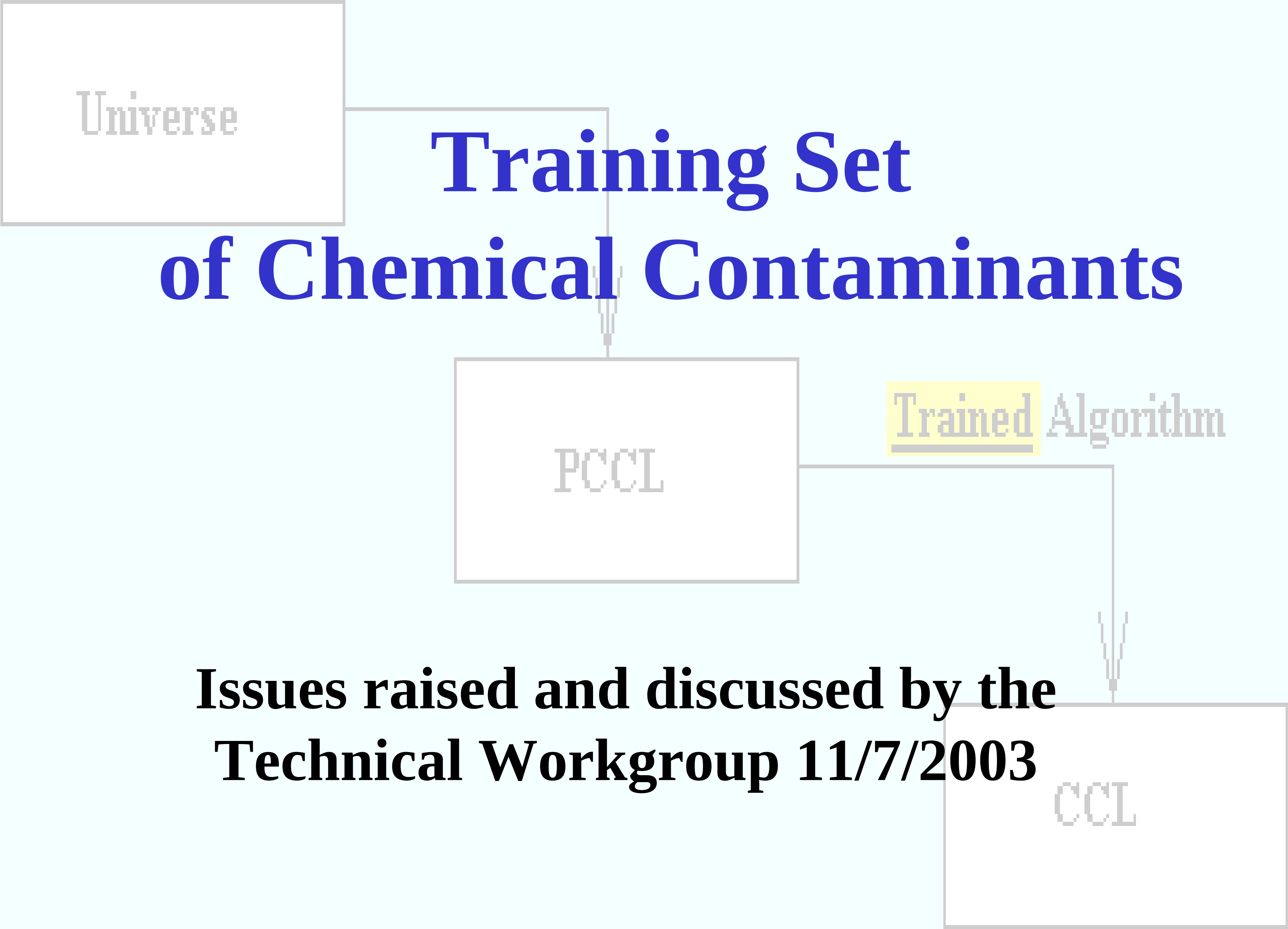
Training Set of Chemical Contaminants



Issues raised and discussed by the Technical Workgroup 11/7/2003

Where were we?

- 2001 NRC Report – recommended automated process, prototype classification, training contaminants with “obvious” decisions
- January 2003 – *Matrix of Decision Method Characteristics* discussed different approaches, their characteristics, and how they compared w.r.t. transparency, the role of expert judgment, flexibility of updating, etc.
- Spring 2003 - An exercise that used a small training set showed that a number of prototype classification algorithms (CART, logistic regression, neural net, and MARS) could work.

Where are we now?

- Considerable development in areas that could support any classification approach:
  - Defining the universe and obtaining data/info
  - Universe → PCCL screening
  - Attribute scoring protocols
- The technical group found training set development is much more complex and important than expected.
  - For example, we believe the training set should include some difficult (gray area) decisions, whereas NRC recommended using only contaminants with obvious decisions.
  - Could continue with a limited diagnostic exercise if this is seen as an effective use of Technical Workgroup resources.

Issues for NDWAC:

1. In light of new concerns, consider if prototype classification is the right way to go.
2. Provide some guiding principles for attribute scoring and decision-making
3. Consider the added value of moving ahead with a limited "diagnostic" training exercise (to be completed by the January meeting) rather than focusing on issues 1 and 2 above.

Transparency issues with the training set-based classification approach

- Involves difficult judgments about which training contaminants are listed and which are not.
- Assumes "correct" decisions can be identified
  - Perspectives on health, resources, and Agency mission vary over time, so today's "correct" decision may be incorrect tomorrow.
  - Convening a team to meet and agree about these decisions is challenging.
- The ultimate rule derived by the algorithm may be transparent, but the decisions associated with training contaminants are what led to the rule. Those decisions and their rationale aren't transparent.

What has the technical team learned about these considerations?

- While historical decisions may have been correct at the time they were made, they don’t all appear to be decisions we would make today.
- The training set must include contaminants for which decisions are difficult (“gray” area contaminants). The importance of “gray” area as well as clearly list & don’t list contaminants can be revealed by a “diagnostic exercise” (to be discussed later).
- Principles are needed to unify the scoring protocols for chemicals and microbial contaminants if they are to be treated together.

What has the technical team learned? – cont’d

- Running the algorithms and interpreting the results will not take great effort.
- The greatest effort will be developing the training set.
  - Gathering all supporting data / information about the contaminants
  - Attribute scoring
  - Assigning correct decisions to the training contaminants
- Principles are needed for attribute scoring, selecting a training set, and identifying correct decisions for the training contaminants.
- The decision maker’s value judgments (used in assigning decisions to training contaminants) are not transparent.

What are the technical team's recommendations for the NDWAC Workgroup? (3 questions / issues)

1. Consider alternatives to the prototype classification approach, including:
   - Rule-based system (experts construct / encode a rule for classifying or ranking contaminants, such as was specified for Universe → PCCL)
   - Multi-attribute utility (an expert-selected function that translates attribute scores to a single measure of strength. Can be used to rank contaminants.)
   - Others

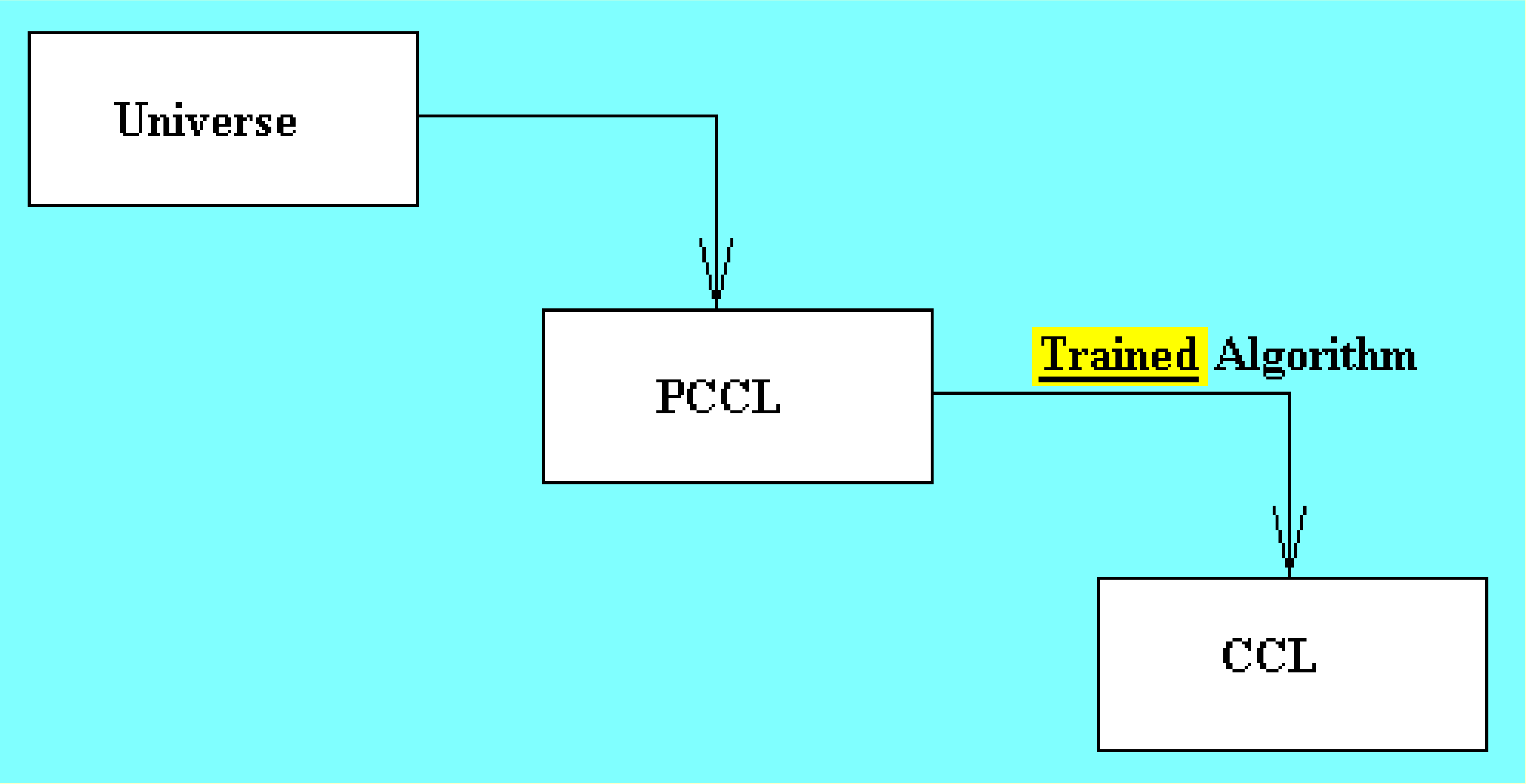
What are the technical team's recommendations for the NDWAC Workgroup? (3 questions / issues)

2. Whatever the recommended approach, provide guiding principles for:
   - attribute scoring,
   - decision making (deciding what gets listed)
   - integrating microbial / chemical approaches
3. Provide guidance about how and whether to proceed with further training set development. (Next several slides.)

Training Set Issues (remaining slides)

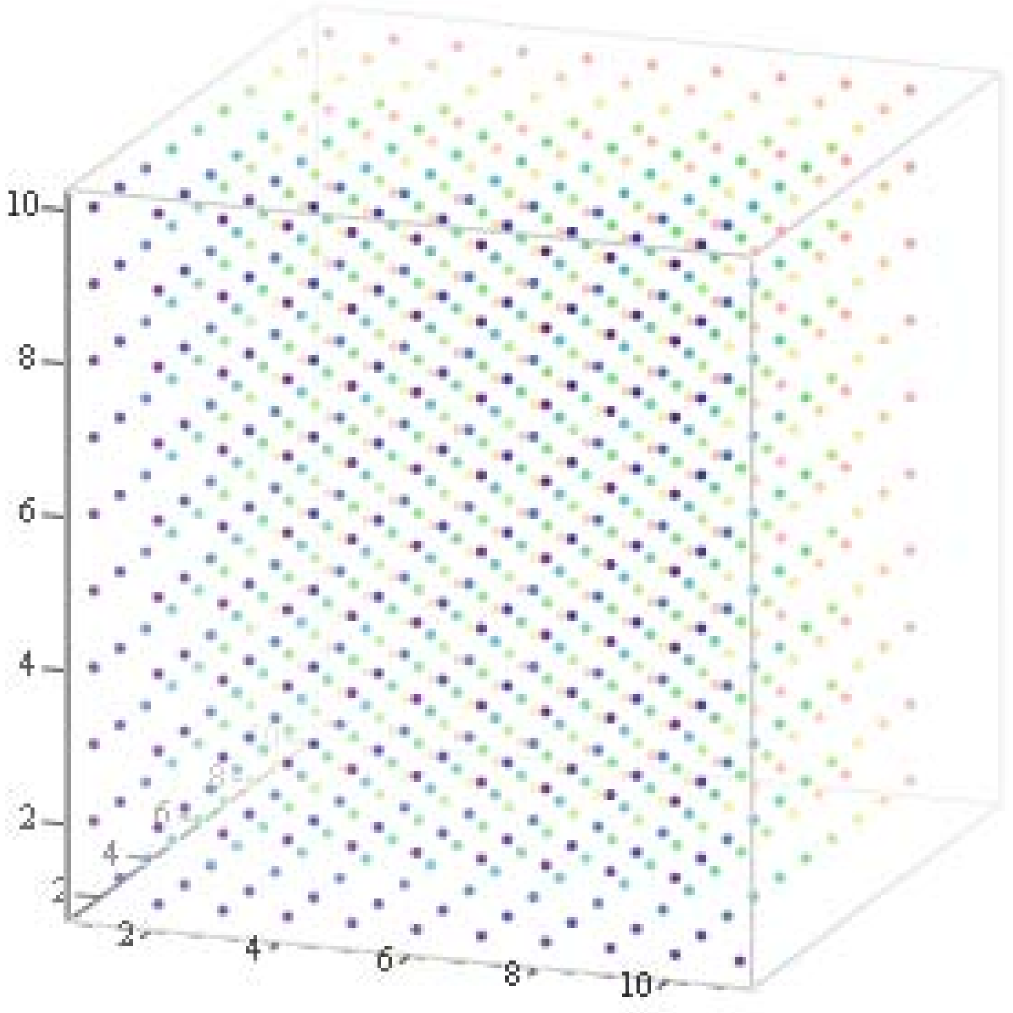
- Some terminology / graphical perspective
- What is a training set? A good training set?
- How big is a good training set?
- What is the technical team's current process for developing a training set?
- What can be learned in a diagnostic exercise (by the January meeting)?
- What could alternatively be done with the same resources?

The algorithm used to decide which PCCL contaminants move to the CCL must be trained. This is where training data come in.



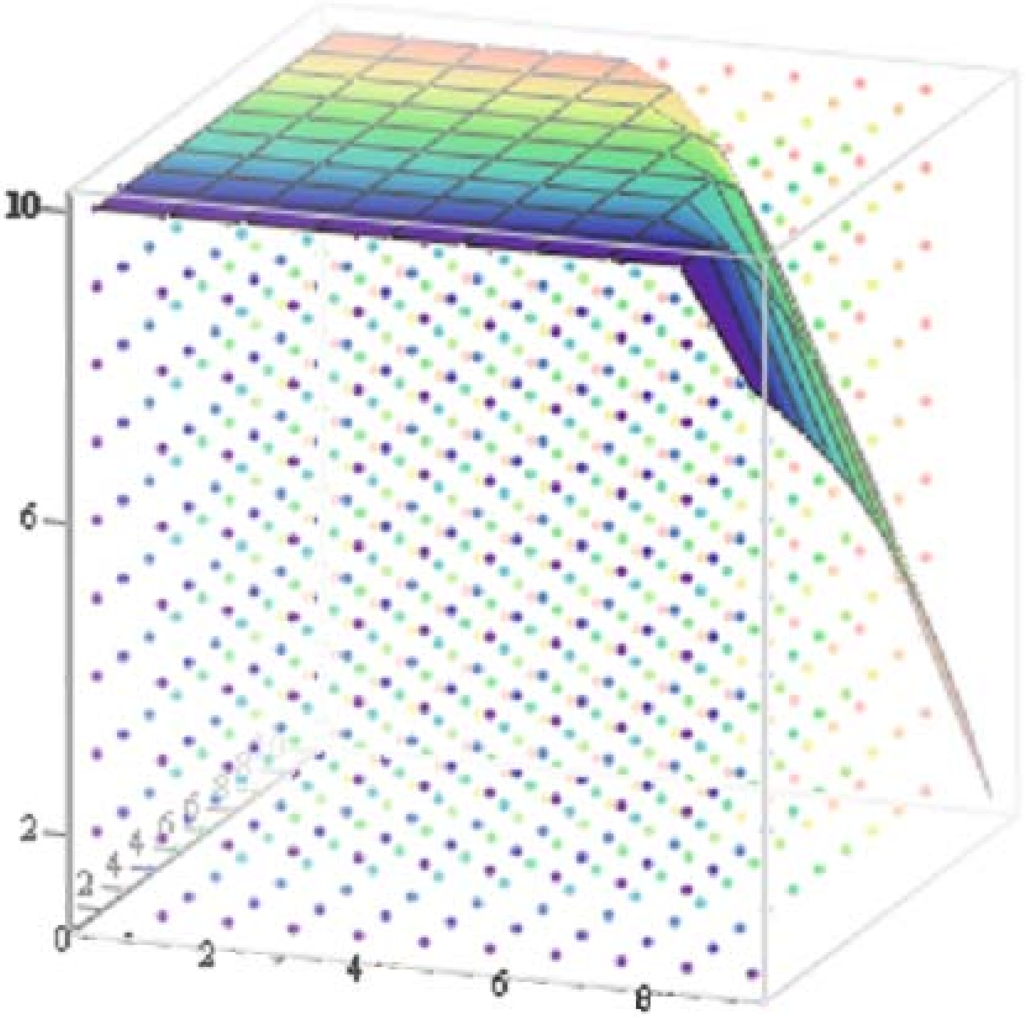
Some Terminology and Graphics

1. **Attribute Space**
   - Set of all possible combinations of attribute scores.
   - 5 dimensional (as we have 5 attributes).
   - Hard to display, so we'll limit to 3 for now.



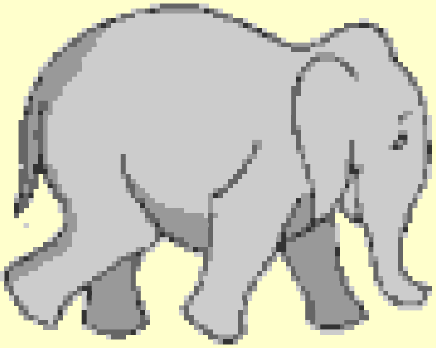
Terminology and Graphics, cont’d

2. **Discriminant Function**
   - Divides “List” from “Don’t List” contaminants.
   - We don’t know precisely where it is.
   - The classification algorithm will search for a good solution (one that minimizes decision errors)



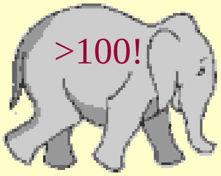
What is a training set?

- **Contaminants** (real or contrived), accompanied by **data/information** on occurrence and health effects.
- Complete sets of **scored attributes** for all contaminants.
- **Decisions** = Designation of “list” or “don’t list” for each contaminant.



What is a **GOOD** training set?

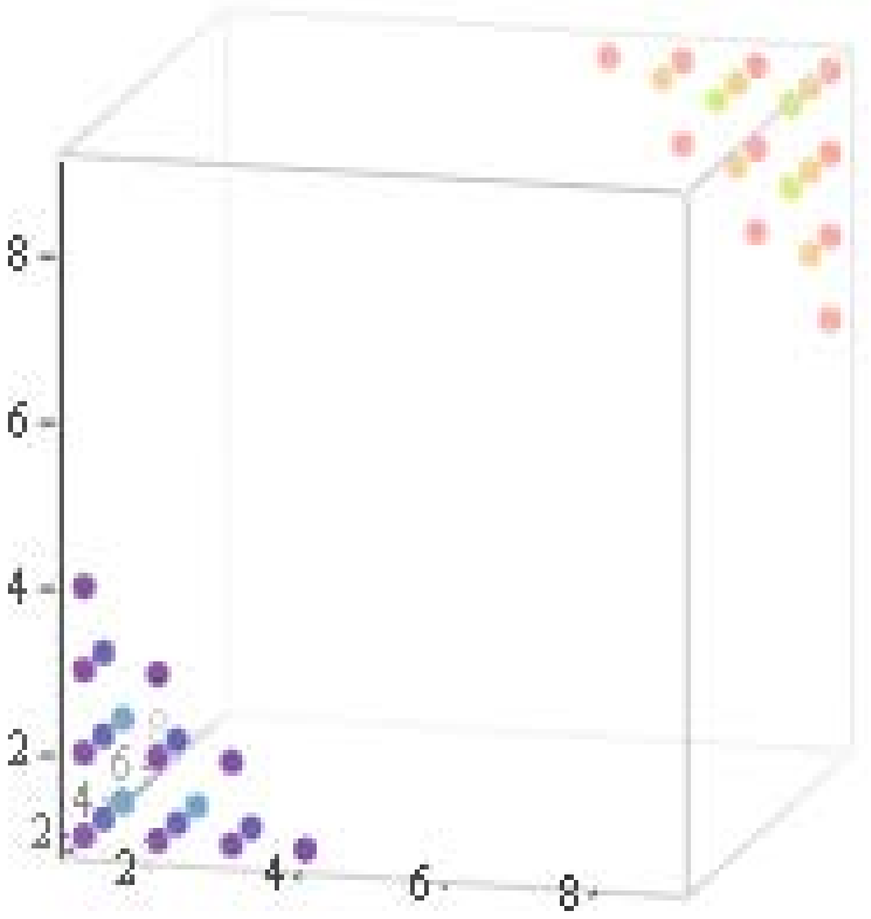
- **Its decisions are "correct:"**
  - Based on consideration of current values (perspectives on health, resources, and Agency mission)
  - Made by group of experts acting as though their decisions were final (and recognizing that further analyses will be made through research and regulatory determinations)
  - Based on consensus or majority opinion after thoughtful consideration of available info/data for the contaminant



- **It covers the full range of PCCL contaminants in terms of attributes and scores.** (Need to train the algorithm to make good decisions "everywhere," not only where it is easy for experts.)
- **It is sufficiently large for training and validation.**
  - Number should probably be greater than 100.
  - Smaller sets would not provide good coverage and may lead to unacceptable misclassification rates (false positive & false negative)

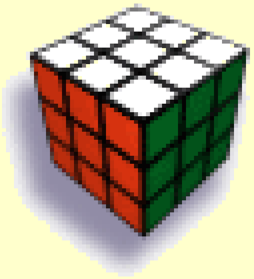
“No Brainers” = Poor Coverage

(algorithm could put discriminant just about anywhere)



Why at least 100?

- *Craig Stow and Song Qian* suggest 100 or more may be needed (based on experience with the algorithms and earlier practice runs).
- *Jeff Rosen* stated a “rule of thumb” for discriminant analysis is to have at least 30 times the number of attributes (at least 150).
- *Mike Messner* noted that a 3*3*3*3*3 Latin hypercube has 243 cells (known as cubicles to EPA folks).



- **A diagnostic exercise was proposed to help us better understand how performance (error rates) is related to training set size. We’d also learn other things from the exercise.**

Recommended Process for Developing a Good Training Set

1. Establish principles for scoring, training, and decision making
2. Determine draft attribute scoring protocols (should be close to final protocols)
3. **Build training set by iterative process (diagnostic exercise)**
4. Develop formal Data Quality Objectives
5. Reduce number of algorithms if appropriate
6. Finalize the attribute scoring protocols.
7. Develop the “real” training set.
8. Train the algorithm.

Step 3 – Build the training set by iterative process

The diagnostic exercise:

- a pilot run of training data
- requires development of a draft training set (complete with attribute scores, supporting data/information, and decisions)
- similar to the "toy" exercise conducted spring 2003, but with more-refined scoring protocols, decision-making (not just historical decisions), and training set

Step 3 – Build the training set by iterative process

a) Conduct **diagnostic exercise** to learn and obtain better information regarding:

- How to make decisions for training set contaminants
- How to characterize error rates (and decide what can be tolerated).
- Learn how to interpret algorithm output for decision making
- Relate error rates to training set size, distribution in attribute space, proximity to discriminant surface (in gray area), and other features.
- Understand how the scoring rules can influence performance (sensitivity).
- Determine which attributes are most influential (could be misleading if protocols are far from final form).

b) Conduct additional diagnostics if necessary

Concerns with running the Diagnostic Exercise now

- Scoring for chemicals and microbials should be based on common principles. To date, no such principles have been developed.
- Attribute scoring protocols haven’t been finalized.
  - What we learn may be sensitive to changes in the protocols.
  - Resources spent in conducting the exercise could be better spent finalizing the protocols.
- The means of expressing and estimating decision errors (algorithm performance) need to be developed.
- The tolerances for decision errors need to be systematically developed (consider following EPA’s DQO process).
- The time and energy devoted to building the exercise’s training set could perhaps be better spent on other efforts such as developing principles for scoring and decision-making.

So where are we now?

- We could start the diagnostic exercise now – if advised that it is a good idea – and have results by next NDWAC Workgroup meeting.
- Much to learn about how many contaminants will ultimately be needed, how they should be selected, and how correct decisions will be made.
- Concerned about potential pitfalls of starting now and diverting resources away from other needed work.

- **Look again at the key question and associated pros and cons (of running the diagnostic exercise now versus later).**

Should we move ahead with a diagnostic training exercise now (or pause to complete some other work first)?

- **What can we learn from the exercise now?**
  - Opportunity to learn about the competing algorithms
    - Interpreting algorithm output
    - Algorithm performance
  - Opportunity to learn about processes that inform the algorithm, including:
    - Performance and training set (size and dispersion) requirements
    - Attribute scoring
    - Decision making
  - *NOTE 1: Results may be misleading (sensitive to changes in attribute protocols)*
  - *NOTE 2: Running the algorithms and interpreting results will be relatively easy once training set is in place. Most resource-intensive will be generating the training set and decisions.*

Should we move ahead with a diagnostic training exercise now (or pause to complete some other work first)?

- **What could alternatively be done now (through next meeting)?**
  - Develop principles for scoring and decision making
  - Refine the scoring and decision making protocols
  - Discuss what information and analyses are needed to inform NDWAC's recommendation on the approach.
  - Begin to develop data quality objectives (e.g., define the decision and consider relative value of avoiding false negative versus false positive decisions)

What are the technical team's recommendations for the NDWAC Workgroup?

1. Consider alternatives for PCCL → CCL
   - Prototype classification approach, including:
   - Rule-based system (experts construct / encode a rule for classifying or ranking contaminants, such as was specified for Universe → PCCL)
   - Multi-attribute utility (an expert-selected function that translates attribute scores to a single measure of strength. Can be used to rank contaminants.)
   - Others?

What are the technical team's recommendations for the NDWAC Workgroup?

2. Whatever the recommended approach, provide guiding principles for:
   - expert intervention and judgment (the role of experts in PCCL → CCL processes)
   - attribute scoring
   - transparency
   - decision making (e.g., deciding list / not list for individual training contaminants)
   - integrating microbial / chemical approaches
3. Provide guidance about how and whether to proceed with further training set development.