THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Big Data with ADAMS

What the heck is ADAMS?

Peter Reutemann

About myself

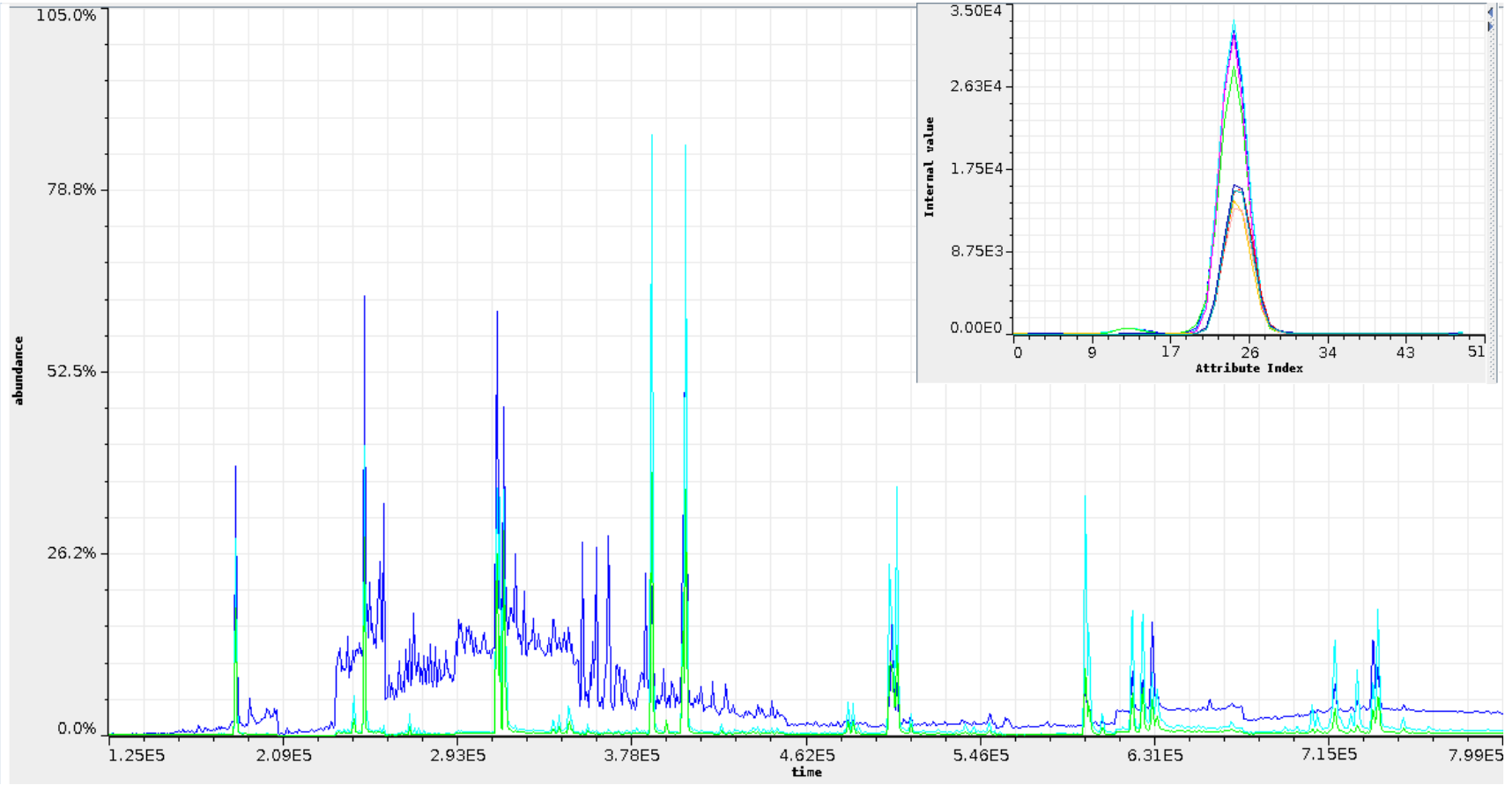
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- University of Waikato, NZ - since 2003
- Senior Research Programmer
- Lead developer of ADAMS
- Other projects
  - WEKA
  - WEKA MOOC
  - python-weka-wrapper
  - MEKA
- More info

  http://www.cms.waikato.ac.nz/~fracpete/

How things started

- PAH domain (poly-aromatic hydrocarbons)



GC-MS Challenges

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- Multi-dimensional
  (each GC point has 2-dim MS data attached)
- Retention time shifts (x-axis, linear and non-linear)
- Baseline shifts (y-axis)
- Noise (impurities in samples)
- Coelution (shoulder peaks)
- Varying sample rates (x-axis)
- Hardware problems (detector saturation)
- Multiple compounds to predict

Solution

- Extensive pre-processing
- Pre-processing/prediction in parallel for various compounds

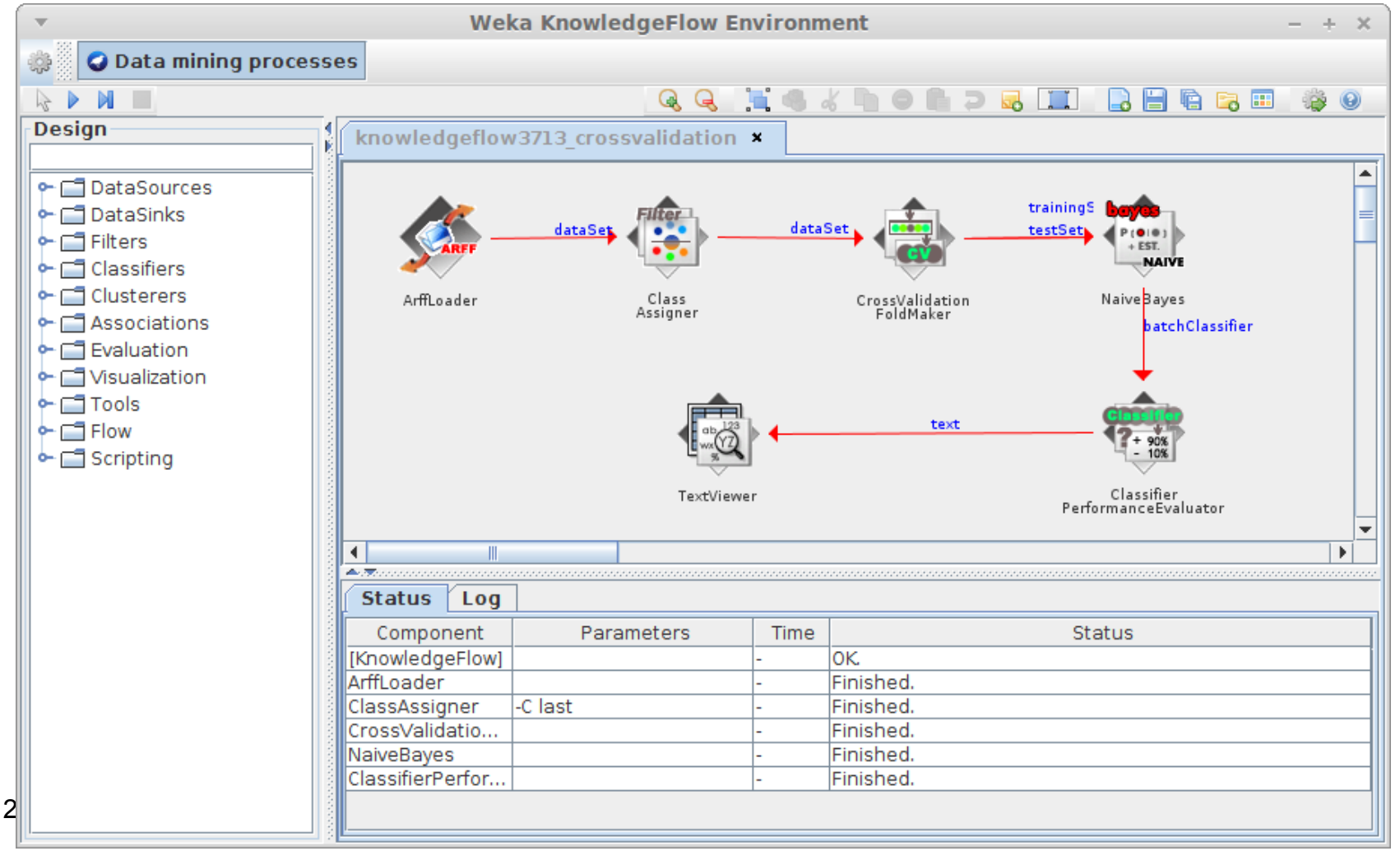
  → Workflow engine, but which one?

Workflow systems

- most workflow systems are “canvas-based”
  - user places operators on canvas
  - manually connects operators
- older system were DAG-based
  - eg using a tree-layout
- following examples
  - cross-validation of classifier on single dataset
  - output of results

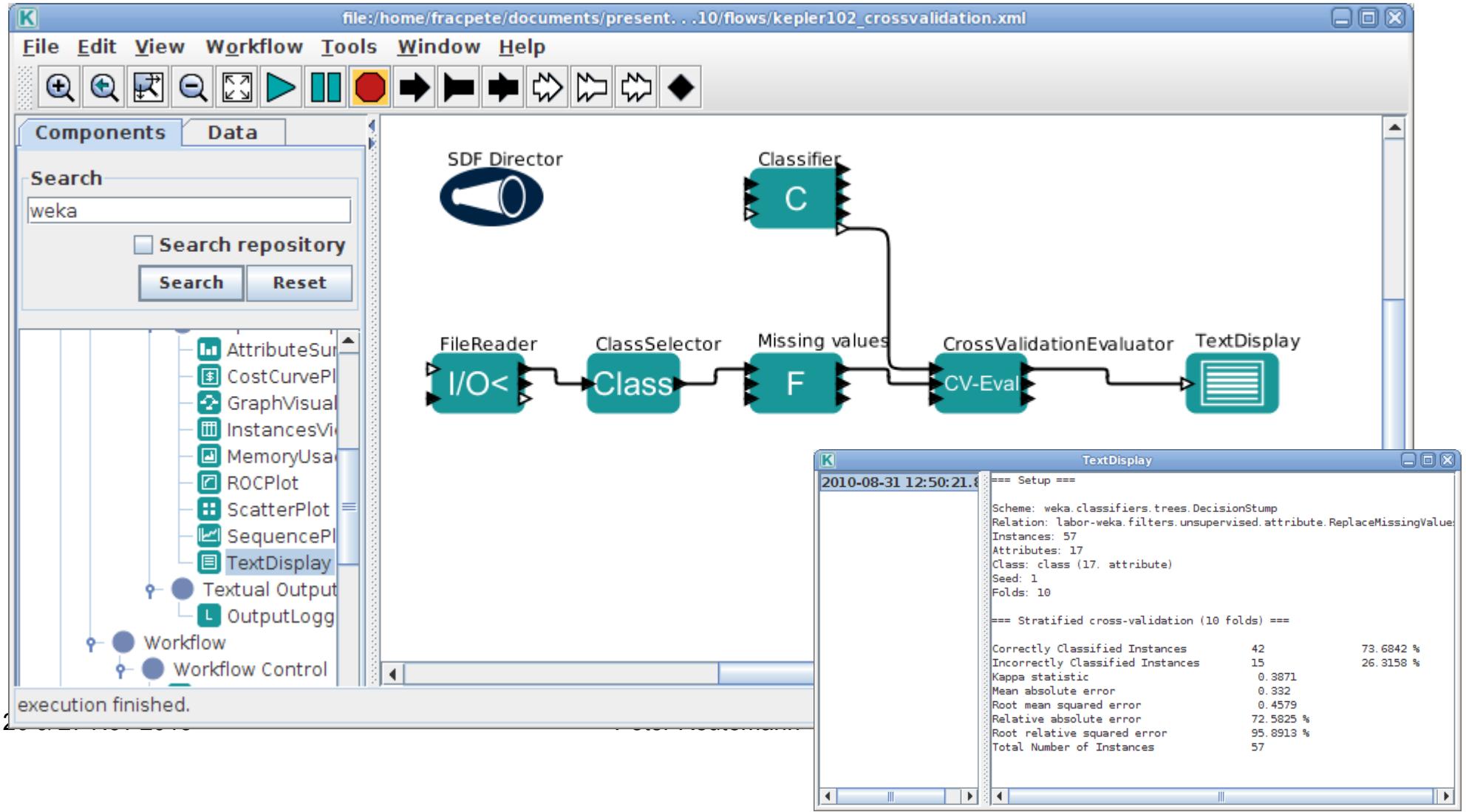
Workflow systems (2)

- WEKA KnowledgeFlow 3.7.13



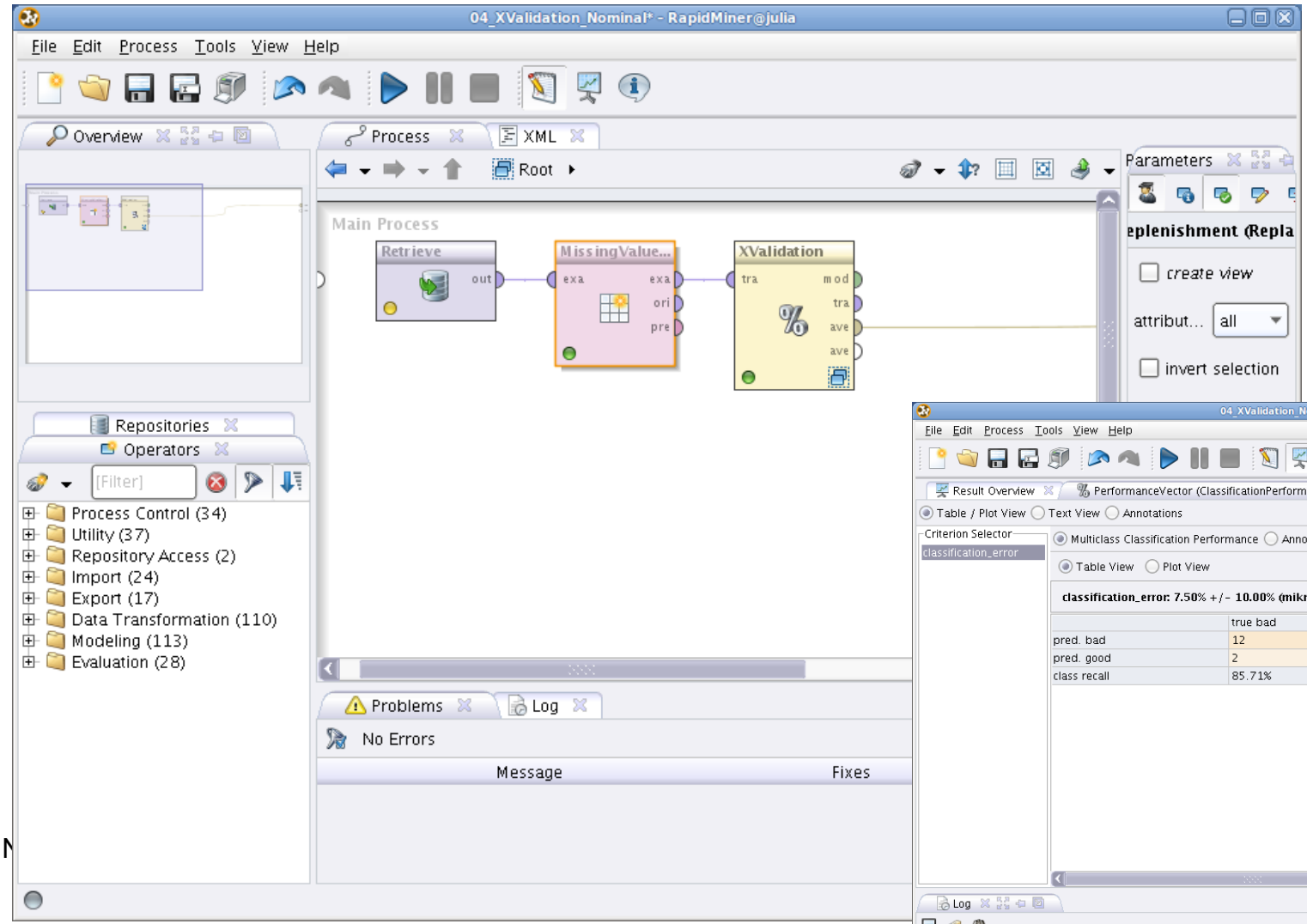
Workflow systems (3)

- Kepler 1.0.2



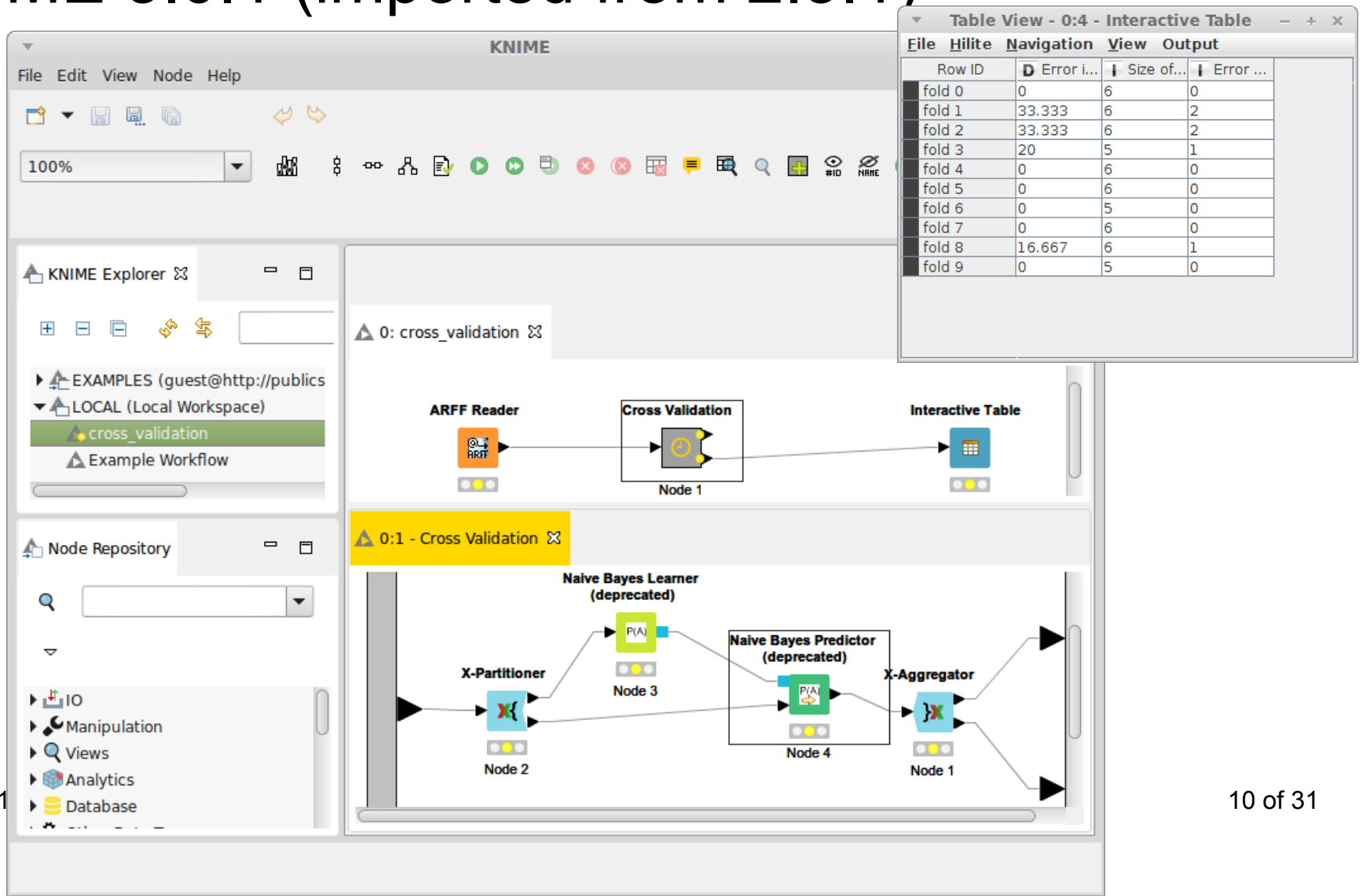
Workflow systems (4)

- RapidMiner 5.0 (“free” 6.5 is useless)



Workflow systems (5)

- KNIME 3.0.1 (imported from 2.3.1)



Why another workflow?

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- KnowledgeFlow too WEKA-centric
- Initial development with Kepler
- Canvas-based set up is tedious
- Only minimal subset of functionality necessary (e.g., no grid computing)
- Primary connection types: 1-to-1 and 1-to-n
  → prototype implementation

What is ADAMS?

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

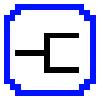
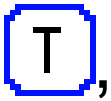
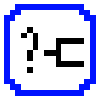
- Java, GPLv3
- Base modules

  access, core, compress, event, excel, gnuplot, groovy, imaging (+boofcv, imagej, imagemagick, openimaj), jython, latex, maps, meta, moa, net, odf, osm, pdf, r, random, spreadsheet, timeseries, twitter, visualstats, weka
- Add-ons modules

  heatmap, image-webservice, jooq, meka, rats, video, webservice, weka-webservice
- Incubator modules

  nlp, jclouds, openstack, openml, javacv, ...

Flow

- Operators are called “actors”
- Actors arranged in tree, no connections
- Actor “handlers” nest other actors
  - e.g., sequence of actors
- Control actors control data flow
  - e.g., branch  , tee  , if-then-else  , switch 
- Input/output defines
  - standalone  , source  , transformer  , sink 

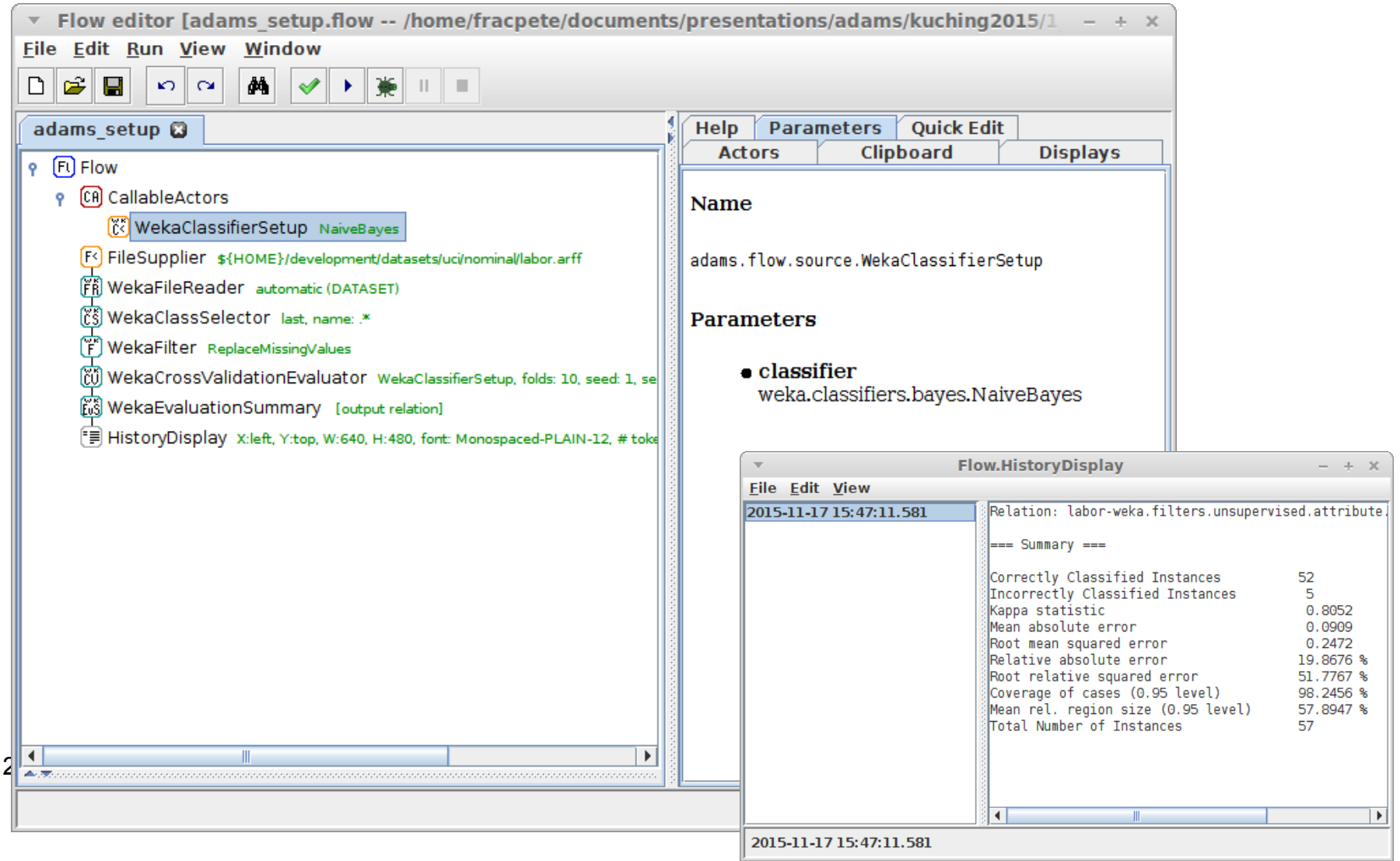
Flow (2)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- Data-driven system, but events possible
- Tree only supports 1-to-n connections
- Simulating n-to-m semantics
  - Containers
  - Variables
  - Internal storage
  - Callable actors

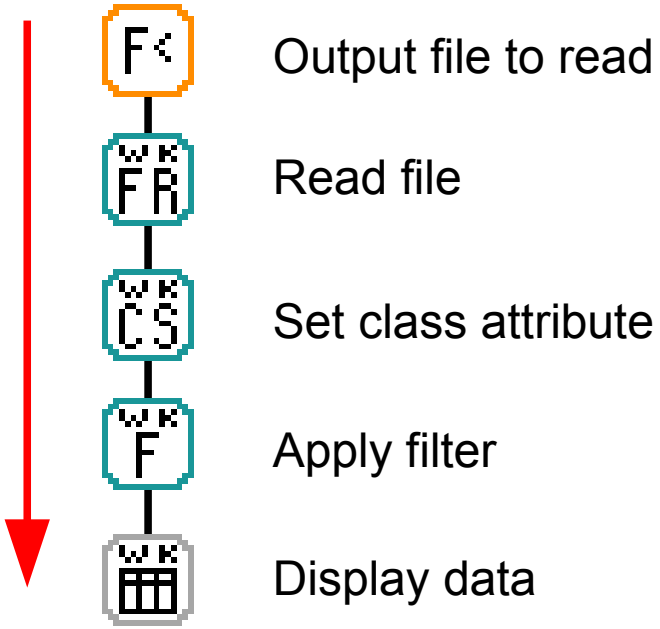
Previous example

- Same workflow in ADAMS



Examples

Execute nested actors one after the other



Output file to read

Read file

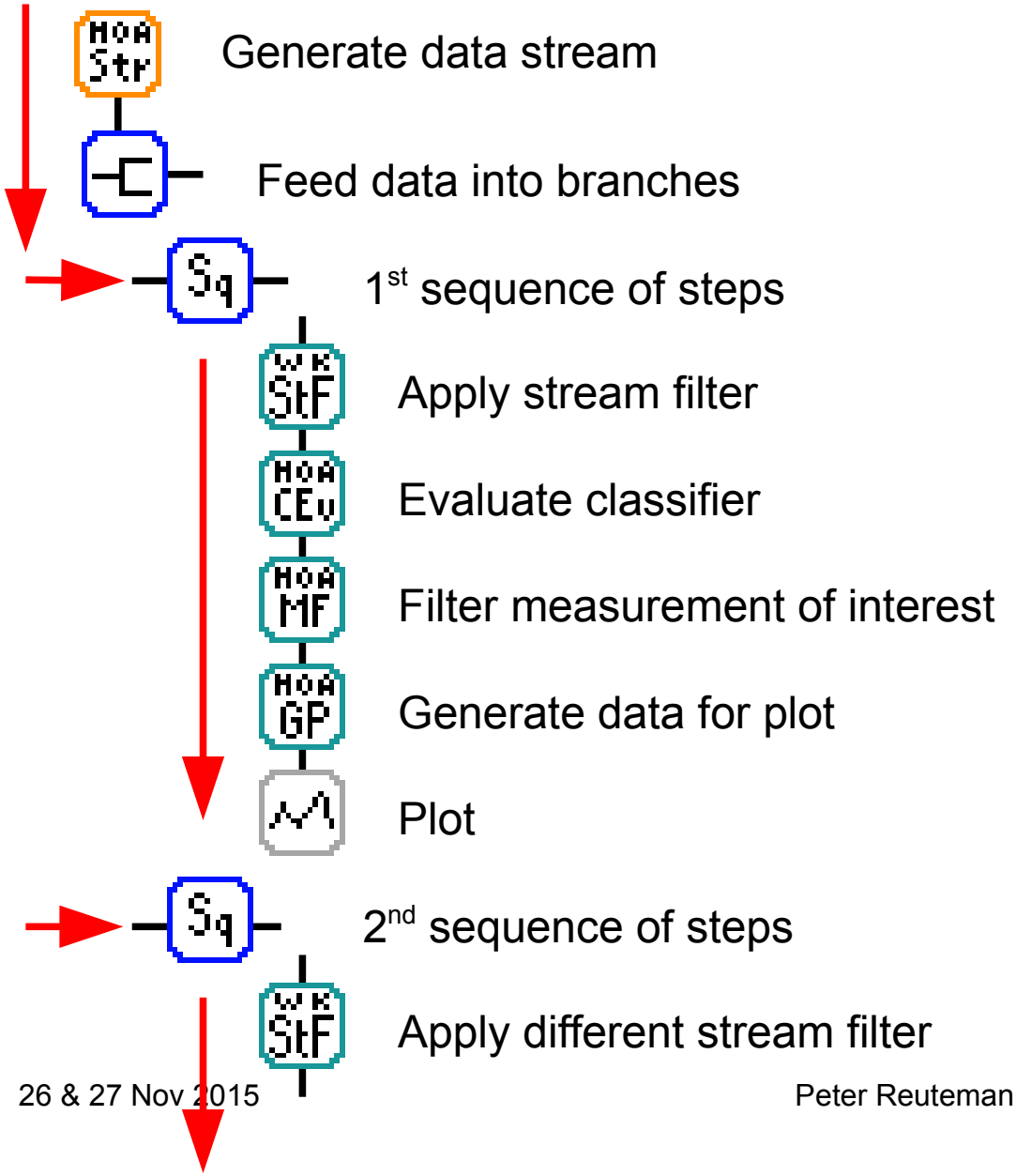
Set class attribute

Apply filter

Display data

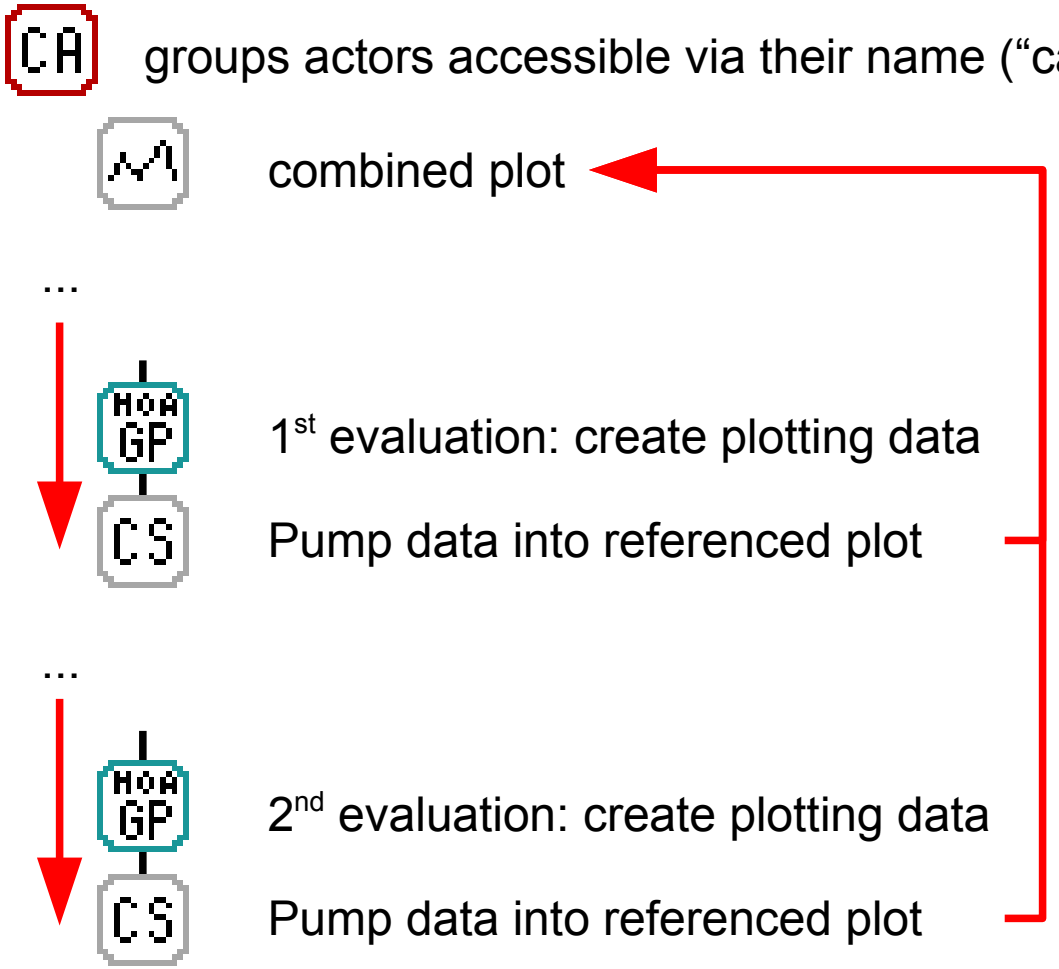
Load dataset, apply filter and display dataset

Examples (2)



Filter data stream in two separate branches with different filters, evaluate classifier and plot metric

Examples (3)



Generate combined plot of two evaluations by using “callable actors” functionality

Tensorflow

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

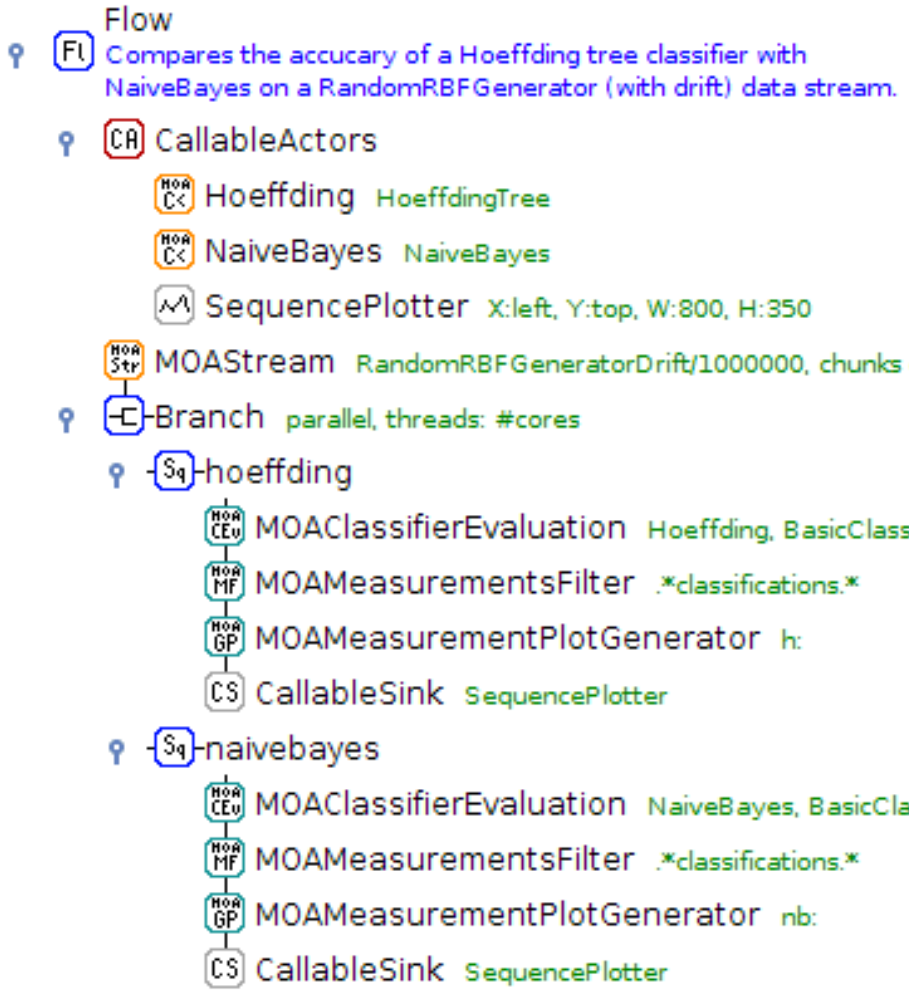
Feature	ADAMS	Tensorflow
Bindings	Java	Python and C++
GUI	✔	
Scripting	Groovy/Jython	Python
GPU		✔
Deep learning		✔
Non-deep learning	✔	
Multi-processor	✔	✔
Parallel execution	✔	✔
Remote execution	✔	
Data flow programming	kind of	✔
Platforms	Linux, Mac, Windows	Linux, Mac
License	GPLv3	Apache 2.0

TensorFlow Disappoints – Google Deep Learning falls shallow [source: KDNuggets]

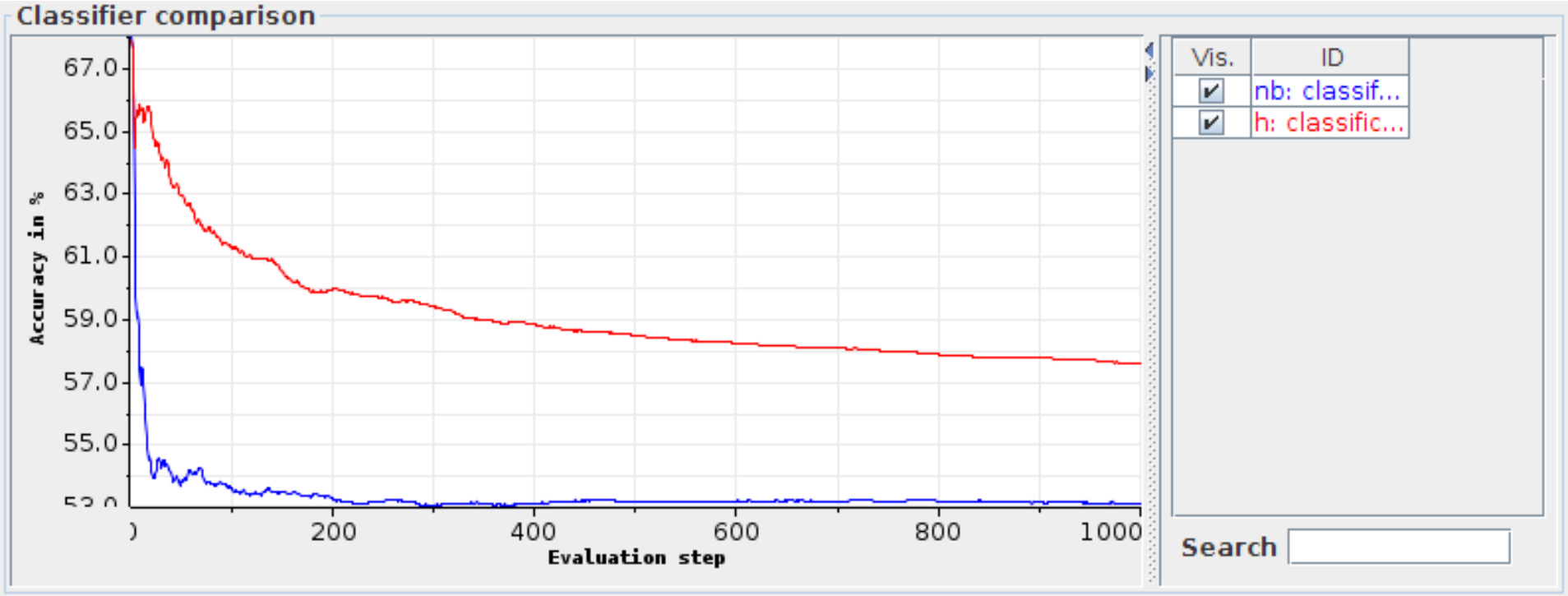
Research (demos)

- Compare two MOA classifiers (drift)
- Compare MOA classifier on different streams
- MOA cluster visualization
- Track mouse in video

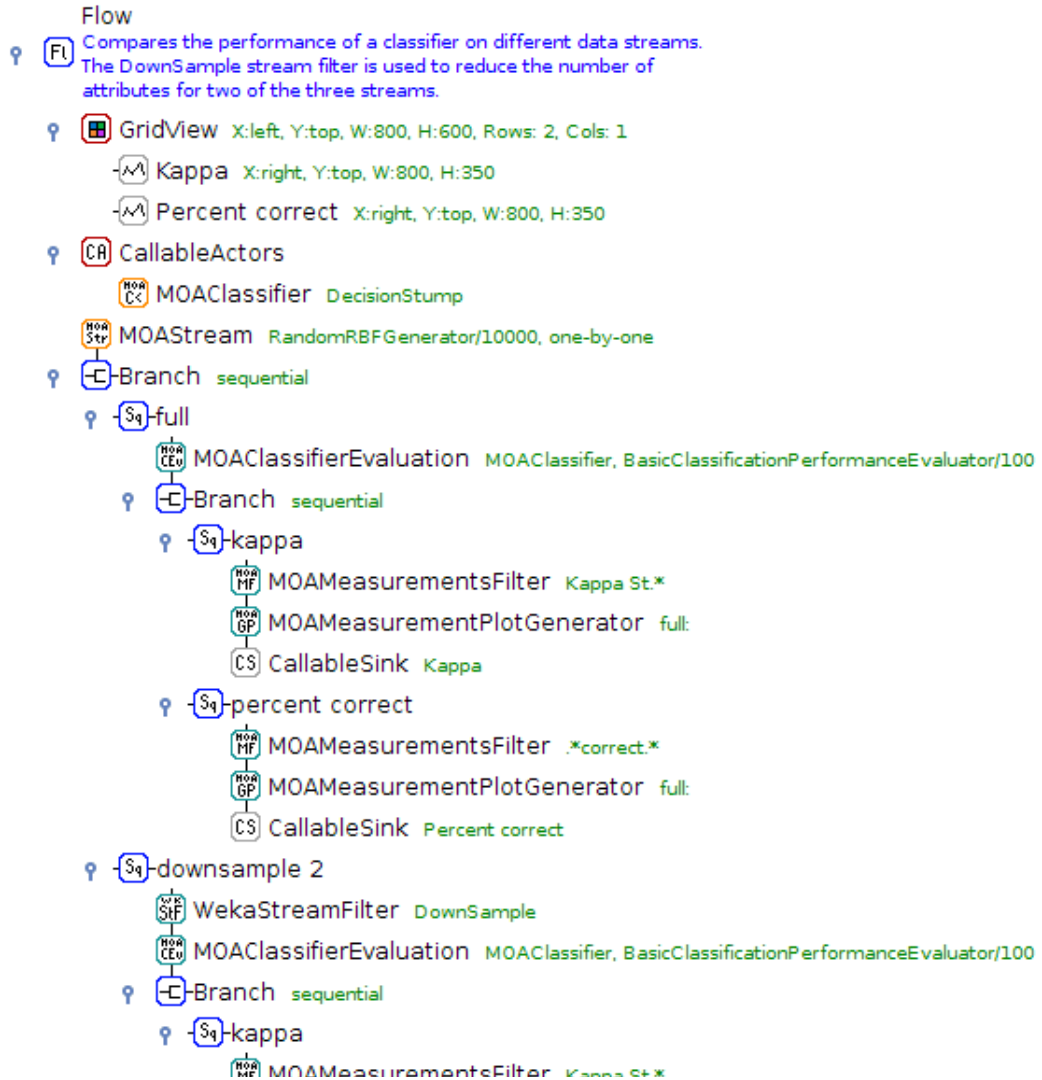
MOA - Drift


Flow
Compares the accucary of a Hoeffding tree classifier with
NaiveBayes on a RandomRBFGenerator (with drift) data stream.
CallableActors
Hoeffding HoeffdingTree
NaiveBayes NaiveBayes
SequencePlotter X:left, Y:top, W:800, H:350
MOAStream RandomRBFGeneratorDrift/1000000, chunks of 1000
Branch parallel, threads: #cores
hoeffding
MOAClassifierEvaluation Hoeffding, BasicClassificationPerformanceEvaluator/1000
MOAMeasurementsFilter .*classifications.*
MOAMeasurementPlotGenerator h:
CallableSink SequencePlotter
naivebayes
MOAClassifierEvaluation NaiveBayes, BasicClassificationPerformanceEvaluator/1000
MOAMeasurementsFilter .*classifications.*
MOAMeasurementPlotGenerator nb:
CallableSink SequencePlotter

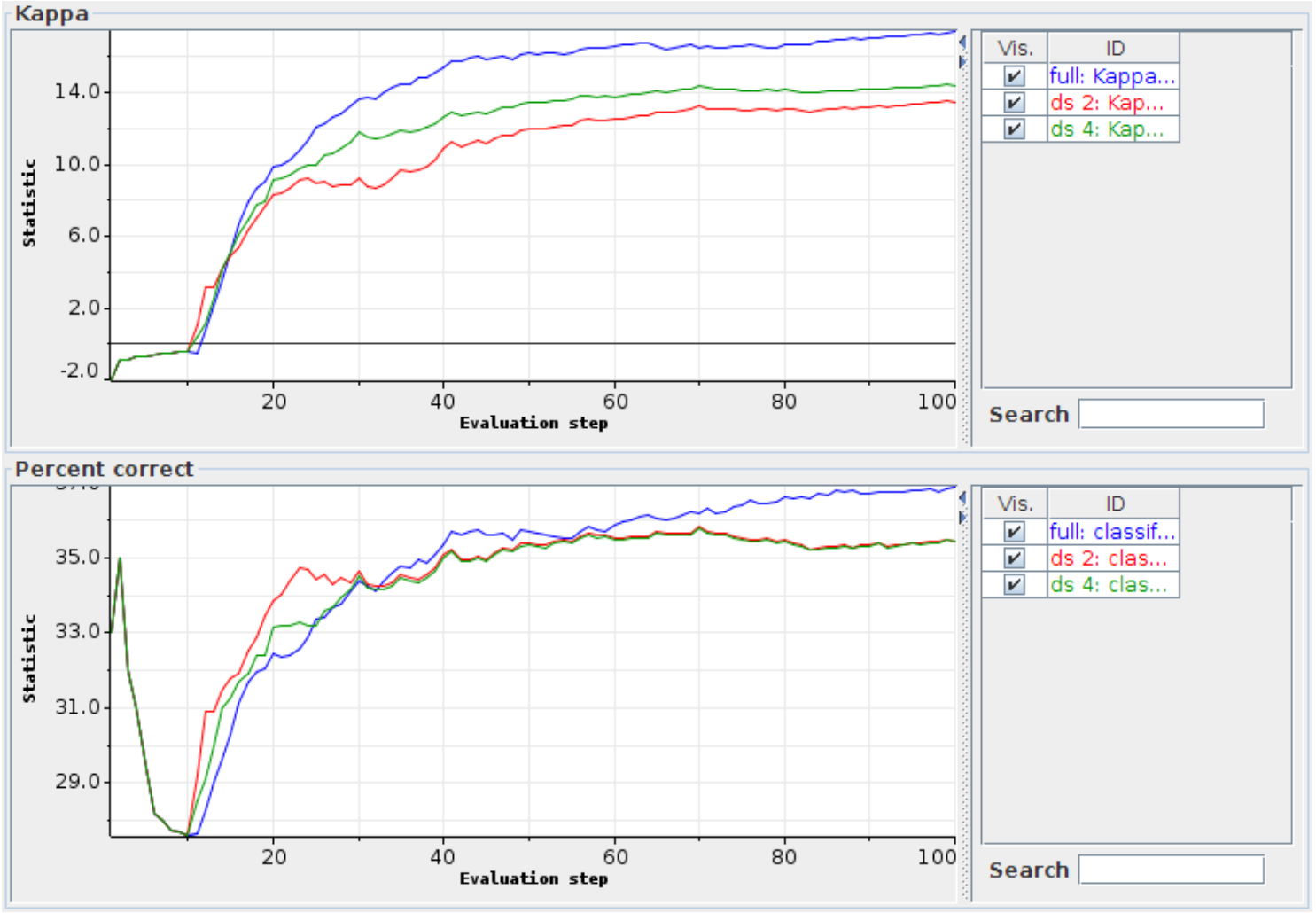
MOA - Drift


Classifier comparison
67.0
65.0
63.0
61.0
59.0
57.0
55.0
Accuracy in %
200
400
600
800
1000
Evaluation step
Vis.
ID
nb: classif...
h: classific...
Search

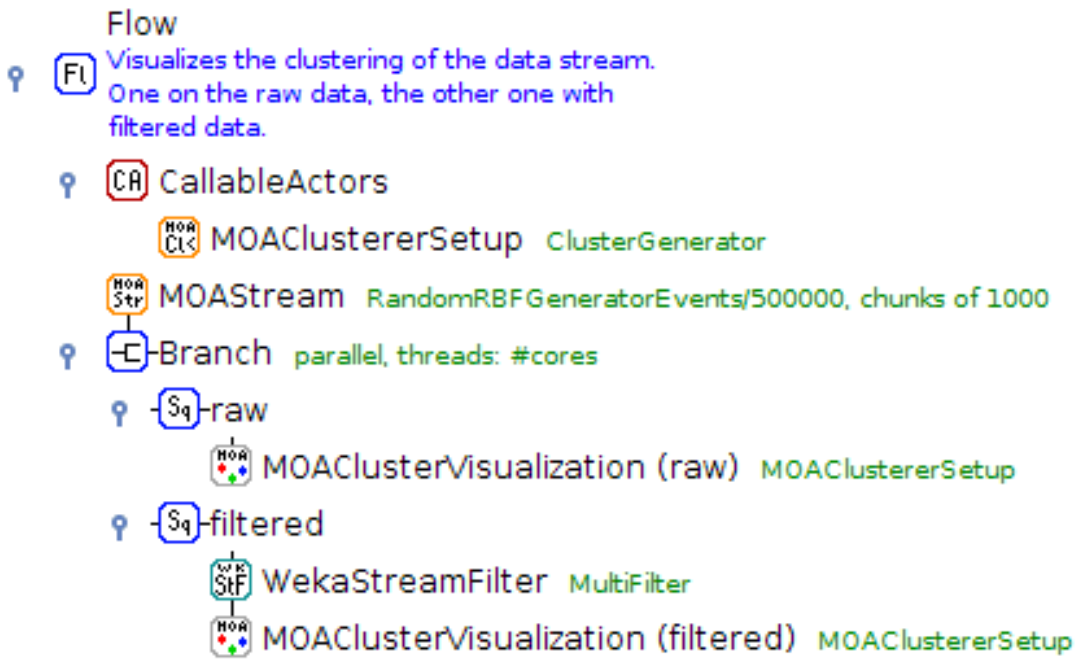
MOA - different streams



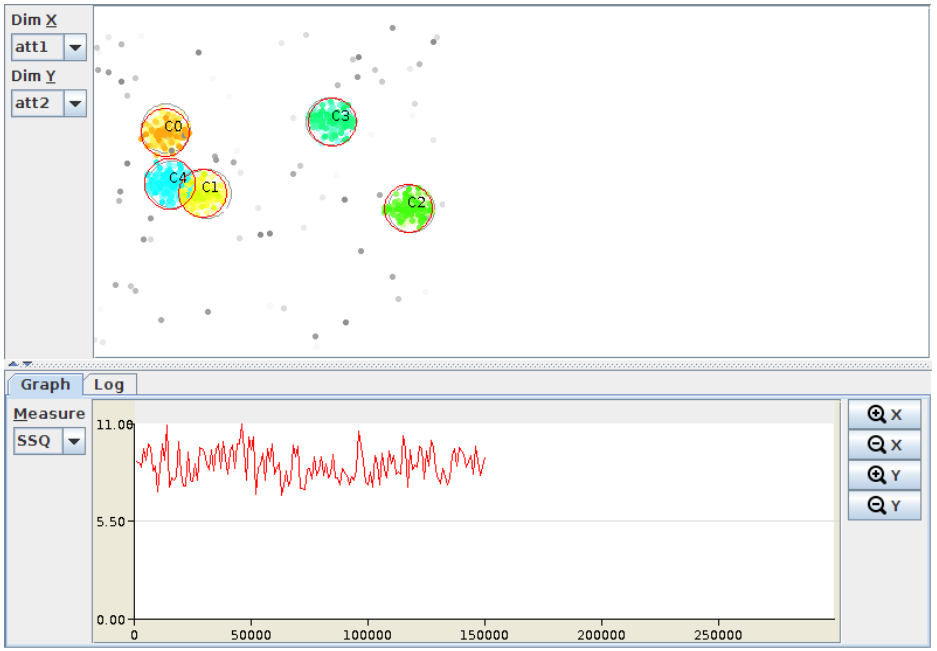
MOA - different streams



MOA - Cluster visualization

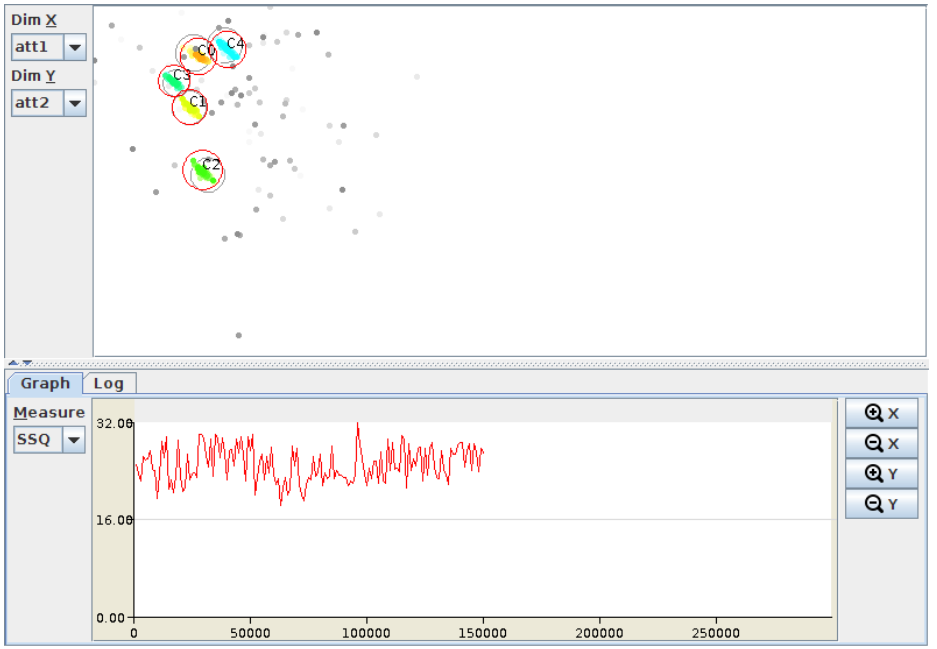


MOA - Cluster visualization

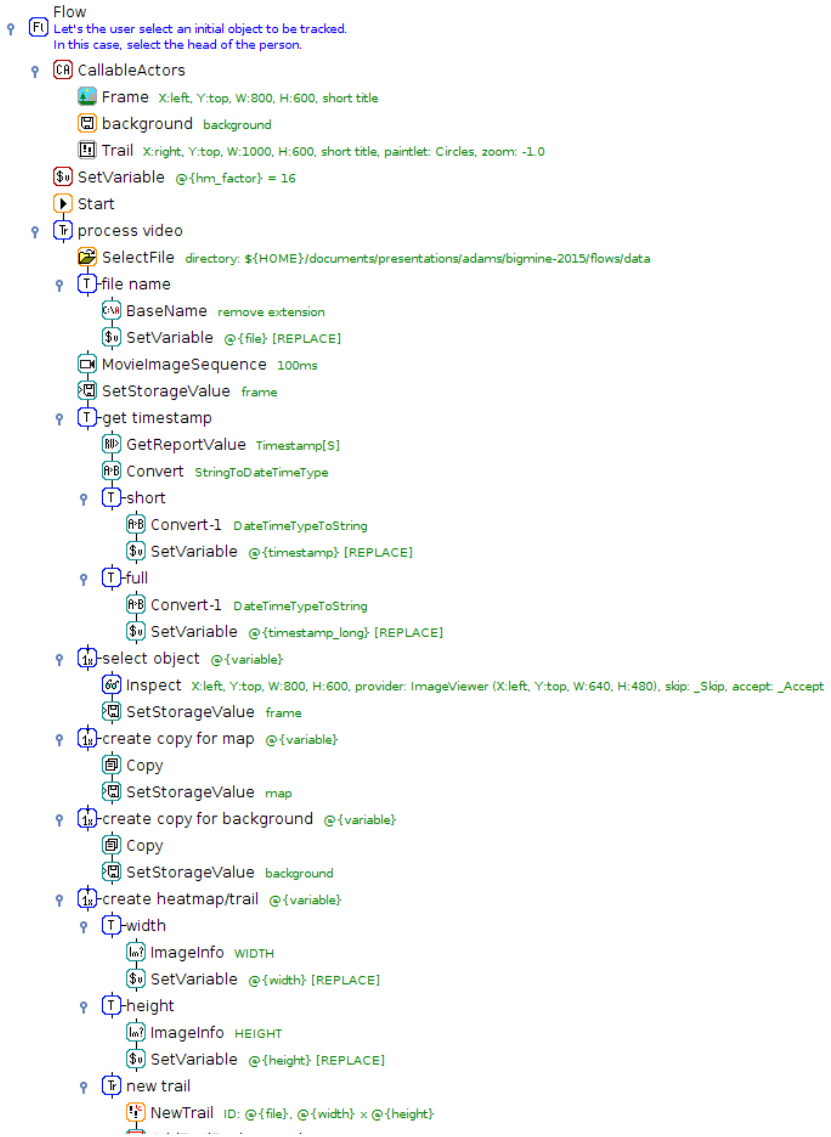


Stream 1

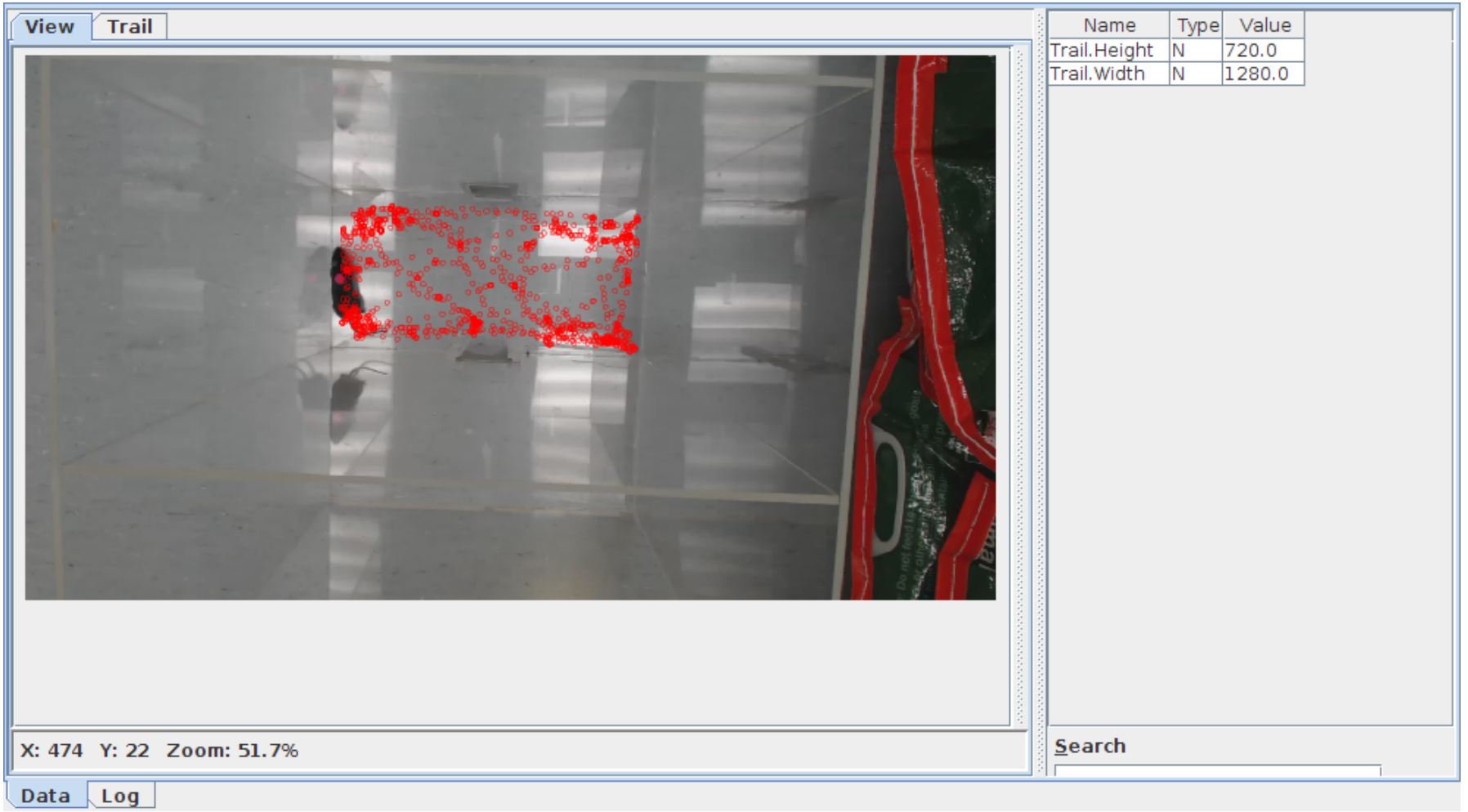
Stream 2



Track mouse


Flow
Let's the user select an initial object to be tracked.
In this case, select the head of the person.
CallableActors
Frame X:left, Y:top, W:800, H:600, short title
background background
Trail X:right, Y:top, W:1000, H:600, short title, paintlet: Circles, zoom: -1.0
SetVariable @{hm_factor} = 16
Start
process video
SelectFile directory: ${HOME}/documents/presentations/adams/bigmine-2015/flows/data
-file name
BaseName remove extension
SetVariable @{file} [REPLACE]
MovieImageSequence 100ms
SetStorageValue frame
-get timestamp
GetReportValue Timestamp[S]
Convert StringToDateTimeType
-short
Convert-1 DateTimeTypeToString
SetVariable @{timestamp} [REPLACE]
-full
Convert-1 DateTimeTypeToString
SetVariable @{timestamp_long} [REPLACE]
-select object @{variable}
Inspect X:left, Y:top, W:800, H:600, provider: ImageViewer (X:left, Y:top, W:640, H:480), skip: _Skip, accept: _Accept
SetStorageValue frame
-create copy for map @{variable}
Copy
SetStorageValue map
-create copy for background @{variable}
Copy
SetStorageValue background
-create heatmap/trail @{variable}
-width
ImageInfo WIDTH
SetVariable @{width} [REPLACE]
-height
ImageInfo HEIGHT
SetVariable @{height} [REPLACE]
new trail
NewTrail ID: @{file}, @{width} x @{height}

Track mouse

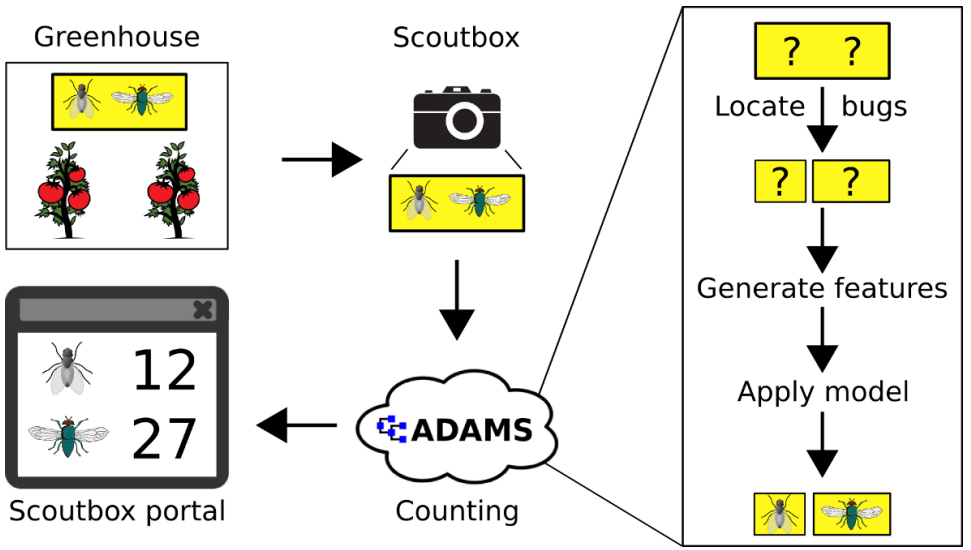

View
Trail
Name
Type
Value
Trail.Height
N
720.0
Trail.Width
N
1280.0
X: 474 Y: 22 Zoom: 51.7%
Search
Data
Log

Industry

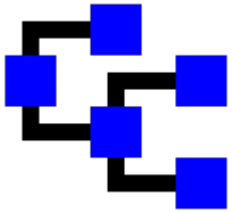
- BLGG - environmental lab in NL
- Spectral analysis
  - XRF: 10,000, MIR: 2,000, NIR: 1,500
- In operation since 2006
- Predictive modelling: soil, plant (~250 models)
- 1,000 to 3,000 samples per day
- Savings due to less wet chemistry
  - USD 18 million to USD 33 million per year

Industry (2)

- Cropwatch BV
  - monitoring trends in insect populations (whitefly, macrolophus, thrips)
  - currently used in greenhouses
- analyzing images of sticky plates



Questions?



https://adams.cms.waikato.ac.nz/

@TheAdamsFlow