PFAS Litigation and Regulatory Developments Conference

HOW AI CAN ADDRESS ENVIRONMENTAL ISSUES

Tom Gulbransen
Battelle – New York
Gulbransen@Battelle.org
516-313-9311

How AI Can Address Environmental Issues

Opportunities to strengthen cases

Threats to e-discovery

Fitness of team & data assets

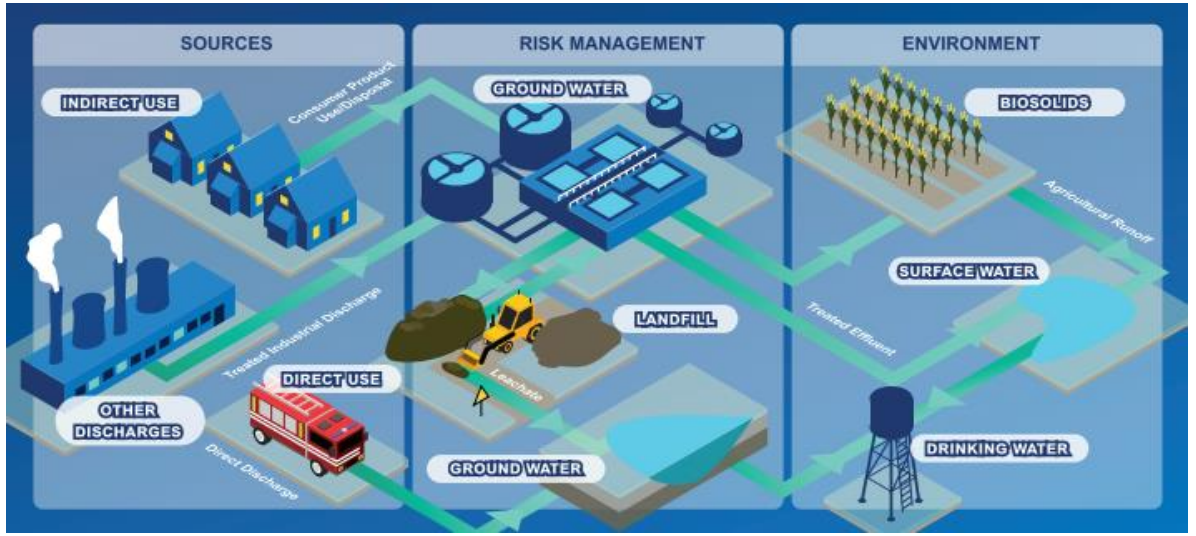
Two examples of Machine Learning

Strategic leadership, Coordination

Speed & scalability via ML tools

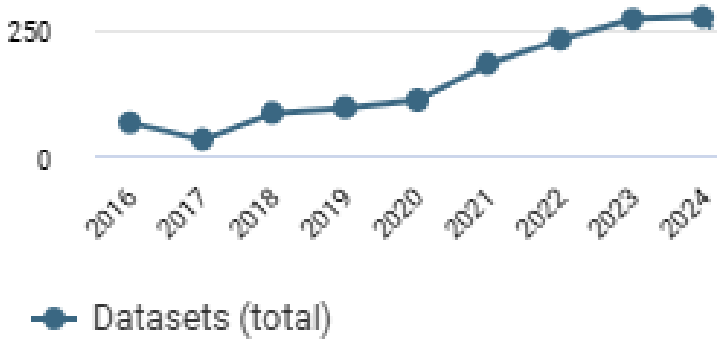
Defensible findings

Focused data management effort, $



BATTELLE

Interested in Machine Learning, yet Anxious



Digital Science & Research Solutions, Inc.

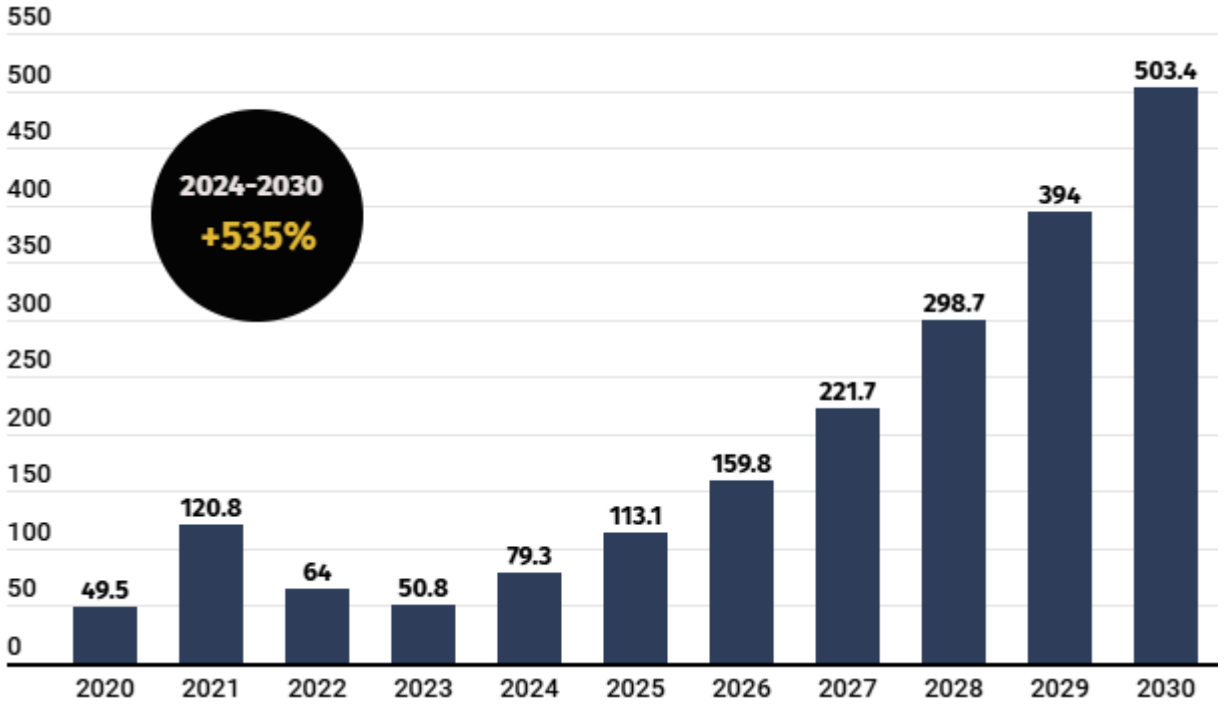
Forecasted increases in:

Markets
Model skills
Data sources, Costs
Information types
Reliance on algorithms
Speed of e-discovery

Concomitant pressures:

Which data are relevant?
What accuracy is sufficient?
Expert knowledge for training
Risks from hallucinations

"Machine Learning Market to Skyrocket by 535% by 2030" (edge-ai.vision.com)



www.statista.com/outlook/tmo/artificial-intelligence/machine-learning/worldwide

Scientific Complexity >> Algorithmic Mystery

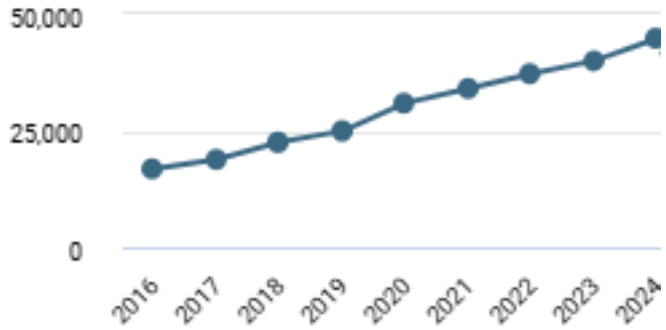
Concomitant pressures:

Which data are relevant?

What accuracy is sufficient?

Expert knowledge for training

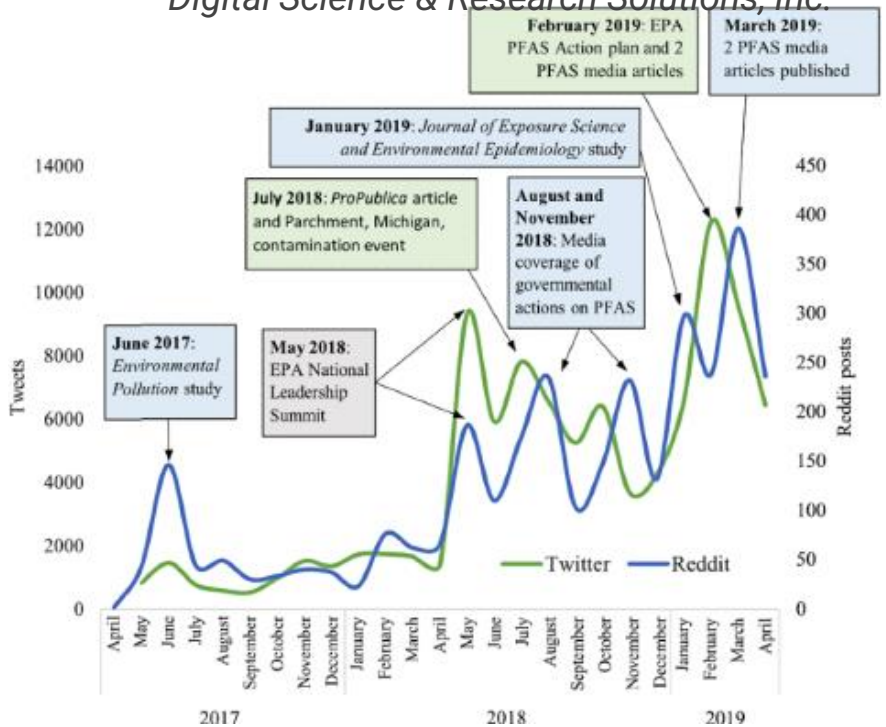
Risks from hallucinations



Digital Science & Research Solutions, Inc.

PFAS vs Other Environmental Contaminants

Attribute	Current State for PFAS	Comparison to Other Contaminants
Class	Family of >4,000 compounds including polymer and non-polymer forms	Larger than most/any that are federally regulated
Properties	Mobile, bioaccumulative, persistent	Greater migration from source area and more difficult to degrade
Environmental behavior	Limited understanding requiring development of new monitoring/modeling approaches	More complex interactions in the environment and unlike that of other contaminants
Health Impacts/Toxicology	Limited number in the class are highly studied and impacts are not consistent across the class	More complex and impacts to exposure less understood
Public engagement	High, with greater awareness and interest in certain states	Affected individuals and communities are more engaged
Regulations	Lack of promulgated standards and inconsistencies across states	Less clarity in requirements for action



J Med Internet Res (2022) vol. 24, iss. 3, e25614

BATTELLE

Learning with Classification vs Generative Transformations

Concomitant pressures:

Which data are relevant?

What accuracy is sufficient?

Expert knowledge for training

Risks from hallucinations

Park v. Kim, No. 22-2057 (2d Cir. 2024)

Justia Opinion Summary

In this case, the United States Court of Appeals for the Second Circuit affirmed the dismissal of Minhye Park's case against David Dennis Kim by the United States District Court for the Eastern District of New York. The District Court dismissed the case due to Park's persistent and knowing violation of court orders, specifically regarding discovery. The Court of Appeals found that Park's noncompliance amounted to "sustained and willful intransigence" despite repeated warnings that continued refusal to comply would result in dismissal.

Additionally, the Court of Appeals addressed the conduct of Park's attorney, Jae S. Lee. Lee cited a non-existent court decision in her reply brief to the court, which she admitted she generated using an artificial intelligence tool, ChatGPT. The court deemed this action as falling below the basic obligations of counsel and referred Lee to the court's Grievance Panel. The court also ordered Lee to provide a copy of the decision to her client. The court emphasized that attorneys must ensure that their submissions to the court are accurate and that they have conducted a reasonable inquiry to confirm the existence and validity of the legal authorities on which they rely.

United States v. Cohen, 18-CR-602 (JMF), 3 (S.D.N.Y. Mar. 20, 2024) ("In support of his motion, Schwartz cited and described three "examples" of decisions granting early termination of supervised release that were allegedly affirmed by the Second Circuit. *See id.* at 2-3 (citing *United States v. Figueroa-Florez*, 64 F.4th 223 (2d Cir. 2022); *United States v. Ortiz* (No. 21-3391), 2022 WL 4424741 (2d Cir. Oct. 11, 2022); and *United States v. Amato*, 2022 WL 1669877 (2d Cir. May 10, 2022)). There was only one problem: The cases do not exist") *(casetext.com/case/united-states-v-cohen-213)*

Classification Relies on Training Data and Validation “Truth” Data

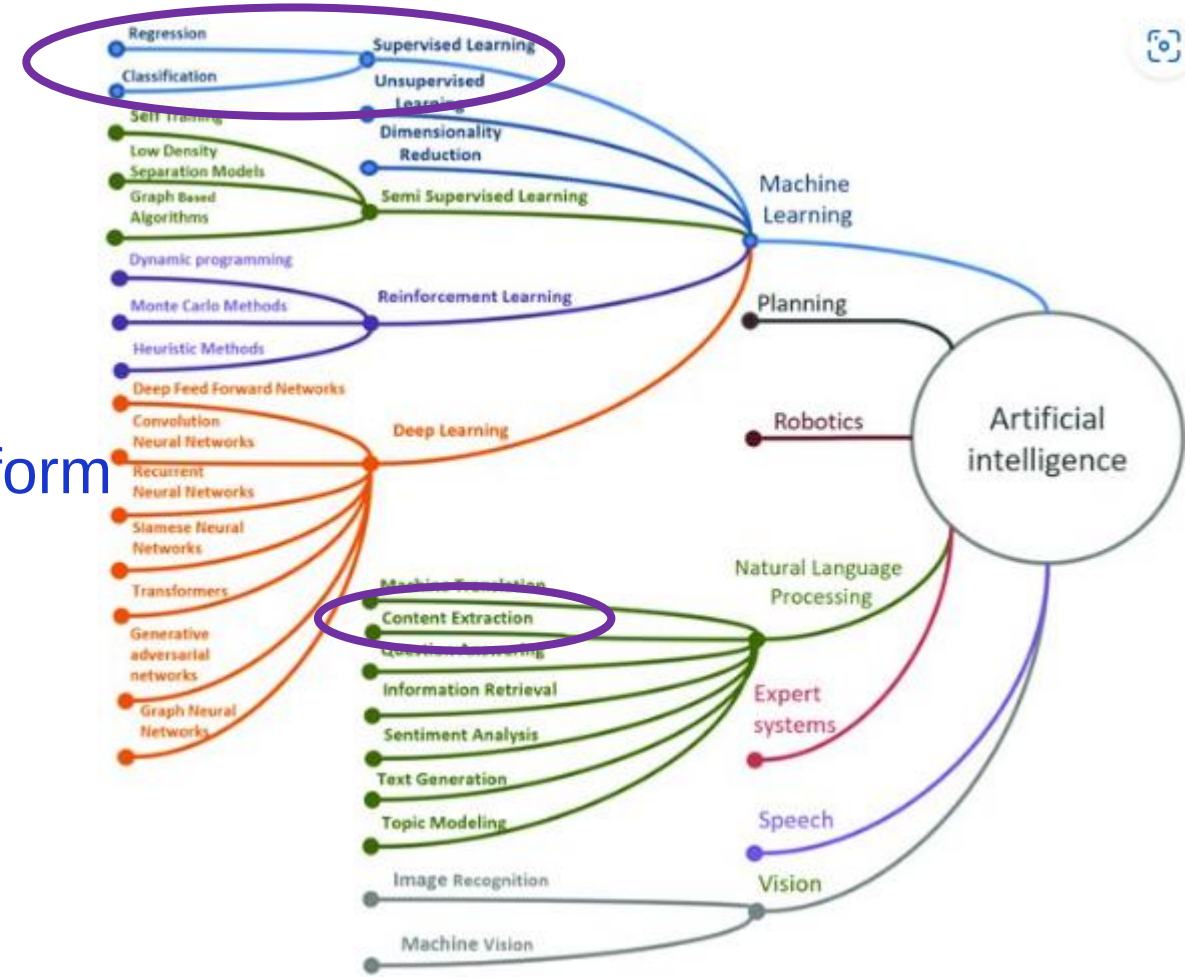
Fitness of ML Methods:

Relevance?

Accuracy? ➔ transparent statistics

Expert training?

No Hallucinations ➔ classify not transform



Mukhamediev et.al. (2022) Mathematics 10(15), 2552

PFAS ML Starts with Data Just like CWA, CERCLA, OPA, MPRSA, RCRA

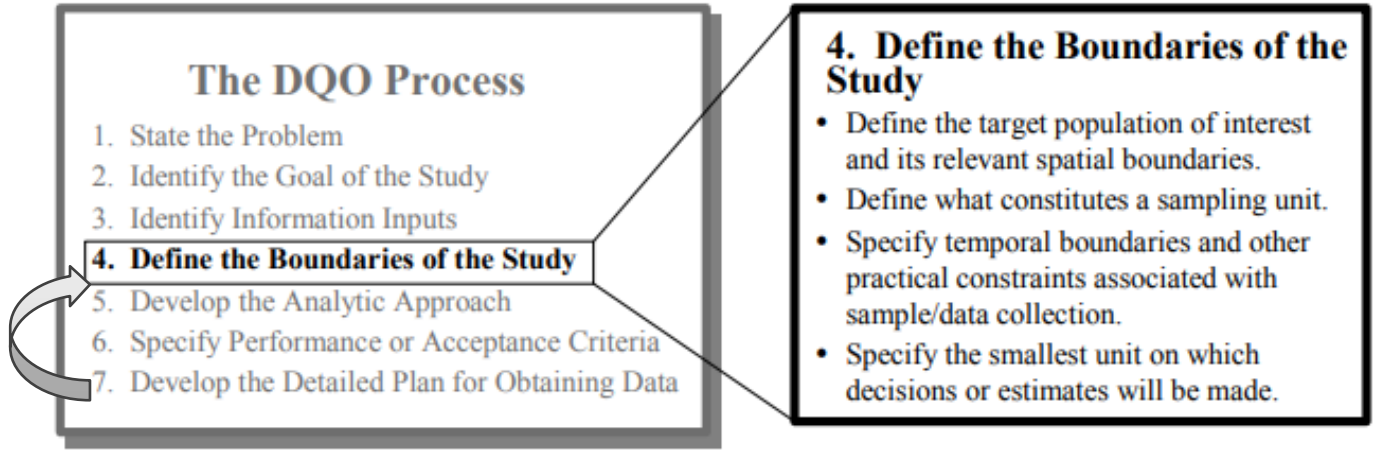
Fitness of ML Methods:

Relevance? ➔ scientific curation

Accuracy? ➔ transparent statistics

Expert training? ➔ guidance of truths

No Hallucinations ➔ classifiers, not transformers



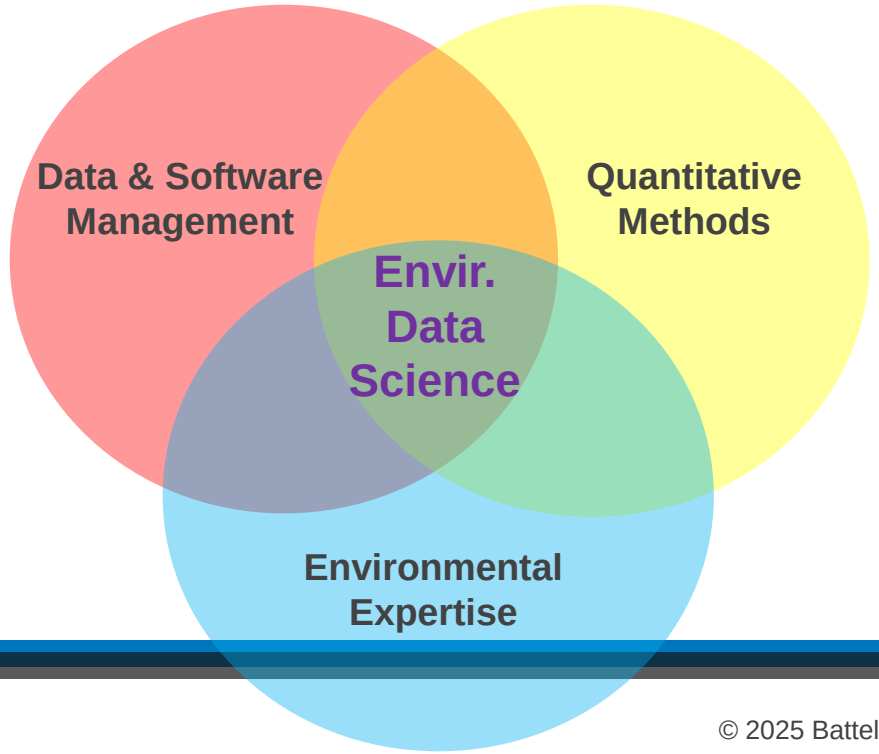
Data Quality Objective Process EPA QA/G4

Are observations/methods comparable? Co-mingling? Ecosystem setting described? Inputs to causation for inference? Data distributions? Outliers? Normalized? Homoskedasticity? Is rich training subset available? How definitive are truths?




Sufficient Data + Judgment of Experts + Suitable Algorithms = ML

Fitness of ML Methods:
Relevance? ➔ scientifically thorough curation
Accuracy? ➔ transparent statistics
Expert training? ➔ guidance of truths
No Hallucinations ➔ classify, not transform

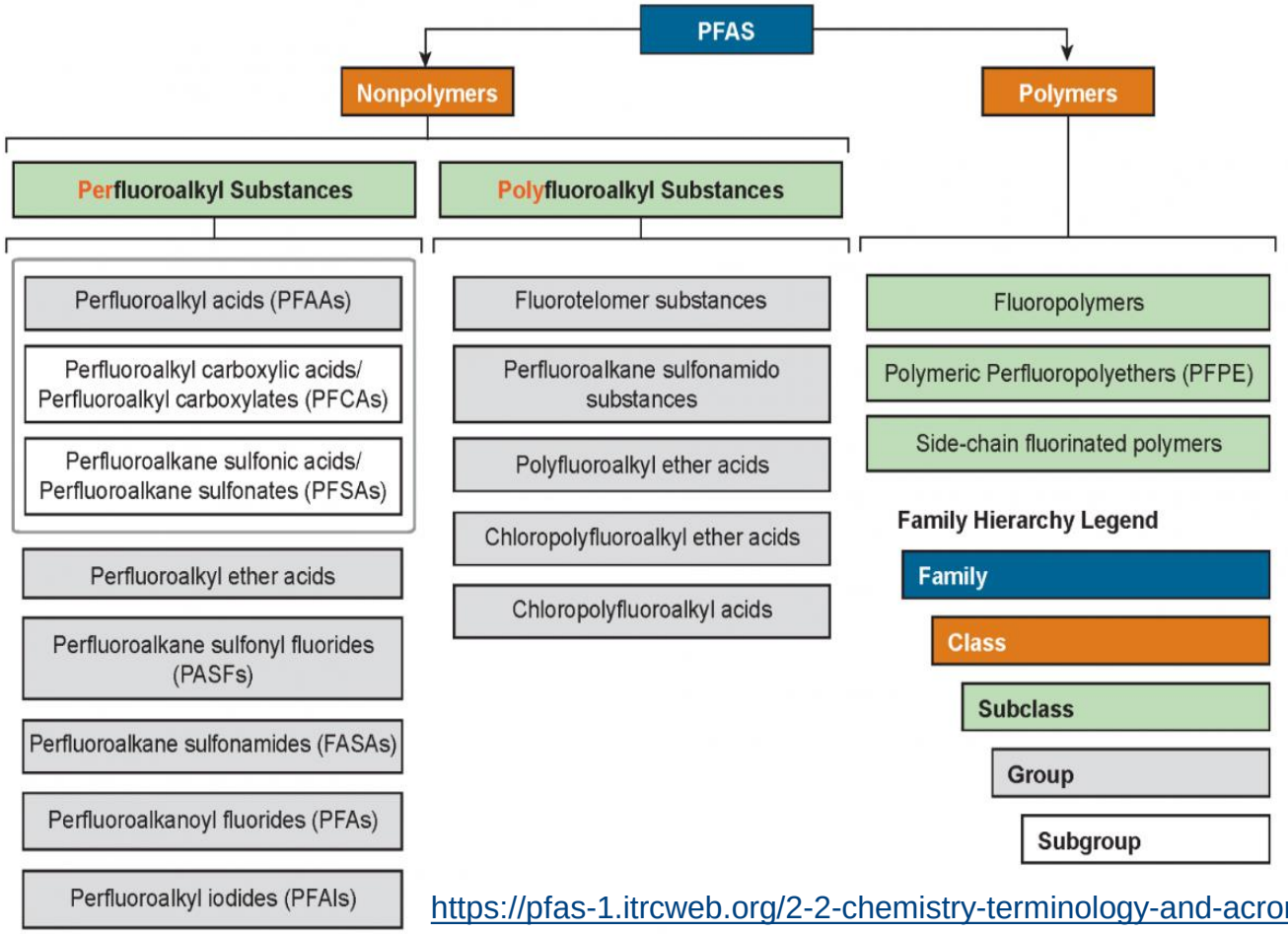


Examples: (1) PFAS Signature® (2) Literature

- PFAS Technical Leader
  - Kavitha Dasu
- Targeted Lab Analyses
  - Battelle Norwell Analytical Laboratory Team
- Analytical Chemists
  - Larry Mullins
  - Cameron Orth
- Data Scientists
  - Dave Friedenberg
  - William White
  - Christopher Scheitlin
  - Allen Chen

BATTELLE

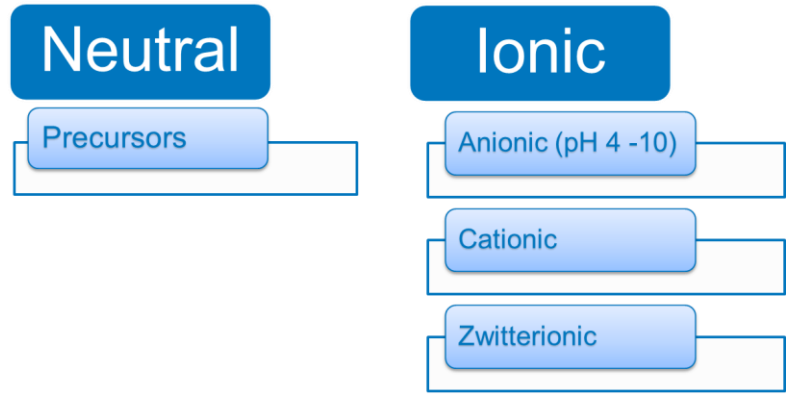
PFAS Source Identification, Mass Balance, Background via Underlying Chemistry



https://pfas-1.itrcweb.org/2-2-chemistry-terminology-and-acronyms/

More than 4700 chemicals in the family

AFFF Chemistry



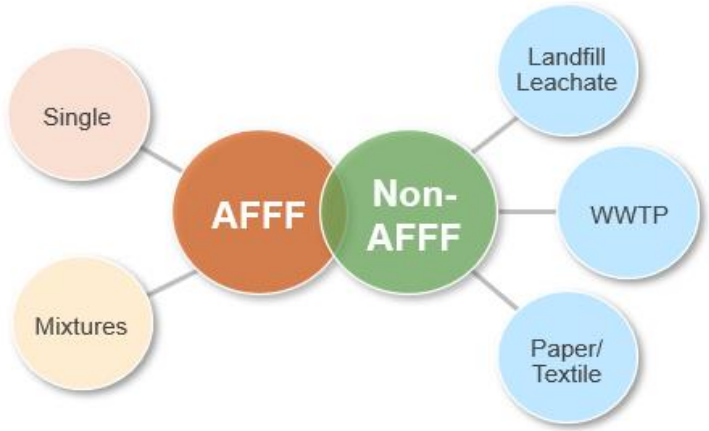
- Complex chemistry
- Changes in formulations
- Mixtures - partitioning behavior

Expertise in precursor chemistry and transformations is key

BATTELLE

PFAS Forensics, Mass Balance, Background Depend on Lab Methods

Approach	Description	Advantages	Disadvantages
Targeted Analysis (EPA Method 1633)	Use ratios of target analytes to identify trends through spider or radial plots, machine learning (ML) and other statistical approaches	• No additional analytical costs	• Non-discriminate – uses common terminal PFAS • Only identifies trends, not sources
Total Oxidizable Precursor (TOP) Assay	Applied to monitor the pre- and post oxidation PFAS analytes	• Limited additional analytical costs • Widely available commercially	• Non-discriminate – uses common terminal PFAS • Not widely accepted method
Non-Targeted Analysis/Suspect Screening Analysis (HRMS/LC-ToF/MS)	Monitor for hundreds of 'uncommon' PFAS	• More PFAS identified allows for more discrimination	• Large volumes of data that are hard to identify trends unless coupled with statistical analysis

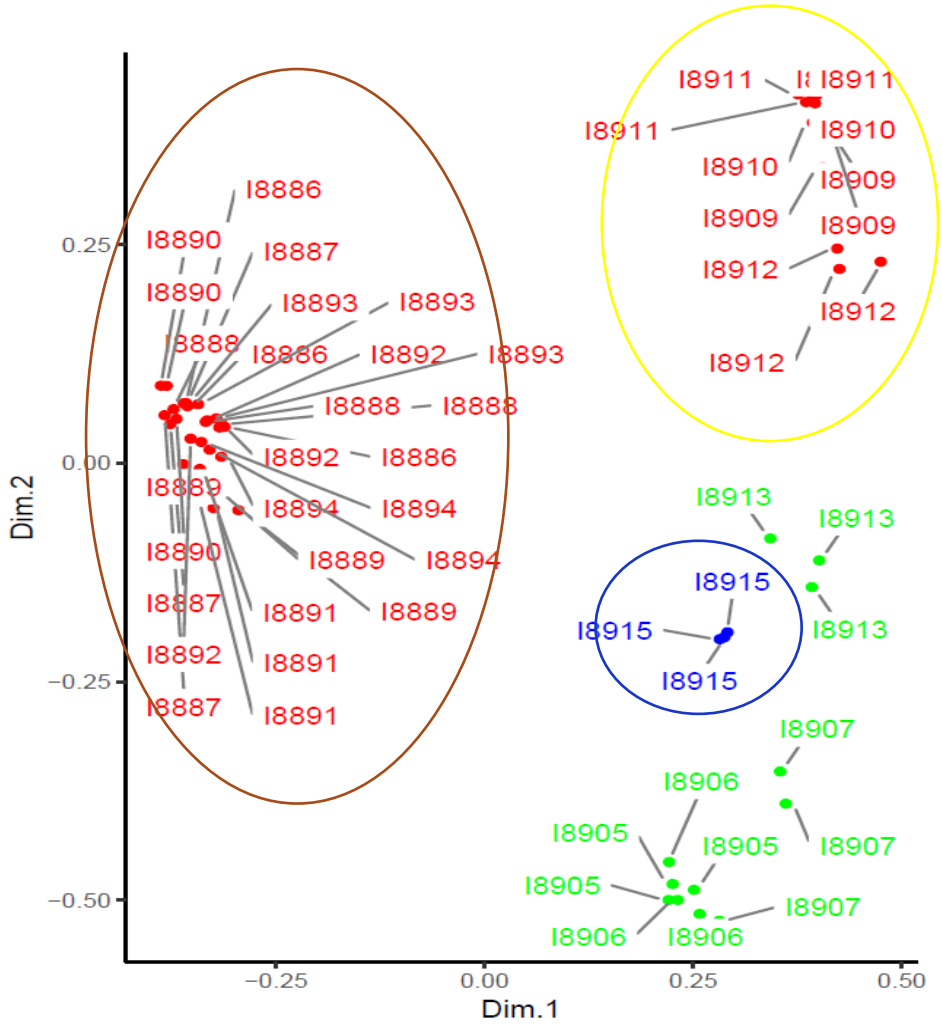


Data Processing
- Data Filtering
- Machine Learning Tool

Data Interpretation
- Other Lines of Evidence

BATTELLE

Final Output Based on Learned Classifications from Training Cycles



Cluster
- a AFFF
- a Non−AFFF
- a Waste Sector

High levels of PFAAs and other transformation products with few detections of FTS, N-SPAmp-FASA, N-SPAmP-FASAPS, N-SP-FASA, N-SHOPAmP-FASAA of both odd and even chain lengths and branched and linear isomers

AFFF related with both ECF and FT chemistry

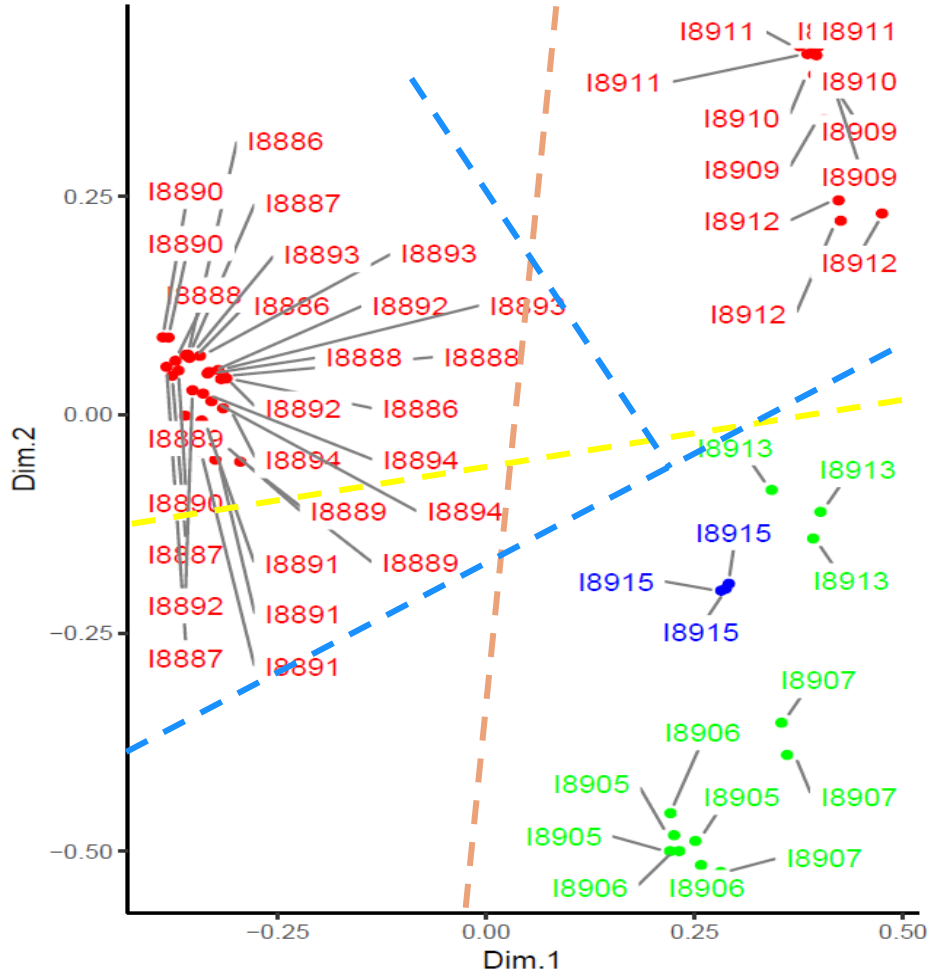
No detections of AFFF related analytes

6:2, 8:2, 10:2 FTABs found only in Waste sector

BATTELLE

To Assist Liability Allocation, Co-mingled Sources May Separate

Training "Learns" by Summing Distances from Various Regression Lines



- Multi-dimensional space, not just X-Y
- Numerous regression types classify observations
- Each cycle's pair-wise classifications are compared to **truth**
- **Gray areas and overlaps warrant expert interpretation**
- **Poor "fit" of new unknown sample can suggest novel source**

BATTELLE

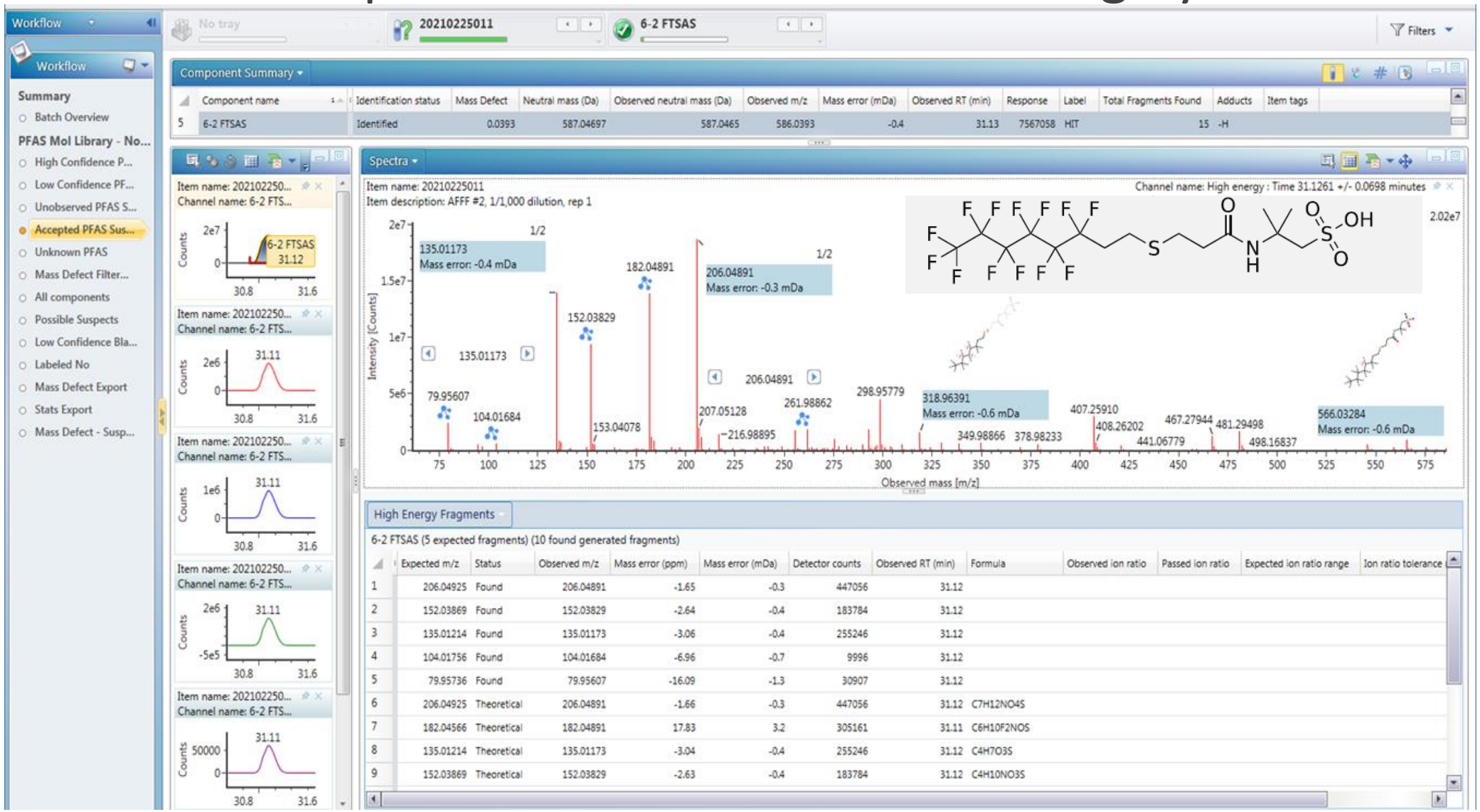
Training Library of Truths Enable Classifier to Discriminate PFAS Sources

- AFFF Formulations
  - ECF based
  - FT based
- AFFF-Impacted matrices
  - AFFF impacted Groundwater
  - WWTP located within AFFF impacted site
  - AFFF impacted biosolids applied soil
  - AFFF used for emergency response
- Industrial Processes
  - Metal Plating
  - Chrome Plating
  - Paper Mill

- Waste Sector
  - Landfill Leachates
  - Municipal WWTP related samples and additives
  - Compost
- Commercial Products
  - Fast Food wrappers
  - Stain resistant carpets
  - Cleaning products
  - Surface protectants

Forensic library is periodically added to as novel sources are confirmed

Example of FTSAS Detection Training by SME



BATTELLE

Literature Search Training and Augmentation of Knowledgebase

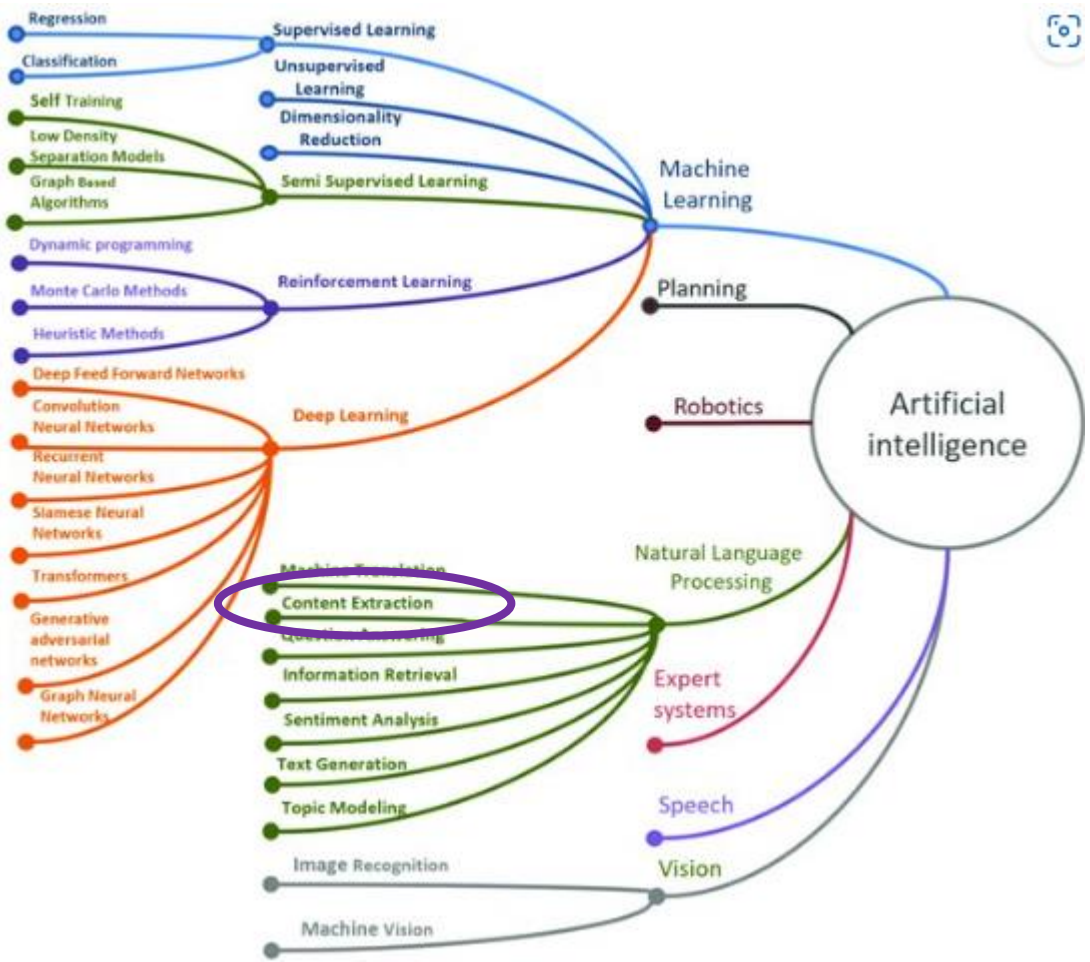
Fitness of ML Methods:

Relevance?

Accuracy? ➔ transparent statistics

Expert training?

No Hallucinations ➔ classify not transform



Mukhamediev et.al. (2022) Mathematics 10(15), 2552

BATTELLE

Milestones in Large Language Models (LLMs)

Foundational models, tailored to predict and generate plausible language

1990s

- Transition from rules-based NLP to statistical / ML models begins

2000s:

- Advancements in compute capability and machine learning complexity (deep learning)
- Increased internet data to train on

2013

- Introduction of Word2Vec (Google) – representation of meaning of words based on surrounding words

2018~2020:

- BERT - Bidirectional Encoder Representations from Transformers, 300M parameters (Google, Oct. 2018)
- BERT-like models, mainly for language understanding tasks

2020~present:

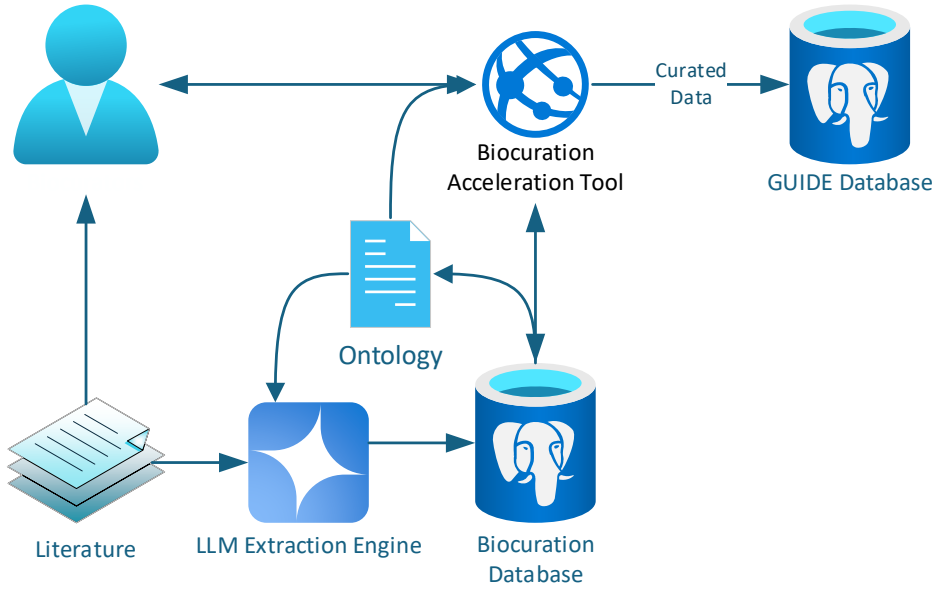
- GPT-1 (Open AI, June 2018), GPT-2 (Feb. 2019), GPT-3 (June 2020)
- GPT-like models, mainly for language generation tasks
- GPT-3.5 (Mar. 2022) 175 B parameters, GPT-4 (Mar. 2023) Super large
- Bard (Google, March 2023)
- LLaMA (Meta, Feb. 2023) 13B, and 65B, LLaMA 2, 7B, 13B, 70B (July 2023), LLaMA 3, 8B, and 70B (Apr. 2024)
- Other open-sourced LLMs (most fine-tuned from LLaMA 2), e.g., Vicuna, UniversalNER, GPT4All

Strategic Literature Search for Medical Countermeasures

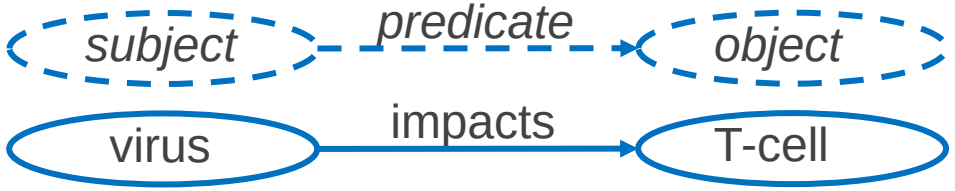
Semi-automated Biocuration Benefits

- Large Language Model (LLM) automates extraction of relevant information from literature
  - Evaluation of LLM performance demonstrated excellent recall – biocurator does not have to manually capture highly detailed information
- Biocuration acceleration tool (BAT) provides structured process **to validate and correct extracted information**
  - Human-machine teaming strategy: Human for quality review and machine for maintaining high attention to voluminous unstructured text
- BAT export of information to GUIDE Database increases standardization of identifier construction

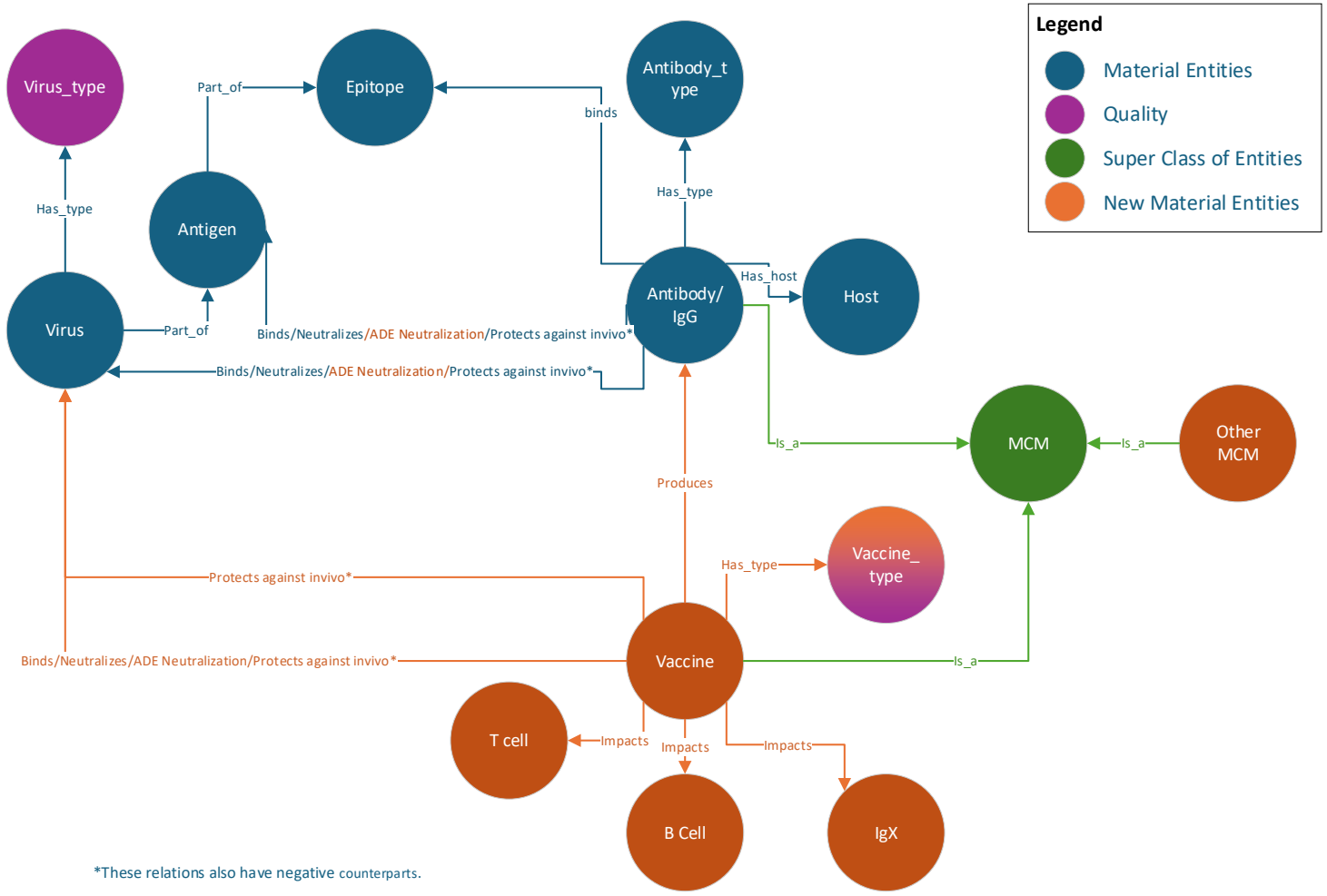
Biocuration Acceleration Pipeline



Resource Description Framework RDFs

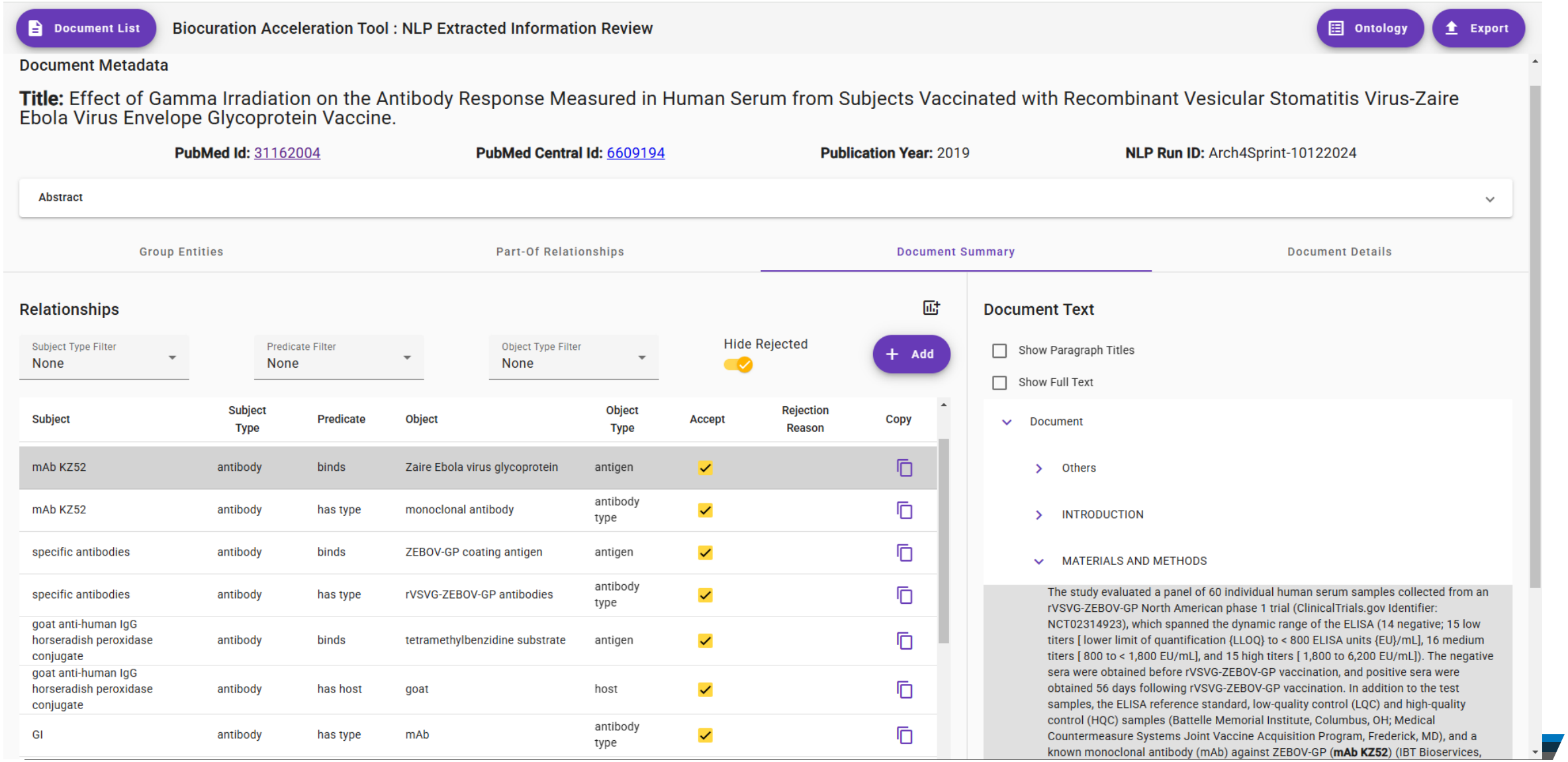


Strategic Literature Search for Medical Countermeasures



- Network of logic to guide search & improve reasoning
- Extracts knowledge from unstructured texts, such as scientific literature, guided by an ontology to standardize recommendations.
- Augments Subject Matter Experts' comprehensive knowledgebase.

Biocuration Acceleration Tool (BAT) Demo



ML Approach = Relevant Data + Expert Judgment + Tools

Fitness of ML Methods

Relevance

Accuracy

Expert training

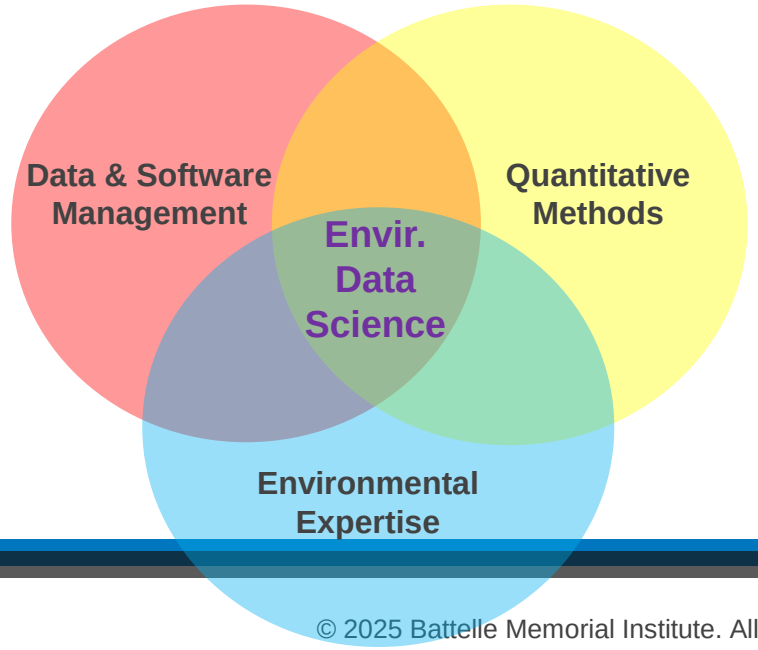
Classifications

Strategic knowledge leadership, Coordination

Speed & scalability via ML tools

Defensible findings

Focused data management effort, $



BATTELLE

THANK YOU

Tom Gulbransen
Gulbransen@Battelle.org
(516) 313-9311

Shalene Thomas
ThomasS3@Battelle.org

BATTELLE

It can be done

www.battelle.org/pfas