iDAVE Web Application

Theory and Application of the Internet Discriminant Analysis Verification Engine for the U-Sourcing Database Project

Martin Robel

02 June, 2010

Lawrence Livermore National Laboratory

LLNL-PRES-428127

Introduction

- ***iDAVE*** (Internet Discriminant Analysis Verification Engine)
- Secure web application
- Predict the source of nuclear material based on its chemical and isotopic concentrations.
- ***iDAVE*** applies Partial Least Squares –Discriminant Analysis (PLS-DA) using data stored in a database.

Many methods of multivariate classification to chose from

- For example:
  - PCA (principal components analysis)
  - LDA (linear discriminant analysis)
  - KNN (k-nearest neighbor)
  - CART (classification and regression tree)
  - PLS-DA (partial least squares –discriminant analysis)

❖ Different problems require different tools.

❖ We chose PLS-DA because it has proven most effective for our data and goals.

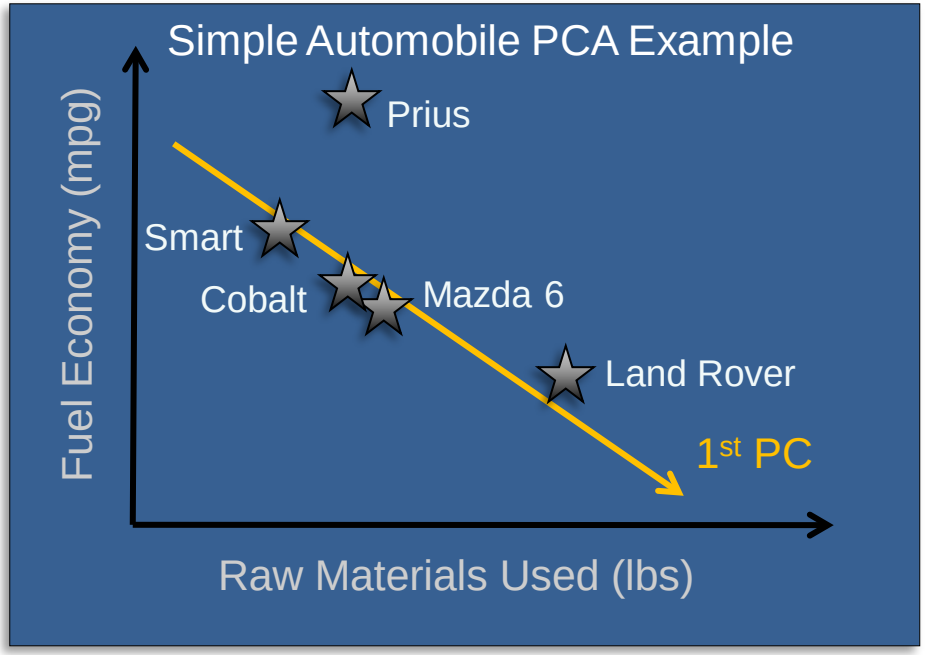
Principal Components Analysis (PCA) reduces dimensions while retaining maximum information (variance)

- PCA exploits correlation in multi-dimensional data
- The basic approach:
  - 1st PC: "If you could have only one dimension to *describe* a distribution (in n-dimensions), which would you chose?"

1st PC: captures greatest variance

Automotive example:

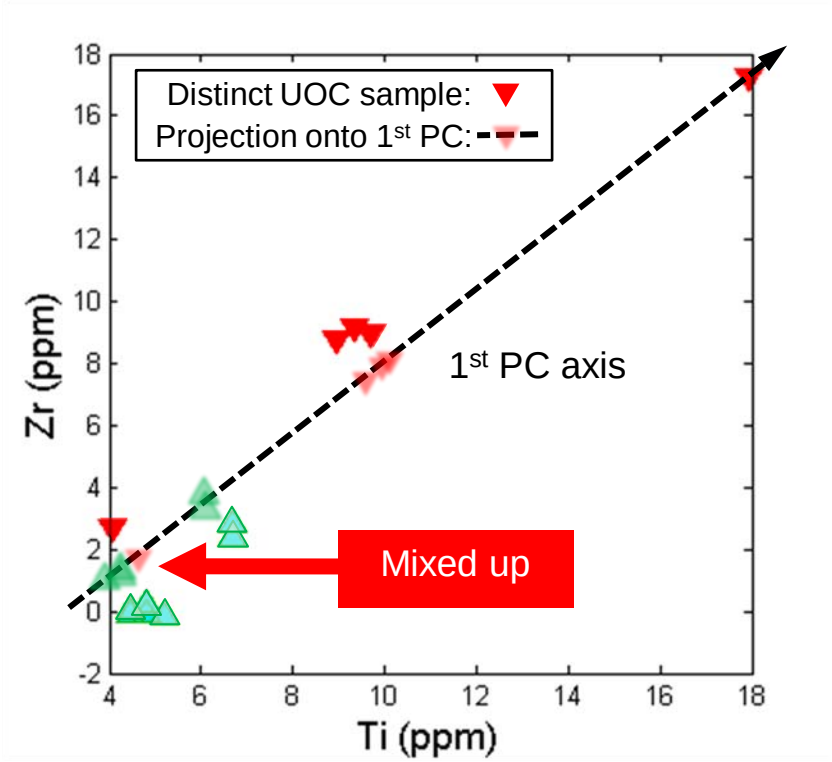
1st PC (Latent Variable 1) = Environmental Impact



PLS-DA is more appropriate for the classification problem.

- The basic approach:
  - 1st Latent Variable (LV) axis: "If you could have only one dimension *to tell two distributions (groups) apart*, …"
  - Conceptually similar to PCA, but different criteria: *maximize ratio of between group variance to within group variance*
  - 2nd LV orthogonal to 1st

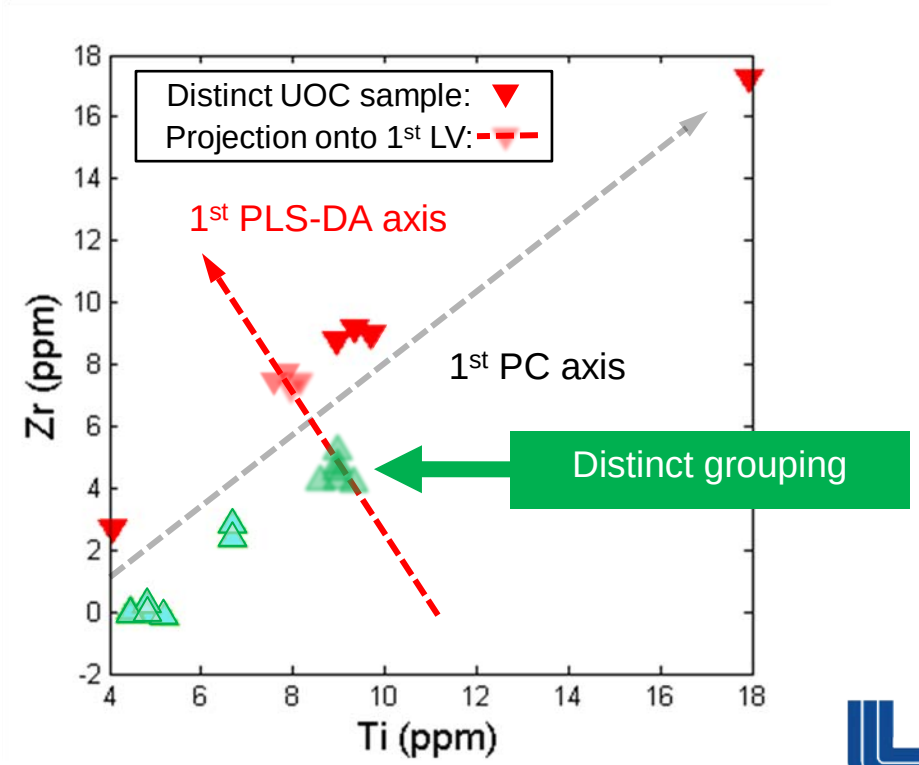
- PCA is optimized for *describing*
- PLS-DA is optimized for *discriminating*



PLS-DA is more appropriate for the classification problem.

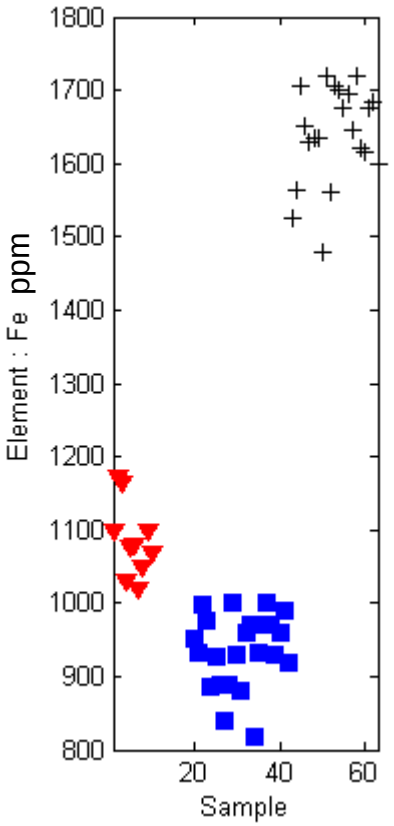
- The basic approach:
  - 1st Latent Variable (LV) axis: "If you could have only one dimension *to tell two distributions (groups) apart*, …"
  - Conceptually similar to PCA, but different criteria: *maximize ratio of between group variance to within group variance*
  - 2nd LV orthogonal to 1st

- PCA is optimized for *describing*
- PLS-DA is optimized for ***discriminating***

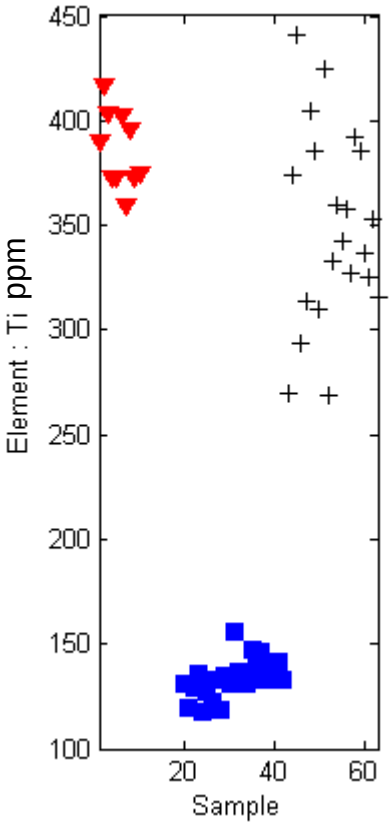


PLS-DA example: obsidian archeological samples from different sources[1] Plotting raw data vs. plotting onto single reduced ("latent") dimension

Univariate discriminant analysis

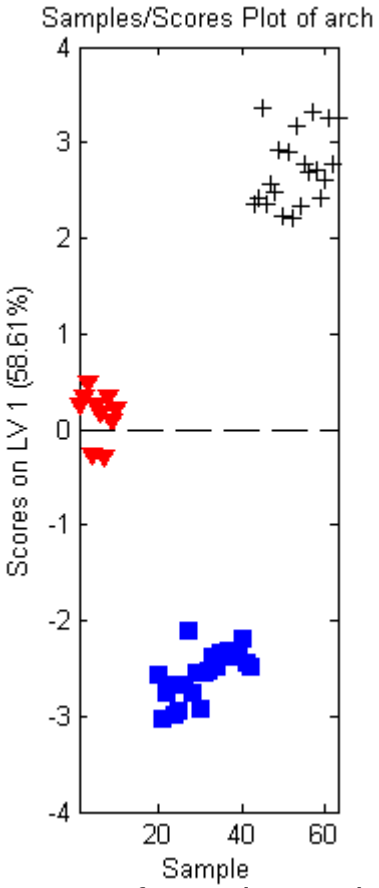


Fe is a good univariate discriminant variable, but cannot fully separate the classes.



Ti is very effective at picking up where Fe leaves off.

PLS -discriminant analysis

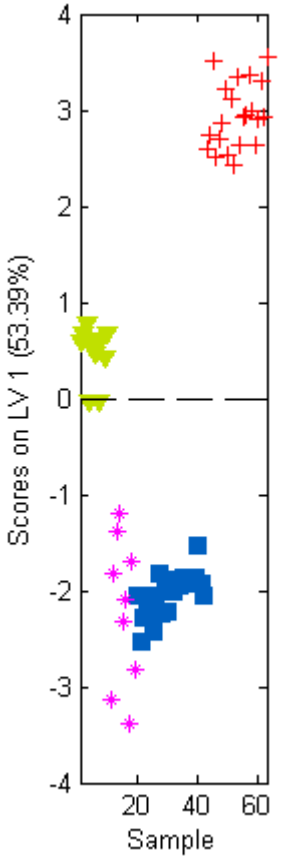


Same data transformed onto the first Latent Variable (LV) in a model with 10 elements.

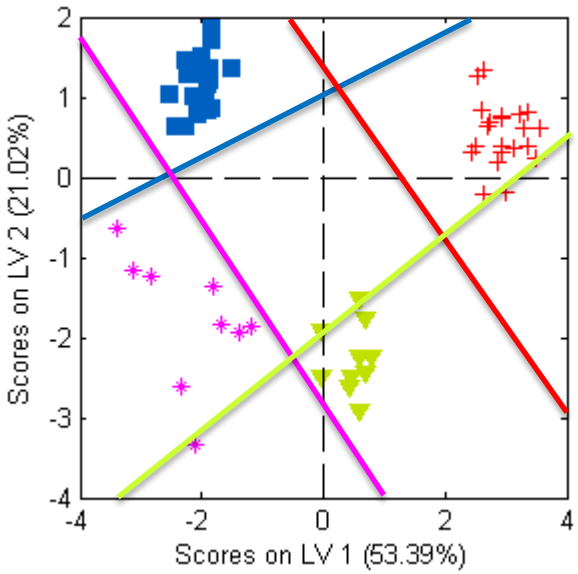
1. BR Kowalski, TF Schatzki, FH Stross. Classification of archaeological artifacts by applying pattern recognition to trace element data. Anal. Chem.; 1972; 44(13); 2176-2180.

PLS-DA example: obsidian archeological samples
Multiple class discrimination

Single discriminant axis for 3 or more classes is often not enough

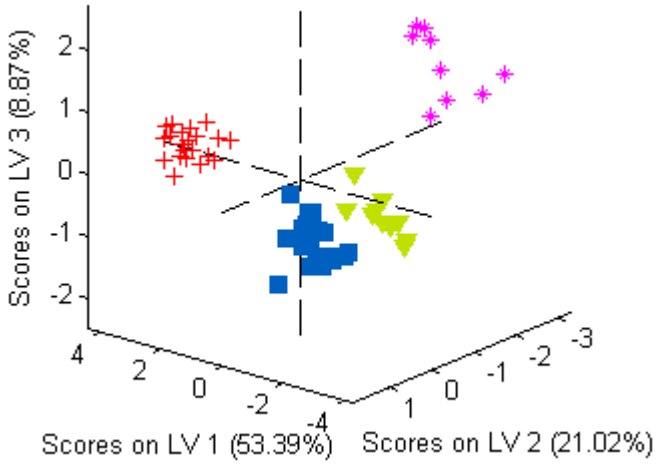


Decision lines in 2D space;



- 4 discriminant axes
- Orthogonal to decision lines.

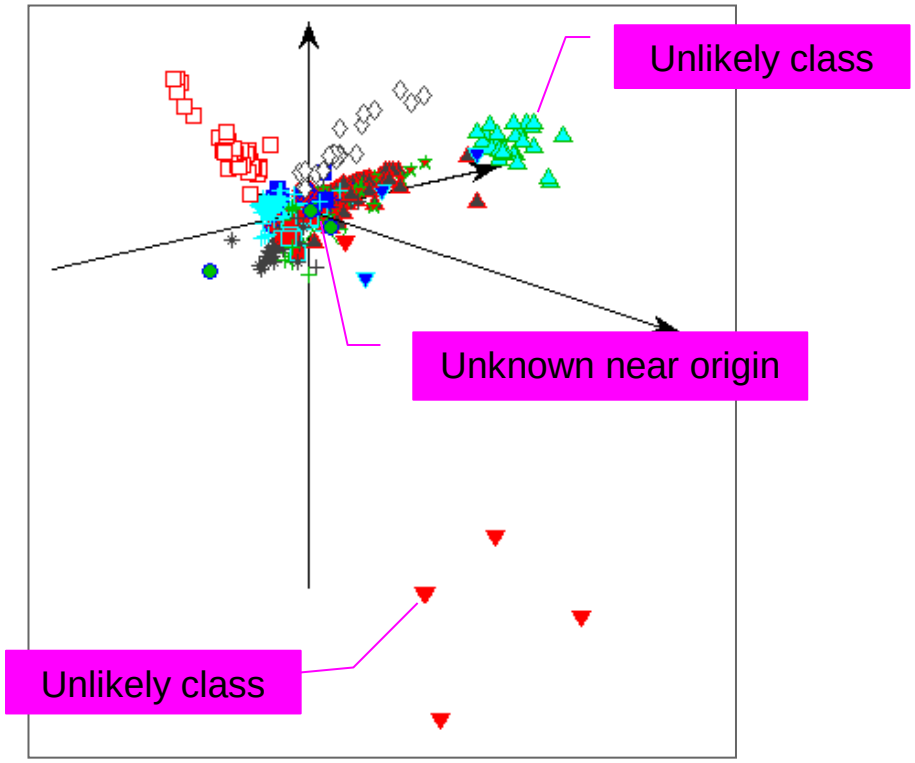
- 3D and decision *planes*
- 100% correct classification



- 4D and above = hyperplanes
- Hard to visualize
- Math is the same

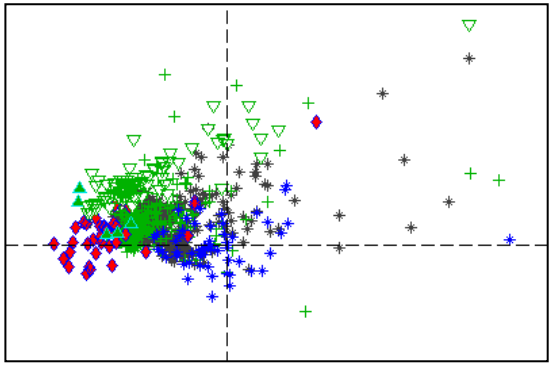
Challenges of UOC data and iterative solution

Even with hyperplanes, highly overlapping UOC data does not separate out perfectly.



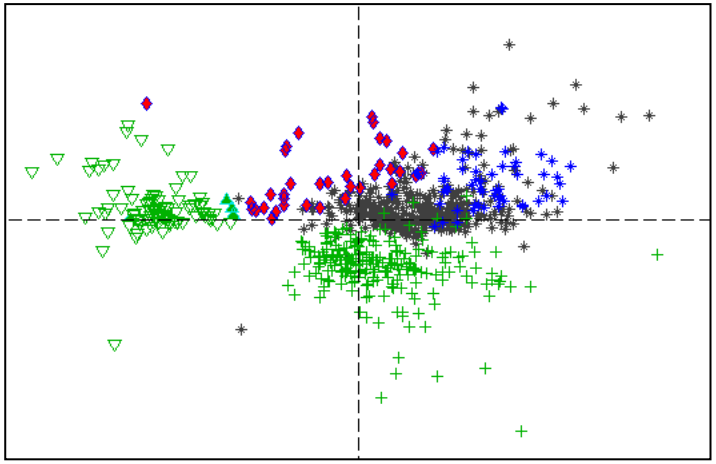
Subset of these data showing before and after removing unlikely classes

Before



Subset of classes in model using all data

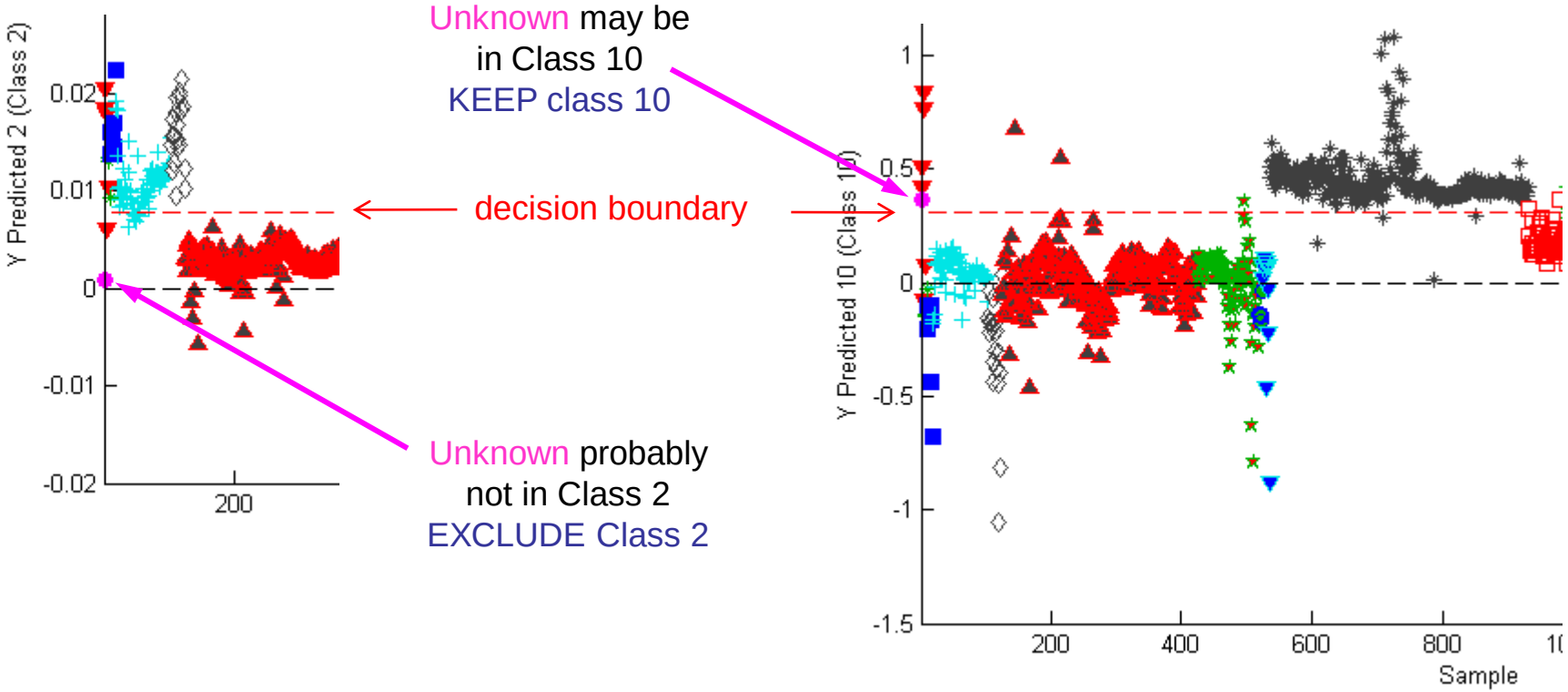
After



Same data in new model -unlikely classes removed

Implementing an objective criteria for iterative removal of outlier classes: achieving a repeatable, objective, and readily automated criteria for class exclusion.

- Test unknown for membership in each class separately
- Sources that have been definitively ruled out using the decision boundary are excluded from the next iteration.



In summary: how iDAVE uses PLS-DA iteratively

1. Build PLS-DA model using all available variables and sources.
2. For each class (source), test if unknown is more like that class or more like everything else using a modified Bayesian decision boundary.
3. Exclude “outlier” classes and rebuild model.
4. Repeat until single class conclusion.

Using iDAVE

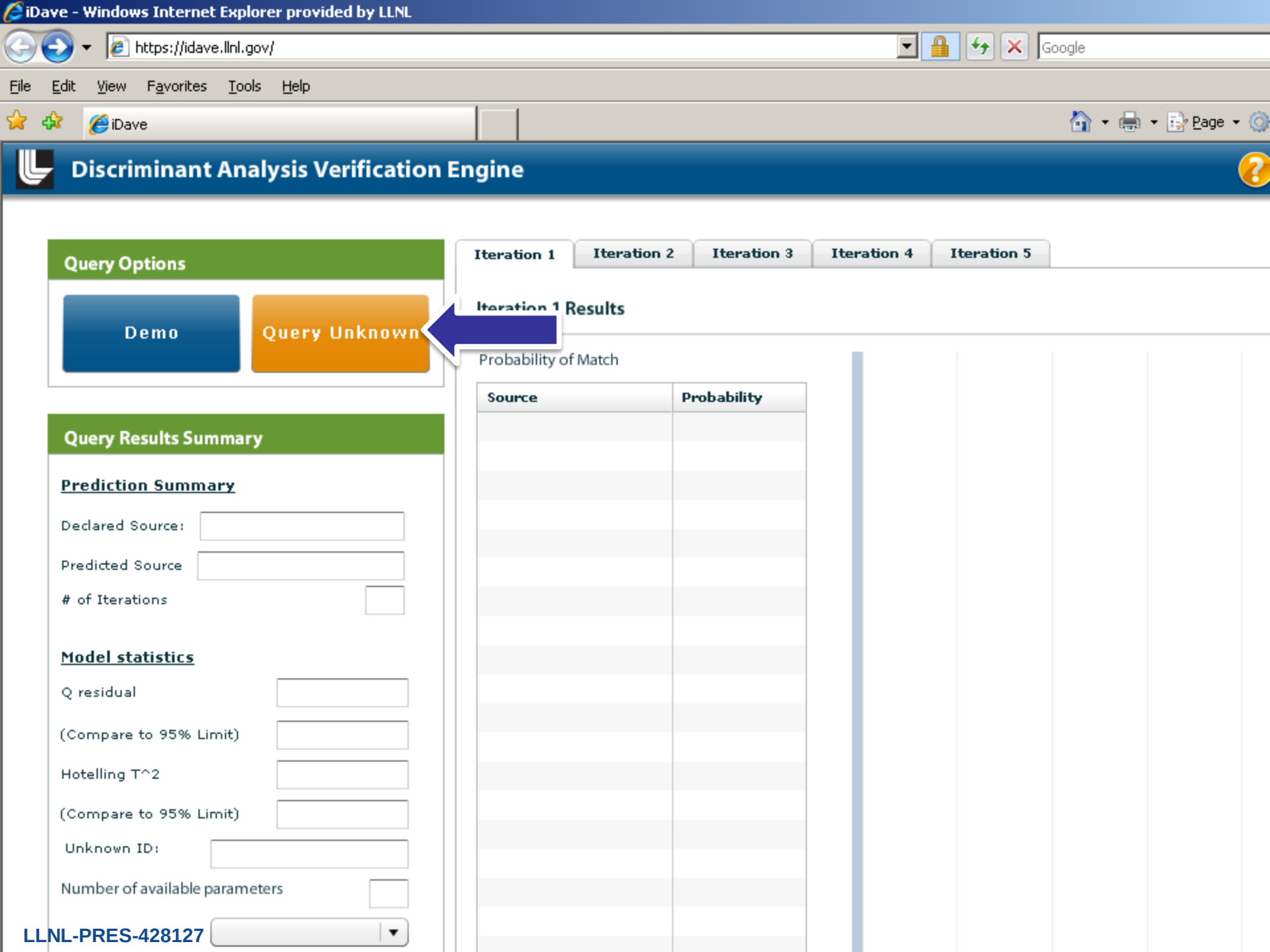
- Two modes of operation
  1. "Query unknown"
     - Input values for your unknown and test it against the current reference database to see what it is most similar to.
  2. "Demo"
     - Perform an *internal validation* one sample at a time to see how well the current model performs for different sources/locations. This is typically referred to as Leave One Out (LOO) validation.

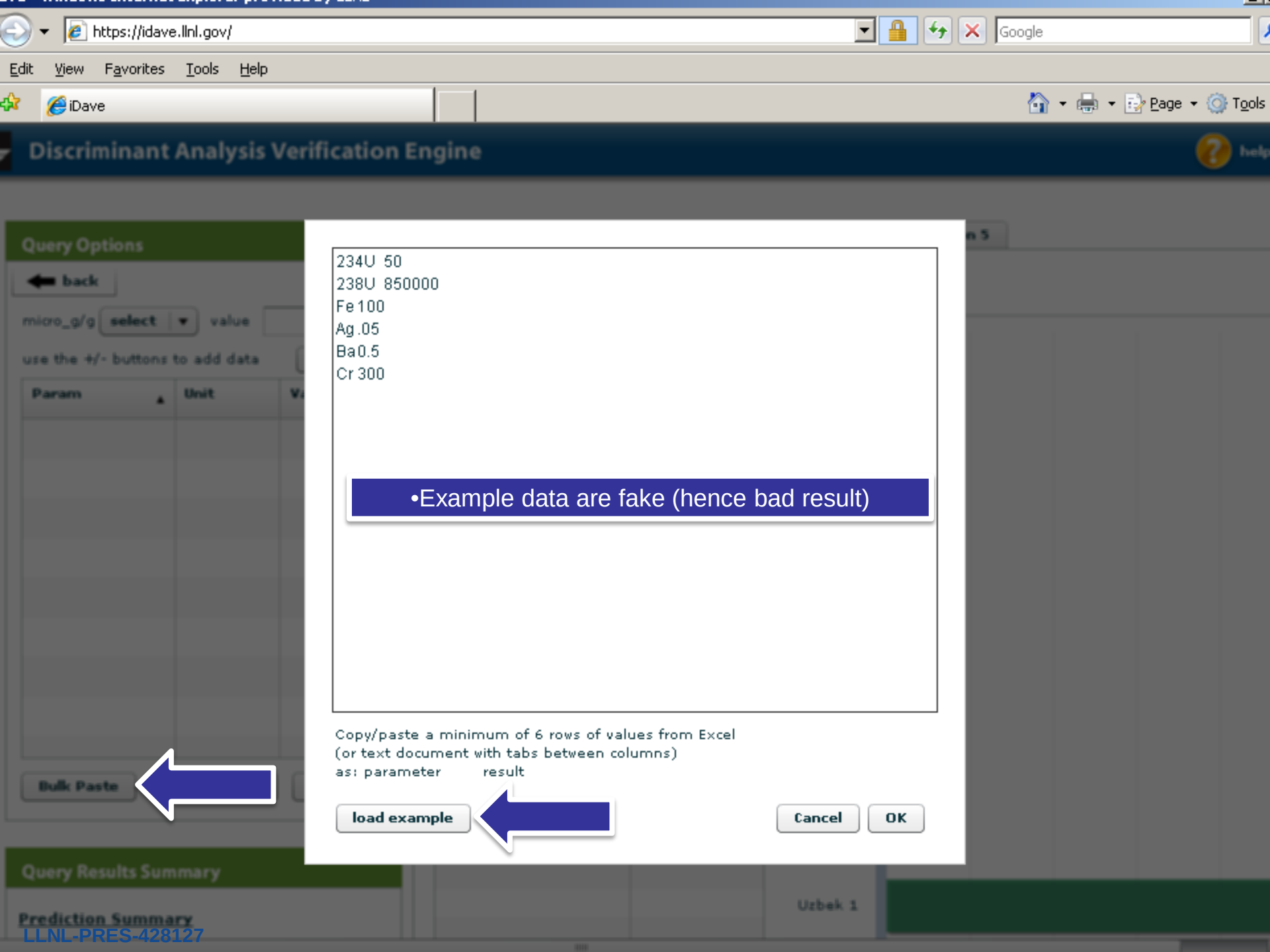
Interpreting results from iDAVE

- iDAVE gives what statisticians call a “prediction.” This is a best guess based on available information.
- We are researching ways to characterize the level of confidence in the “prediction” generated by iDAVE in an appropriate and useful way.

Example using iDAVE to query an unknown sample

- Two ways to upload your data for query
  - One variable at a time (slow)
  - Bulk paste (recommended)


iDave - Windows Internet Explorer provided by LLNL
https://idave.llnl.gov/
Google
File Edit View Favorites Tools Help
iDave
Page
Discriminant Analysis Verification Engine
Query Options
Demo
Query Unknown
Query Results Summary
Prediction Summary
Declared Source:
Predicted Source
# of Iterations
Model statistics
Q residual
(Compare to 95% Limit)
Hotelling T^2
(Compare to 95% Limit)
Unknown ID:
Number of available parameters
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 1 Results
Probability of Match
Source
Probability


https://idave.llnl.gov/
Google
Edit View Favorites Tools Help
iDave
Page Tools
Discriminant Analysis Verification Engine
Query Options
back
micro_g/g select value
use the +/- buttons to add data
Param
Unit
Bulk Paste
234U 50
238U 850000
Fe 100
Ag .05
Ba 0.5
Cr 300
•Example data are fake (hence bad result)
Copy/paste a minimum of 6 rows of values from Excel
(or text document with tabs between columns)
as: parameter result
load example
Cancel
OK
Query Results Summary
Prediction Summary
Uzbek 1

iDave - Windows Internet Explorer provided by LLNL

https://idave.llnl.gov/

File Edit View Favorites Tools Help

iDave

Discriminant Analysis Verification Engine

Query Results Summary

Prediction Summary

Declared Source: my unknown

Predicted Source: Uzbek 1

# of Iterations: 4

Model statistics

Q residual: 6.4263e-029

(Compare to 95% Limit): 7.0506e-031

Hotelling T^2: 1495.8345

(Compare to 95% Limit): 13.2769

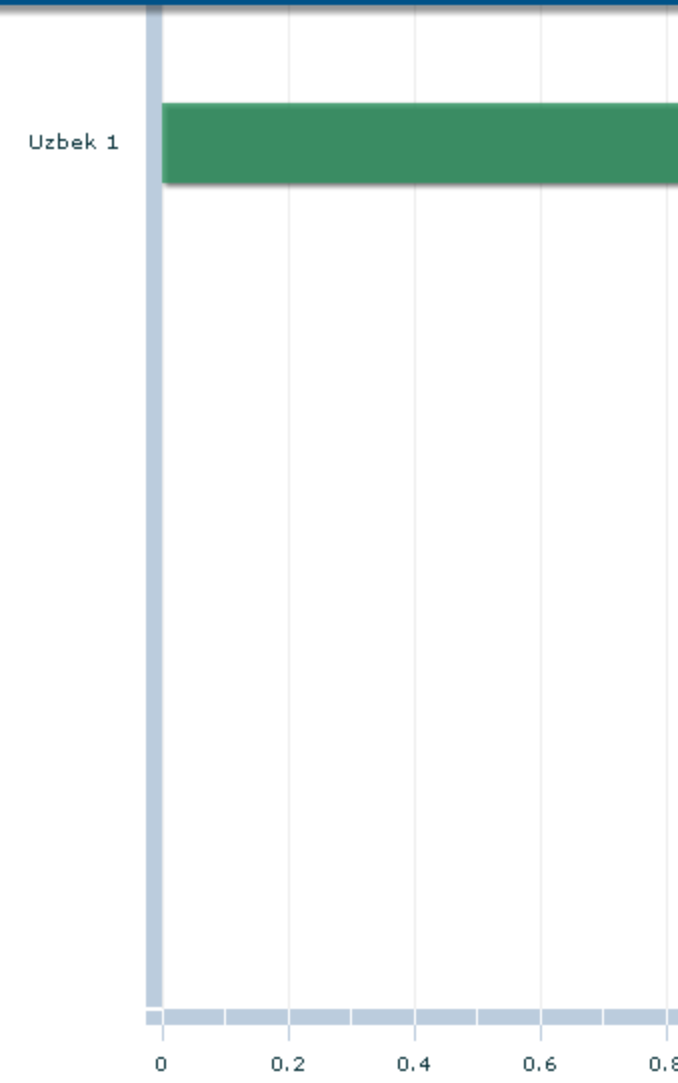
Unknown ID: My Unknown

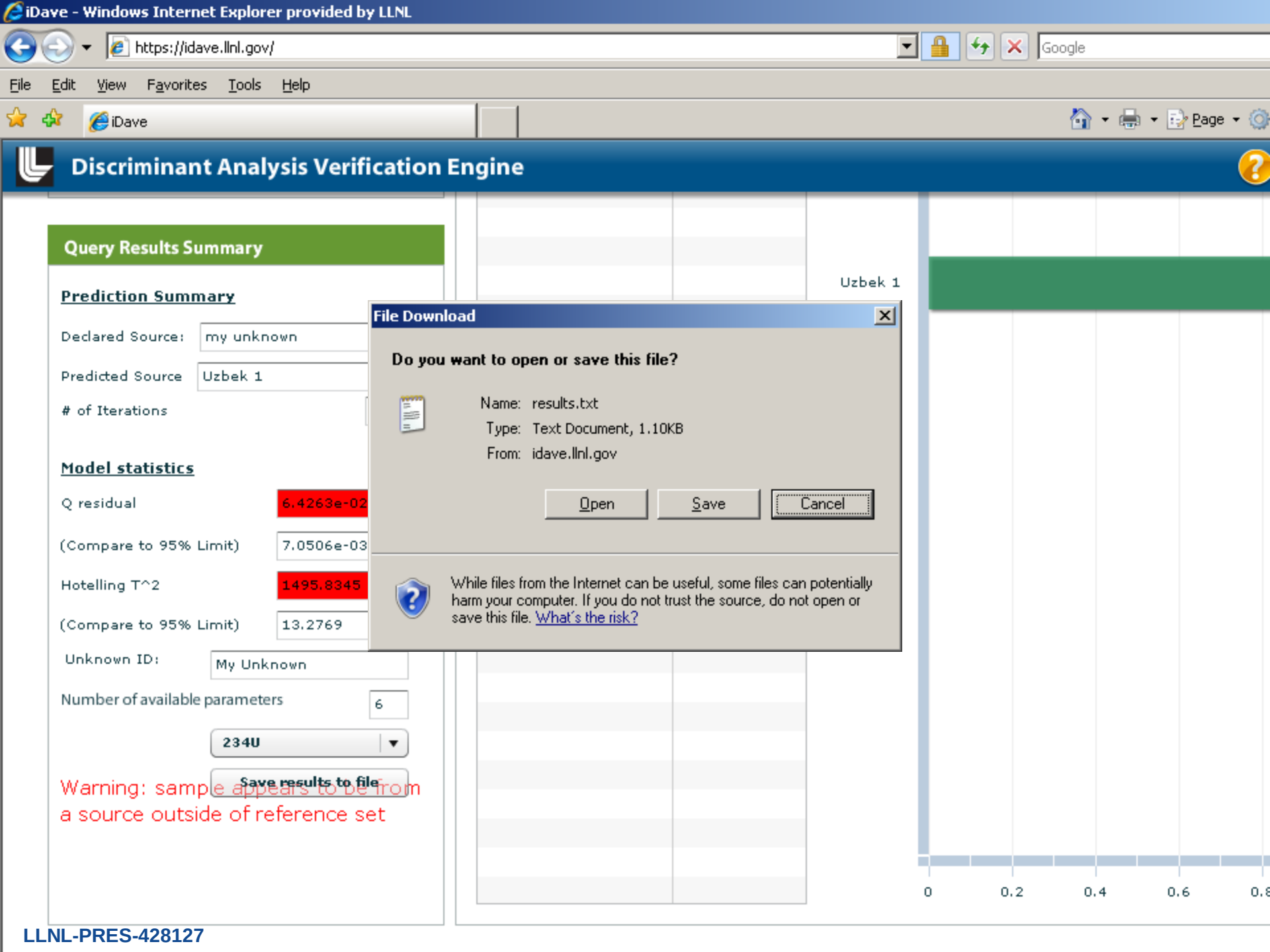
Number of available parameters: 6

234U

Save results to file

Warning: sample appears to be from a source outside of reference set




iDave - Windows Internet Explorer provided by LLNL
https://idave.llnl.gov/
Google
File Edit View Favorites Tools Help
iDave
Page
Discriminant Analysis Verification Engine
Query Results Summary
Prediction Summary
Declared Source: my unknown
Predicted Source Uzbek 1
# of Iterations
Model statistics
Q residual 6.4263e-02
(Compare to 95% Limit) 7.0506e-03
Hotelling T^2 1495.8345
(Compare to 95% Limit) 13.2769
Unknown ID: My Unknown
Number of available parameters 6
234U
Save results to file
Warning: sample appears to be from a source outside of reference set
Uzbek 1
File Download
Do you want to open or save this file?
Name: results.txt
Type: Text Document, 1.10KB
From: idave.llnl.gov
Open
Save
Cancel
While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. What's the risk?
0
0.2
0.4
0.6

Results saved in .txt format

results.txt - Notepad

File Edit Format View Help

```
Unknown ID:My Unknown
Number of available parameters:6

Available Parameters:
234U
238U
Ag
Ba
Cr
Fe

Prediction Summary:
Declared Source:my unknown
Predicted Source:Uzbek 1
Iterations:4

Descriptive Statisticss:
Q Residual for unknown relative to predicted source distribution:6.4263e-029
Q Residual at 95% confidence limit for predicted source distribution:7.0506e-031
Hotelling T^2 for unknown relative to predicted source distribution:1495.8345
Hotelling T^2 at 95% confidence limit for predicted source distribution :13.2769

Results at Each Iteration (Source: Probability):
Iteration1
USA (Area 1):1
Canada (Area 1):0.79786
Canada (Area 2):0.50566
Canada (Area 3):0.26531
Canada (Area 4):0.1308
Australia (Area 1):0.91928
Australia (Area 2):0.44335
Australia (Area 3):0.15767
Kazakstan (Area 1):0.70537
Uzbek 1:0.17859
Namibia (Area 1):0.025798

Iteration2
USA (Area 1):1
Canada (Area 2):1
Australia (Area 1):0.7633
Australia (Area 3):1
Kazakstan (Area 1):1
```

Interpreting Results

- Descriptive statistics: Q residual and Hotelling's T^2
  - Use these relative to 95% values. Not meaningful without context.
- Q residual
  - Measure of variation outside (i.e. not represented by) the model
- Hotelling's T^2
  - Measure of variation within the modeled space (how unusual sample is in the space).
- Number of iterations
  - More iterations means more initial overlap of possible sources.

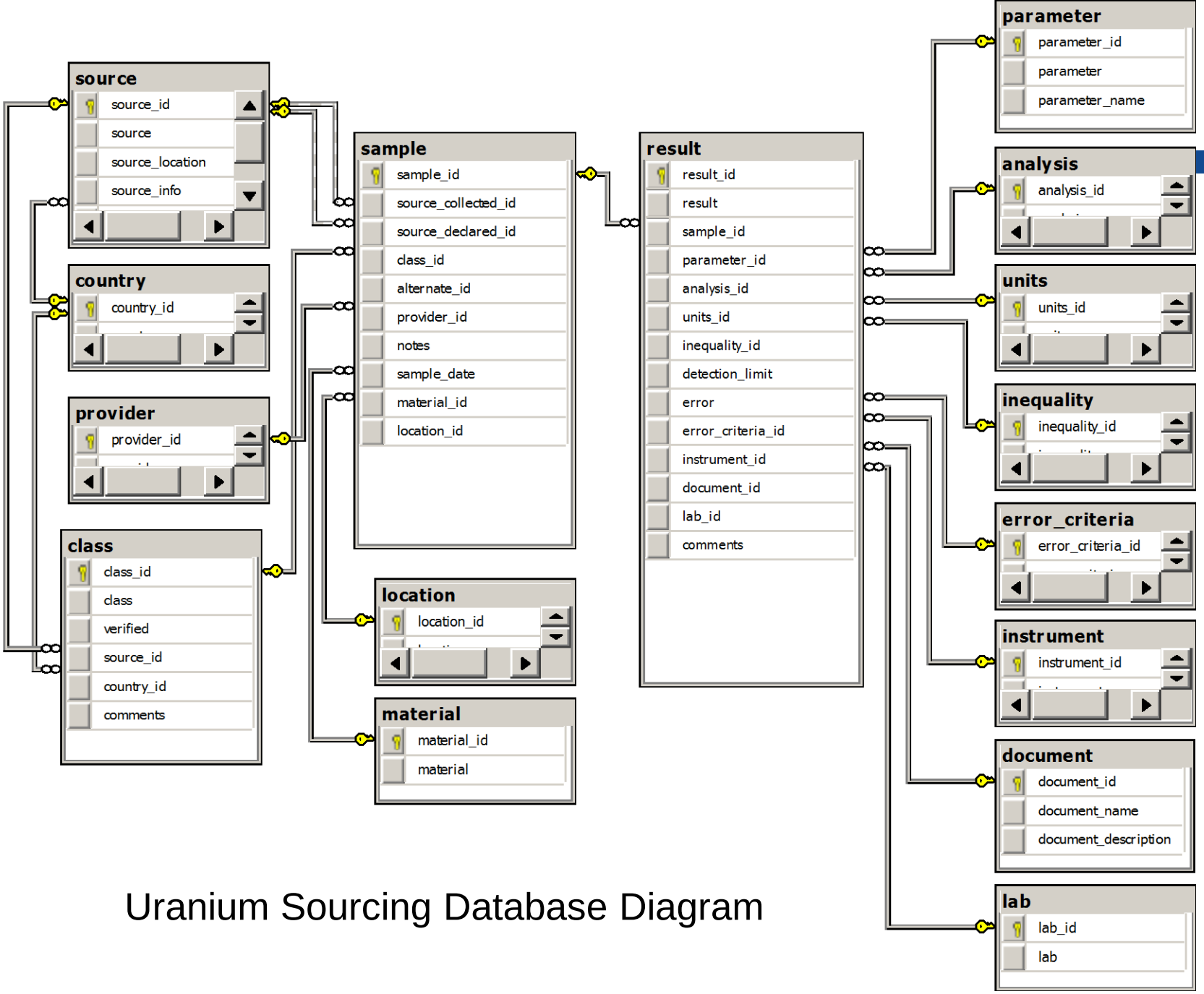
Interpreting Results

- iDAVE tests to see which reference source the unknown is most similar to.
- iDAVE will always make a prediction based on the assumption that the unknown is UOC from one of the reference sources.
- The user must rely on the descriptive statistics and other, uncorrelated information to assess the potential that the unknown is from some outside source.

Database statistics

U Sourcing Database

Samples:	1907
Sources:	111
Parameters currently measured:	65
(includes trace elements, isotopes, U-compound)	
Number of distinct results (measurements) :	62,041



Uranium Sourcing Database Diagram

iDAVE made possible by

Vision

- Ian Hutcheon
- Mike Kristo

Technical development/support

- Martin Robel: MATLAB and database development
- Justin Shinn: Website development
- Greg White: Network and hardware
- And many others

Programmatic support

- NNSA NA-243, Office of Nuclear Verification
- This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.