

Building M2K MIR/MDL Applications in the D2K/T2K Framework

IMIRSEL provides an environment for implementing MIR algorithms based on NCSA's Data To Knowledge (D2K) framework for machine learning. IMIRSEL has augmented D2K with MIR-specific M2K modules for feature extraction and classification and so on. Researchers can augment M2K by adapting their codes to D2K/M2K's simple plug-in API.

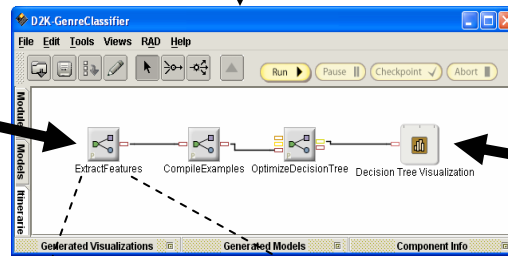
Developing M2K applications in D2K entails assembling processing modules into an *itinerary* characterizing the data flow between modules. Itineraries can then be run as stand-alone applications on clusters of machines.

Rapid Prototyping

Design MIR applications in an interactive, graphical environment

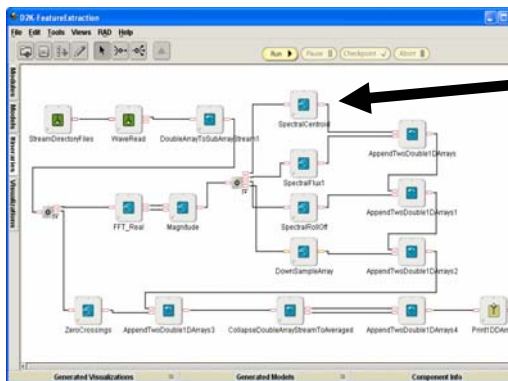
D2K provides an integrated development environment (IDE) including tools for browsing and configuring modules, testing and debugging itineraries, and viewing generated visualizations.

Modules
Link modules together to implement MIR algorithms



Visualization
View results using visualization modules integrated into itinerary

Nesting
Modules can encapsulate complex sub-applications










Parallelism
D2K manages execution of modules on multiple CPU's

Once an itinerary has been developed, it can be used as a module in any other itinerary, allowing for applications of arbitrary complexity.

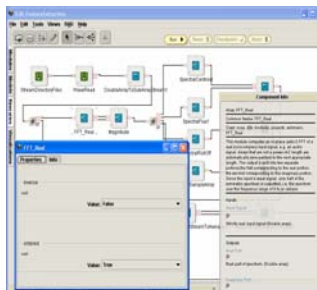
D2K can run any module in parallel with any other module using any number of CPU's. Whether or not to split the execution of a module across CPU's is configurable on a module-by-module basis.

D2K Module Types

-  **Input**
Fetch data from a data source
-  **Output**
Produce output to be exported/stored
-  **Data preparation**
Prepare data for processing
-  **Compute**
Extract features, build models
-  **Parallel compute**
Perform computation using multiple CPU's
-  **User interface**
Get user feedback
-  **Visualization**
Produce graphics showing results

Cross-platform

D2K is written in Java, so it runs on all major platforms. For incorporating legacy applications, it can be interfaced to non-Java implementations such as C/C++ and Matlab. D2K has been successfully used to perform massively parallel data analysis on multi-Gigabyte datasets.



Customization

D2K provides graphical configuration and inline documentation tools for managing complex application development processes and sharing code.