The Consortium for Enabling Technologies and Innovation

Virtual Summer Meeting for Young Researchers

Identifying Data Streams Valuable to Collaborative Intelligence

Jordan Stomps

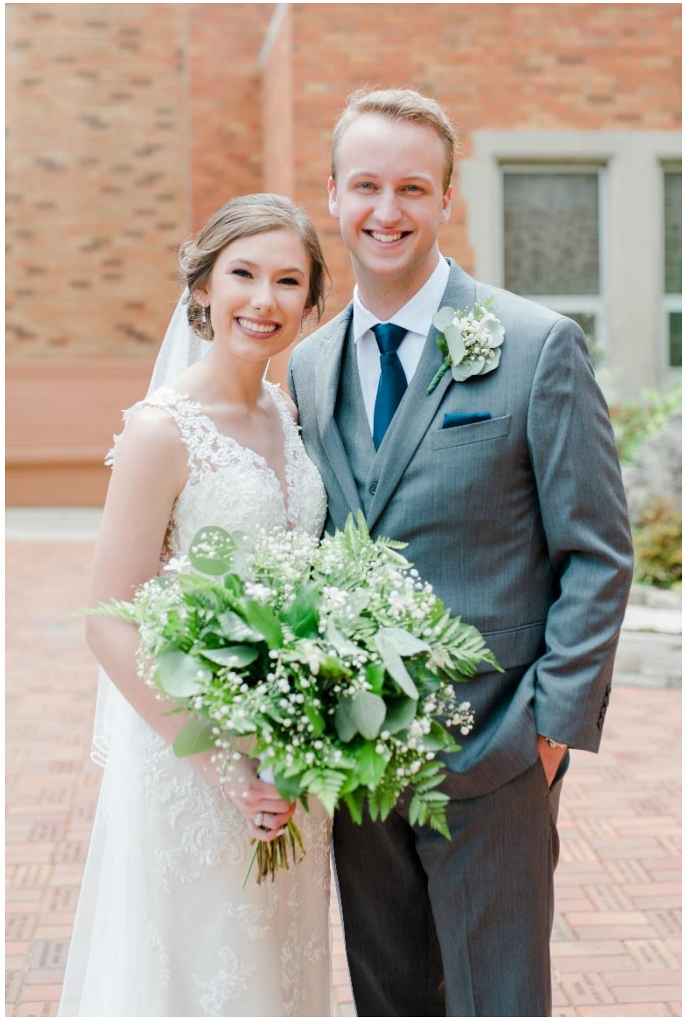
University of Wisconsin-Madison

07-07-2020

National Nuclear Security Administration

About Me

- Rising 2nd-year PhD Student
- Advisor: Paul Wilson, University of Wisconsin-Madison
- Background: B.S. Physics, Michigan State University
- Research Interest: Machine Learning, Data Science, and Fuel Cycle Modeling Applications to Nuclear Nonproliferation.
- Fun Fact: Newly married as of a month ago!



Background

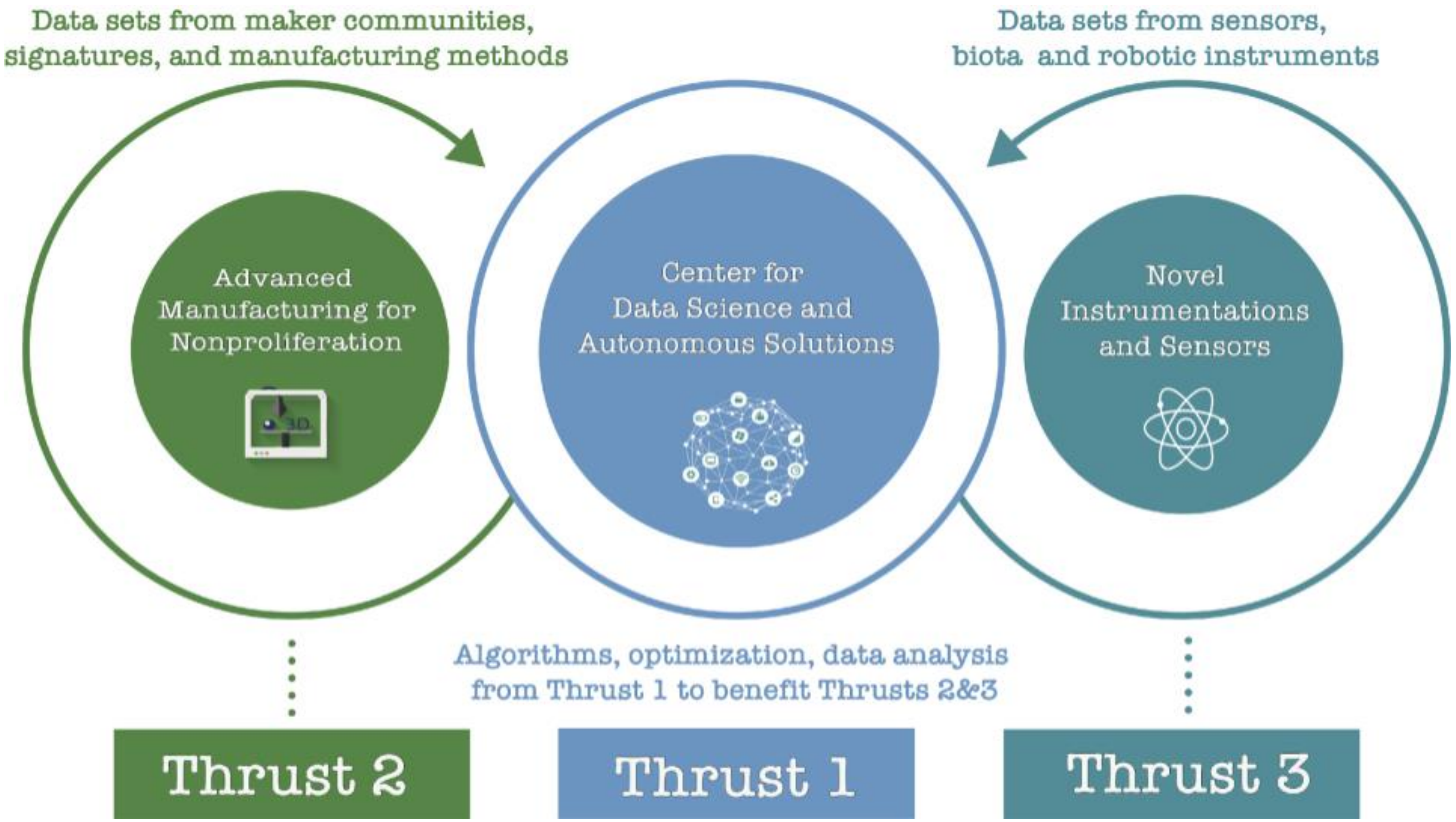
What Is the Consortium for Enabling Technology and Innovation?

Mission: direct the multidisciplinary research and innovation that enable the technologies that support the NNSA, to train the next-generation of human capital, and to bridge the gap between the university basic research and national laboratories' mission-specific applications.

What is Thrust Area 1?

Computer & Engineering Sciences for Nonproliferation

Using advanced computing techniques to develop data collection algorithms



Data sets from maker communities, signatures, and manufacturing methods

Data sets from sensors, biota and robotic instruments

Advanced Manufacturing for Nonproliferation

Center for Data Science and Autonomous Solutions

Novel Instrumentations and Sensors

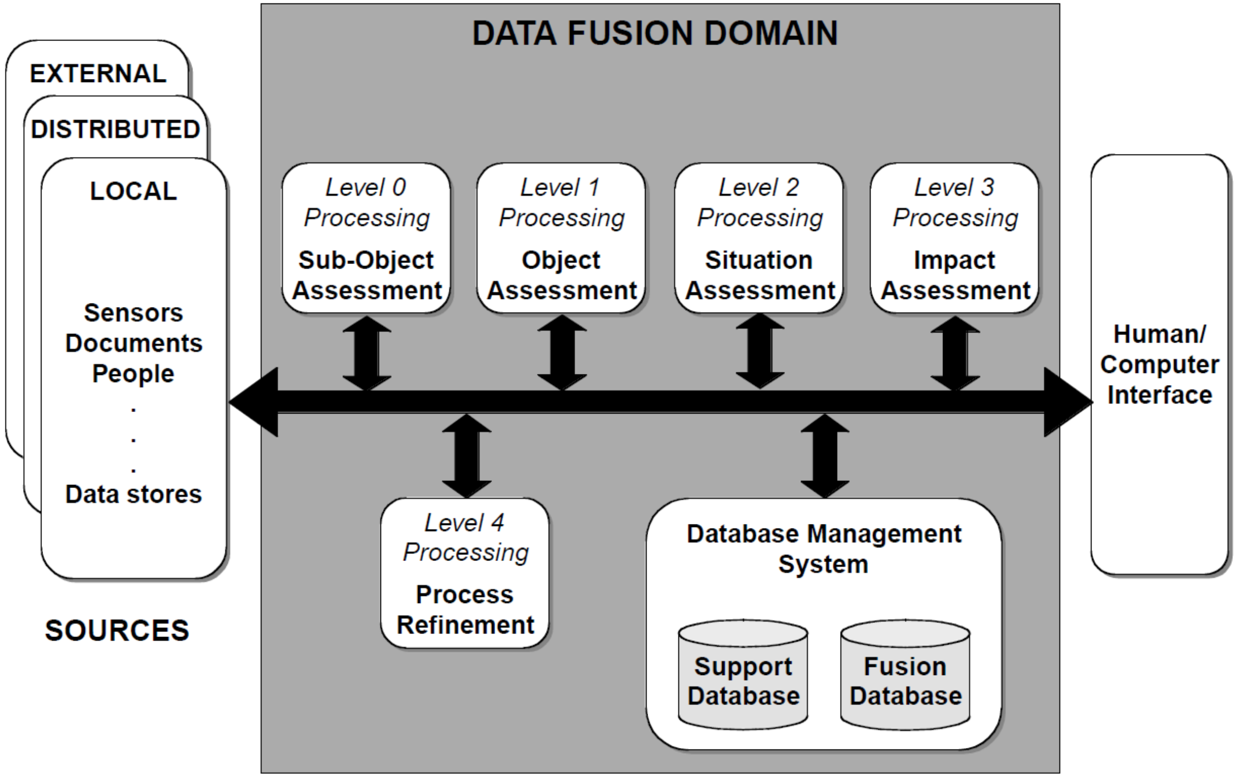
Algorithms, optimization, data analysis from Thrust 1 to benefit Thrusts 2&3

Thrust 2

Thrust 1

Thrust 3

PROPOSAL TO FUNDING OPPORTUNITY ANNOUNCEMENT DE-FOA-0001875



doi.org/10.1117/12.341367

Data Fusion

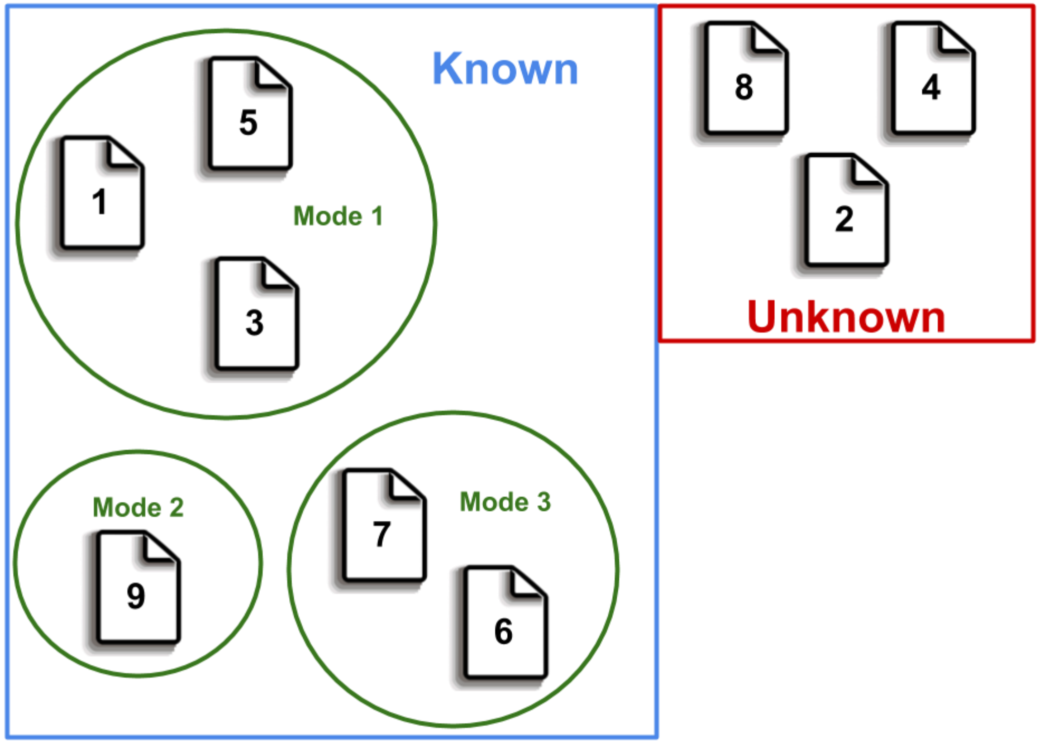
Definition:

The process of combining data or information to estimate or predict entity states.

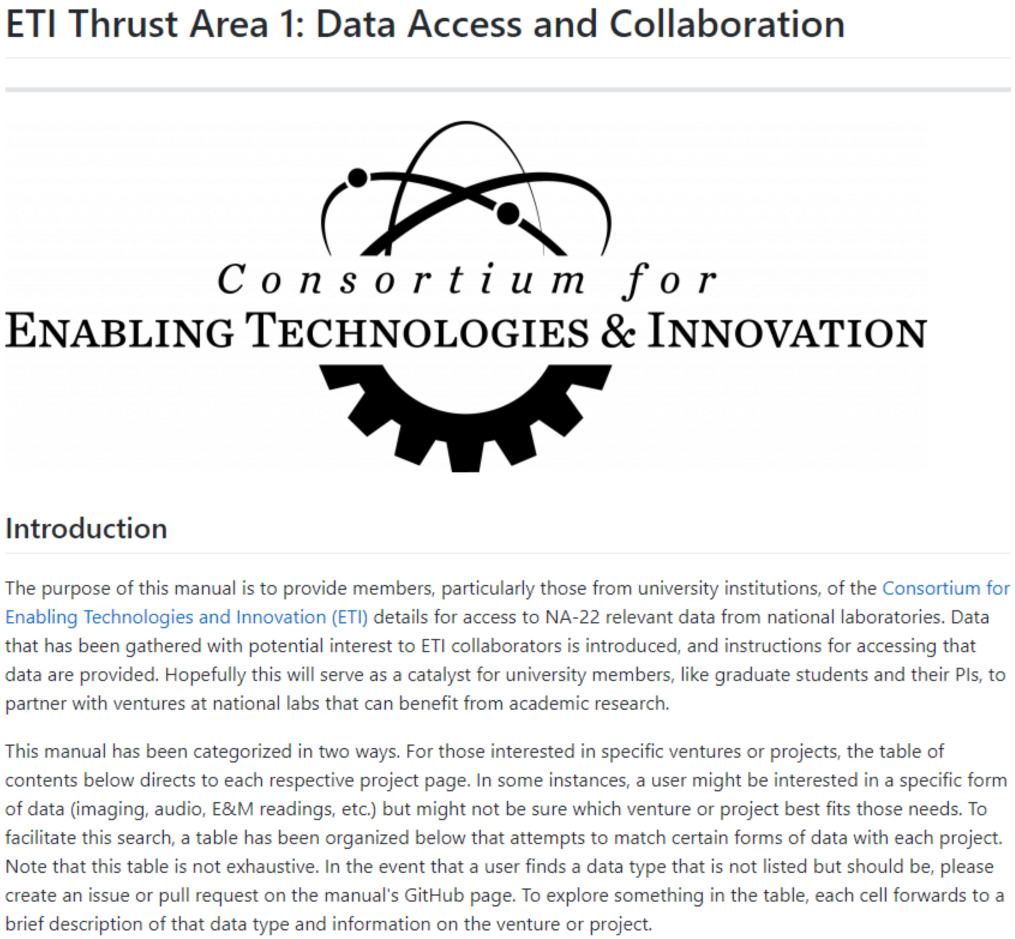
Data Fusion Framework:

1. Identify a focused purpose,
2. Facilitate user understanding/communication,
3. Permit comparison and integration,
4. Promote expandability, modularity, and reusability,
5. Promote cost-effective system development,
6. Apply to the required range of situations.

Problem



Solution



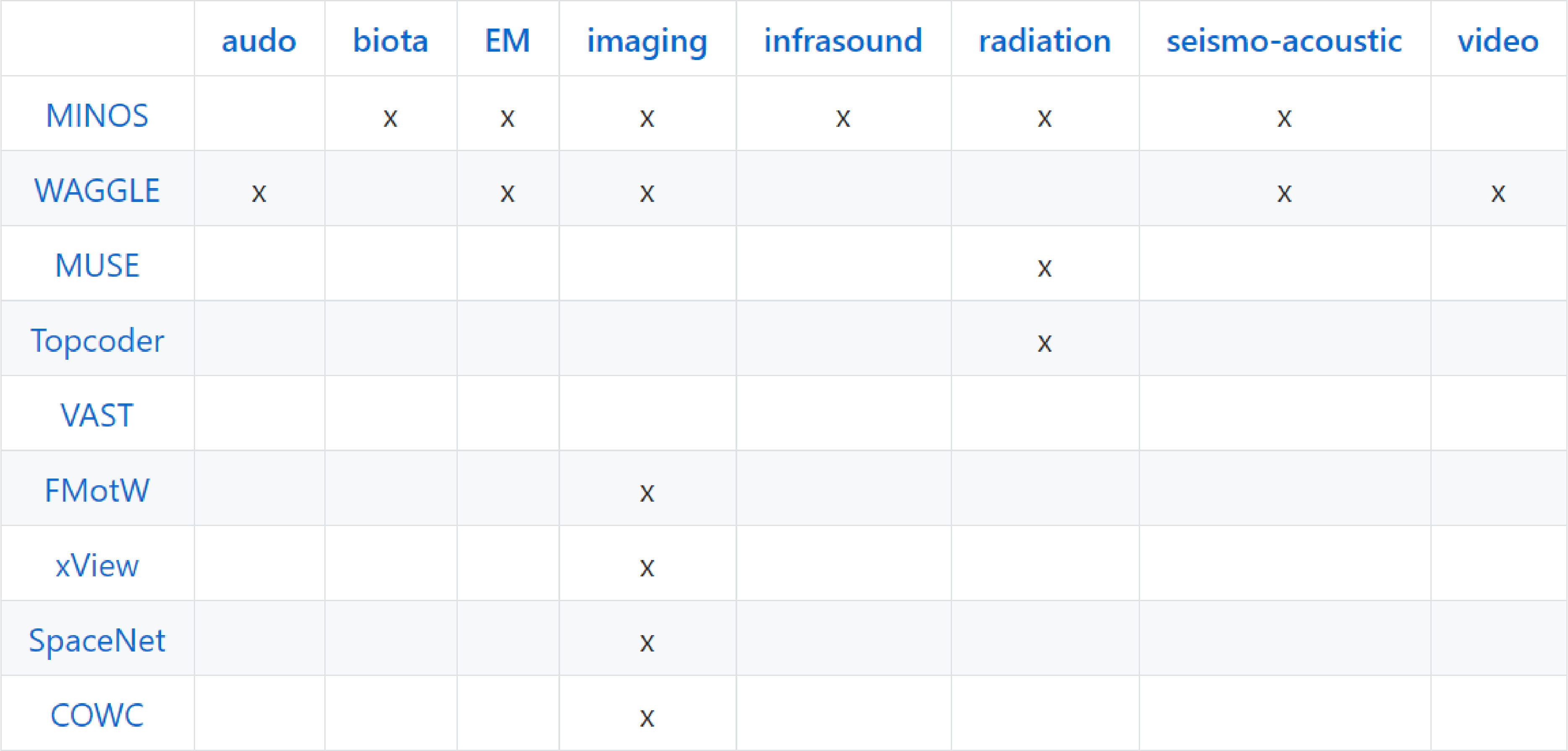
ETI Thrust Area 1: Data Access and Collaboration

Introduction

The purpose of this manual is to provide members, particularly those from university institutions, of the Consortium for Enabling Technologies and Innovation (ETI) details for access to NA-22 relevant data from national laboratories. Data that has been gathered with potential interest to ETI collaborators is introduced, and instructions for accessing that data are provided. Hopefully this will serve as a catalyst for university members, like graduate students and their PIs, to partner with ventures at national labs that can benefit from academic research.

This manual has been categorized in two ways. For those interested in specific ventures or projects, the table of contents below directs to each respective project page. In some instances, a user might be interested in a specific form of data (imaging, audio, E&M readings, etc.) but might not be sure which venture or project best fits those needs. To facilitate this search, a table has been organized below that attempts to match certain forms of data with each project. Note that this table is not exhaustive. In the event that a user finds a data type that is not listed but should be, please create an issue or pull request on the manual's GitHub page. To explore something in the table, each cell forwards to a brief description of that data type and information on the venture or project.

Data Streams Table



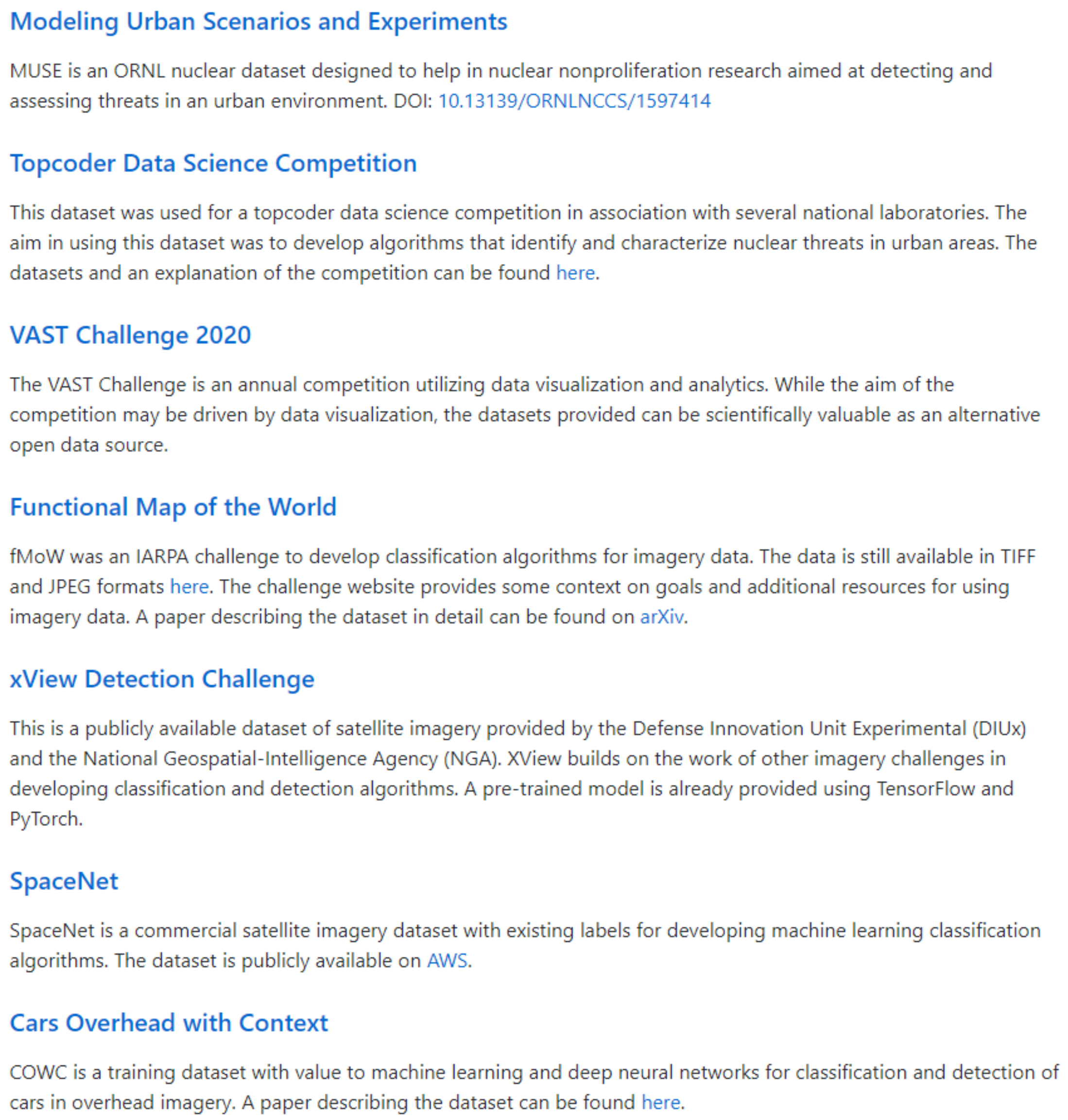
	audo	biota	EM	imaging	infrasound	radiation	seismo-acoustic	video
MINOS		x	x	x	x	x	x	
WAGGLE	x		x	x			x	x
MUSE						x		
Topcoder						x		
VAST								
FMotW				x				
xView				x				
SpaceNet				x				
COWC				x				

Highlighted datasets

Challenge: Providing transparent documentation of NNSA data streams while abiding by privacy/security measures.

Alternative Route: Use datasets that are open-source/openly-available.

- These can be used for training, diagnostics, practice, testing, etc.



Modeling Urban Scenarios and Experiments

MUSE is an ORNL nuclear dataset designed to help in nuclear nonproliferation research aimed at detecting and assessing threats in an urban environment. DOI: 10.13139/ORNLNCCS/1597414

Topcoder Data Science Competition

This dataset was used for a topcoder data science competition in association with several national laboratories. The aim in using this dataset was to develop algorithms that identify and characterize nuclear threats in urban areas. The datasets and an explanation of the competition can be found here.

VAST Challenge 2020

The VAST Challenge is an annual competition utilizing data visualization and analytics. While the aim of the competition may be driven by data visualization, the datasets provided can be scientifically valuable as an alternative open data source.

Functional Map of the World

fMoW was an IARPA challenge to develop classification algorithms for imagery data. The data is still available in TIFF and JPEG formats here. The challenge website provides some context on goals and additional resources for using imagery data. A paper describing the dataset in detail can be found on arXiv.

xView Detection Challenge

This is a publicly available dataset of satellite imagery provided by the Defense Innovation Unit Experimental (DIUx) and the National Geospatial-Intelligence Agency (NGA). XView builds on the work of other imagery challenges in developing classification and detection algorithms. A pre-trained model is already provided using TensorFlow and PyTorch.

SpaceNet

SpaceNet is a commercial satellite imagery dataset with existing labels for developing machine learning classification algorithms. The dataset is publicly available on AWS.

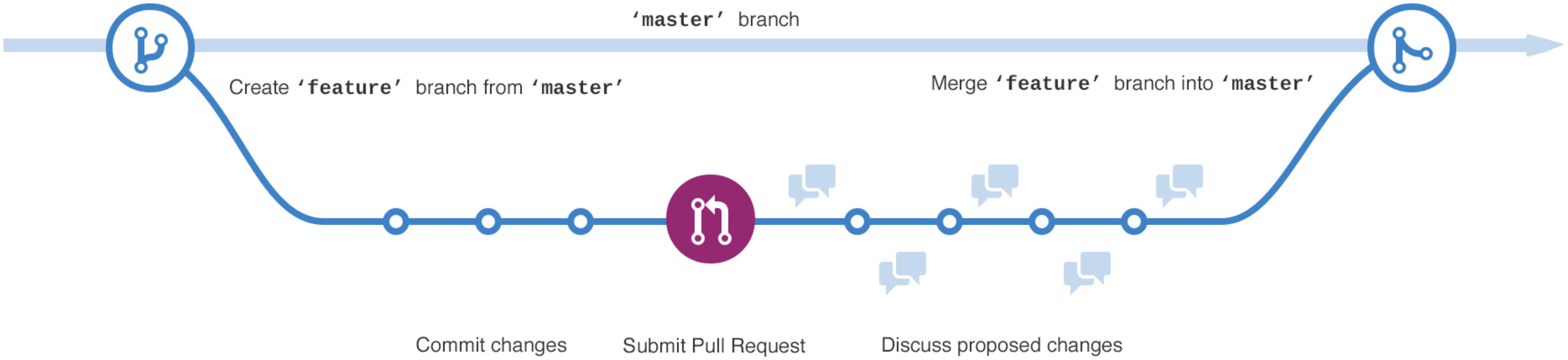
Cars Overhead with Context

COWC is a training dataset with value to machine learning and deep neural networks for classification and detection of cars in overhead imagery. A paper describing the dataset can be found here.

Collaboration

- Eventually someone will identify a previously “unknown” data source.
- The manual lives on GitHub to be open-source, transparent and accessible.

→ Open a new issue or pull request to add information for new data sources.


‘master’ branch
Create ‘feature’ branch from ‘master’
Merge ‘feature’ branch into ‘master’
Commit changes
Submit Pull Request
Discuss proposed changes

guides.github.com/introduction/flow/

Conclusion

- Disparate data streams and sources need to be identified for their value to ETI.
- Identified data needs to be documented in a centralized location.
- Open-source methodology allows ETI collaborators to share newly identified data with others.
- The ETI.data_manual is designed for new and experienced researchers.

View the manual now!

Next Steps

- Continue identifying data streams.
- Encourage dissemination.

Or request access to edit!

Acknowledgement

This material is based upon work supported by the Department of Energy/National Nuclear Security Administration under Award Number DE-NA0003921.