Leveraging Integrated Datasets to Understand the Origin of Groundwater Contamination

Athena Nghiem, Benjamín Bostick and Elizabeth Shoenfelt, Columbia University, LDEO.

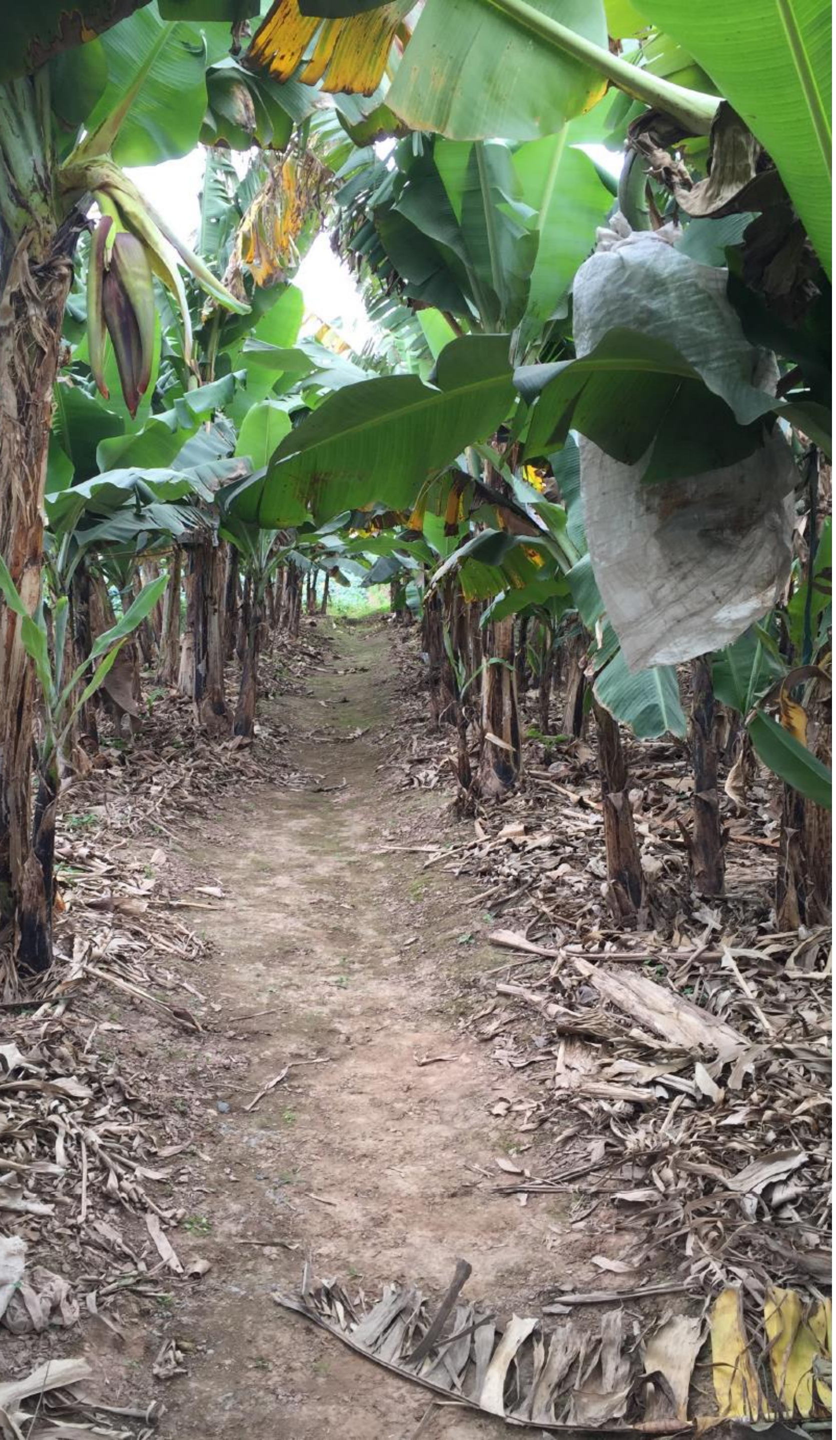
Mindy Erickson, USGS)

Sarah Thompson, NSLS-II)

Tracy Punchon, Dartmouth College

Jon Chorover, Univ. Arizona

José Manuel Cerrato, Univ. New Mexico



Overview

What

Redox processes to a large extent control subsurface groundwater composition.

Why

In redox processes that release As, a key process is the reduction of solid mineral substrates like iron oxides that host As. ***Although we know a lot about groundwater composition from analyzing well water, we know little about solids because they are hard to see and not often analyzed at any field site.***

Bringing Data Science to the Table

We can ***characterize sediment redox states and deposition history using X-ray absorption***, and larger integrated datasets of groundwater arsenic to learn about these solid transformations more broadly, including in the US

FAIR Synchrotron-Based Data Access, and Workflows

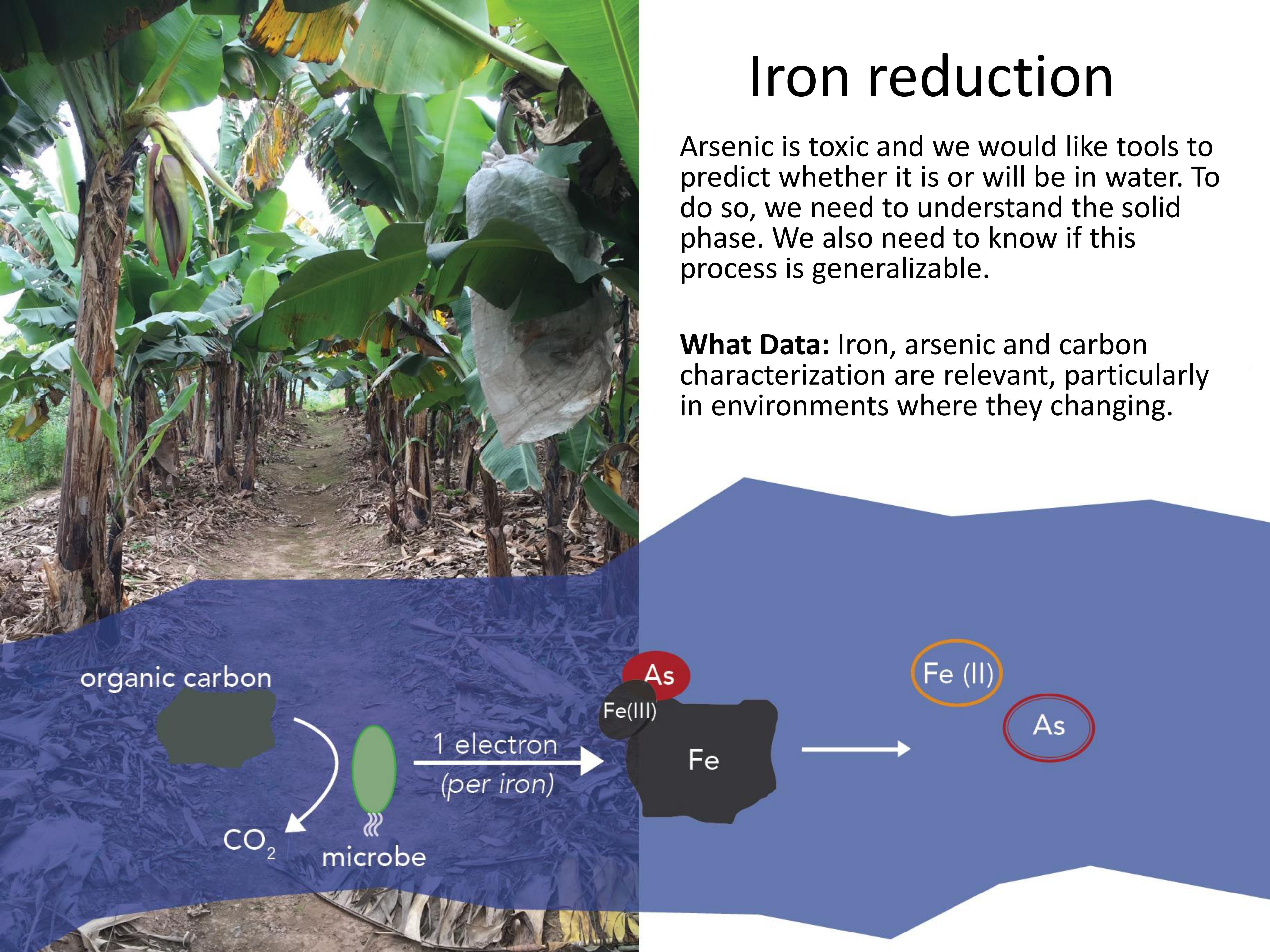
To best use synchrotron-based spectroscopic data, and thus to understand the chemical speciation, and thus the fate, transport and toxicity of arsenic, we need automated, integrated, improved, and standardized methods of data analysis.

- **Integrating Environmental Analyses of biological and environmental samples analyzed using synchrotron-based methods.** This starts at the integration of data from different students and field projects within a group and extends across collaborators and institutions.
- **Integrating reference spectra to create a uniform set for Iron, Sulfur and Arsenic Reference Spectra (others in progress)**
- Developing statistical approaches of determining reference integrity and data quality
- **Creating unsupervised spectral analysis tools that leverage these reference datasets to simplify the complexity of environmental complexity.**
- **Integrating intercompatible formats to ensure uniform data collection in and analysis with existing software.**
- Creating web-based workflows that (a) are easy to use, (b) ensure uniformity in data analysis, (c) permit direct data upload to a database, and (d) conduct automated QA/QC checks on submitted data to better understand. Submission is currentlyboth voluntary and tedious, but probably needs to change to be most useful.

Iron reduction

Arsenic is toxic and we would like tools to predict whether it is or will be in water. To do so, we need to understand the solid phase. We also need to know if this process is generalizable.

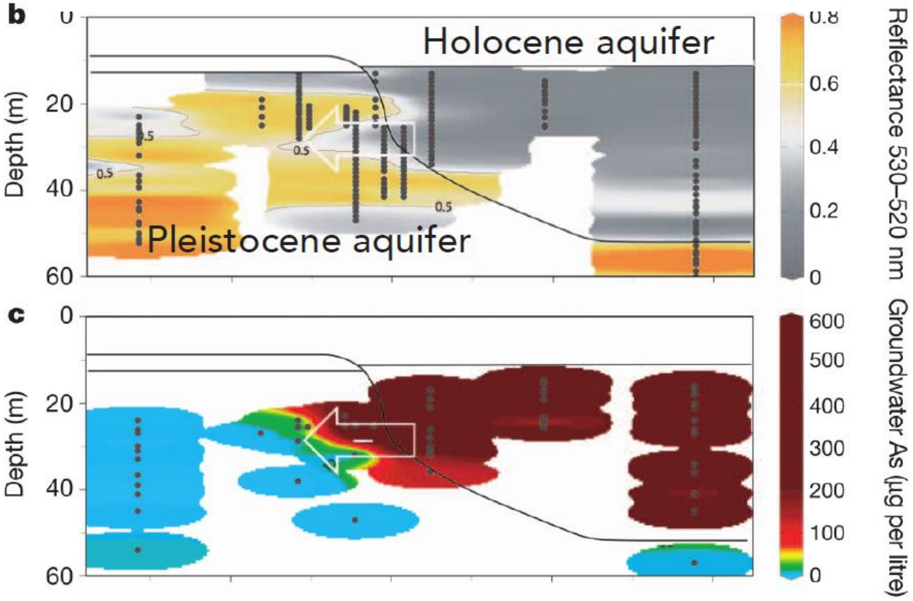
What Data: Iron, arsenic and carbon characterization are relevant, particularly in environments where they changing.



Where

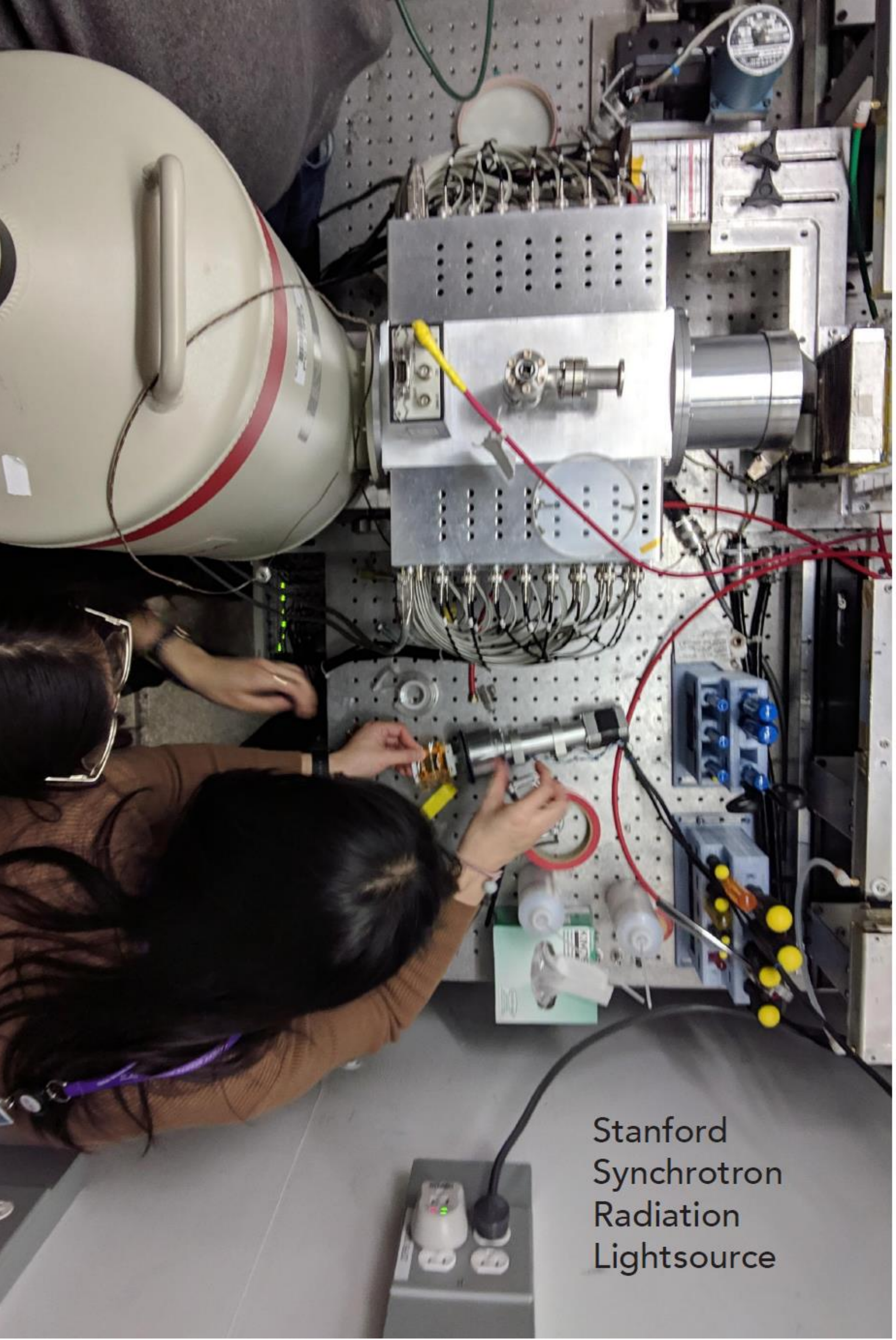
We do not have dense data in most places. Need to develop method with dense dataset.

Field site in Vietnam (ear Hanoi), but with the ability to link the processes here to the other areas (Bangladesh, the US)



(van Geen et al., 2013)

how



Stanford Synchrotron Radiation Lightsource

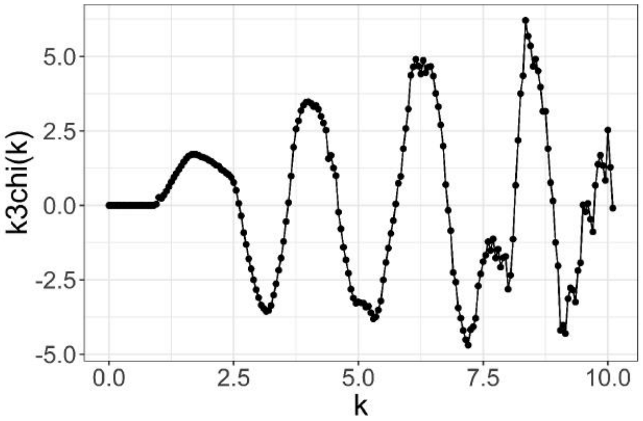
how

X-ray absorption spectroscopy of Fe, a measure of the chemical environment of Fe.

Fitting complex data for small datasets requires linear combination approaches (a lot of knowledge)

Often, we lack some of the information needed for classification. ***Unsupervised approaches*** are effective as datasets become larger. (Individual projects, grad students → unified datasets that can grow and be amended)

Metadata: WWWWWH the sample was collected from, water data from the site, other information. High standard to be useful data.

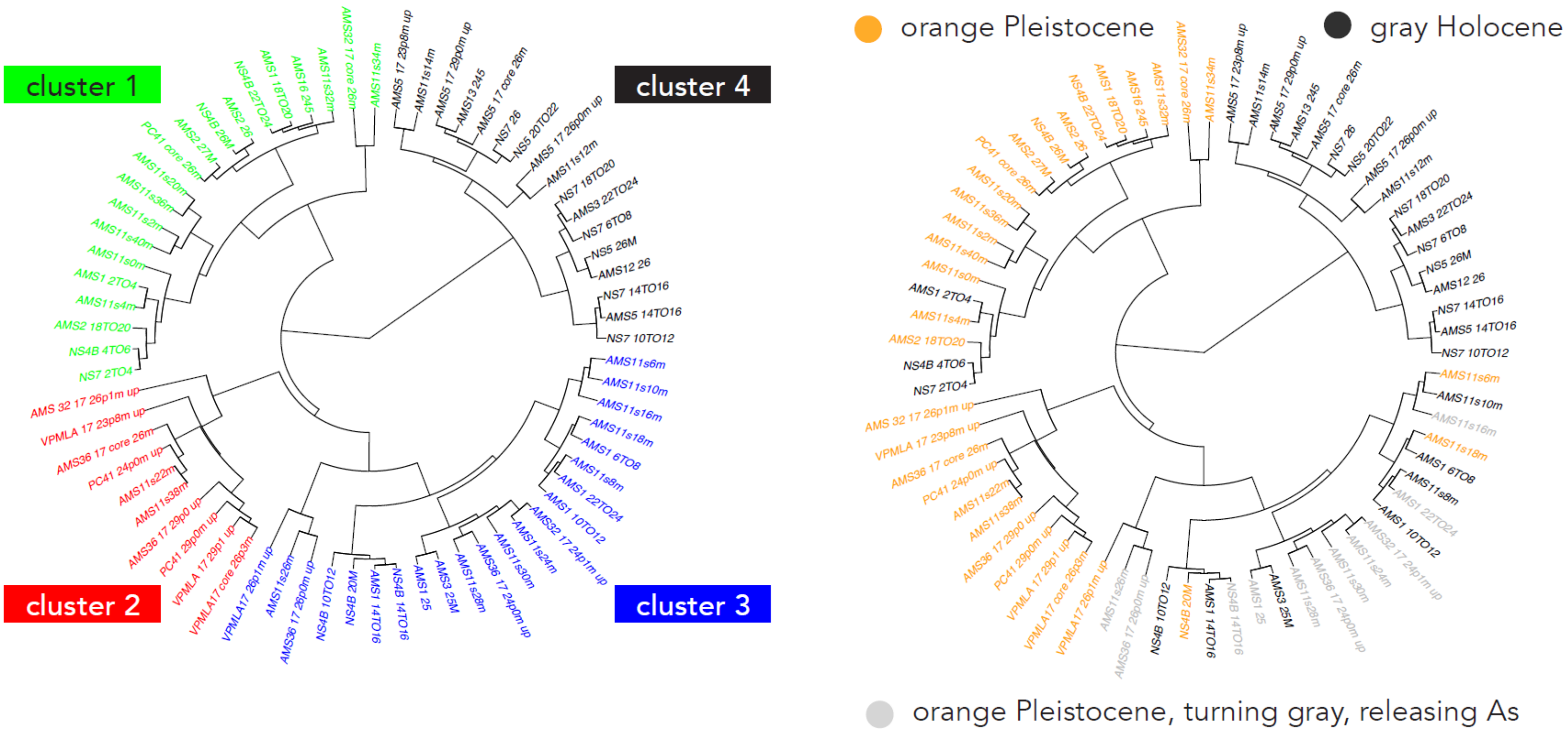


Nghiem et al. *ES&T Letters*, 2021

figure 1 unsupervised cluster analysis distinguishes between Holocene and Pleistocene sediments, including varying As-releasing redox states

colored by
hierarchical cluster analysis clusters

colored by **field classifications**



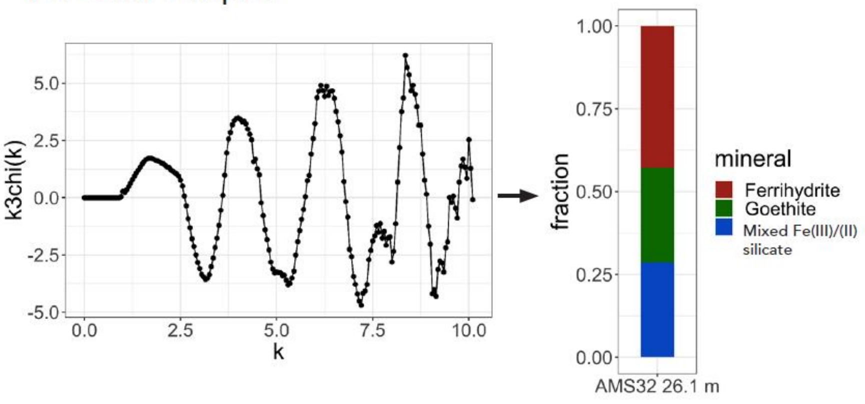
Nghiem et al. *ES&T Letters*, 2021

how

4. LINEAR COMBINATION FITTING

Use 10 Fe standards: ferrihydrite, goethite, hematite, green rust sulfate, magnetite, mixed Fe(III)/(II) silicate, biotite, siderite, mackinawite, pyrite to fit sample spectra.

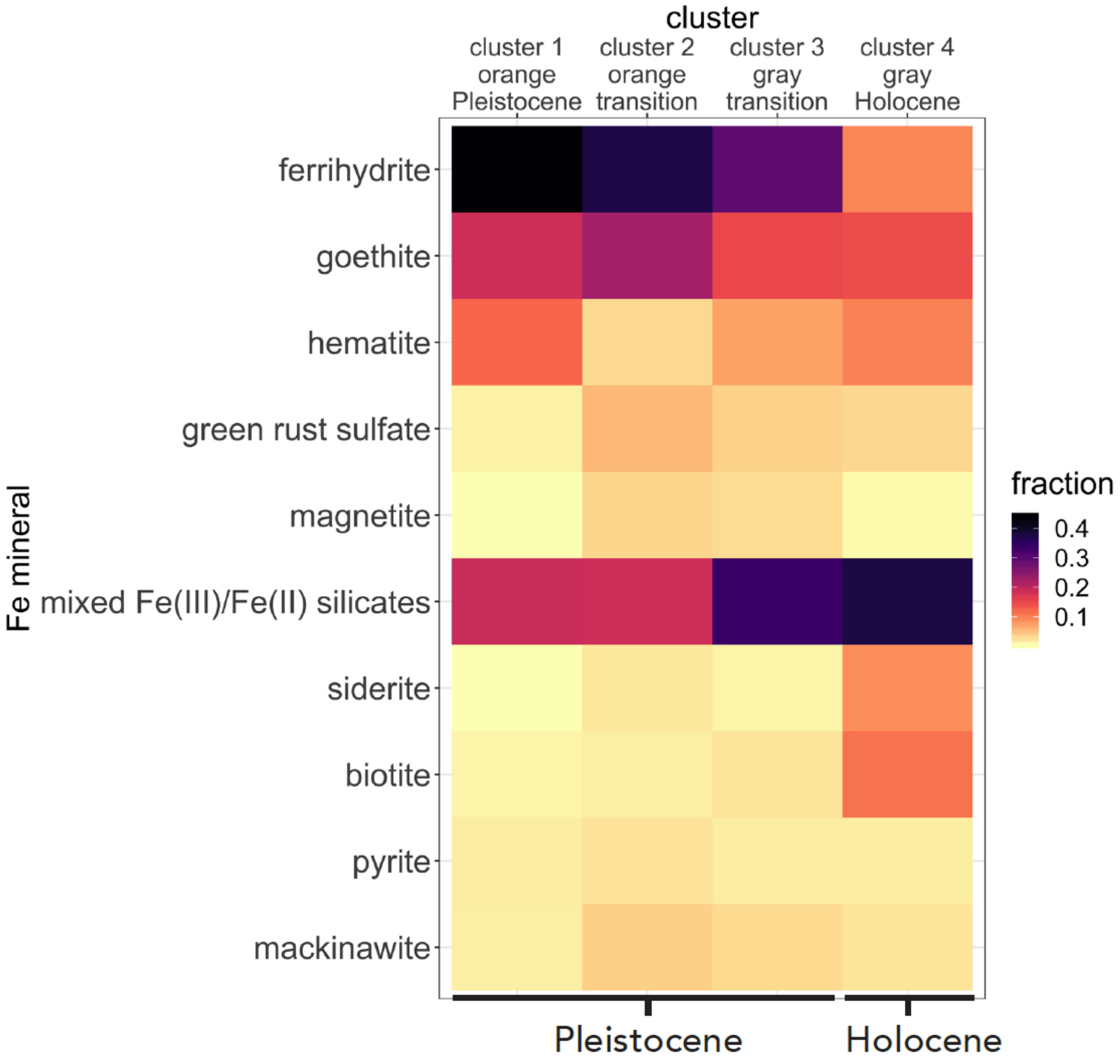
For each sample,



Nghiem et al. *ES&T Letters*, 2021

figure 2

clusters are meaningful: oxidized minerals and minerals undergoing reduction found in Pleistocene clusters, reduced minerals in Holocene cluster



Need ancillary data to determine which of the reference materials should be used for fitting. And you need a complete set of possibilities since you might not have the proper ones.

Integrated datasets of references are very useful. Shared references critical because the references are more complex than you expect them to be.

Nghiem et al. *ES&T Letters*, 2021

how

Match the sediment cluster with mineralogy changes in groups (not individual samples) and pair sediment environment/cluster to aqueous composition

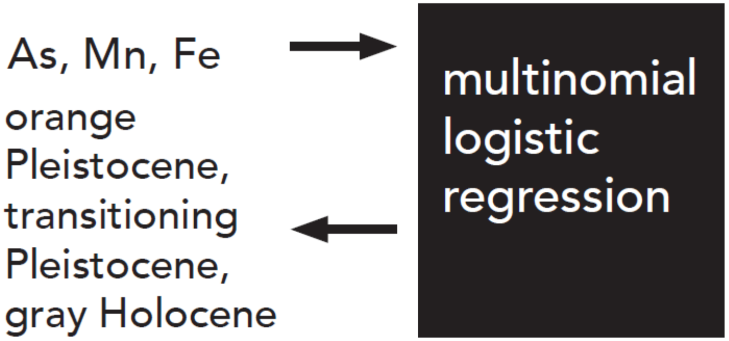
Data: Need integrated data of other types for comparison and scaling activities. In this case, >1M water samples from around the world.

Nghiem et al. *ES&T Letters*, 2021

how

6. CLASSIFICATION MODEL

Run and cross-validate supervised classification model to classify sediment by redox sensitive aqueous measurements



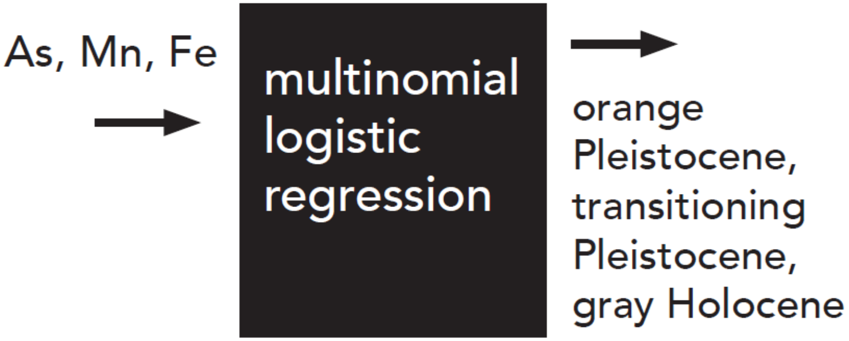
As, Mn, Fe

orange Pleistocene, transitioning Pleistocene, gray Holocene

multinomial logistic regression

7. APPLY CLASSIFICATION MODEL

Apply it to USGS Powell Center compiled database aqueous measurements of As, Mn, Fe (n= 16294)



As, Mn, Fe

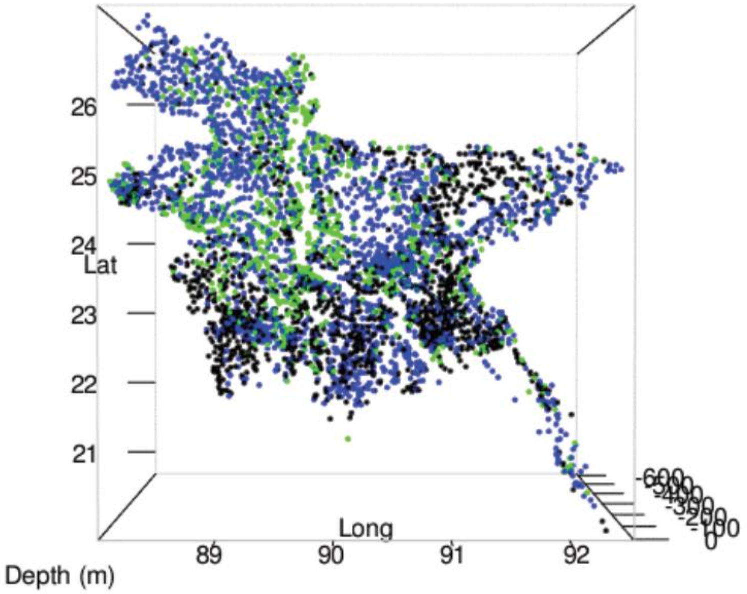
multinomial logistic regression

orange Pleistocene, transitioning Pleistocene, gray Holocene

Nghiem et al. *ES&T Letters*, 2021

figure 4

sediment classification of Pleistocene clusters matches previously identified Pleistocene terraces



cluster 1 & 2	orange Pleistocene
cluster 3	transitioning Pleistocene
cluster 4	gray Holocene

Barind Tract

Madhupur terrace

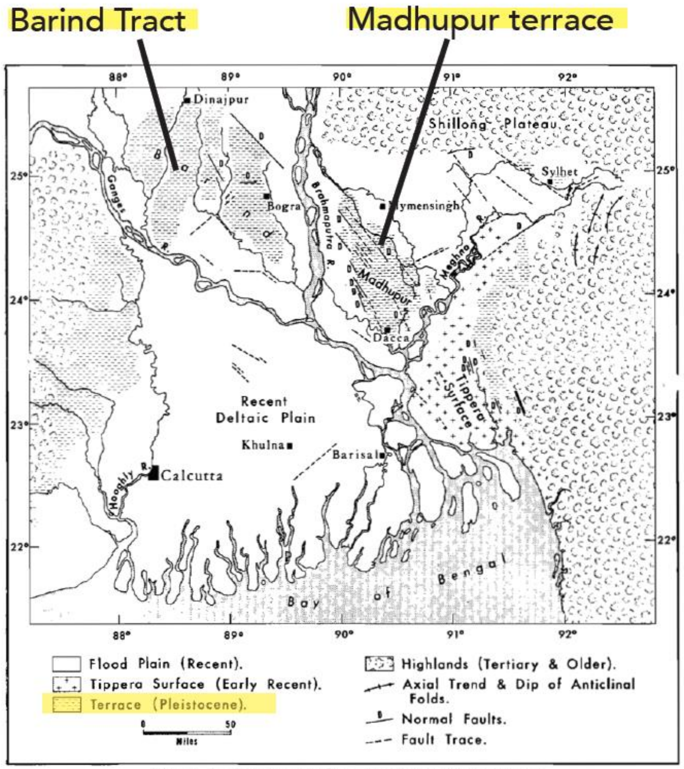


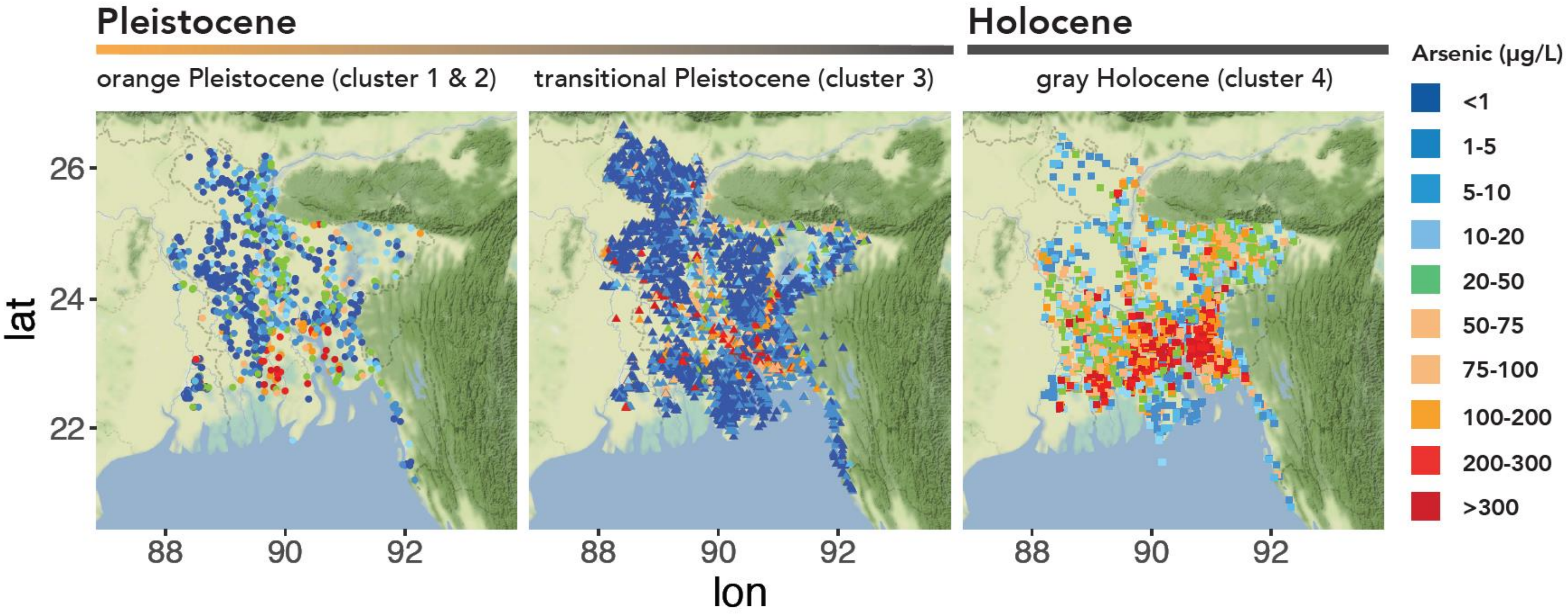
FIGURE 3.—QUATERNARY GEOLOGY OF THE BENGAL BASIN

(Morgan and McIntire, 1959)

Nghiem et al. *ES&T Letters*, 2021

figure 5

High aqueous arsenic is found in Holocene sediments and in a significant number of converted Pleistocene sediments undergoing reduction. The widespread occurrence of this group is worrisome.



Nghiem et al. *ES&T Letters*, 2021

Overall Results

Iron mineralogical structure useful to classify sediments into meaningful and useful categories relevant to geology, hydrology and public health.

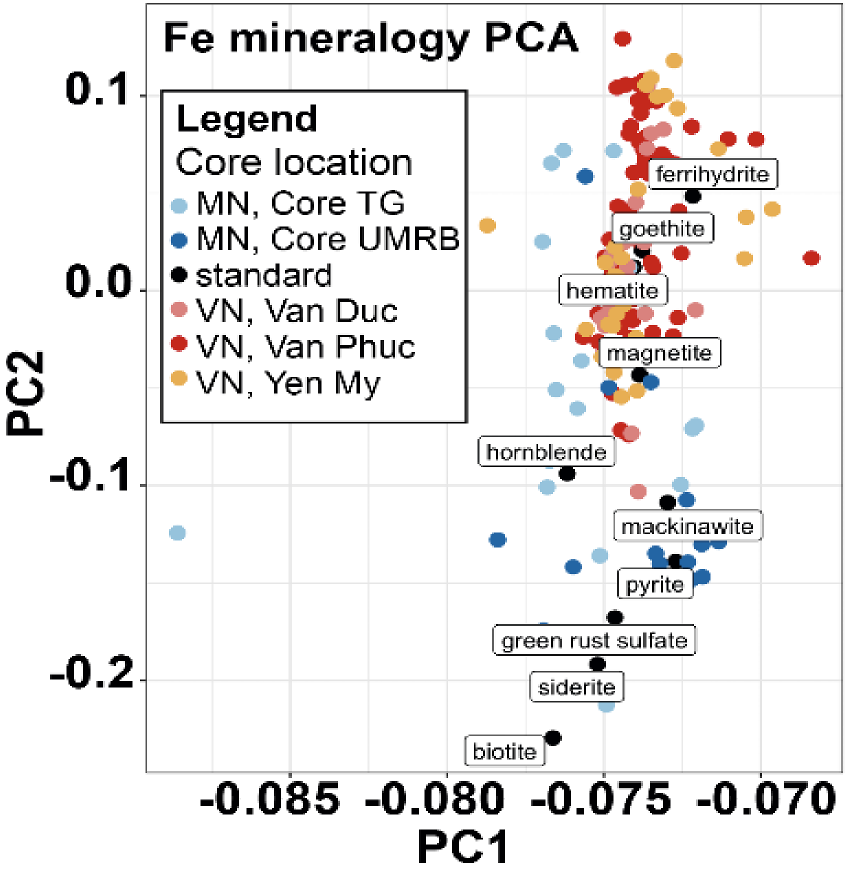
Challenge

How far can we apply unsupervised method?

Nghiem et al. *ES&T Letters*, 2021

Figure 6. Extending the data to Minnesota

Comparison of Vietnam (“VN”) and glacial aquifers from Minnesota (“MN”) sediment core Fe mineralogy based on PCA. More oxidized abundant Fe(III) are on the top, while reduced sulfidic phases are at the bottom of the reaction coordinate. Data from Nghiem 2020, Nicholas 2017and other data integrated in EUC.



Nghiem et al. *In prep.*