated tools to quantify the confidence DoD should have in SBS research and claims requires…

Creating algorithms that can quantify confidence …

- With results equal to or better than the best expert methods
- With explainable and tailorable outputs

Developing approaches for expert scoring of SBS studies for algorithm training/test set …

- With sufficient speed and accuracy
- With ability to understand basis for scores

Preparing a curated (selected and organized) dataset of diverse SBS literature…

- At a rate that can train/test effective machine learning algorithms
- With sufficient diversity while being machine-readable

Empirically testing the R&R of a representative subset of studies…

- At a rate sufficient to provide assurance of expert accuracy
- That reflects different content, authors, journals

Why now?

"Weak Signals" for Algorithms to Exploit

Expert Predictions at Scale

Open Research and Replication Platforms

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Mechanics

- Program Structure
- Technical Areas
- Evaluation and Performance Metrics
- Teams and Teaming
- Proposal Details

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Program structure

- The SCORE program will be divided into three Technical Areas (TAs) with an independent Test and Evaluation (T&E) team providing oversight. The three TAs are:
  - **TA1: Data**
  - **TA2: Experts**
  - **TA3: Algorithms**

- Proposals to any of the TAs must address the full program timeline, however TA3 teams will officially start work after Month 6 in Phase 1
- Proposers should structure their proposals with Phase 1 as the **base period** and Phase 2 as an **option** for funding
- Please note that to avoid conflicts of interest, no person or organization may be a performer for more than one TA, whether as a prime or as a sub-contractor

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Program Phases

Phase 1: Data +Experts
18 months

Phase 2: Data + Algorithms
18 months

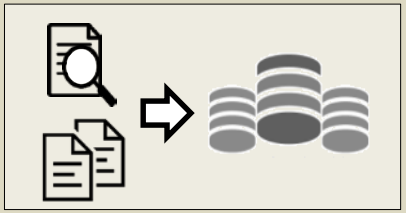
1Q FY19

Program Kickoff

3Q FY20

Phase 1 Outcome

SCORE Common Task Framework (CTF)
Create datasets of SBS research with validated Confidence Scores



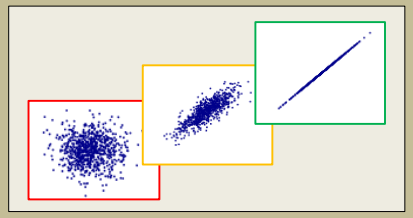
GOALS

- 6K+ curated SBS articles in CTF for training/testing algorithms
- Understanding of expert processes and signals

1Q FY22

Phase 2 Outcome

SCORE Algorithms
Build algorithms that rapidly and interpretably assign expert-like Confidence Scores



GOALS

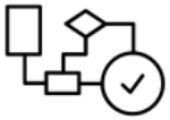
- Trustworthy algorithms that convincingly overlap with experts' judgments
- Algorithms that enhance DoD use of SBS research

SCORE will combine data, experts, and algorithms to create a systematic approach for developing Confidence Score technologies

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Technical Areas

SCORE will develop and test new capabilities to rapidly and accurately estimate the Reproducibility and Replicability (R&R) of SBS research claims



TA1 (Data) Teams will:

- Curate SCORE datasets for TA2 and TA3 teams
- Empirically evaluate representative sample of studies to test accuracy of TA2 methods
- Test TA3 algorithms' ability to update and detect gaming efforts

TA2 (Experts) Teams will:

- Assign CSs to all TA1 datasets via "expert" crowd-sourcing methods
- Be ≥80% accurate in predicting TA1 R&R empirical evaluations in each phase
- Capture signals that experts use to assign confidence levels

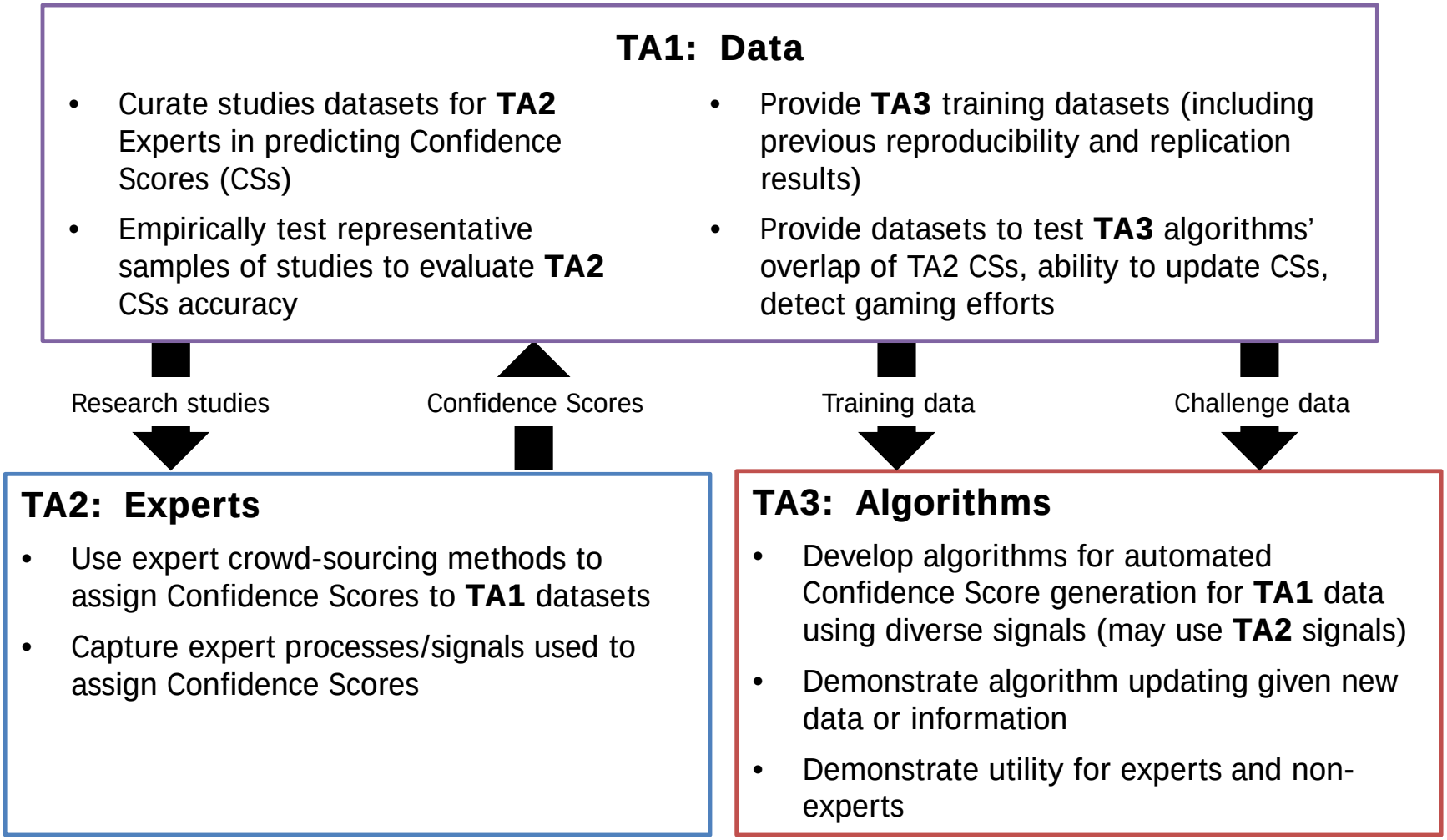
TA3 (Algorithms) Teams will:

- Create algorithms that assign CSs to TA1 test datasets that correlate with best TA2 team CSs
- Demonstrate usability of algorithms/systems for DoD SBS consumers

DISTRIBUTION STATEMENT A. Approved for public release.

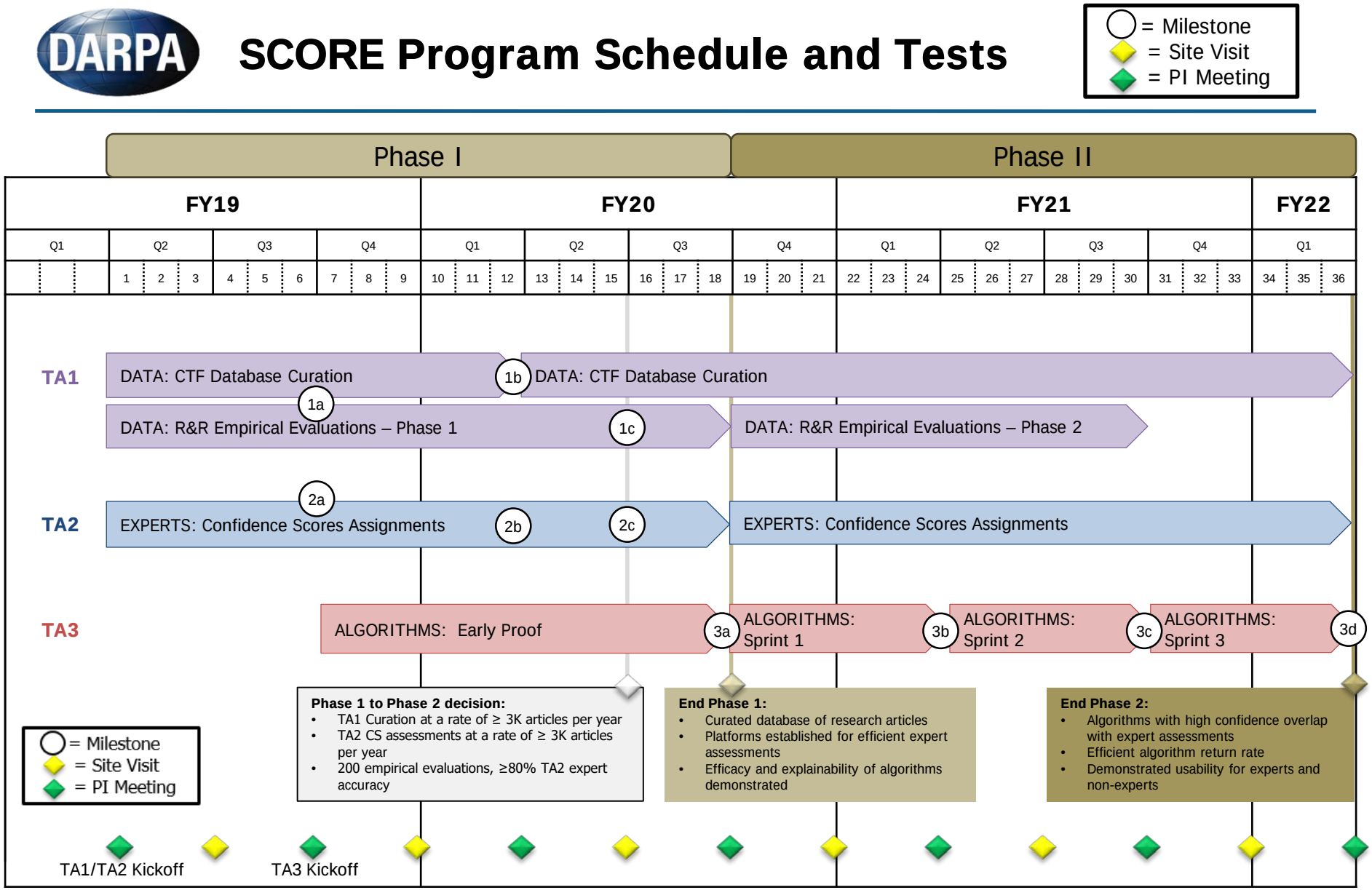
SCORE Technical Areas

SCORE is a two-phase program built around three technical areas



DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Program Schedule and Tests



Please see Figure 3 and Tables 1-2 in the BAA

SCORE Mid-term and Final Exams

	Metric	SOA	Phase 1	Phase 2	Outcome
TA1 Data	Curation rate	?	>3K per year	>3K per year	CTF Datasets for SCORE
	R&R Empirical Evaluation	100 studies /12 months	200 studies /15 months	200 studies /12 months	
TA2 Experts	CSs Assignment rate	?	>3K per year	>3K per year	Accurate Confidence Scores
	Accuracy	75%	80%	>80%	
TA3 Algorithms	Scoring rate	N/A	1 study /hour	1 study /30 minutes	SCORE Algorithms
	Correlation with TA2	N/A	Demonstration of efficacy and explainability	75/85/95%	

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Proposal Tips

Read the BAA! (If the BAA differs from this presentation, be guided by the BAA)

- If in doubt, address the Heilmeier Catechism
- Don't overlook mandatory inclusions as highlighted by the BAA – a great idea can be sunk by ignoring the details
- Present a compelling, innovative approach that isn't addressed by current state of the art - describe how it will advance the science, provide new capabilities, and positively impact DoD
- Back up your ideas and technical approaches (e.g., theoretical arguments, models, past results, new data)
- Provide quantitative metrics and milestones to assist DARPA in evaluating feasibility and transparency of proposed work
- Where possible, go open-source. If you can't, provide strong justification.
- Don't forget to address risks! "Hope is not a management strategy."

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE encourages multidisciplinary teaming!

- DARPA highly encourages – and will facilitate – teaming. See BAA, VIII.B.
- Teaming Profiles are due June 12, 2018 no later than 4:00pm Eastern
- Consolidated teaming profiles will be sent via email to the proposers who submitted a valid profile

- However… DARPA will attempt to update the consolidated teaming profiles with submissions past the due date
- Interested parties can still submit a one-page profile including the following information to SCORE@darpa.mil:
  - Contact information
  - Proposer's technical competencies.
  - Desired expertise from other teams, if applicable

- **Complete teaming information is not required for abstract submission**

DISTRIBUTION STATEMENT A. Approved for public release.

But!...

Specific content, communications, networking, and team formation are the sole responsibility of the participants.

Neither DARPA nor the DoD endorses the information and organizations contained in the consolidated teaming profile document, nor does DARPA or the DoD exercise any responsibility for improper dissemination of the teaming profiles.

SCORE Key Dates

BAA Published	Anticipated June 12, 2018
Teaming Profiles Due	June 12, 2018
Proposers Day	June 8, 2018
Abstracts Due (TA1 and TA2)	June 18, 2018
Abstracts Due (TA3)	November 1, 2018
FAQ Submissions Due	July 24, 2018
Proposals Due (TA1 and TA2)	July 31, 2018
Proposals Due (TA3)	December 12, 2018

Please refer to the BAA for any changes in dates

DISTRIBUTION STATEMENT A. Approved for public release.

Intellectual Property

- Data sharing and collaboration are key aspects of this program
- Therefore, intellectual property rights asserted by proposers are strongly encouraged to be aligned with open source regimes
- **See Section VI.B in the BAA for further information**

Proposal Abstracts

Proposers are **highly encouraged** to submit an abstract

- Submit to https://baa.darpa.mil/ (do not submit via email) – see BAA Section IV.E.1 for details
- DARPA will respond to abstracts with a statement as to whether DARPA is interested in the idea
  - While it is DARPA policy to attempt to reply to abstracts within thirty calendar days, proposers may anticipate a response within approximately two weeks
- Regardless of DARPA's response to an abstract, proposers may submit a full proposal
- Abstracts will be reviewed in the order they are received
- DARPA will review all full proposals submitted using the published evaluation criteria and without regard to any comments resulting from the review of an abstract
- Complete teaming information is not required for abstract submission

DISTRIBUTION STATEMENT A. Approved for public release.

SCORE Evaluation Criteria

- **Review and Selection Process:** DARPA will conduct a scientific/technical review of each conforming proposal. Proposals will not be evaluated against each other since they are not submitted in accordance with a common work statement.

- **Evaluation Criteria**: Proposals will be evaluated using the following criteria, listed in descending order of importance:
- **(a) Overall Scientific and Technical Merit**;
- **(b) Potential Contribution and Relevance to the DARPA Mission;**
- **(c) Cost Realism**

(See BAA Section V. A. for specific details on each criterion)

References

Camerer et al. (2016). “Evaluating replicability of laboratory experiments in economics.” *Science*, 351(6280).

Camerer et al. (unpublished). “Evaluating the Replicability of Social Science Experiments in *Nature* and *Science*.”

Chang, Andrew C., and Phillip Li (2015). “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ”Usually Not”,” *Finance and Economics Discussion Series* 2015-083. Washington: Board of Governors of the Federal Reserve System.

Dreber, Anna et al. (2015). “Using predication markets to estimate the reproducibility of scientific research.” *PNAS,* 112(50.

Ioannidis, John & T. D. Stanley, Hristos Doucouliagos (2017). “The Power of Bias in Economics Research.” *The Economic Journal,* 127(605).

Klein, Rick & Michelangelo Vianello, Fred Hasselman, and Brian Nosek (unpublished). “Many Labs 2: Investigating Variation in Replicability Across Sample and Setting.” *Advances in Methods and Practices in Psychological Science.*

Kovanis M, Porcher R, Ravaud P, Trinquart L (2016). “The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise.” *PloS ONE,* 11(11).

Maket, Matthew & Jonathan Plucker, Boyd Hegarty (2012). “Replications in Psychology Research.” *Perspectives on Psychological Science*, 7(6).

Martin, GN & Clarke R (2017). “Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices.” *Frontiers in Psychology*,

Munafo, Marcus et al. (2017). “A manifesto for reproducible science.” *Nature Human Behavior*, 0021.

Neuliep, JW & Crandall R (1990). “Editorial bias against replication research.” *Journal of Social Behavior & Personality*, 5(4).

Nosek, Brian & Jeffrey Spies, Matt Motyl (2012). “Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability.” *Perspectives on Psychological Science*, 7(6).

Nosek, Brian et al. (2015). “Estimating the reproducibility of psychological science.” *Science,* 349(6251).

Smith, Richard (2006). “Peer Review: a flawed process at the heart of science and journals.” *Journal of the Royal Society of Medicine,* 99(4).

Steen RG, Casadevall A, Fang FC (2013). “Why Has the Number of Scientific Retractions Increased?” *PLoS ONE,* 8(7).

Szucs, Denes & John Ioannidis (2017.). “Empirical assessment of publiched effect sizes and power in the recent cognitive neuroscience and psychology literature.” *PLoS Biology,* 15(3).

Veldkamp CLS, Nuijten MB, Dominguez-Alvarez L, van Assen MALM, Wicherts JM (2014). “Statistical Reporting Errors and Collaboration on Statistical Analyses in Psychological Science.” *PLoS ONE,* 9(12).

DISTRIBUTION STATEMENT A. Approved for public release.

Thank you

DISTRIBUTION STATEMENT A. Approved for public release.