Integrating High-Performance Computing and Machine Learning within the Army Veterinary Service to improve Surveillance of Companion Animal Disease within the Department of Defense

Association for Veterinary Informatics

Sheldon Waugh, MSc, PhD
One Health Division
Veterinary Services and Public Health Sanitation Directorate
Army Public Health Center

Learning Objectives and Outline

- Background and GPAWSS
- MPL, NLP and current progress

- Inform Army Public Health Course attendees about GPAWSS and its objectives
- Educate attendees on the basics of companion animal surveillance data collection in the DoD
- Describe the current initiative to apply Natural Language Processing on current veterinary clinical data

What Do The Regulations Say?

- The Army is the lead service for veterinary public health and animal health services
  - Responsibility to champion biosurveillance efforts to support One Health initiatives, improving Service Member, family, and veteran health across the Joint Force
  - Army Regulation (AR) 40-905 tasks Veterinary Corp Officers (VCOs) to “conduct disease surveillance programs for DOD-owned and Government-owned (non-DOD) animals”

Animal Disease Surveillance in DoD

Government Owned Animals (GOAs)

Privately Owned Animals (POAs)

Animal Disease Surveillance in DoD

- Centralized disease surveillance in Privately and Government Owned Animals (POAs and GOAs) has been non-existent
  - Prior to the Remote Online Veterinary Record (ROVR): constrained by paper records or private commercial software with no central data-sharing capabilities
  - Knowledge gap of overall burden, distribution, risk factors, and potential impact of diseases

- ROVR gives us the capability to pull data centrally
  - Limitations still exist
  - Need specific guidance to ensure accurate data capture

GPAWSS Background

GPAWSS

Government and Privately-owned Animal
Worldwide Surveillance System

GPAWSS Overview Video

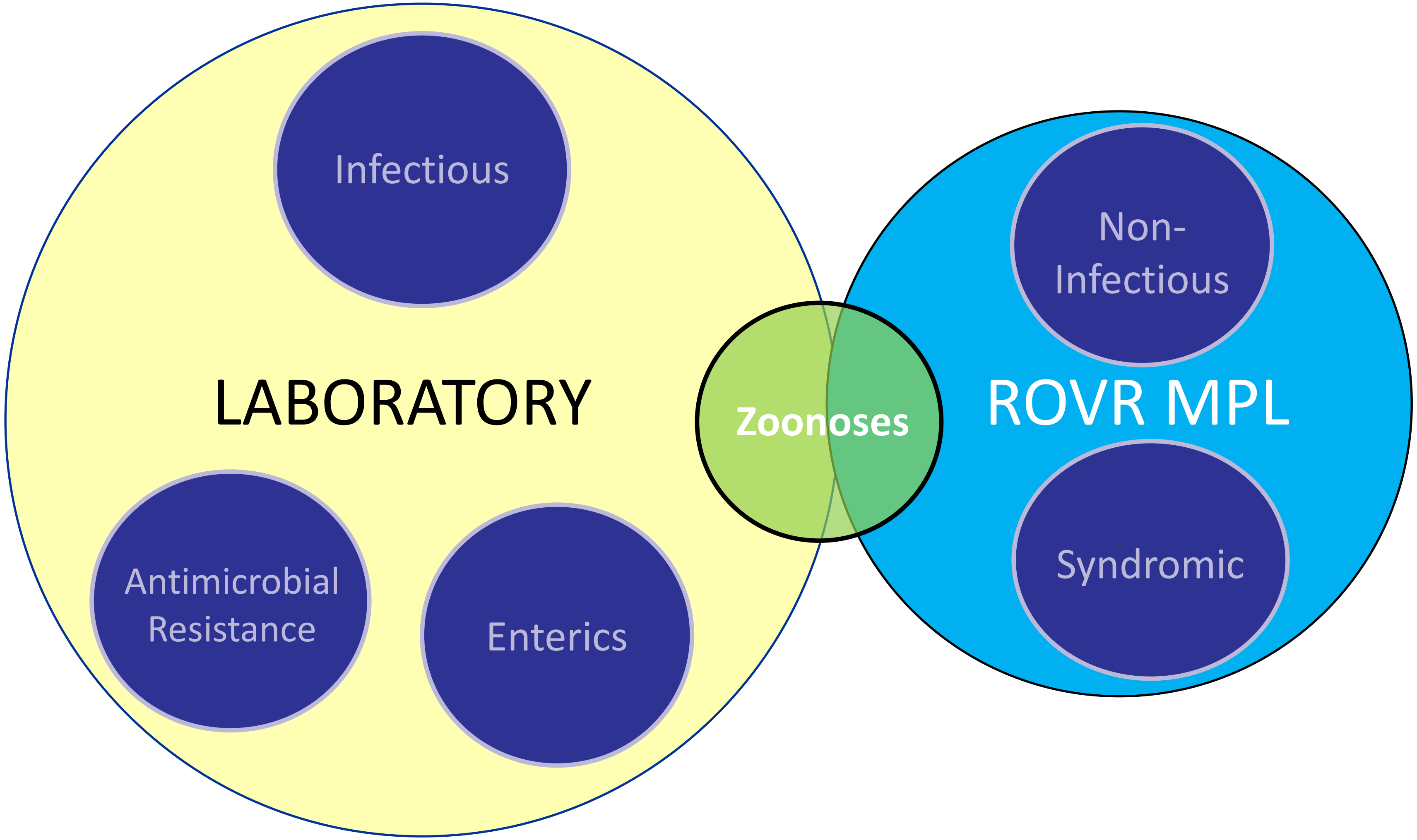
- Show 7-minute GPAWSS Overview Video.

- The video can be found on the GPAWSS milSuite site:
  - https://www.milsuite.mil/book/community/spaces/armyveterinaryservices/one-health/gpawss

GPAWSS Launch Timeline

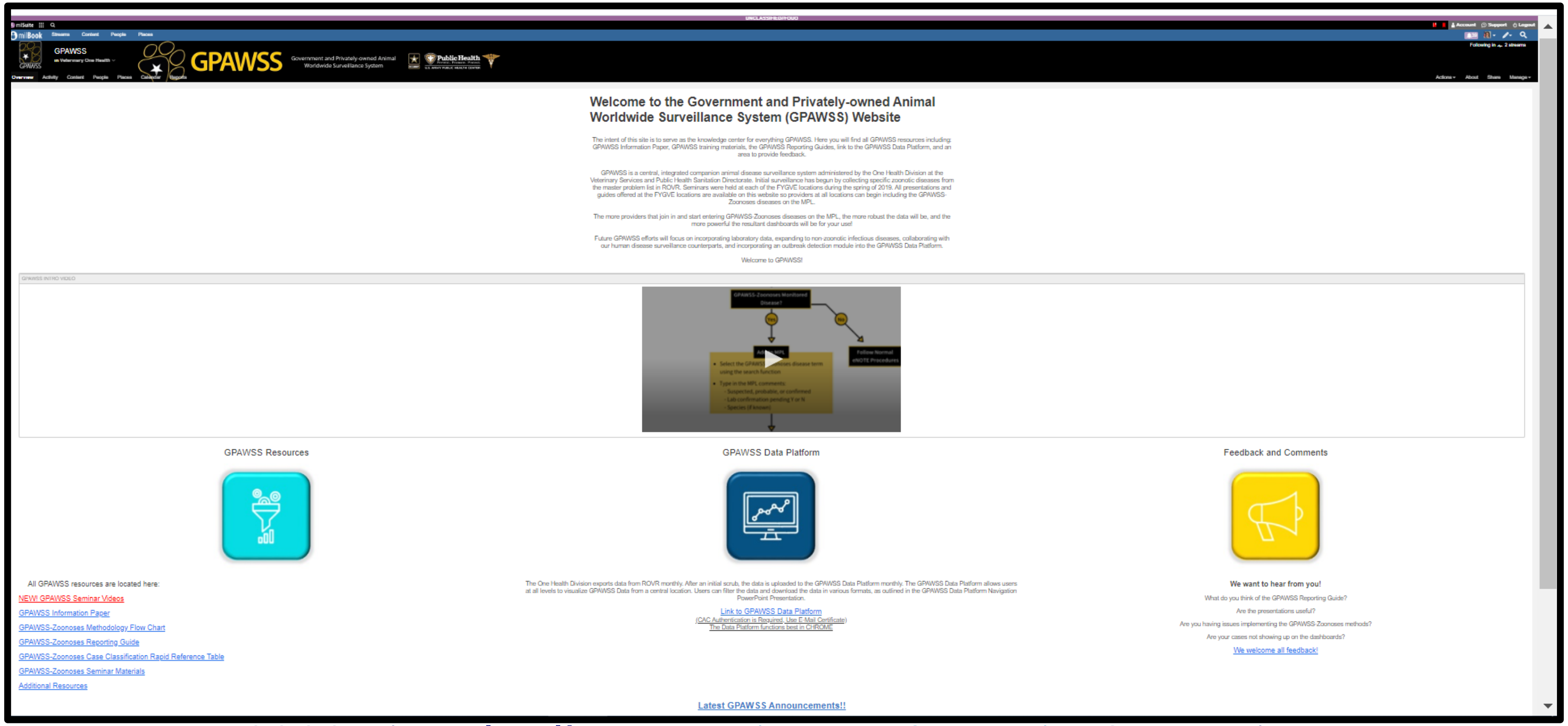
- January2019
  - GPAWSS Information paper sent to Public Health Region Senior Veterinary Leaders and Public Health Activity Commanders
- February – May 2019
  - One Health Division presented the GPAWSS Seminar to all 7 FYGVE locations
- August 2019
  - GPAWSS Seminar at Army Public Health Course
- January 2020
  - Video tutorials corresponding to each GPAWSS Seminar presentation added to GPAWSS milSuite site

GPAWSS Data Sources



Where do I find Information on GPAWSS Today?

https://www.milsuite.mil/book/community/spaces/armyveterinaryservices/one-health/gpawss



GPAWSS-Zoonoses Methodology

GPAWSS-Zoonoses Data Source

- GPAWSS-Zoonoses currently **relies solely** on **data captured from the MPL**

- MPL in ROVR is the **only data point that**:

Corresponds to a potential diagnoses

+

Can be downloaded centrally with a non-registry report

+

Can be entered in a systematic way

Why is the MPL the GPAWSS Primary Data Source?

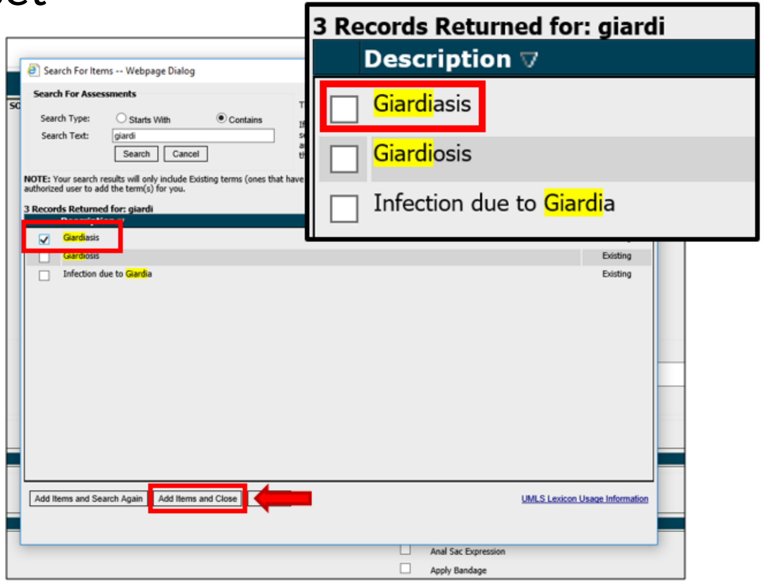
The master problem list (MPL) is the only disease diagnosis data point in ROVR that can be:

- Input by VTF personnel via a pick list (non-free text entry)
  - Systematic data entry
- Exported via an existing central report that is NOT a registry report (i.e., will not crash ROVR accidentally)
- Easily accessed by a central reporting element

Importance and Limitations of Using MPL Search

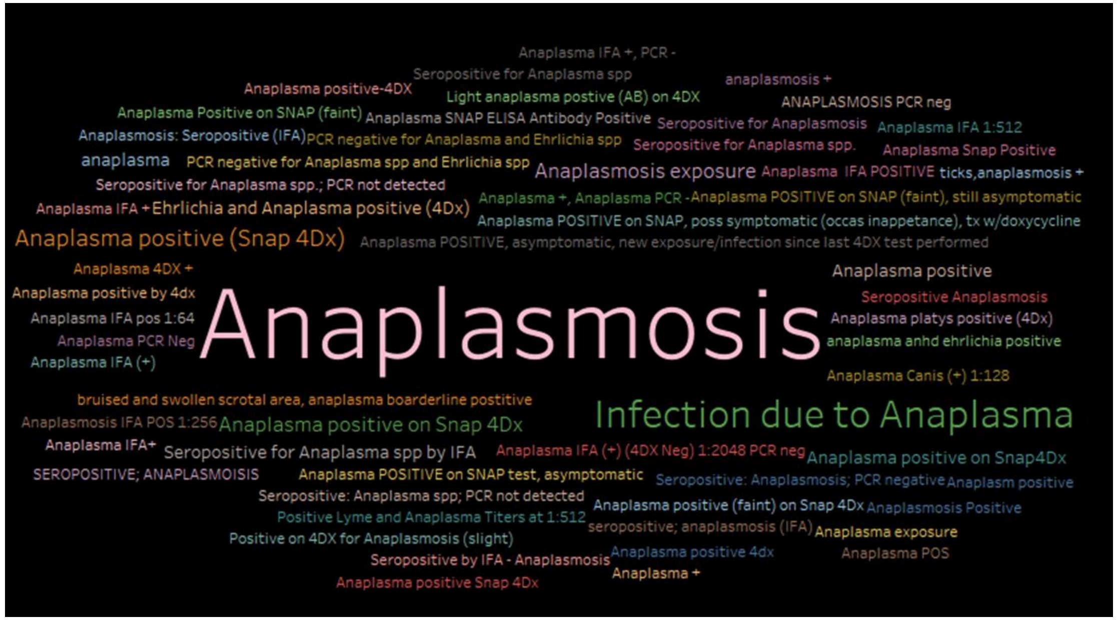
Importance of Using the MPL Search Function and Selecting the Exact Search Term

- Standardized disease diagnosis
- Reduction of data entry and analysis errors
- Reduction in time required to aggregate diagnoses and conduct quality assurance and quality control of data set



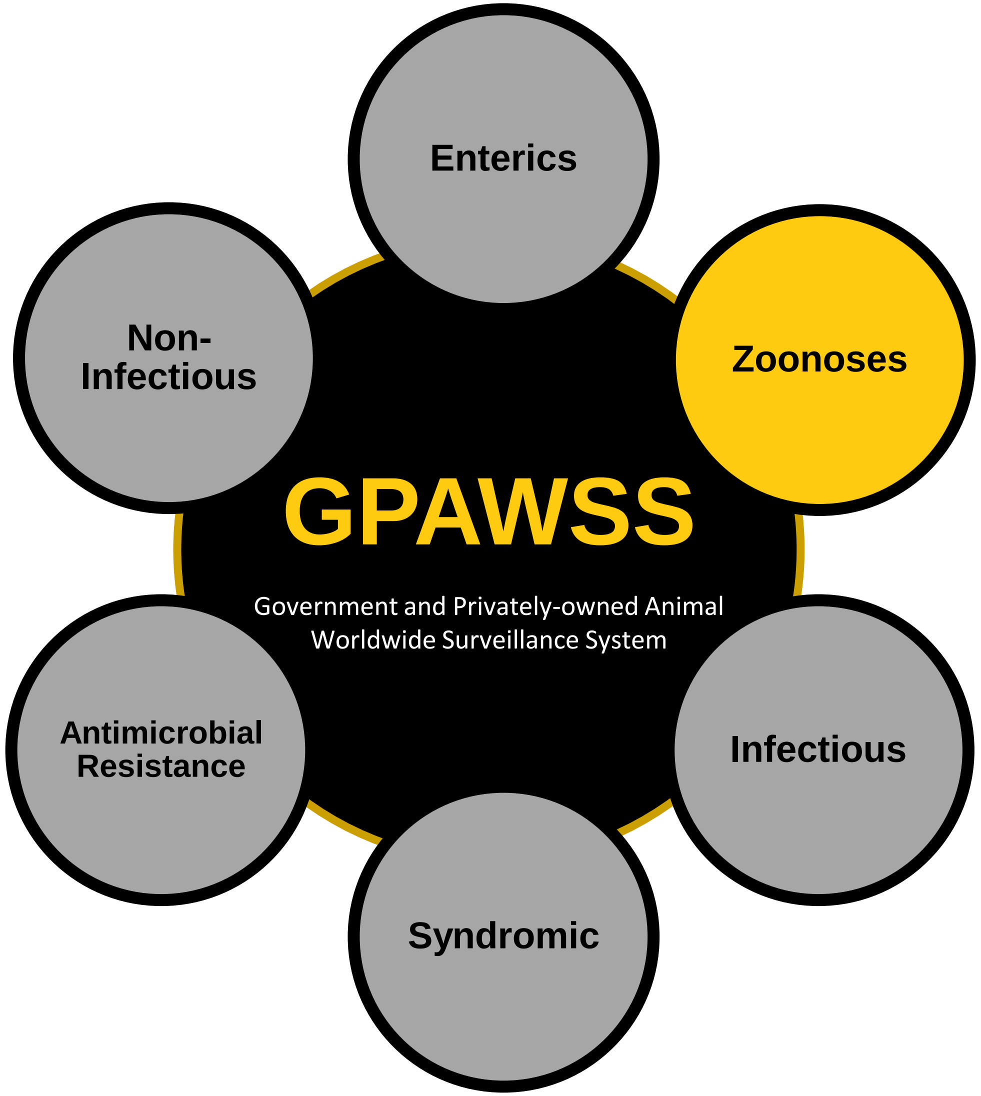
Limitations of Using the MPL as the GPAWSS Data Source

- Underreporting of suspected and diagnosed diseases and injuries
- MPL has the capability to accept free text entries resulting in multiple different entries for the same diagnosis

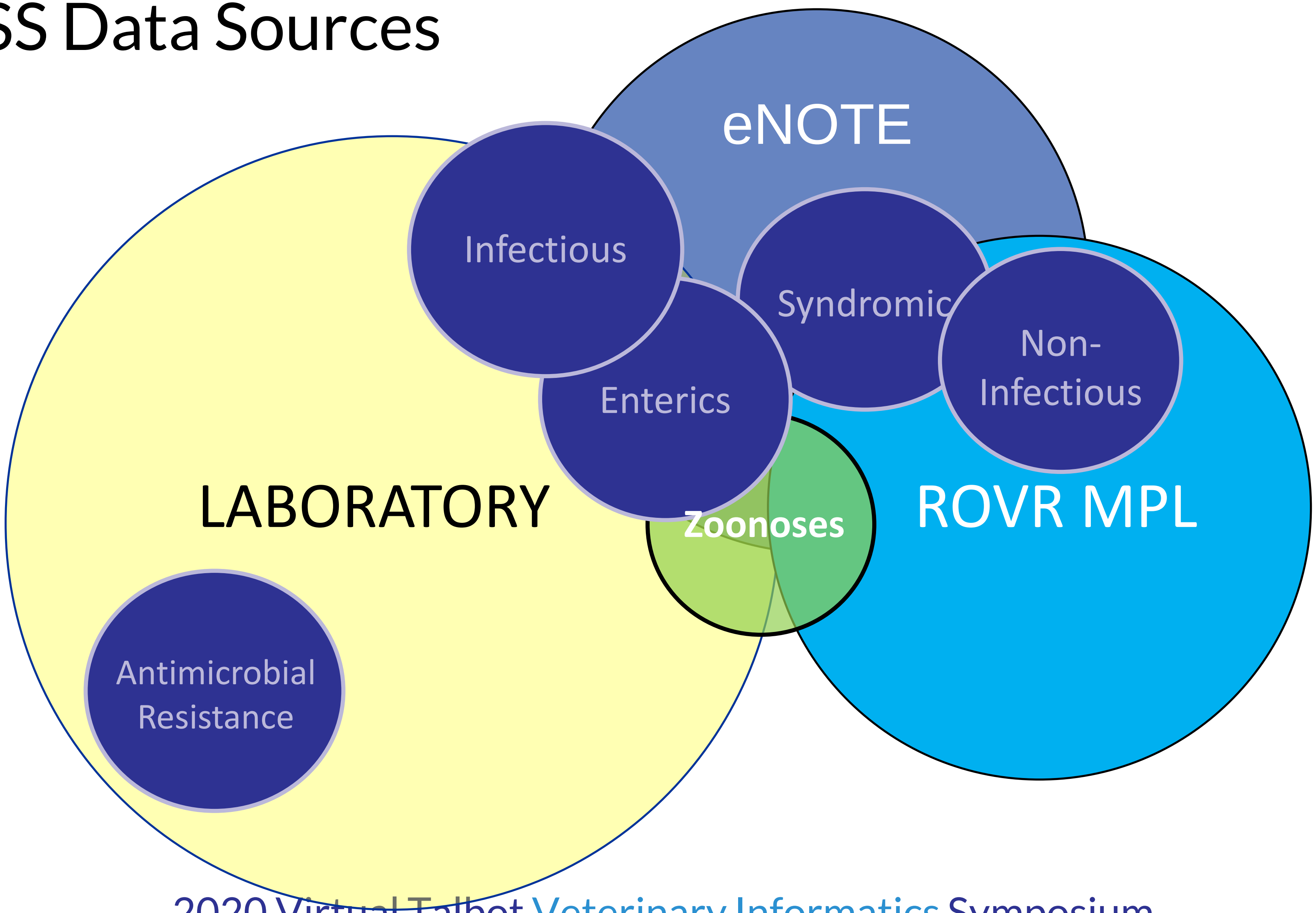


MPL, NLP and Current Progress

GPAWSS Programs



GPAWSS Data Sources



Two Big Questions:

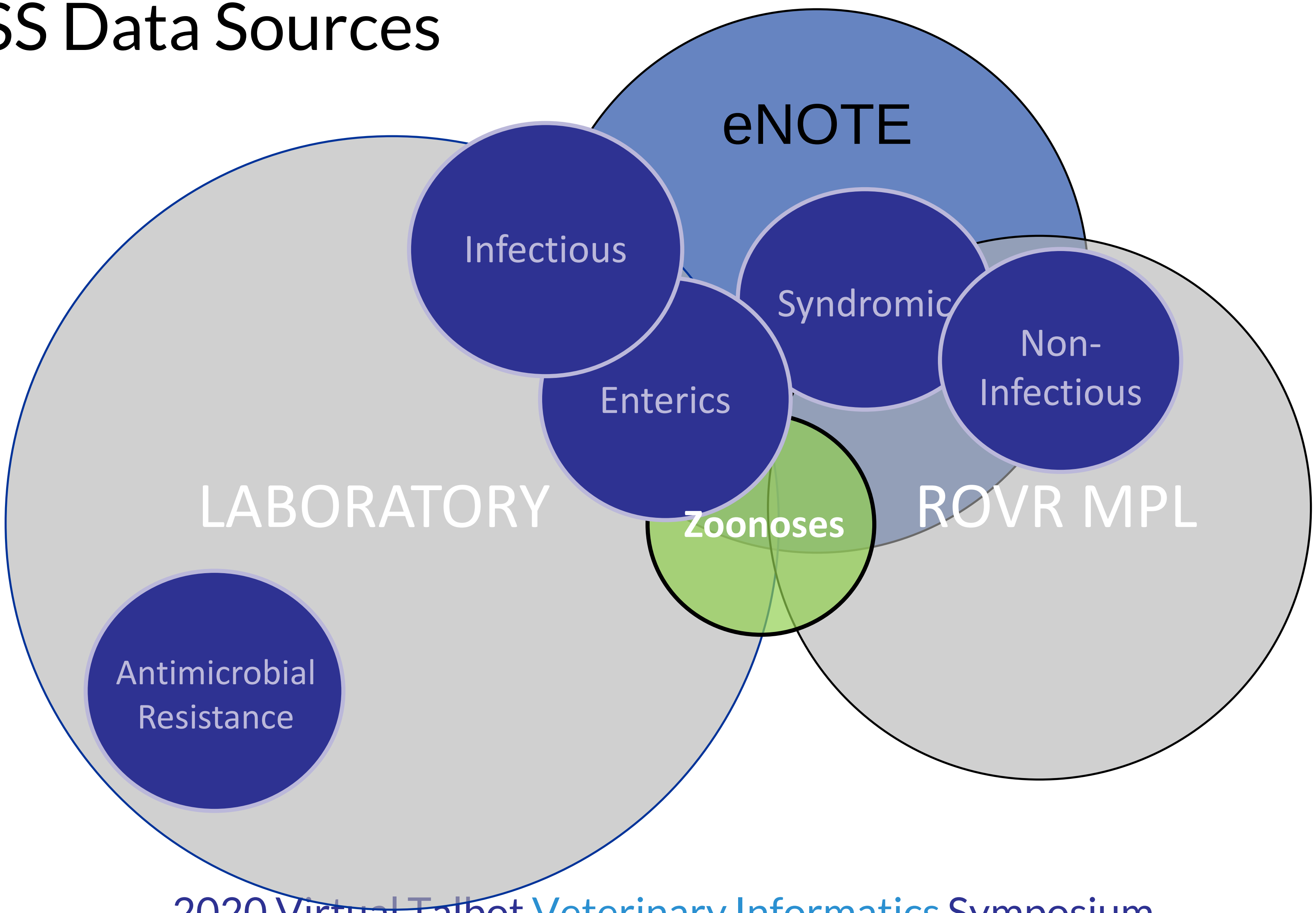
What are we going to do with the Data?

How are we going to store it?

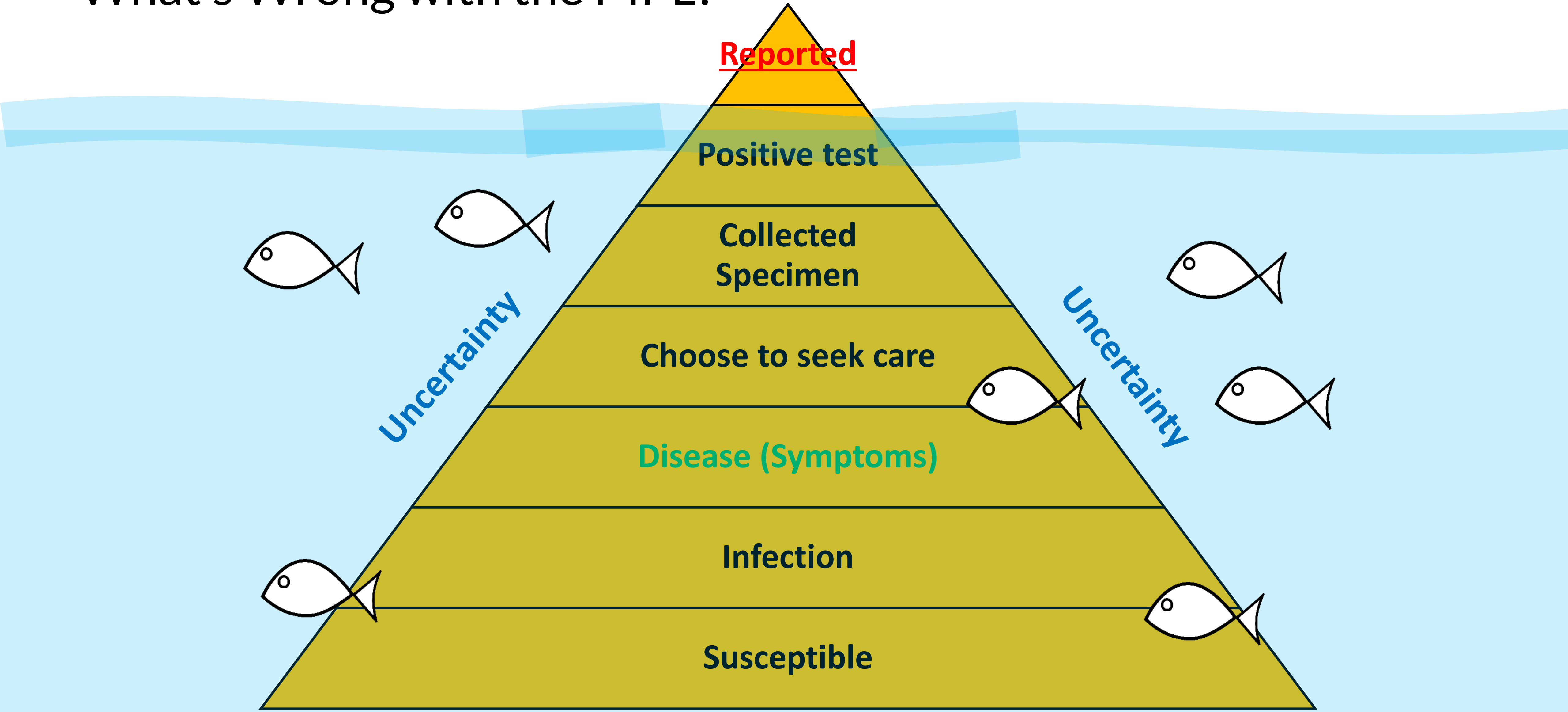
What are we going to do with the Data?

How are we going to store it?

GPAWSS Data Sources



What's Wrong with the MPL?



What's Wrong with the MPL?

#DRAINTHEOCEAN

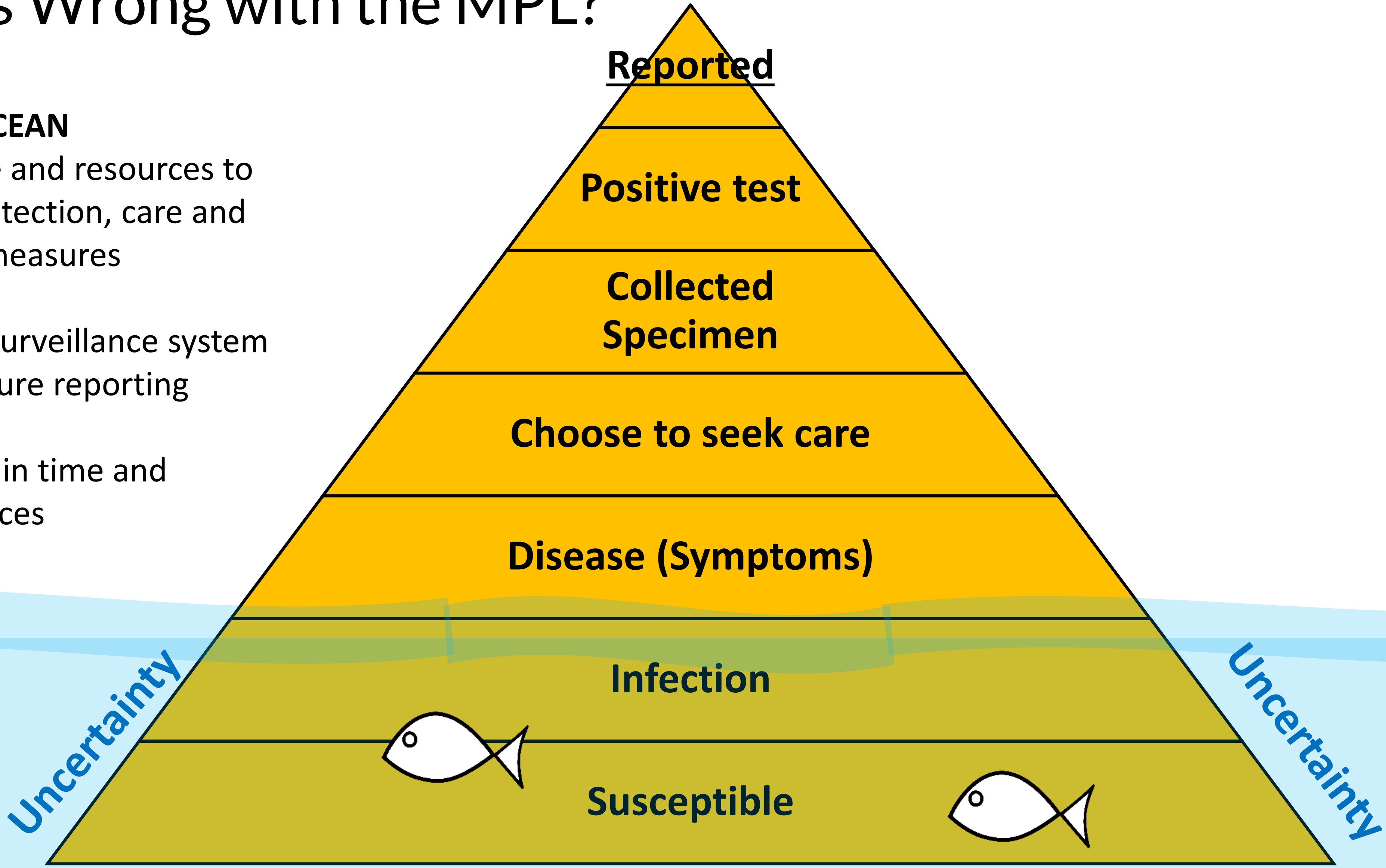
- Spend time and resources to improve detection, care and reporting measures

Pros:

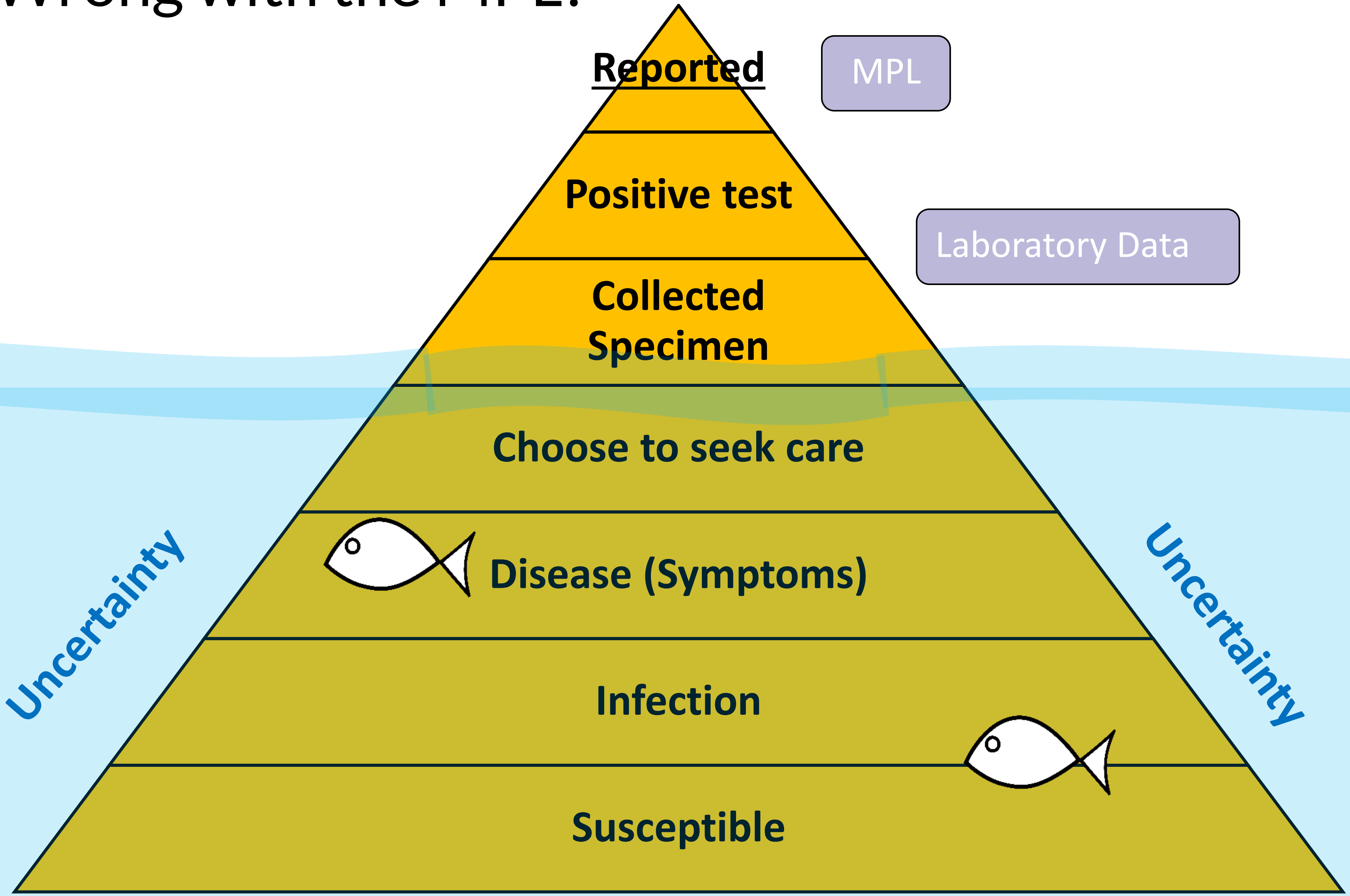
  - Valid Surveillance system for future reporting

Cons:

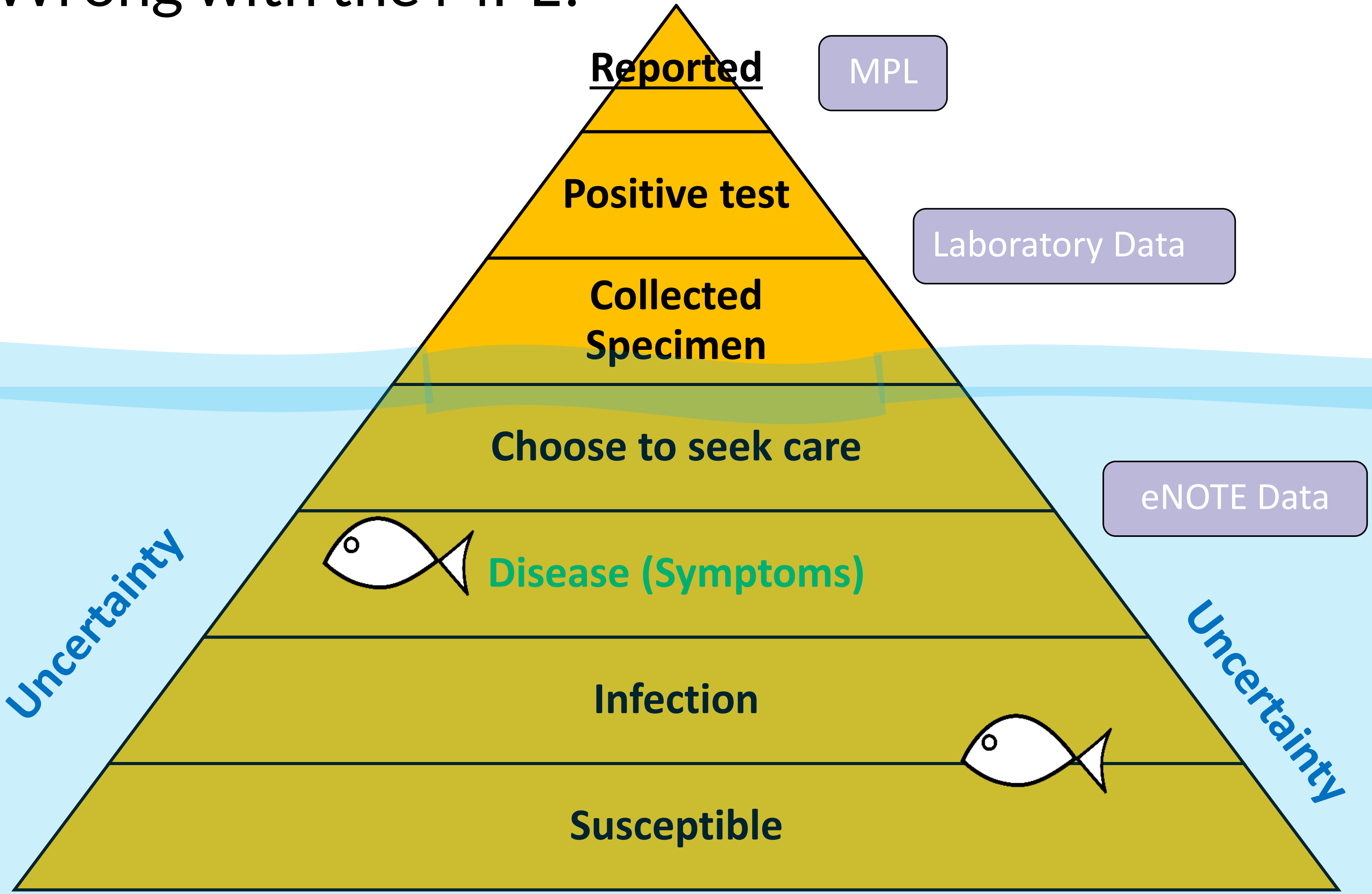
  - **Costly** in time and resources



What's Wrong with the MPL?

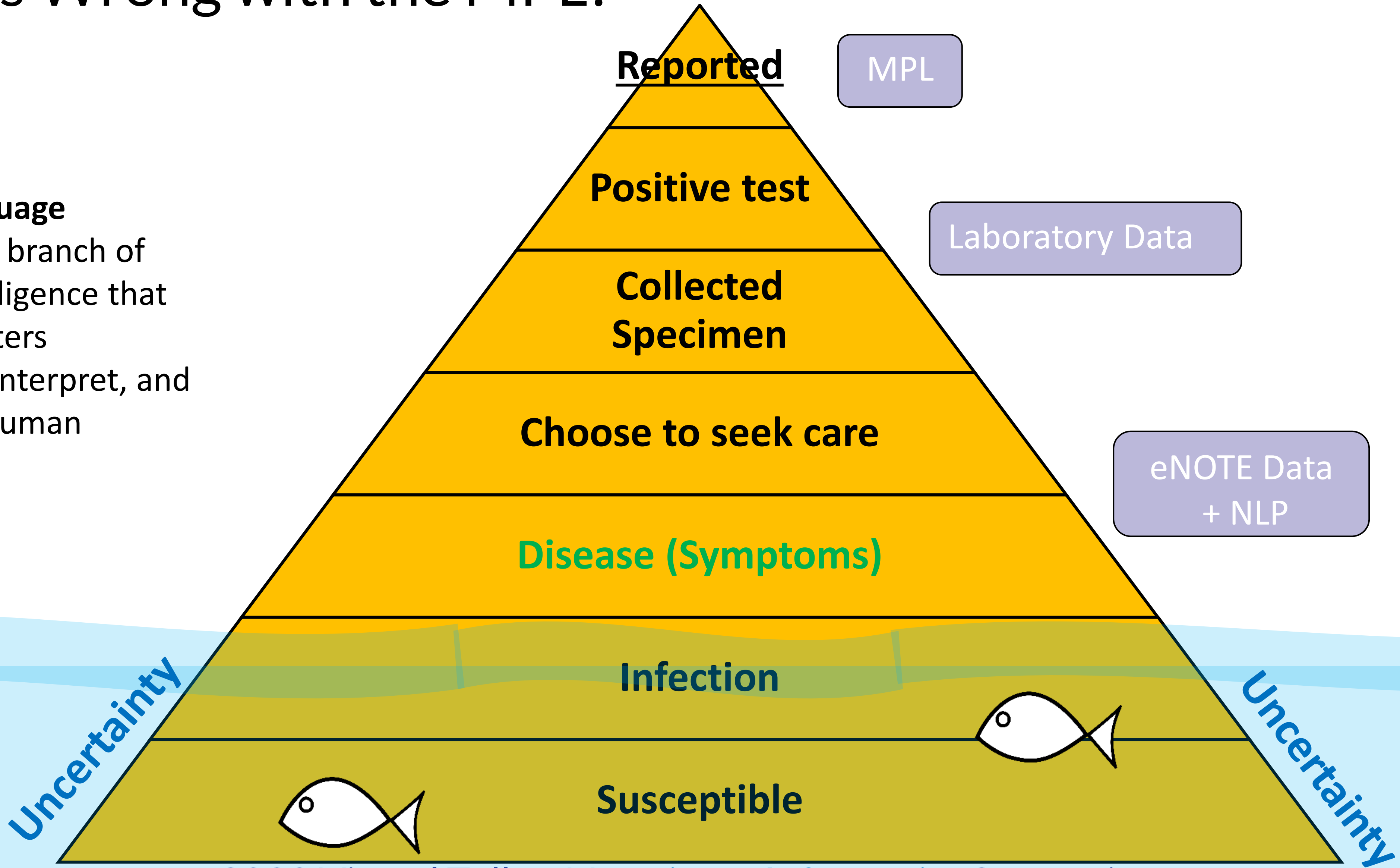


What's Wrong with the MPL?



What’s Wrong with the MPL?

Natural Language Processing: a branch of artificial intelligence that helps computers understand, interpret, and manipulate human language.



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

VetTag: improving automated veterinary diagnosis coding via large-scale language modeling

8 May 2019

Yuhui Zhang[1], Allen Nie[2], Ashley Zehnder[2], Rodney L. Page[3] and James Zou[2,4]

Unlike human medical records, most of the veterinary records are free text without standard diagnosis coding. The lack of systematic coding is a major barrier to the growing interest in leveraging veterinary records for public health and translational research. Recent machine learning effort is limited to predicting 42 top-level diagnosis categories from veterinary notes. Here we develop a large-scale algorithm to automatically predict all 4577 standard veterinary diagnosis codes from free text. We train our algorithm on a curated dataset of over 100 K expert labeled veterinary notes and over one million unlabeled notes. Our algorithm is based on the adapted Transformer architecture and we demonstrate that large-scale language modeling on the unlabeled notes via pretraining and as an auxiliary objective during supervised learning greatly improves performance. We systematically evaluate the performance of the model and several baselines in challenging settings where algorithms trained on one hospital are evaluated in a different hospital with substantial domain shift. In addition, we show that hierarchical training can address severe data imbalances for fine-grained diagnosis with a few training cases, and we provide interpretation for what is learned by the deep network. Our algorithm addresses an important challenge in veterinary medicine, and our model and experiments add insights into the power of unsupervised learning for clinical natural language processing.

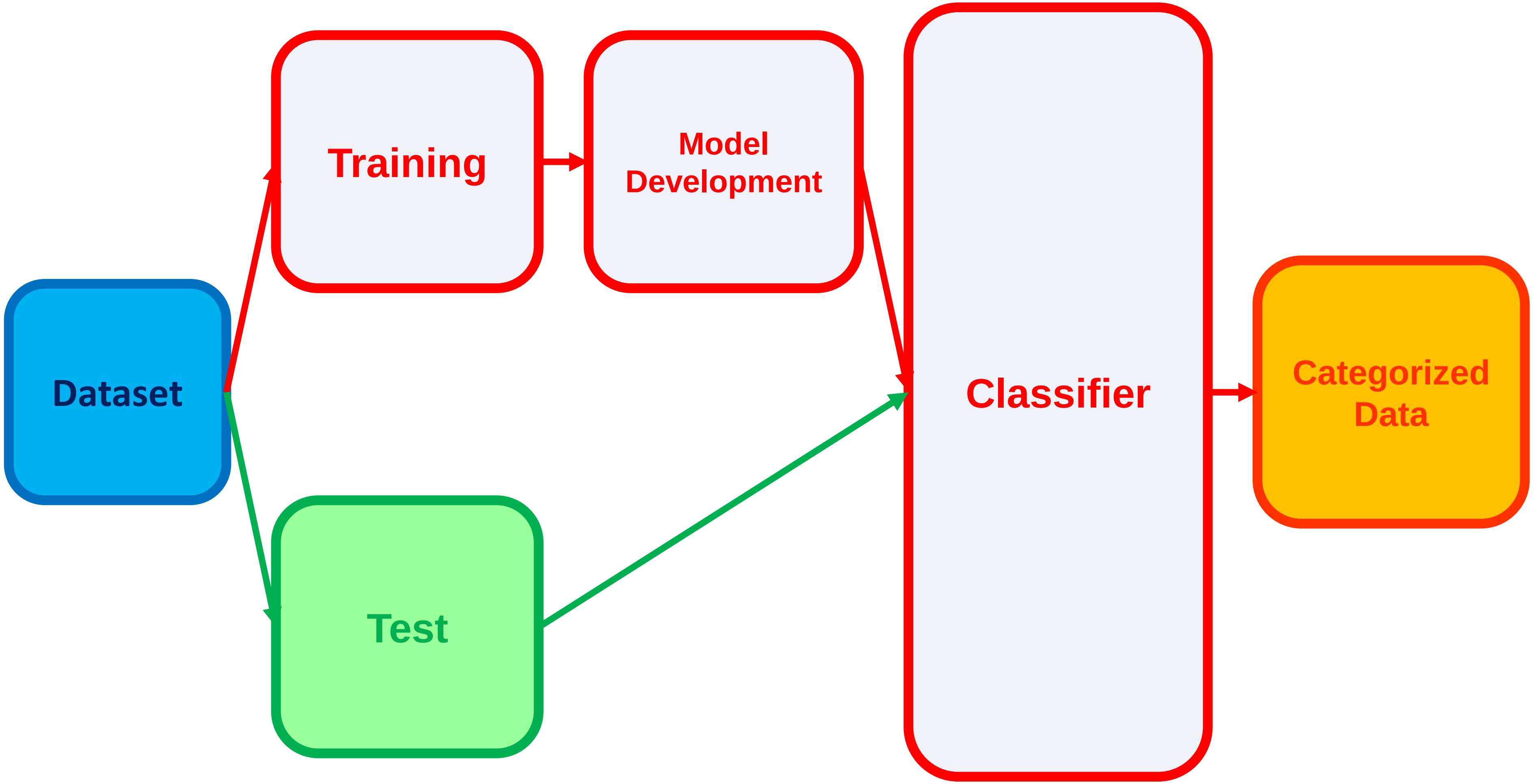
npj Digital Medicine (2019)2:35 ; https://doi.org/10.1038/s41746-019-0113-1

INTRODUCTION

Large-scale electronic health records (EHR) can be a powerful resource for patient care and research. There have been many exciting efforts applying machine learning to human medical records—e.g. predicting in-hospital mortality, 30-day unplanned and evaluated on clinical notes gathered from the same hospital as well. Veterinary notes have different styles and vocabulary, and its diagnosis codes use a terminology framework different from humans. Therefore an automated veterinary coding algorithm is needed. Moreover, due to the lack of general coding practice in

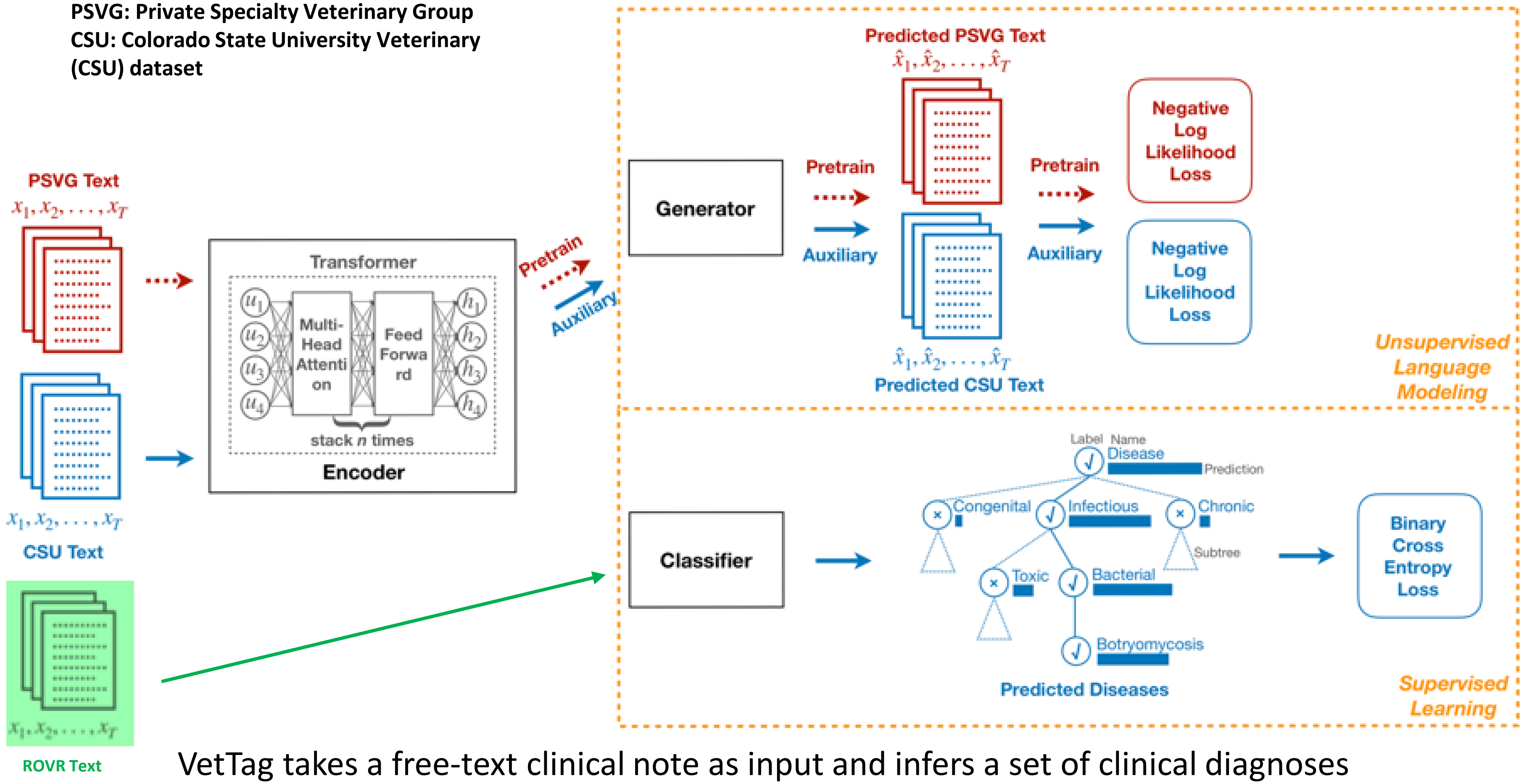
Zhang, Yuhui, et al. "VetTag: improving automated veterinary diagnosis coding via large-scale language modeling." *npj Digital Medicine* 2.1 (2019): 35.

Basics of ML Model Building



NLP Methodology: VetTag + ROVR eNOTE data

PSVG: Private Specialty Veterinary Group
CSU: Colorado State University Veterinary (CSU) dataset



VetTag takes a free-text clinical note as input and infers a set of clinical diagnoses from the note. The inferred diagnosis is in the form of SNOMED-CT codes and each note can be associated with multiple codes if the patient has several diagnoses.

Zhang, Y., Nie, A., Zehnder, A., Page, R. L. & Zou, J. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. *npj Digital Medicine* **2**, 35 (2019).

ROVR eNOTE Data

Variables

-ID
-Facility
-Date
-Chief Complaint
-Subjective information
-Objective information
-Assessment information
-Plan information

Limitations

-Requires Admin Pulls
-LARGE DATASETS
    -~1.7M observations
    -8 Variables
    -Average width of 175 to 195 (Very conservative estimate)

~ 1.7 to 1.9 GB file + 27mb
Monthly Refresh

https://www.stata.com/support/faqs/data-management/datacalc.cgi?nobs=6000000&nvars=8&width=35

What are we going to do with the Data?

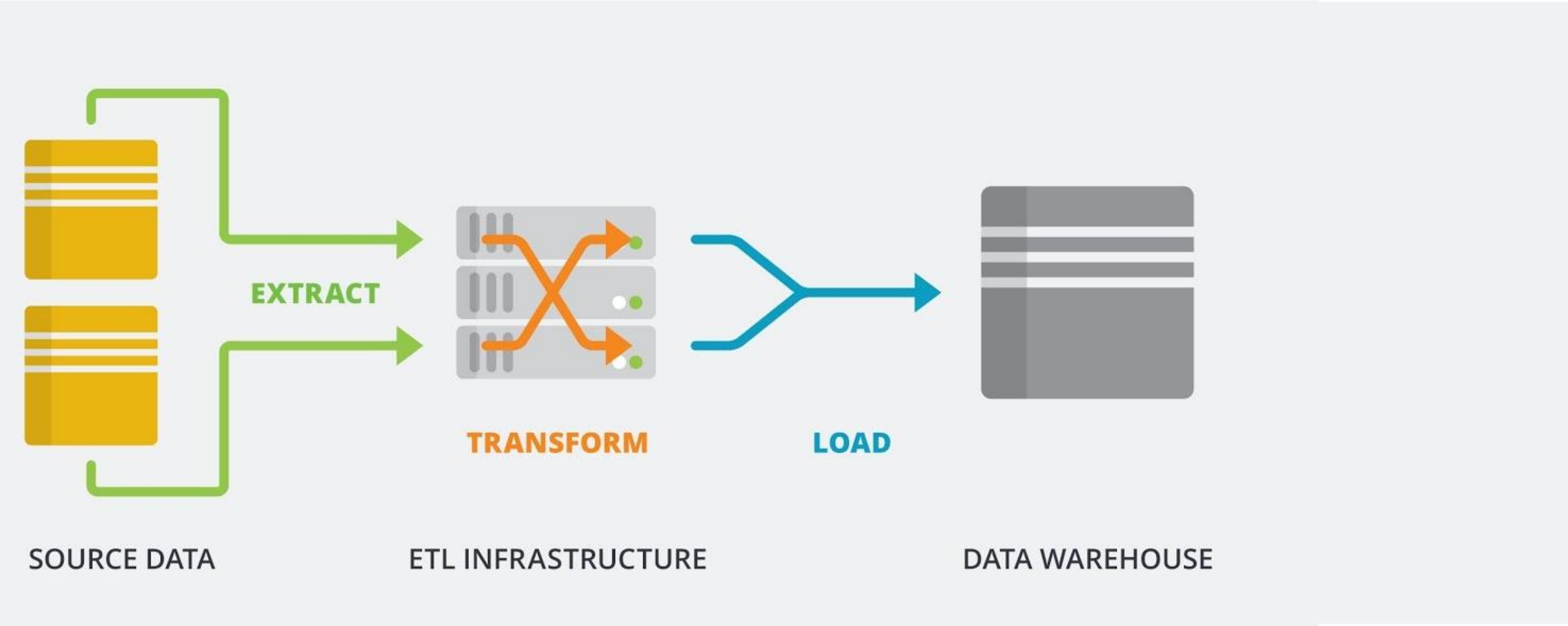
How are we going to store it?

What is the solution?

DATA LAKE

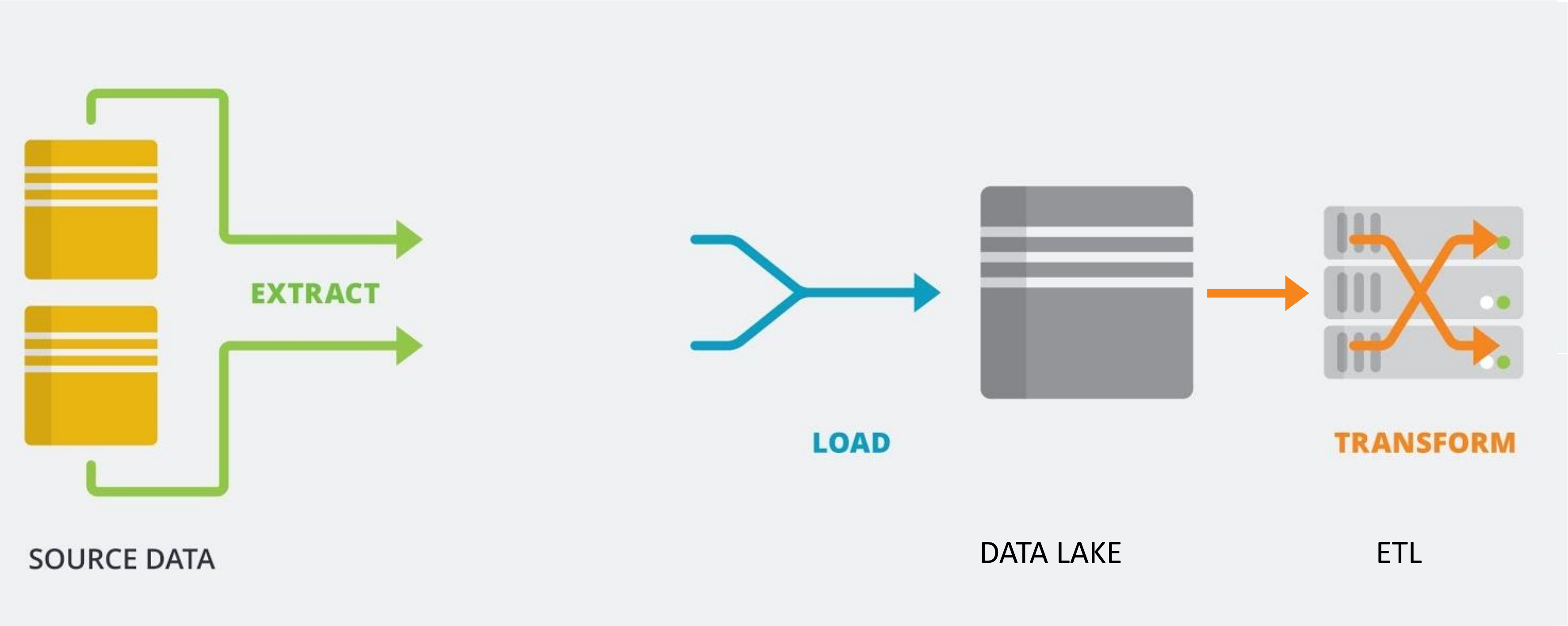
- Accepts structured & unstructured data
- Organizes with appropriate metadata tags; allows a user to search by topic to discover data streams within the directorate
- Takes a decoupling approach that allows for much needed flexibility within the environment that can change as innovation advances
- Analytics can be operationalized DIRECTLY WITHIN the data lake, eliminating the need to export data into another platform to perform analyses

How are data lakes different?



https://indatalabs.com/blog/data-warehouse-concepts

How are data lakes different?



https://indatalabs.com/blog/data-warehouse-concepts

Where?

- Data Storage and Analytical Solutions provided at the ERDC Supercomputing Resource Center (DSRC)
- One of the five DoD DRSCs operated by the DoD High Performance Computing Modernization Program (HPCMP)

DoD Supercomputing Resource Centers (DSRCs)				
AFRL DSRC	ARL DSRC	ERDC DSRC	Navy DSRC	MHPCC DSRC
Air Force Research Laboratory WPAFB, OH	Army Research Laboratory Aberdeen Proving Ground, MD	Army Engineer Research and Development Center Vicksburg, MS	Navy DoD Supercomputing Resource Center Stennis Space Center, MS	Maui High Performance Computing Center Kihei, Maui, HI

Low Barrier to Implementation

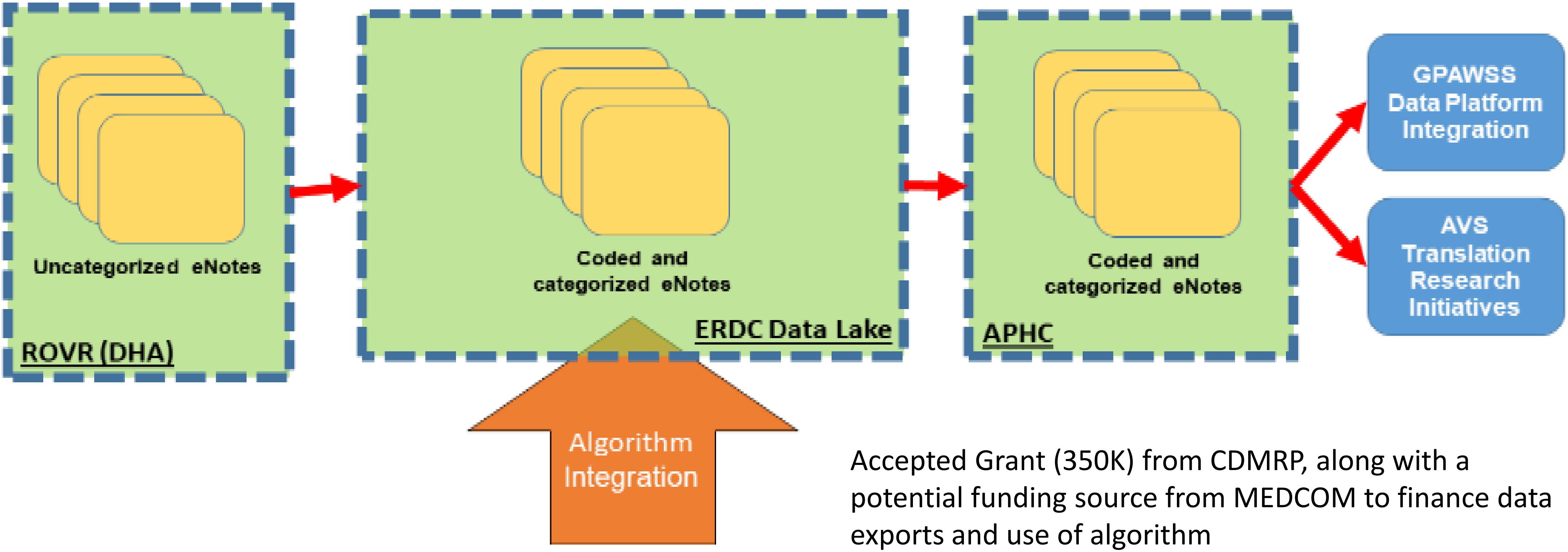
Hosting platform is ***already approved*** for use of DoD data (Majority of our data are HIPPA exempt)

The monetary investment is ***NONE*** compared to current database spending and continual contractor bids necessary to update existing structure for improved use or implementation of new data streams

Improves the outdated data warehousing problems that silo information within organizations and have a high barrier of entry for collaboration or access

Implements an ***ease of collaboration and use of downstream visualization and analytical capabilities*** (i.e. artificial intelligence)

Workflow and Next steps



Accepted Grant (350K) from CDMRP, along with a potential funding source from MEDCOM to finance data exports and use of algorithm

Pilot Study

Pilot Study:

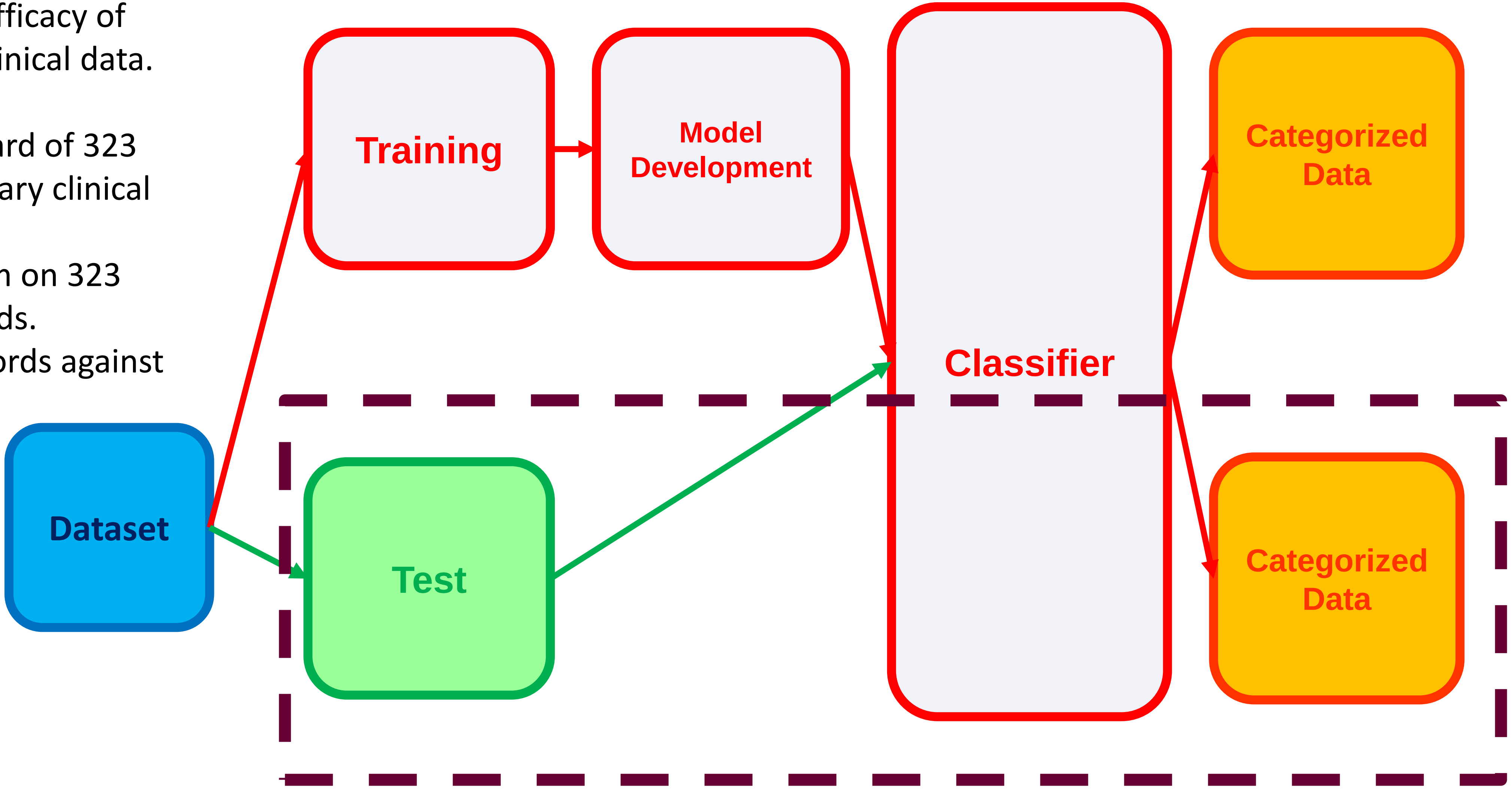
Objective: To test the efficacy of VetTag on our eNOTE clinical data.

Methods:

- Develop a Gold Standard of 323 manually coded veterinary clinical records.
- Apply VetTag algorithm on 323 veterinary clinical records.
- Test VetTag coded records against

Results:

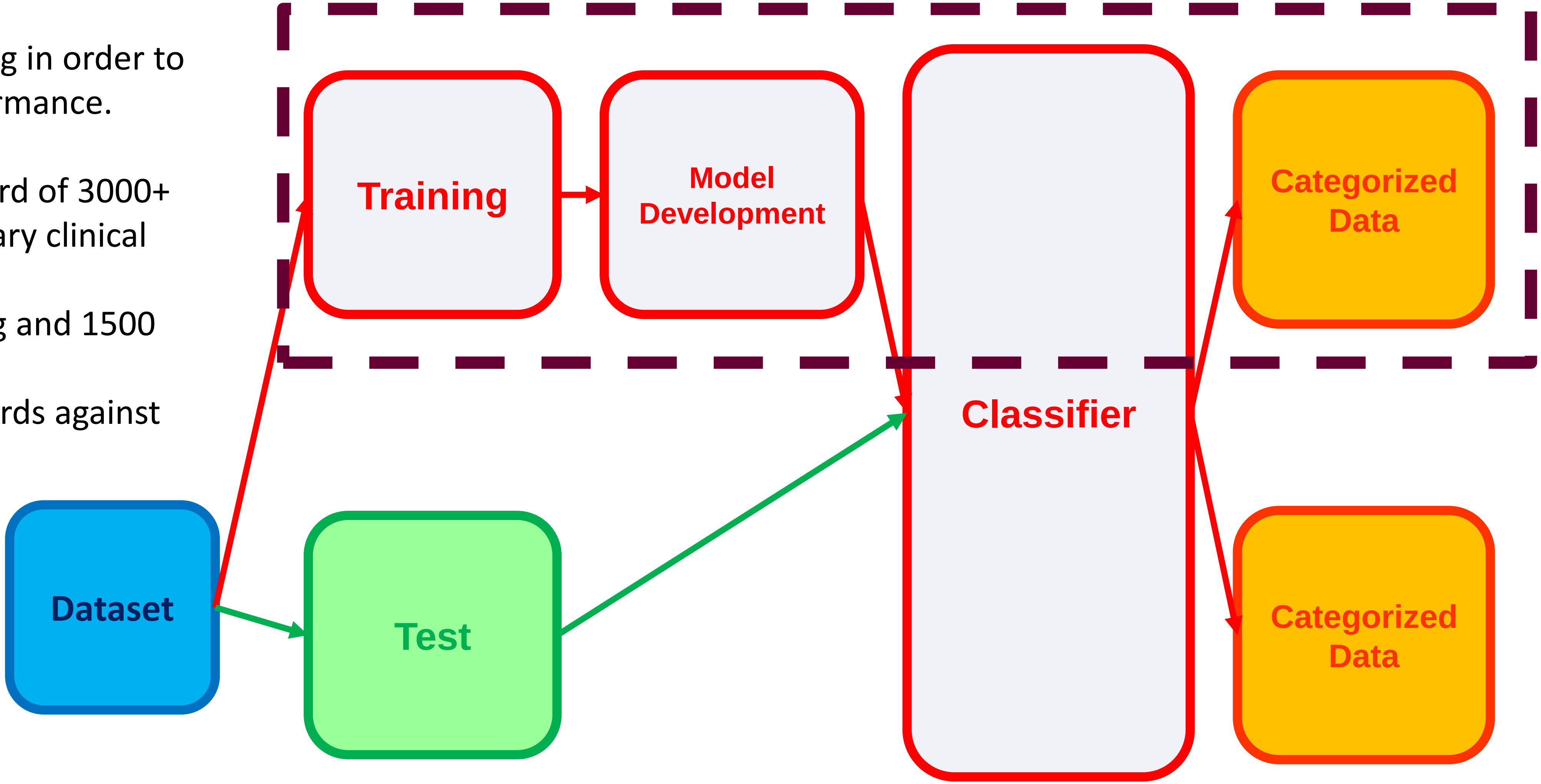
- 47% Precision
- 18% Recall



Model Retraining

Model Retraining:
Objective: Retrain VetTag in order to increase classifier performance.
Methods:
- Develop a Gold Standard of 3000+ manually coded veterinary clinical records.
- Utilize transfer learning and 1500 Gold Standard records.
- Test VetTag coded records against remaining 1500 records

Results:
- Ongoing



Questions?

Contact the One Health Division team usarmy.apg.medcom-aphc.mbx.gpawss-feedback@mail.mil

AVI Association for Veterinary Informatics

#BetterDataSavesPets

Education Opportunities:

- Annual Talbot Symposium
- VMX 2021
- Virtual Education, TBA

Follow AVI on LinkedIn & Twitter:

- @avinformatics
- #Talbot20
- #BetterDataSavesPets