Large Scale Biological Data Handling and Analysis Using Open Source Alternatives

Dr Karol Kozak

ETH Zurich

May 2009, Paris OME

Large Scale Experiments and Informatics

Large Scale Experiments and Informatics

Instrument management

Large Scale Experiments and Informatics

Instrument management

Data acquisition

Large Scale Experiments and Informatics

Instrument management

Data acquisition

Image processing

Large Scale Experiments and Informatics

Instrument management

Data acquisition

Image processing

Normalization, QC

Large Scale Experiments and Informatics

Instrument management

Data acquisition

Image processing

Normalization, QC

Data storage

Large Scale Experiments and Informatics

Instrument management

Data acquisition

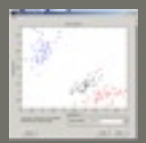
Image processing

Normalization, QC

Data storage

Archiving

Large Scale Experiments and Informatics



Instrument management

Data acquisition

Image processing

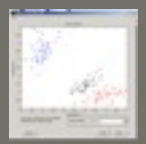
Normalization, QC

Data storage

Archiving

Data mining

Large Scale Experiments and Informatics



Instrument management

Data acquisition

Image processing

Normalization, QC

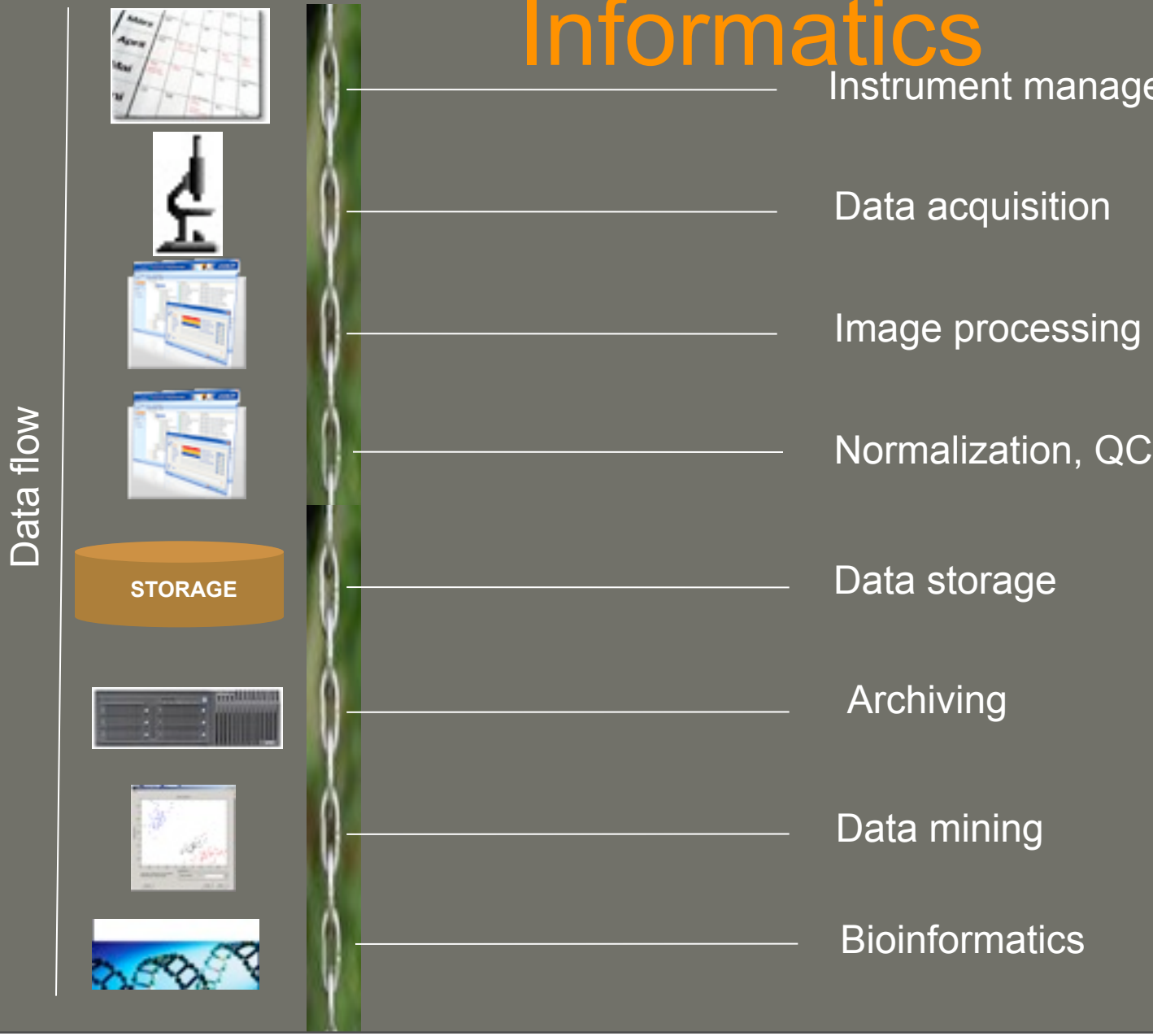
Data storage

Archiving

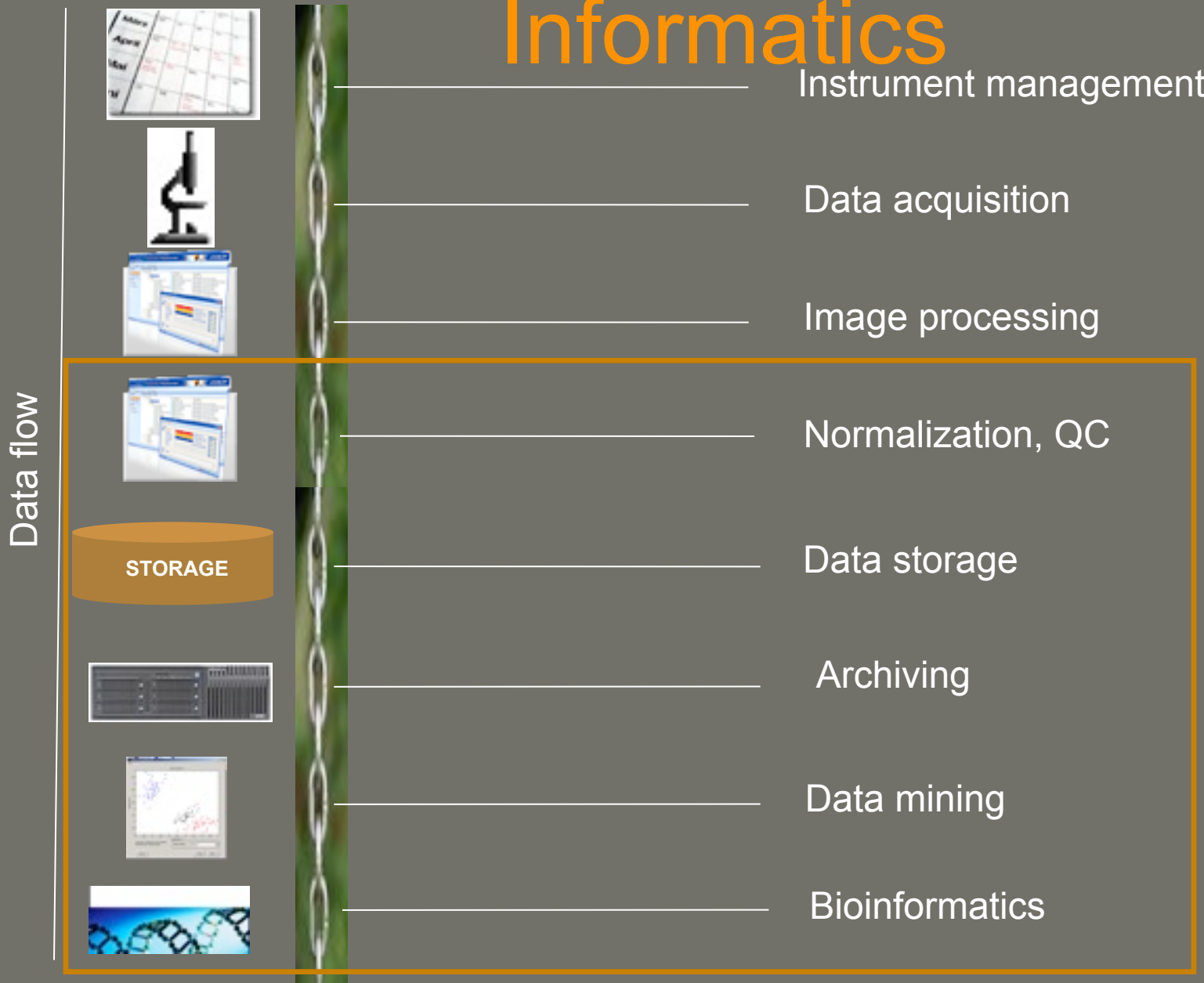
Data mining

Bioinformatics

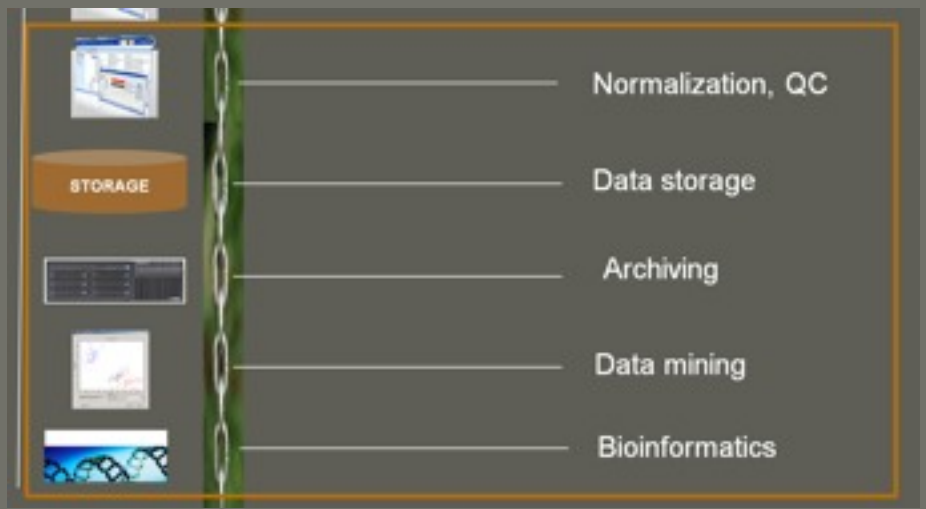
Large Scale Experiments and Informatics



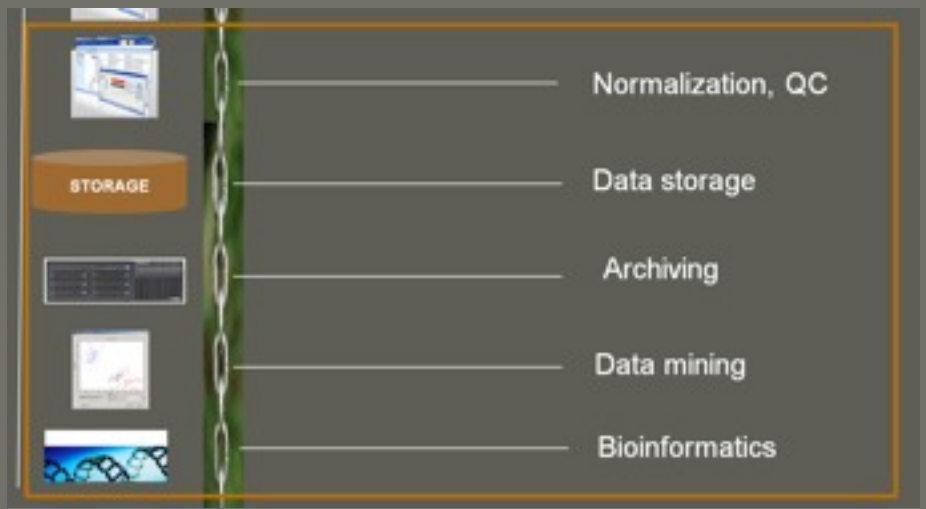
Large Scale Experiments and Informatics



Detection points



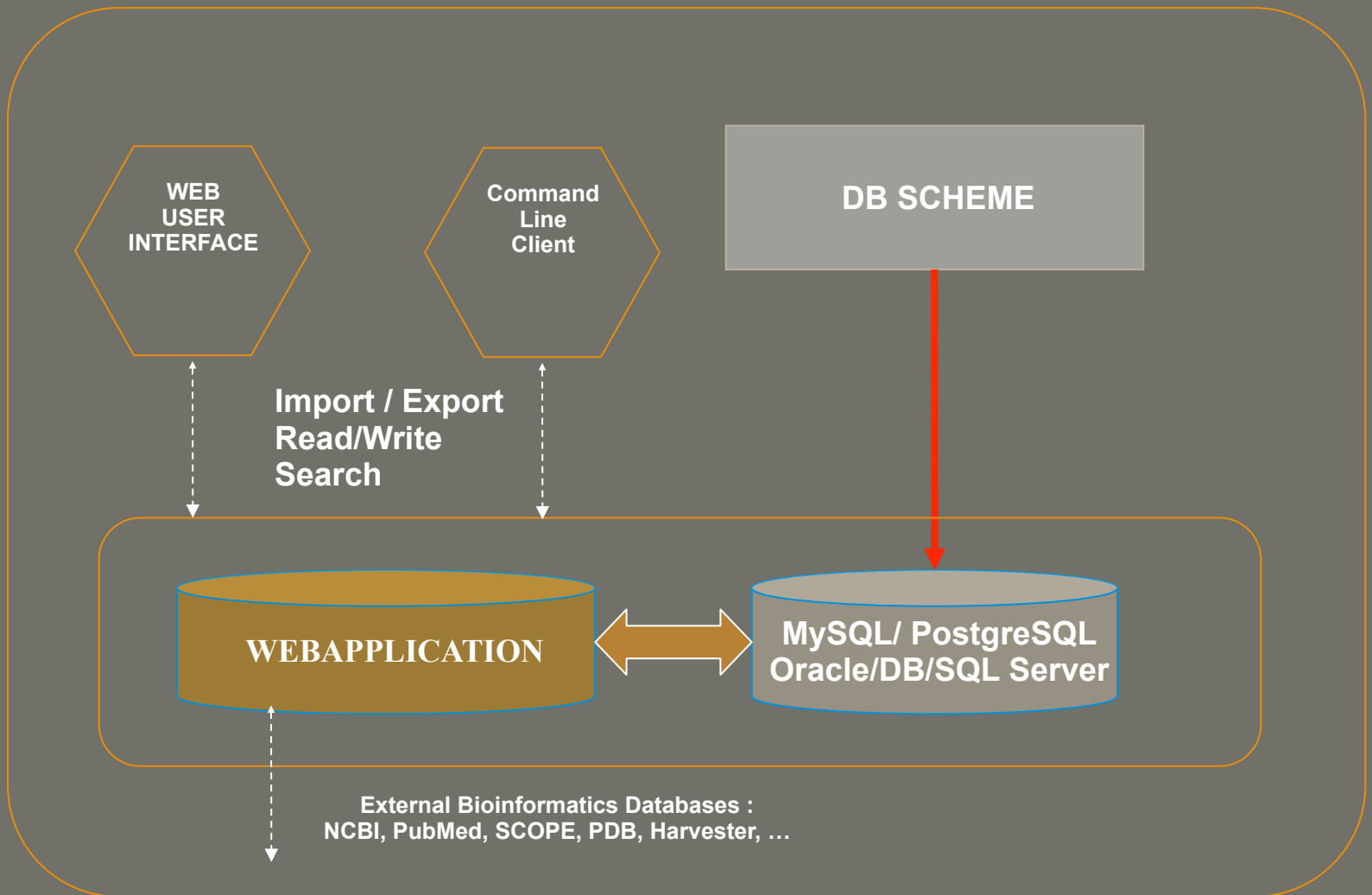
Detection points



Data Automation:

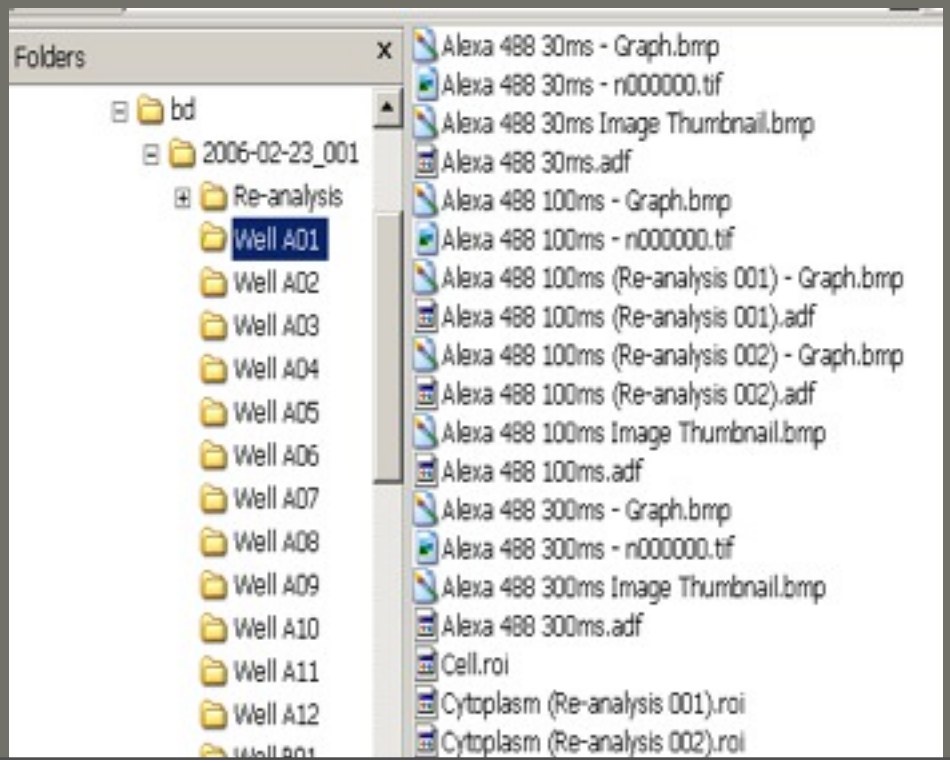
- Library Handling
- Robotics
- Microscopy
- Image processing
- Cell reliability
- Hit Definition

Database architecture



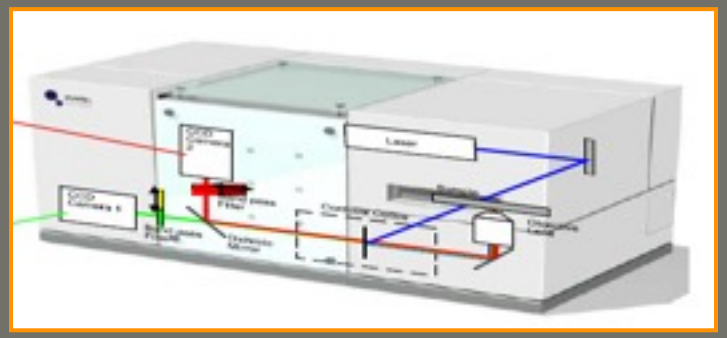
STORAGE

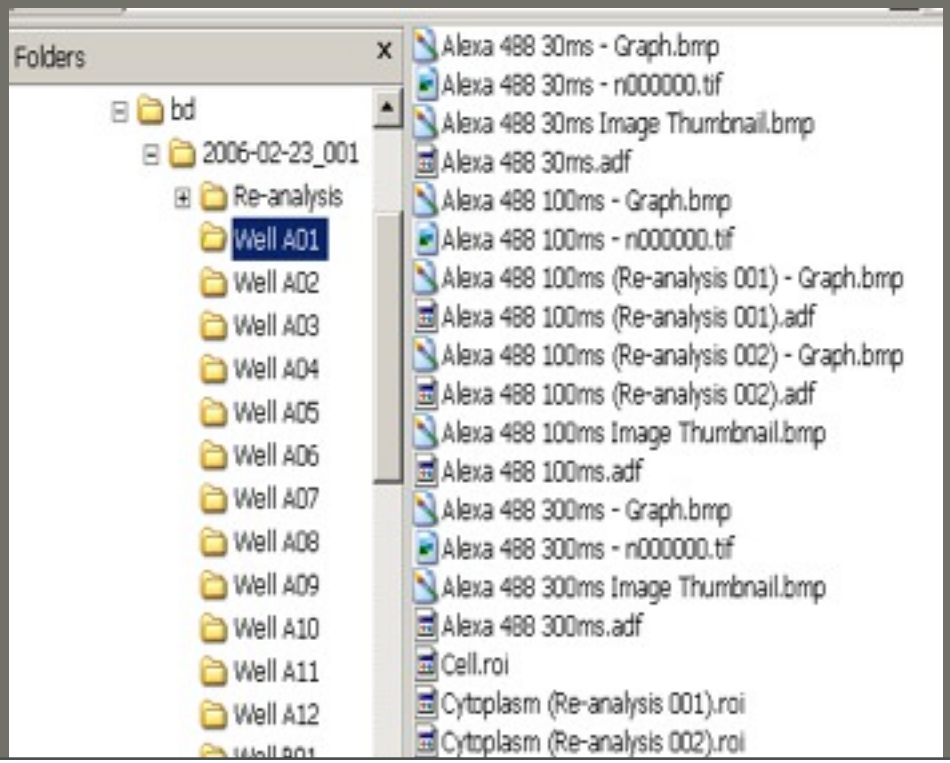
Link to microscope

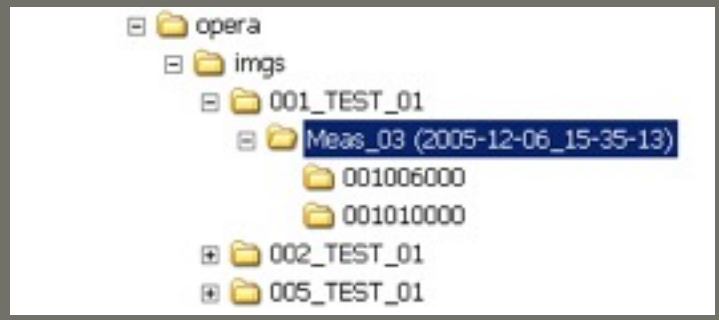


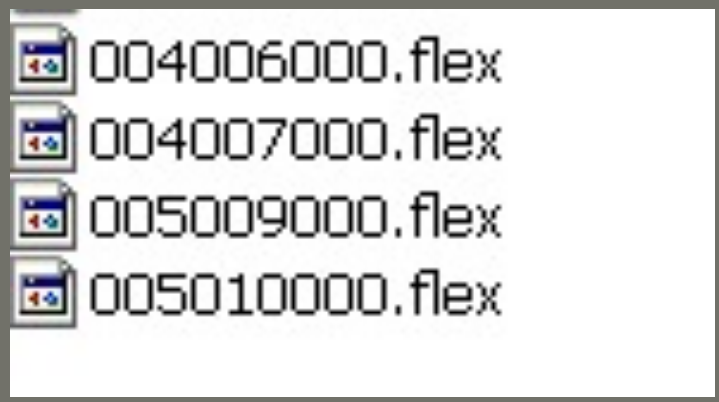
STORAGE

Link to microscope



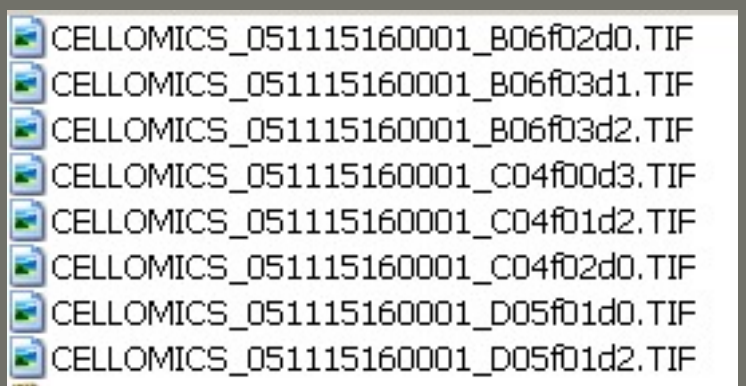






STORAGE

Link to microscope



STORAGE

Link to microscope

- thermo
  - experiment1
    - platebarcode

- MD_MICRO
  - experiment1
    - 2008-06-28_001barcode

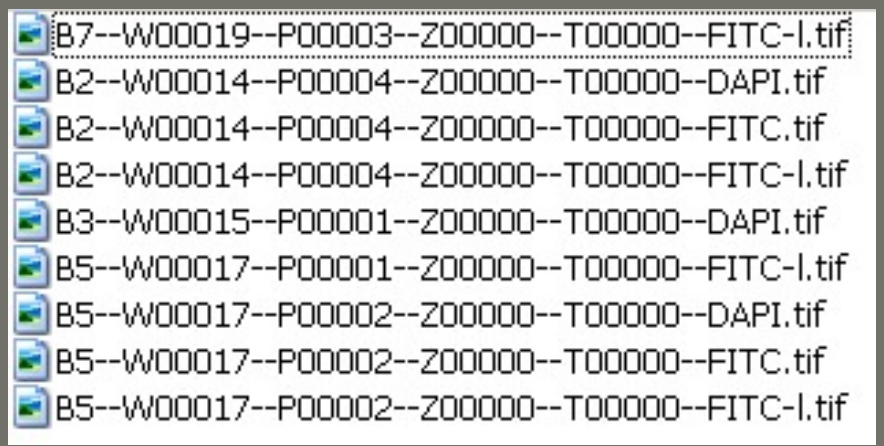
CELLOMICS_051115160001_B06f02d0.TIF
CELLOMICS_051115160001_B06f03d1.TIF
CELLOMICS_051115160001_B06f03d2.TIF
CELLOMICS_051115160001_C04f00d3.TIF
CELLOMICS_051115160001_C04f01d2.TIF
CELLOMICS_051115160001_C04f02d0.TIF
CELLOMICS_051115160001_D05f01d0.TIF
CELLOMICS_051115160001_D05f01d2.TIF

2008-06-28_000_A01_00_dapi.tif
2008-06-28_000_A01_00_yfp.tif
2008-06-28_000_A02_00_dapi.tif
2008-06-28_000_A02_00_yfp.tif
2008-06-28_000_A03_00_dapi.tif
2008-06-28_000_A03_00_yfp.tif
2008-06-28_000_A04_00_dapi.tif
2008-06-28_000_A04_00_yfp.tif
2008-06-28_000_A05_00_dapi.tif
2008-06-28_000_A05_00_yfp.tif
2008-06-28_000_B02_02_yfp.tif
2008-06-28_000_B02_03_dapi.tif
2008-06-28_000_B02_03_yfp.tif

STORAGE

Link to microscope

Olympus **SCAN R**



Data archiving

~100MB

Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

Data archiving

~100MB

Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

In Screening Process

Results Database

DB

Data archiving

~100MB

Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

In Screening Process

Results Database

DB

Data Management Tool

Data archiving

~100MB
Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

Image processing

In Screening Process

Results Database

DB

Data Management Tool

Data archiving

~100MB
Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

Image processing

In Screening Process

Convert to jpeg
50 x smaller

Results Database

DB

Data Management Tool

Data archiving

~100MB
Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

Image processing

In Screening Process

Convert to jpeg 50 x smaller

Results Database

DB

Data Management Tool

JPEG JPEG JPEG

~2MB

+ numeric results plots

Data archiving

Tape

~100MB

Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

Image processing

In Screening Process

Convert to jpeg 50 x smaller

Results Database

DB

Data Management Tool

JPEG JPEG JPEG

~2MB

+ numeric results plots

Data archiving

Tape

~100MB

Microscope –format

- BD
- ThermoFisher
- Zeiss
- Leica
- Olympus
- MD

- tiff
- lsm
- lei

Image processing

In Screening Process

Convert to jpeg 50 x smaller

Results Database

DB

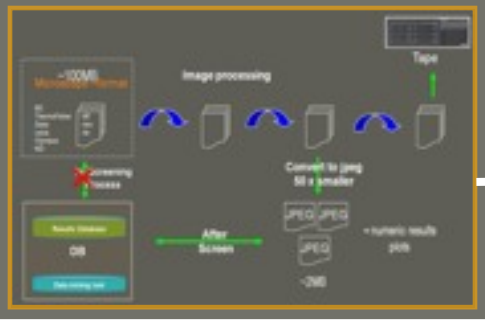
Data Management Tool

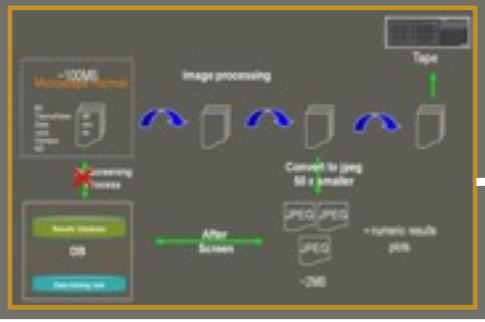
After Screen

JPEG JPEG JPEG

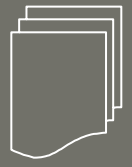
~2MB

+ numeric results plots


~100MB
Image processing
Tape
Convert to jpeg
50 x smaller
JPEG
JPEG
JPEG
DB
After
Screen
~2MB



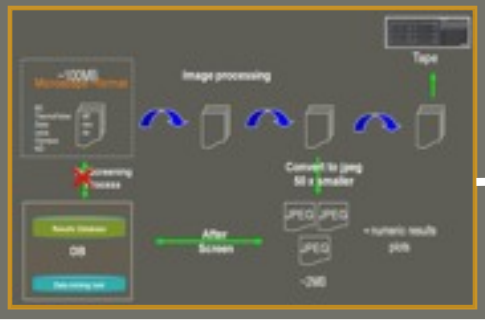
*.tiff



In Screening Process

Results Database

DB



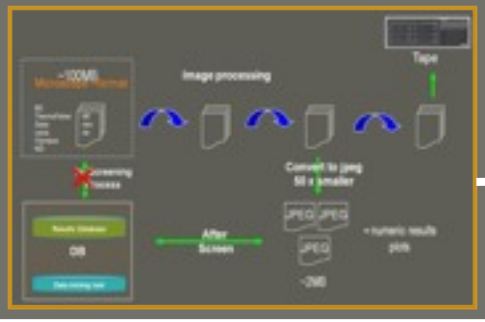
*.tiff

Buffer server

In Screening Process

Results Database

DB



*.tiff

In Screening Process

Results Database

DB

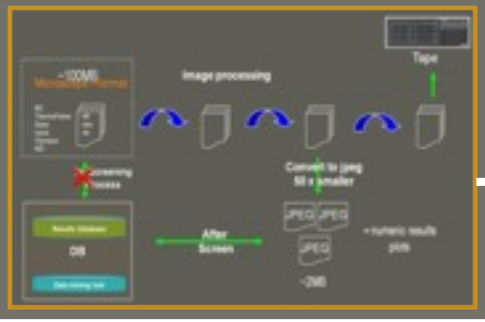
Buffer server

2x25TB LMC (6 weeks)

JPEG JPEG JPEG

+ numeric results plots

~0.5MB



*.tiff

In Screening Process

Buffer server

2x25TB LMC (6 weeks)

Results Database

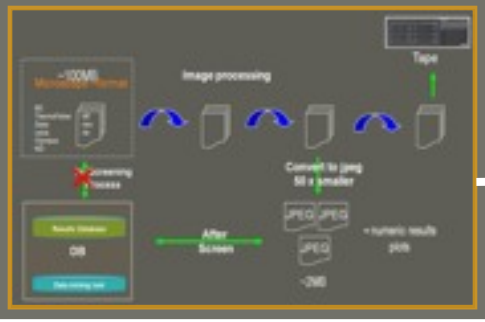
DB

After Screen

JPEG JPEG JPEG

+ numeric results plots

~0.5MB



*.tiff

In Screening Process

Buffer server

2x25TB LMC (6 weeks)

ETH NAS Users Tape

ETH NAS User storage

Results Database

DB

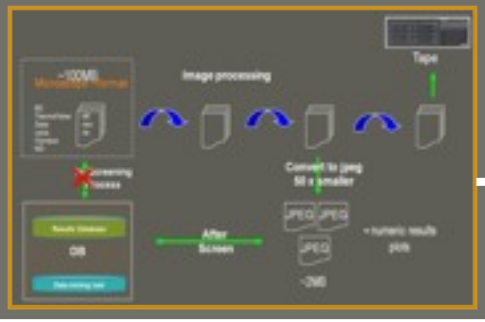
After Screen

JPEG JPEG

JPEG

~0.5MB

+ numeric results plots



*.tiff

LMC interface- user

ETH NAS
Users Tape

Buffer server

2x25TB
LMC (6 weeks)

In Screening
Process

ETH NAS
User storage

Results Database

DB

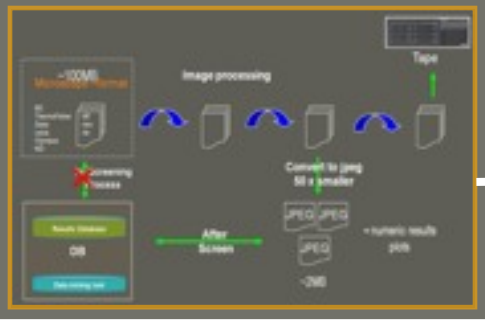
After
Screen

JPEG JPEG

JPEG

+ numeric results
plots

~0.5MB



*.tiff

In Screening Process

Results Database

DB

Buffer server

LMC interface- user

2x25TB LMC (6 weeks)

ETH NAS Users Tape

ETH NAS User storage

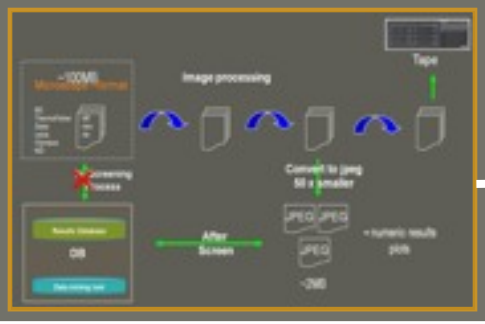
After Screen

JPEG JPEG

JPEG

+ numeric results plots

~0.5MB



*.tiff

LMC interface- user

ETH NAS
Users Tape

Buffer server

2x25TB
LMC (6 weeks)

In Screening
Process

ETH NAS
User storage

Results Database

DB

Data Management Tool

After
Screen

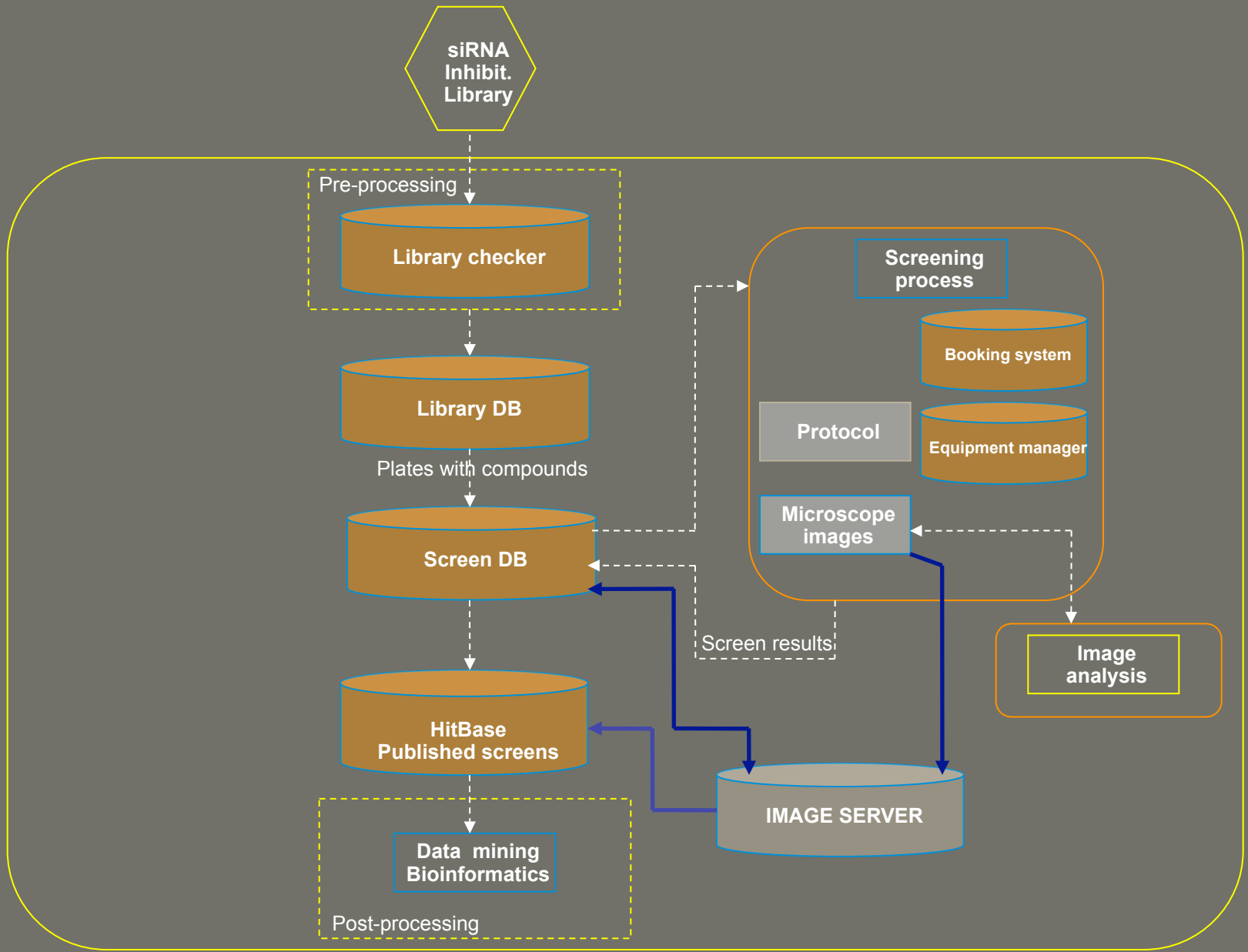
JPEG JPEG

JPEG

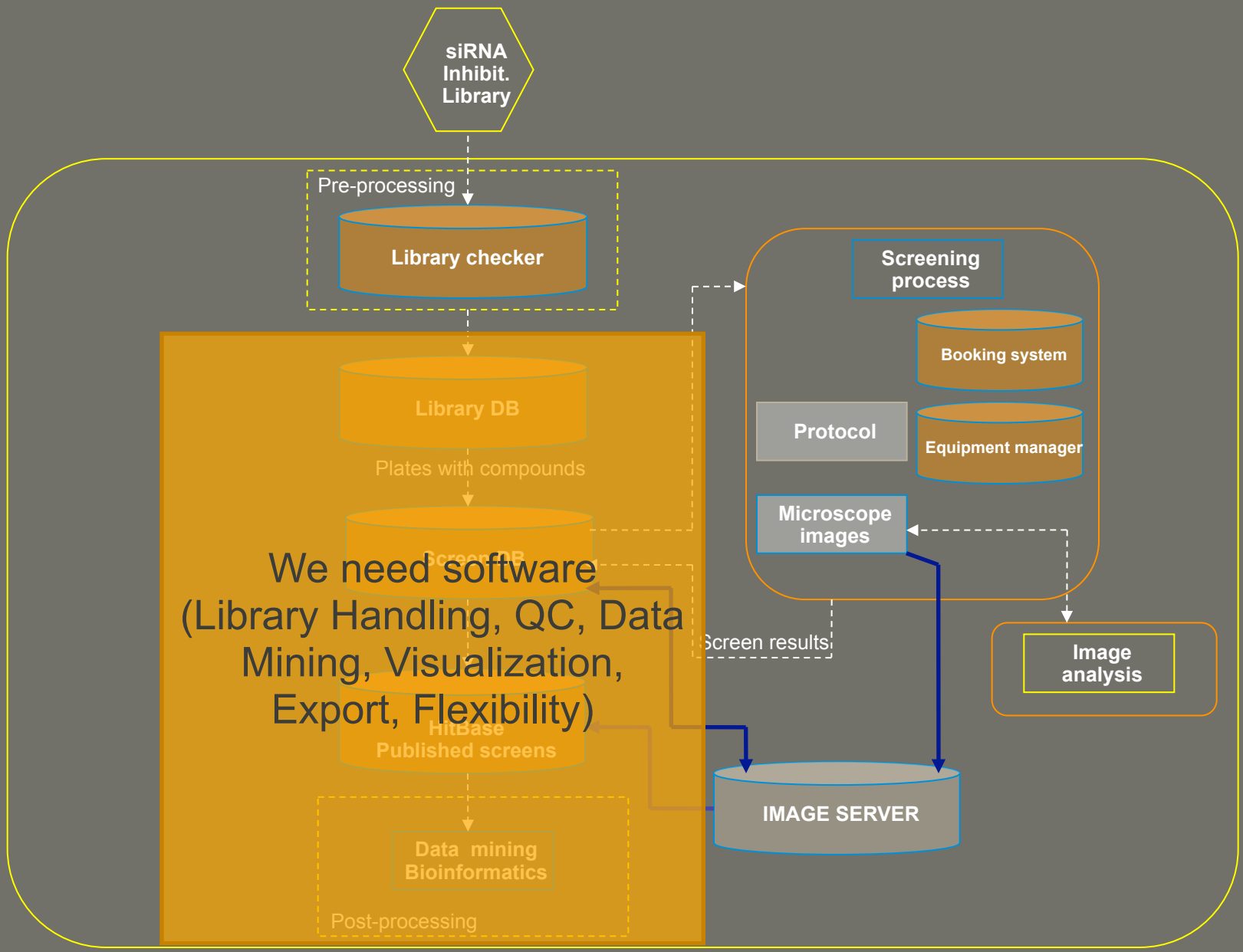
+ numeric results
plots

~0.5MB

Dataflow architecture



Dataflow architecture



HC/DC

High **C**ontent analysis **D**ata **C**hain

Data Handling Unit

based on Eclipse / R / Weka / KNIME

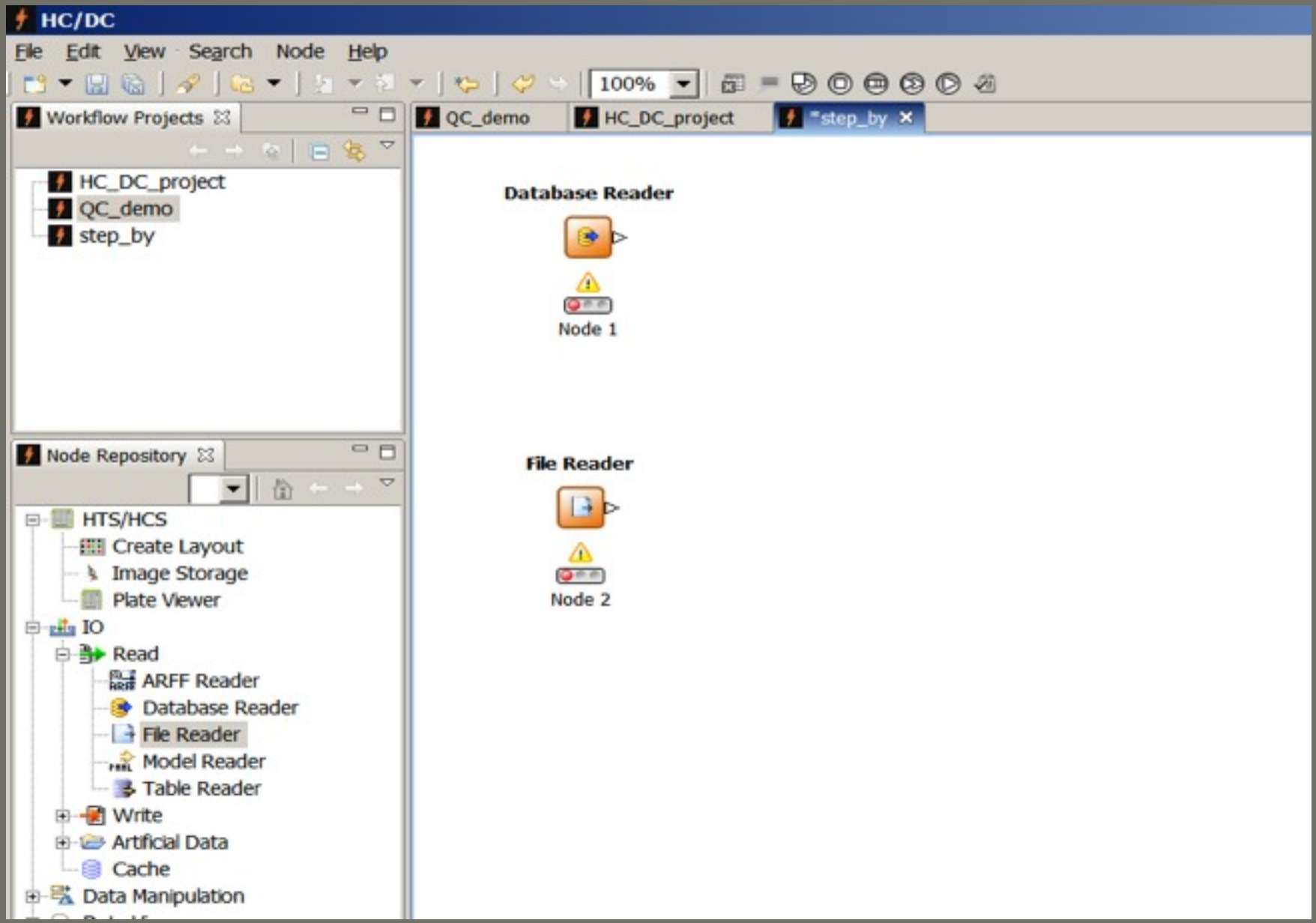
HC/DC

High **C**ontent analysis **D**ata **C**hain

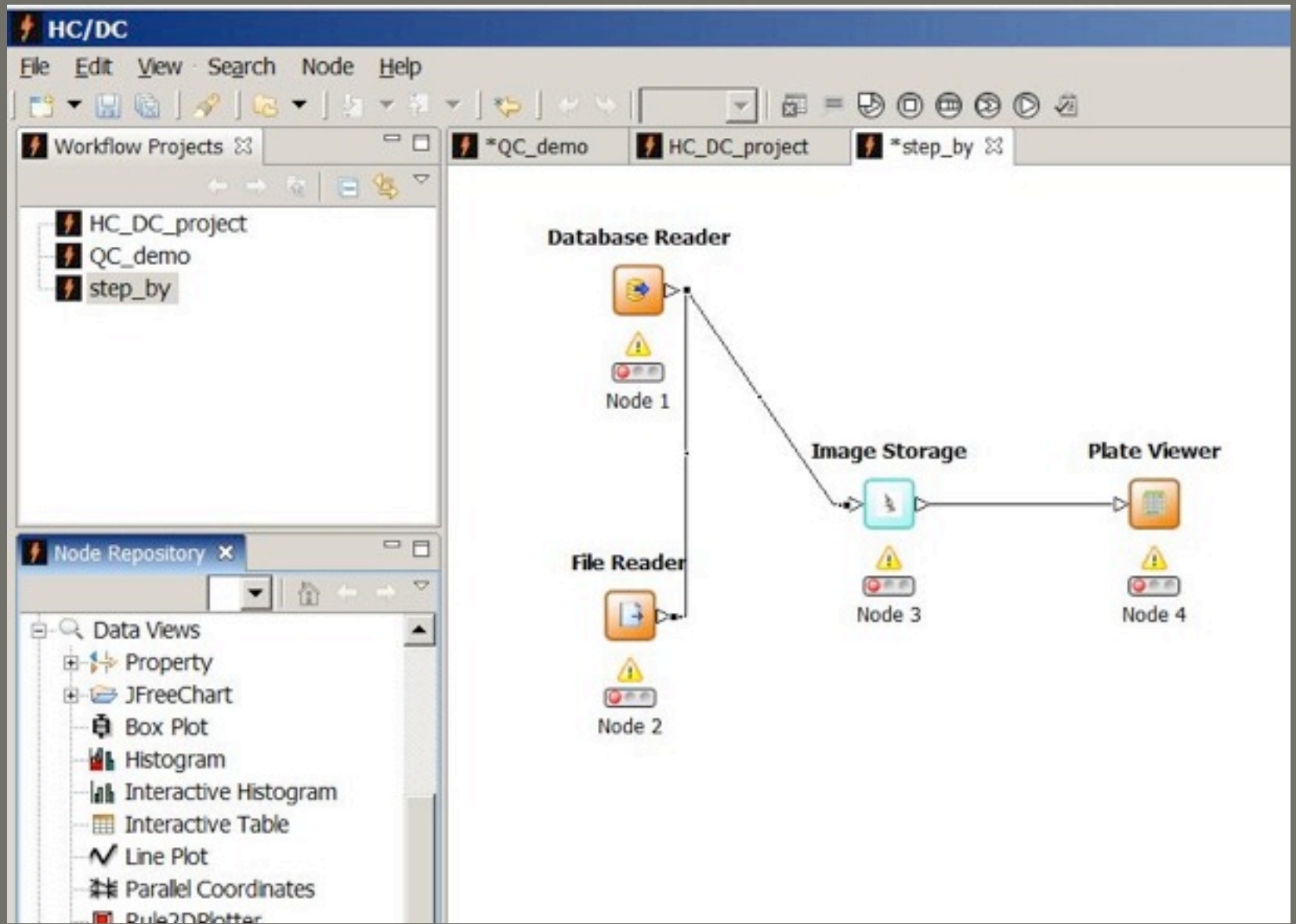
Data Handling Unit

based on Eclipse / R / Weka / KNIME

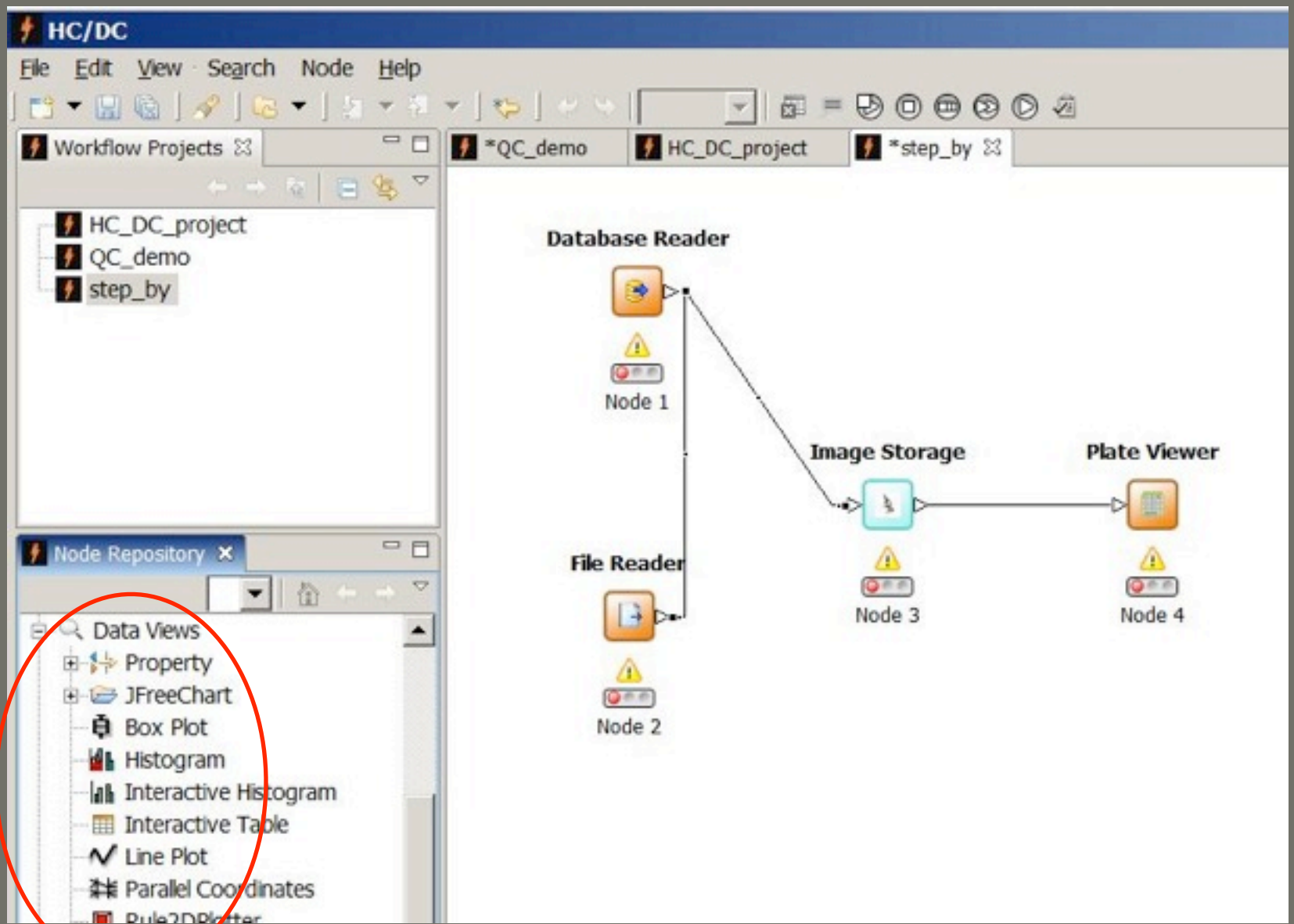
Read data from database



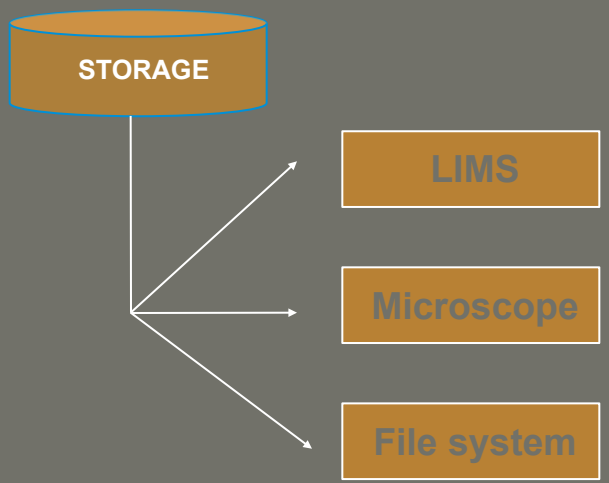
Link with image storage



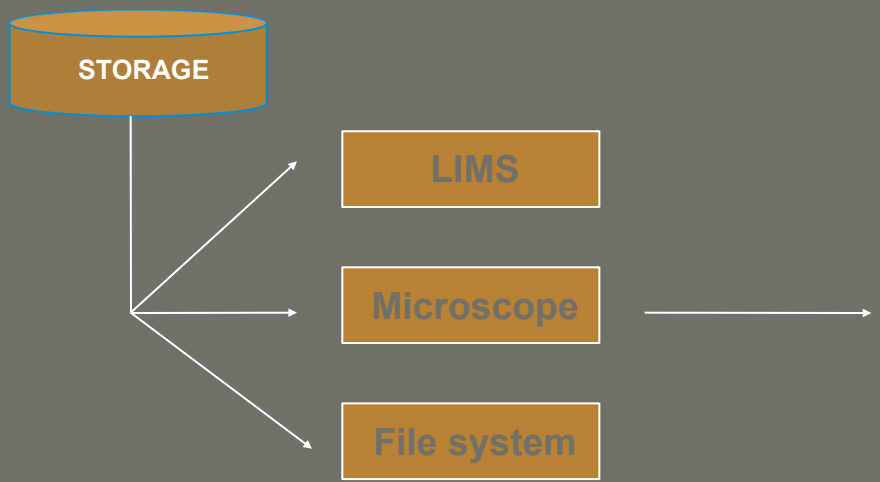
Link with image storage



Data mining



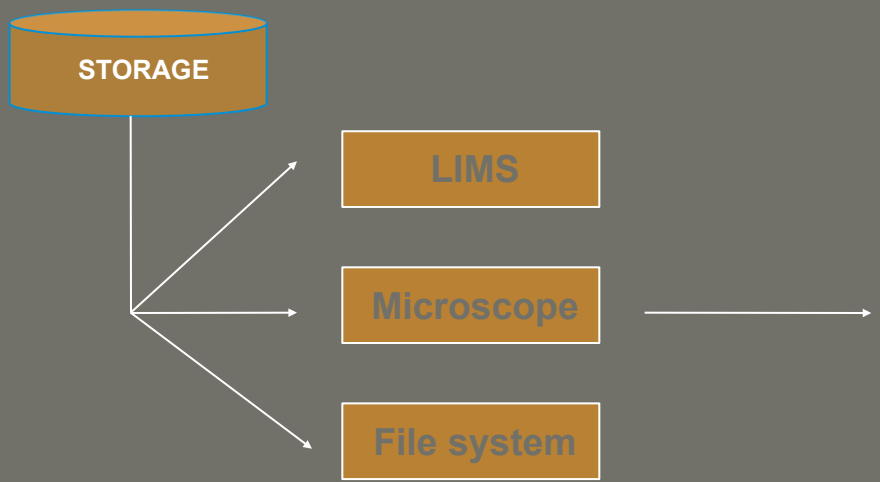
Data mining



Statistics

-One/two parameters
-Normalization
-Correlation
-Compare distribution
-Ex. Statistical tests, Z-score, Z', etc

Data mining



Statistics

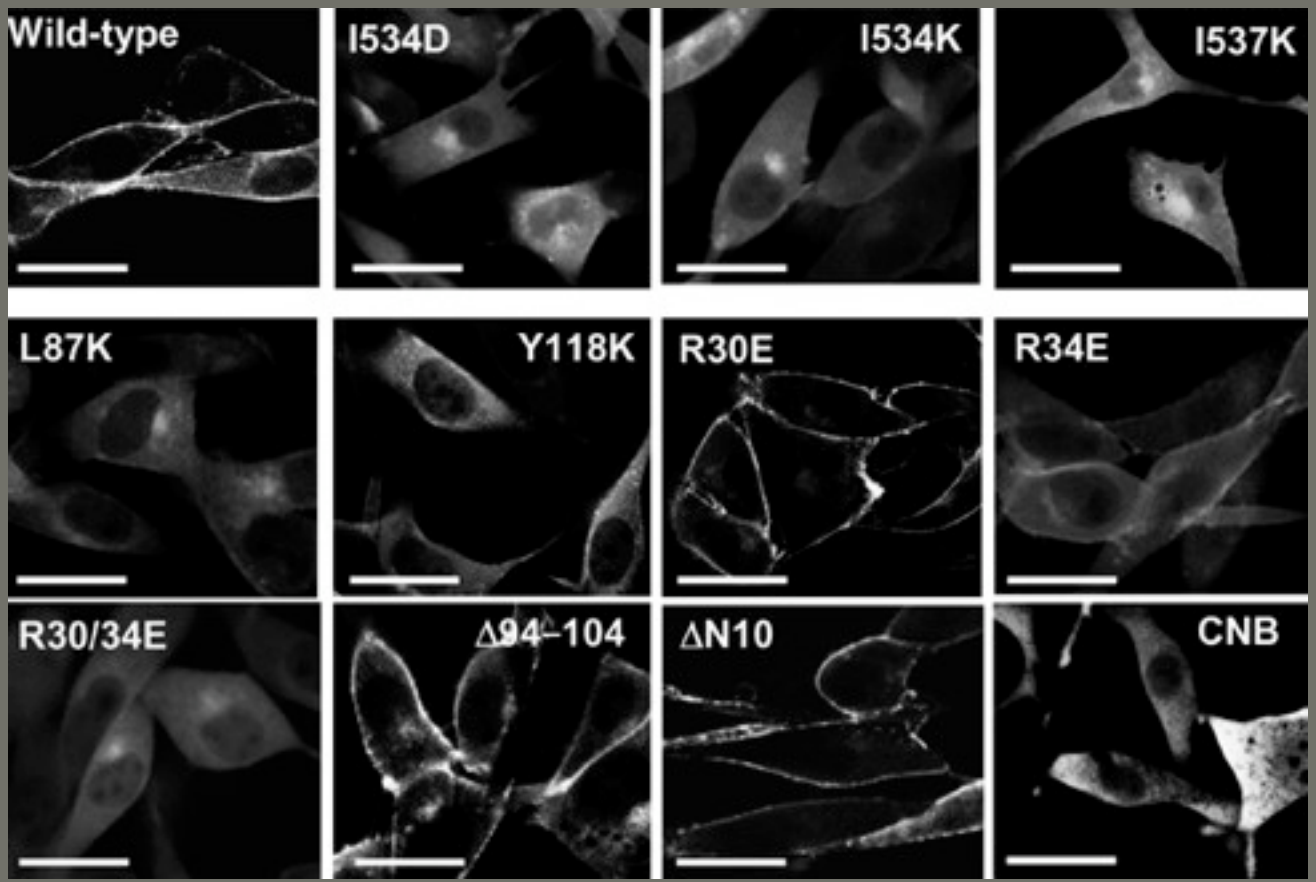
-One/two parameters
-Normalization
-Correlation
-Compare distribution
-Ex. Statistical tests, Z-score, Z', etc

Pattern recognition – Machine Learning

-Multi-parameters
-Clustering
-Dimensionality reduction
-Check parameter importance

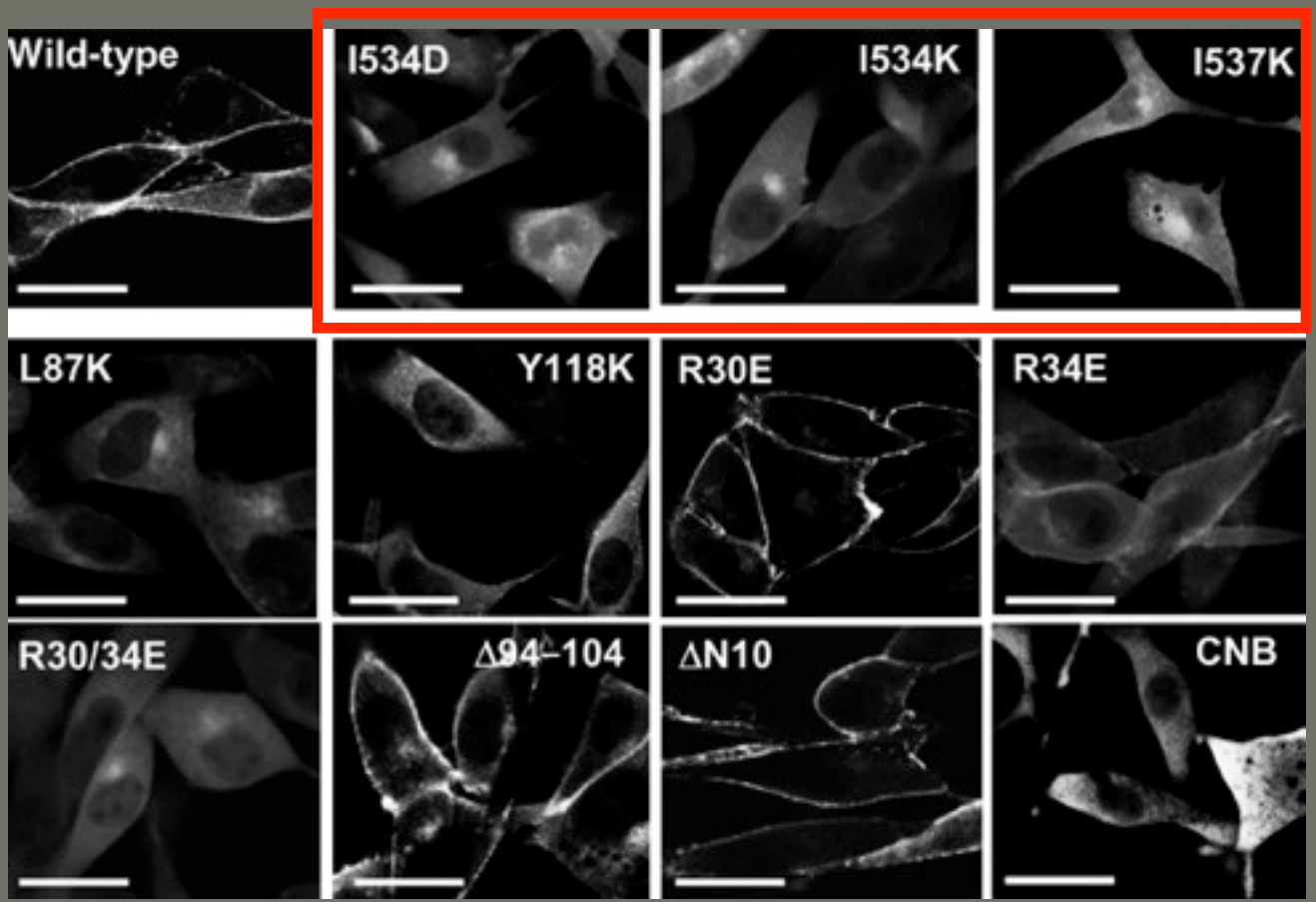
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



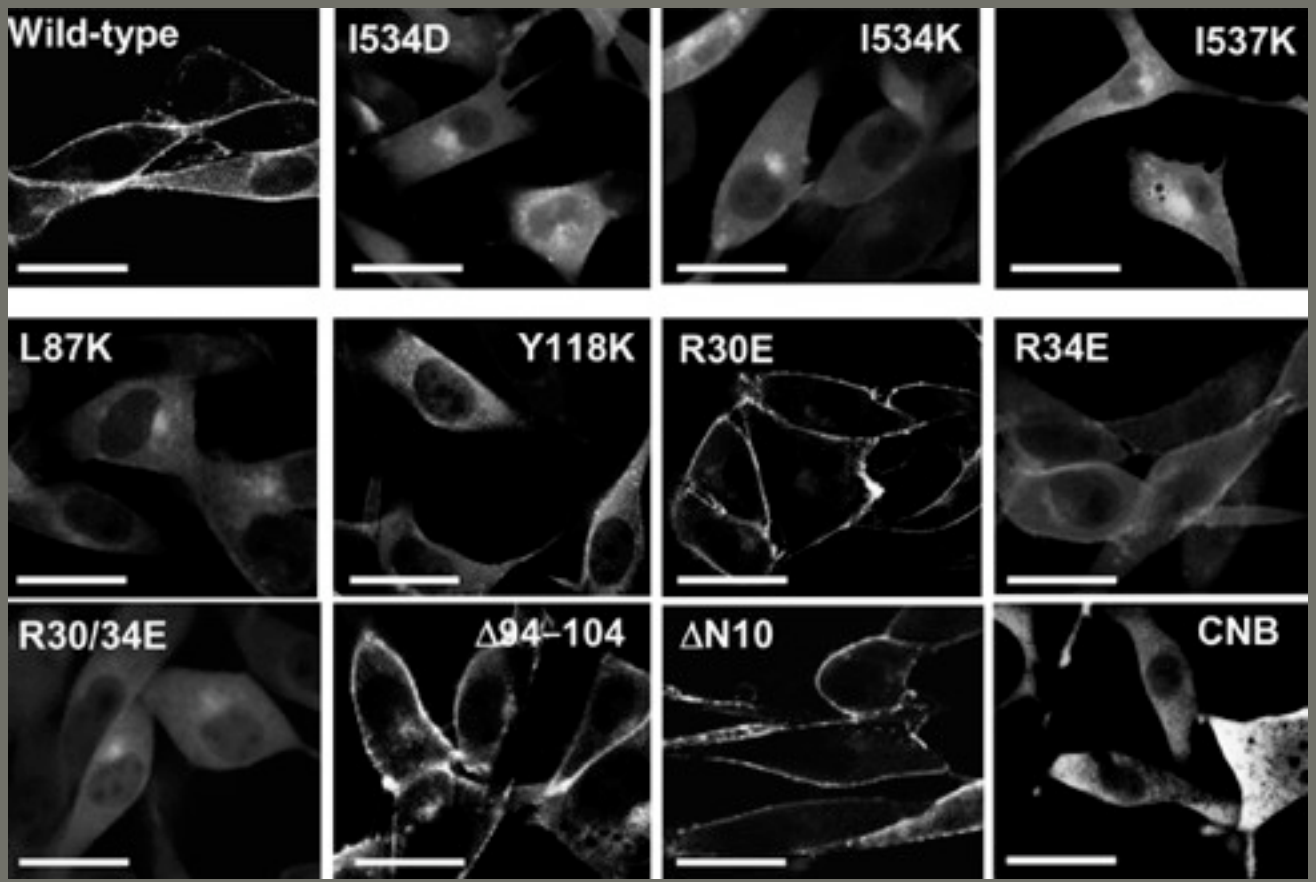
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



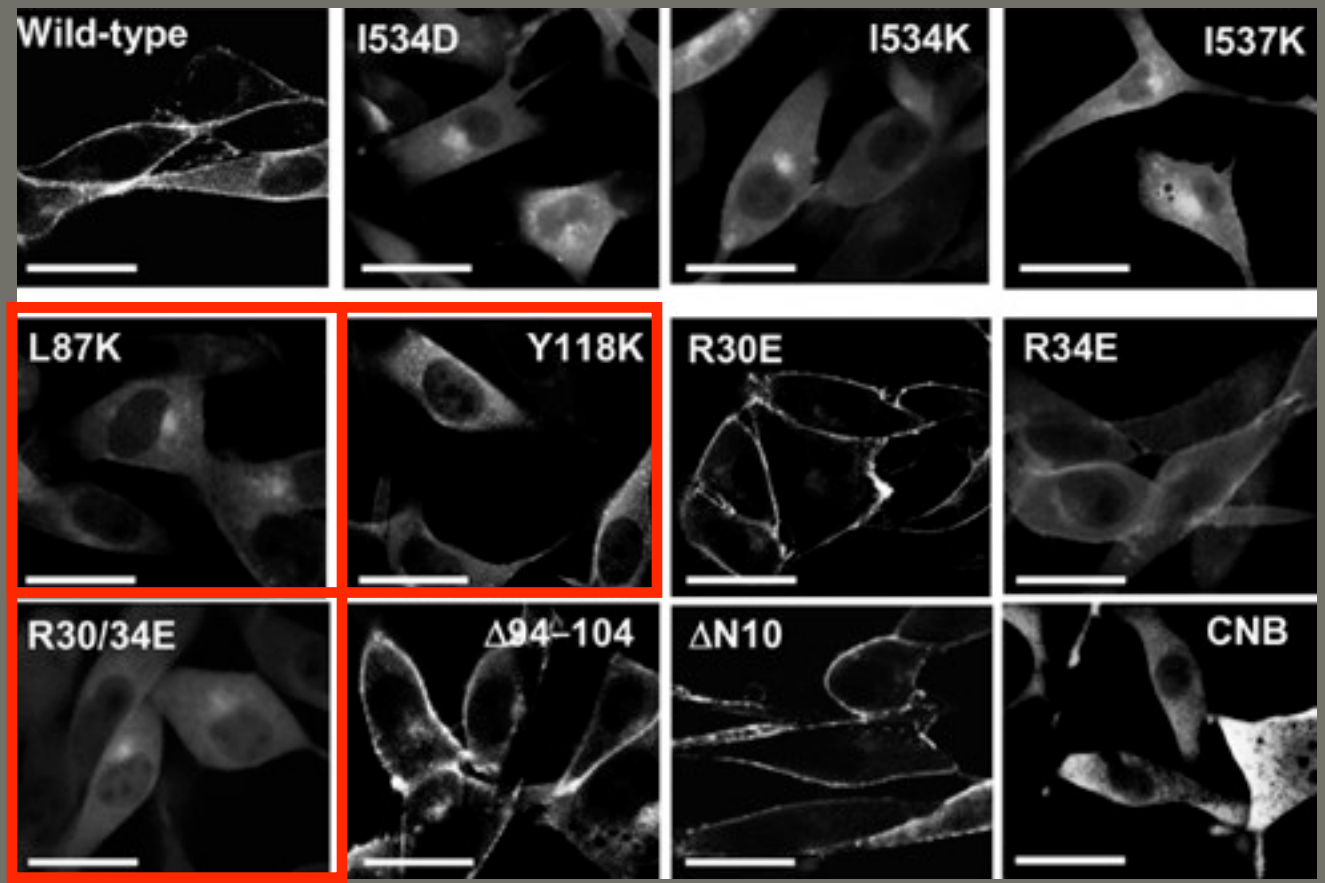
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



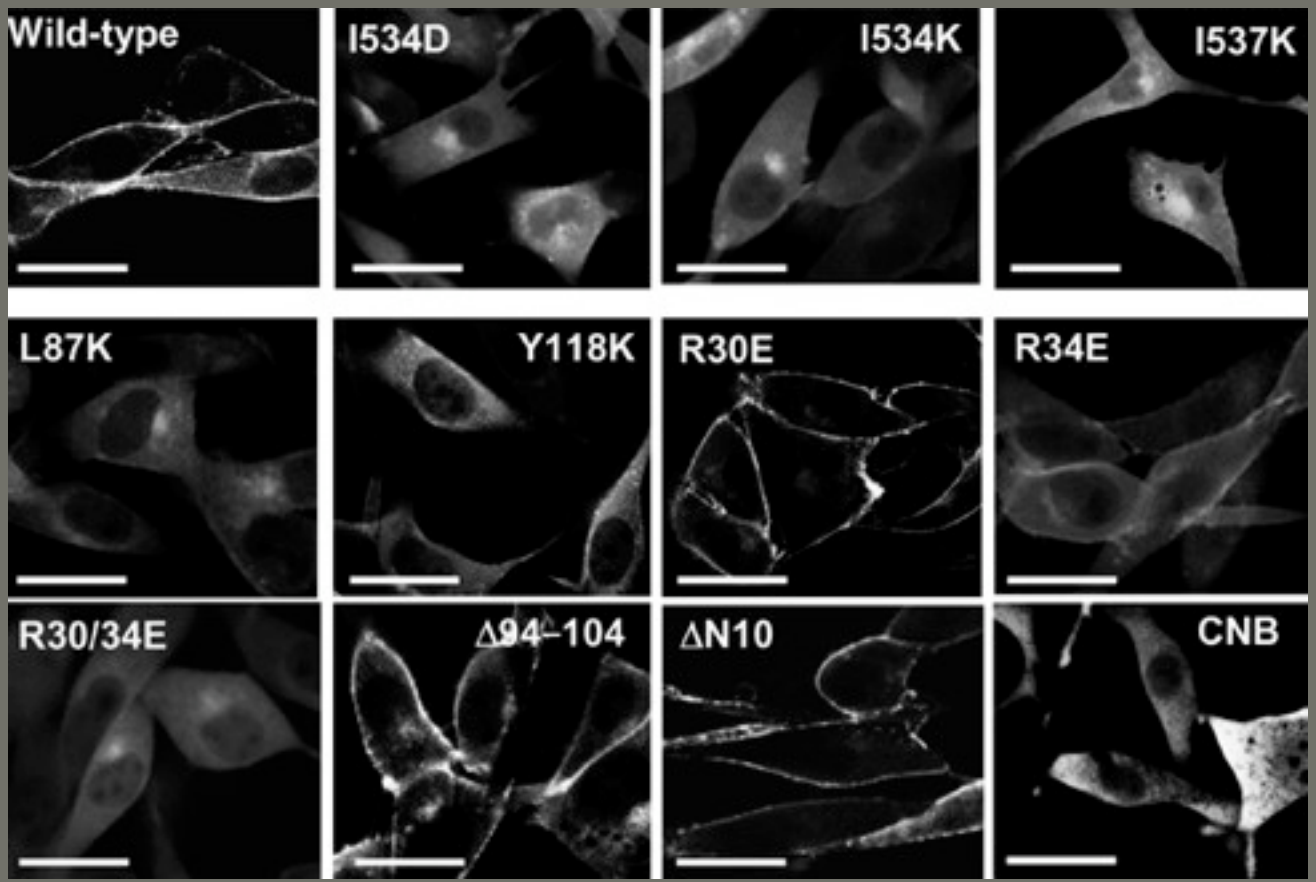
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



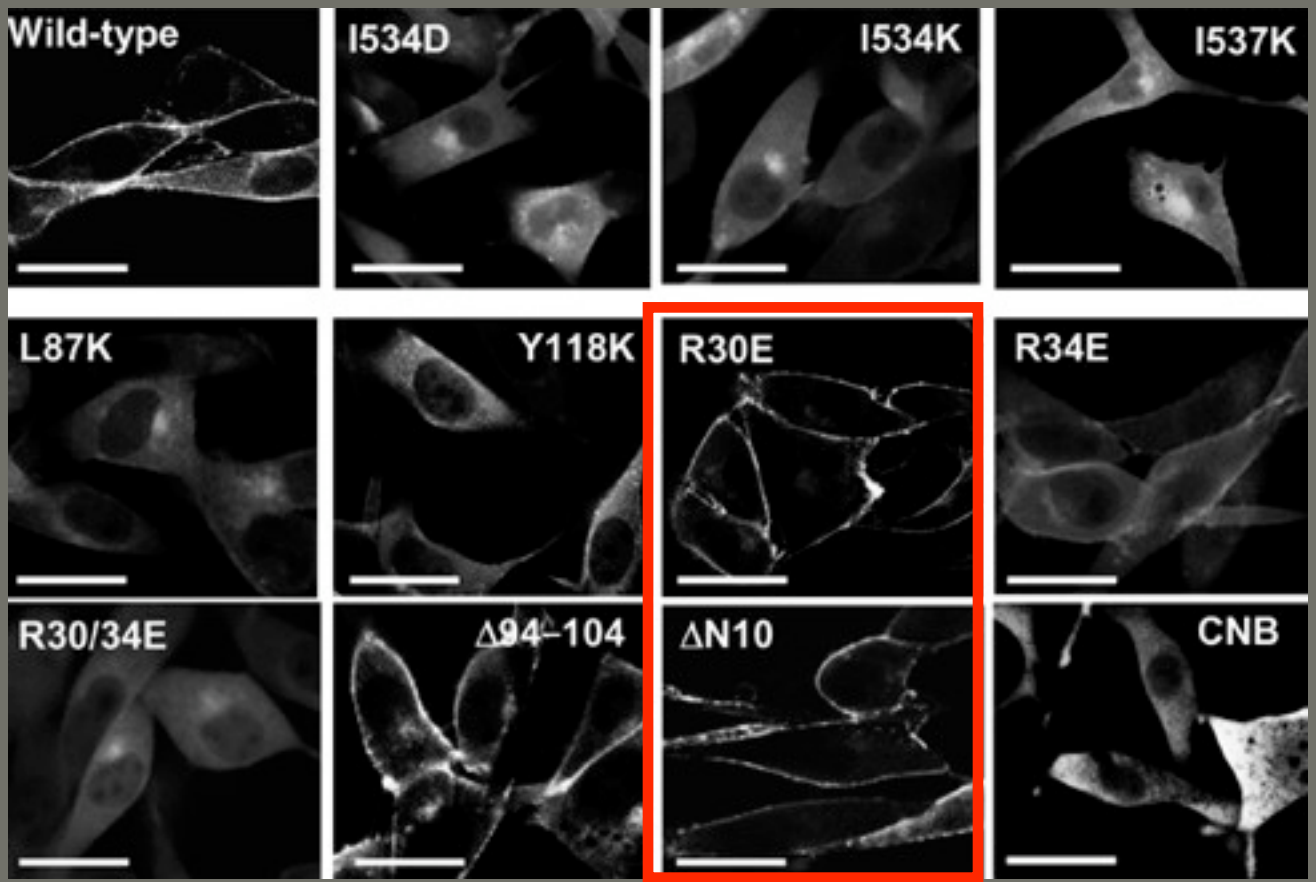
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



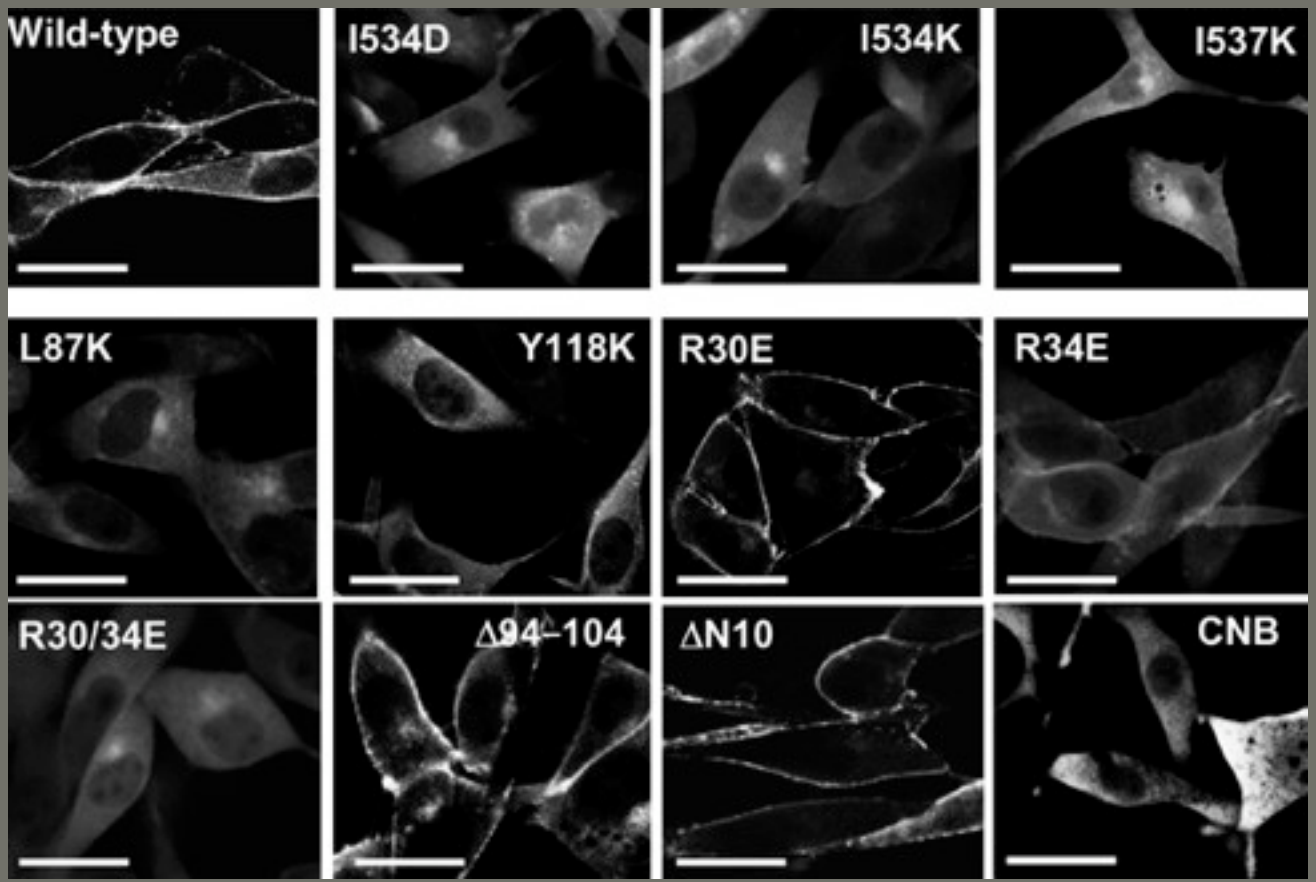
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



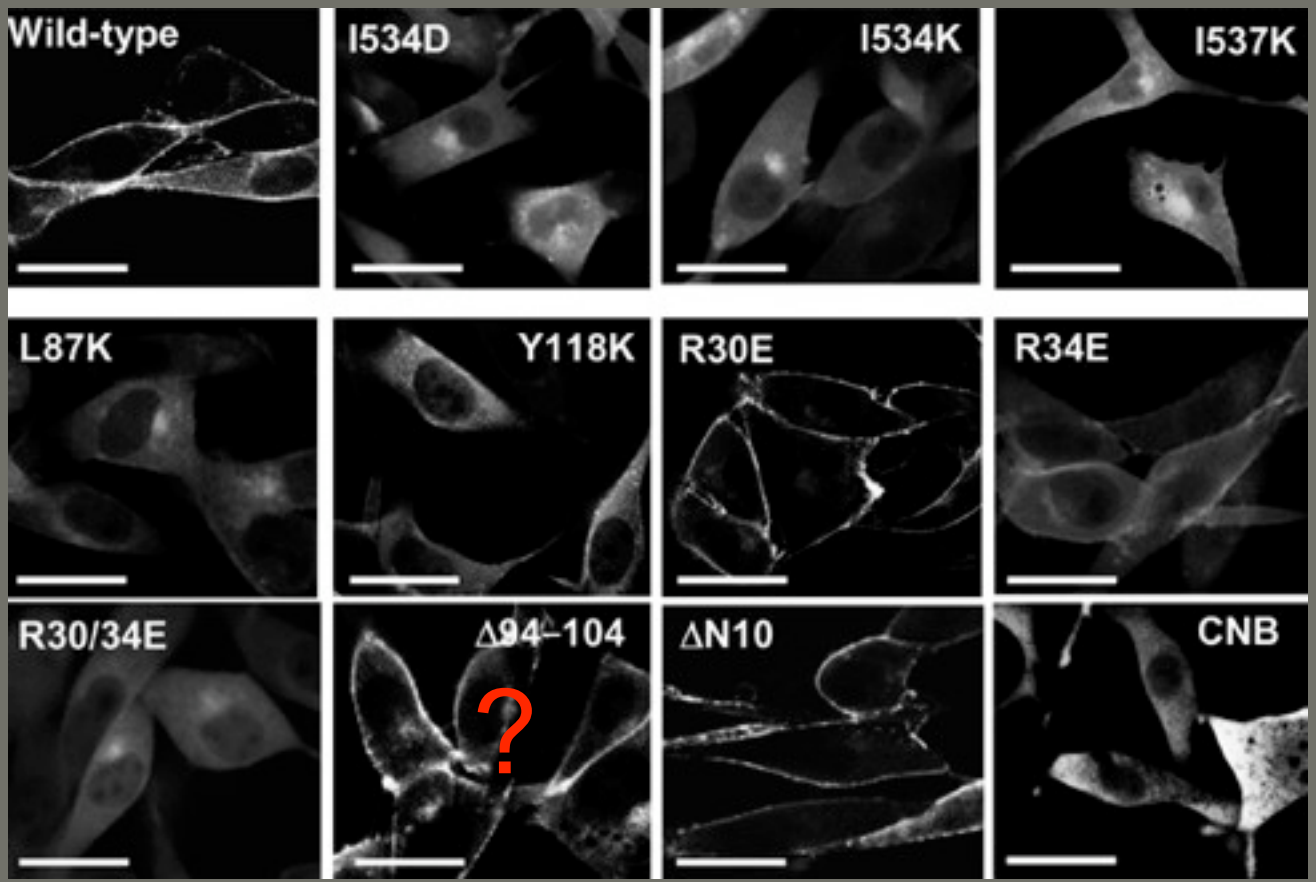
Classification

- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes

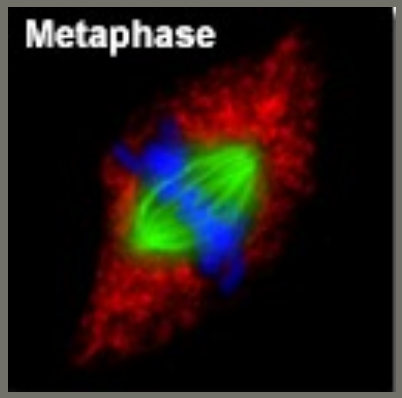


Classification

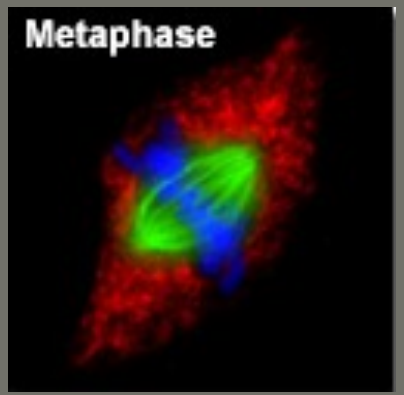
- Unsupervised learning (Cluster analysis, Clustering) seeks to discover the classes



Classification problem

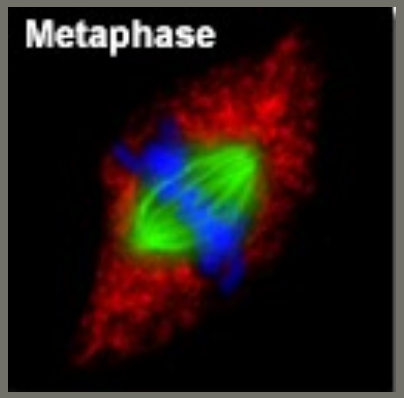


Classification problem



?

Classification problem

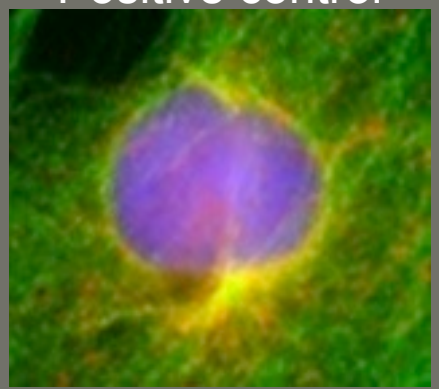


?

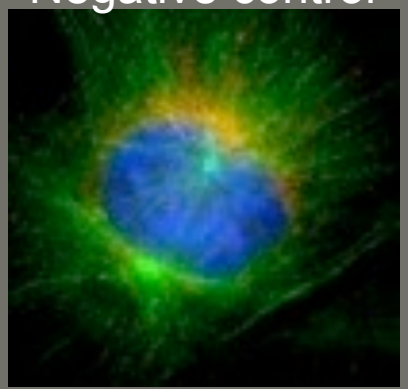
- Supervised learning (Classification) assumes classes are known

2 class problem

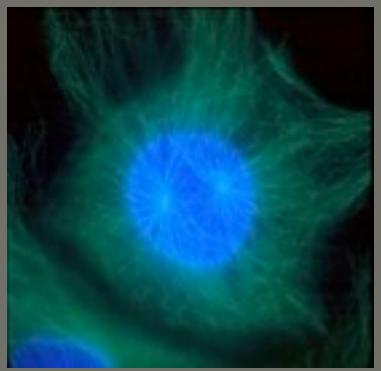
Positive control



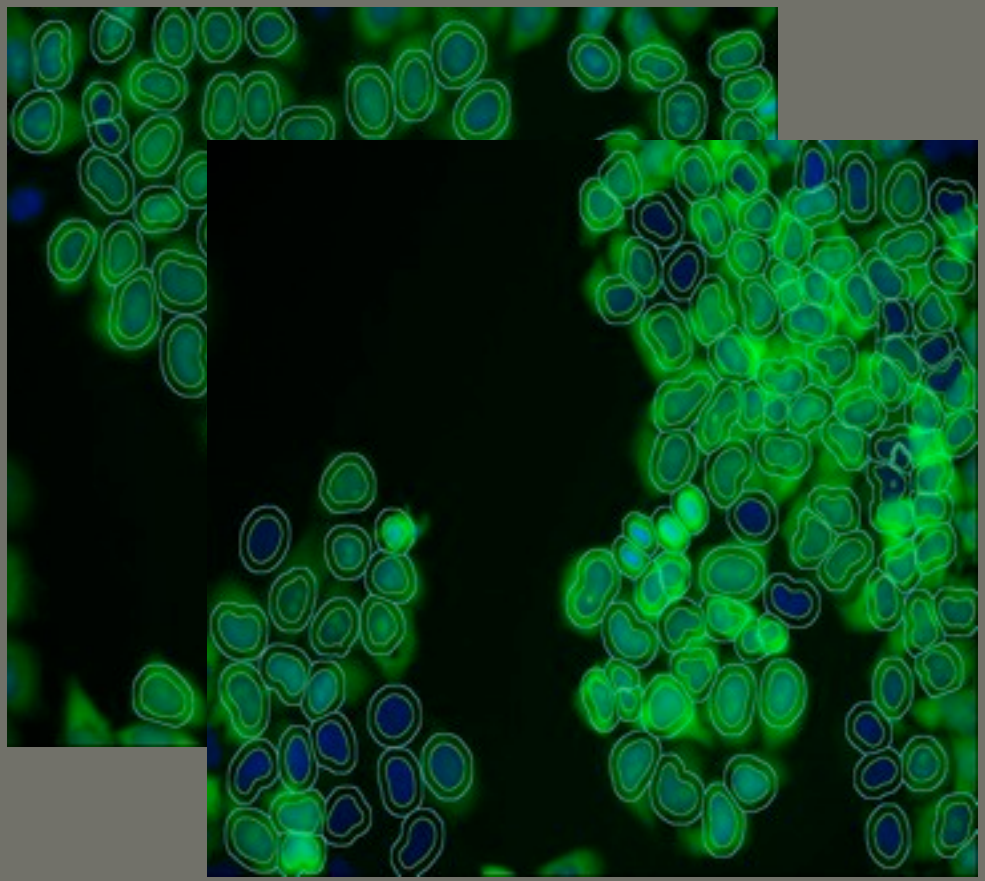
Negative control



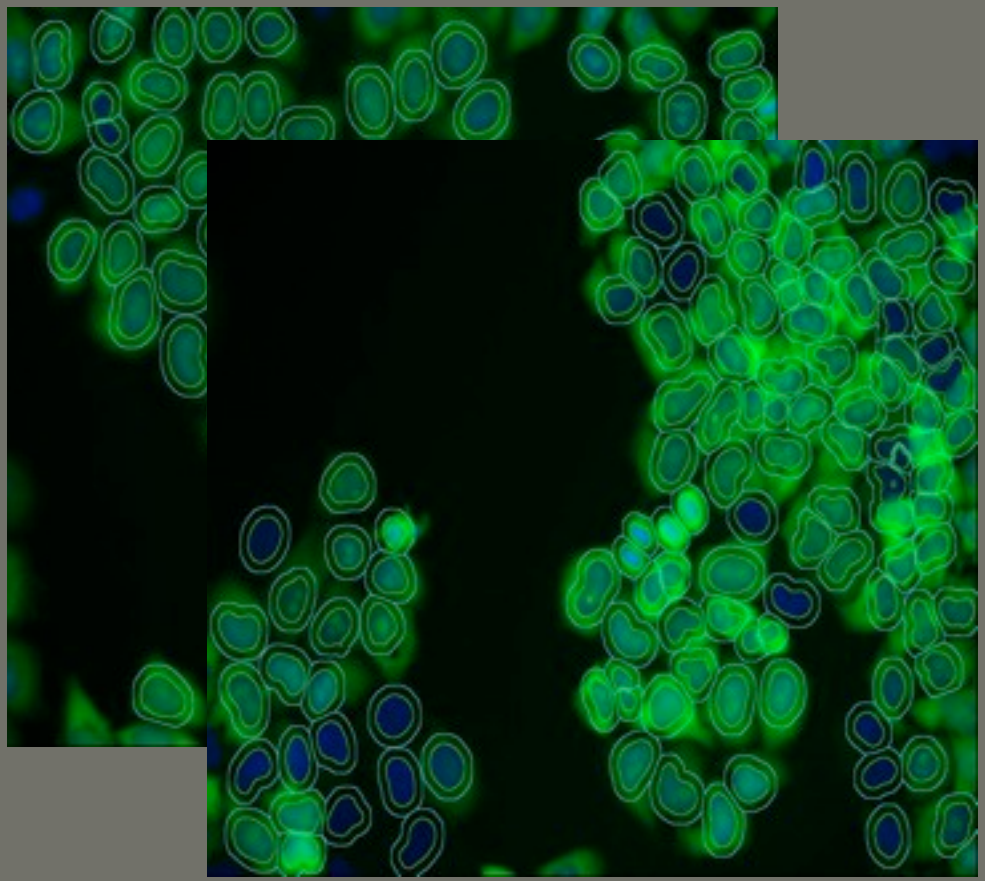
?



Train data for supervised learning

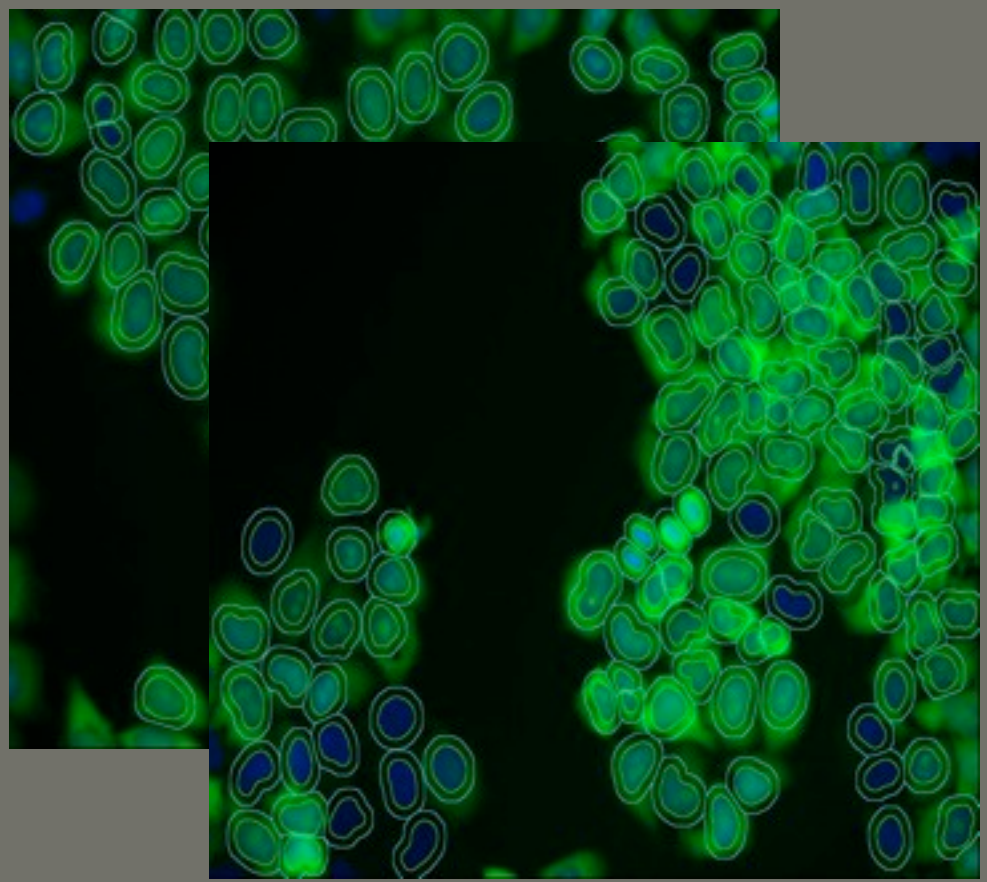


Train data for supervised learning

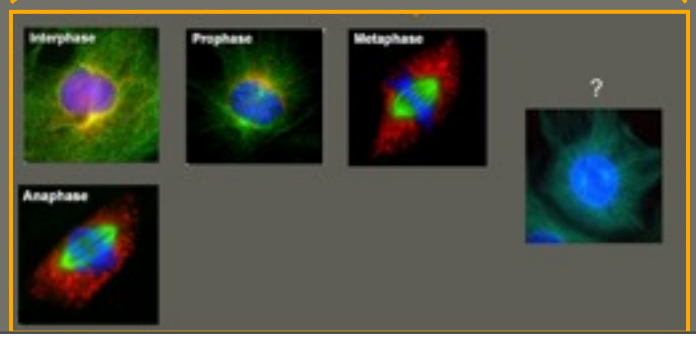


5 people

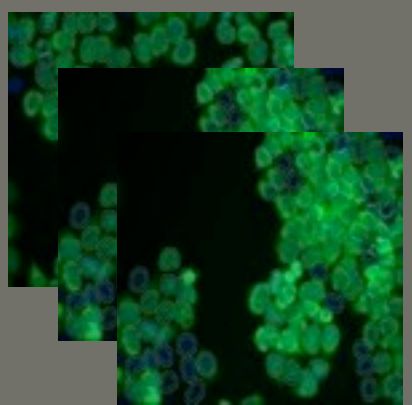
Train data for supervised learning



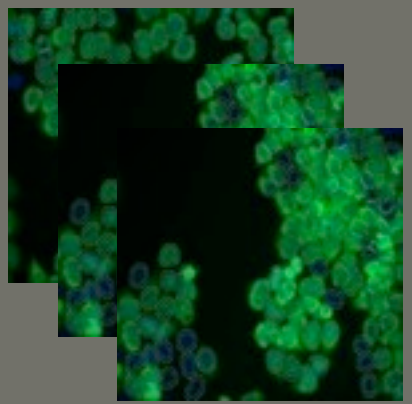
5 people



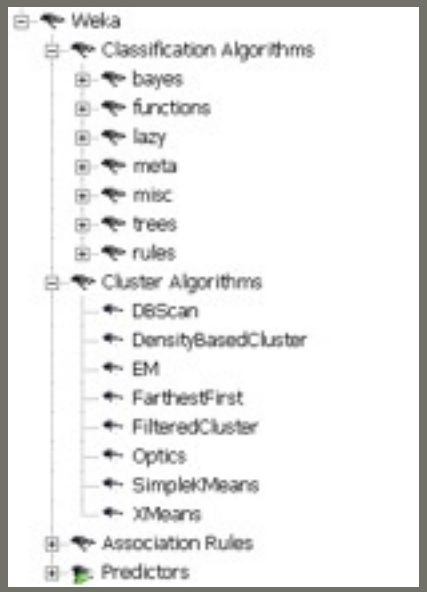
Train data for supervised learning



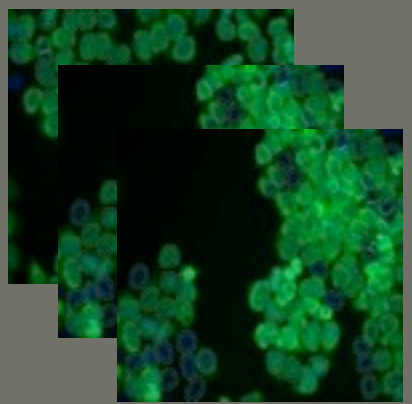
Train data for supervised learning



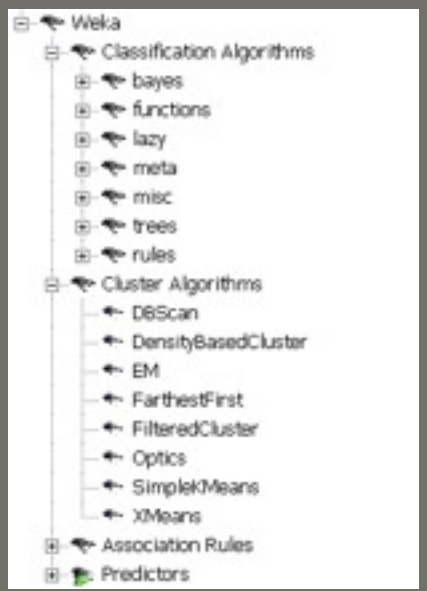
WEKA/R-Project nodes (KNIME)

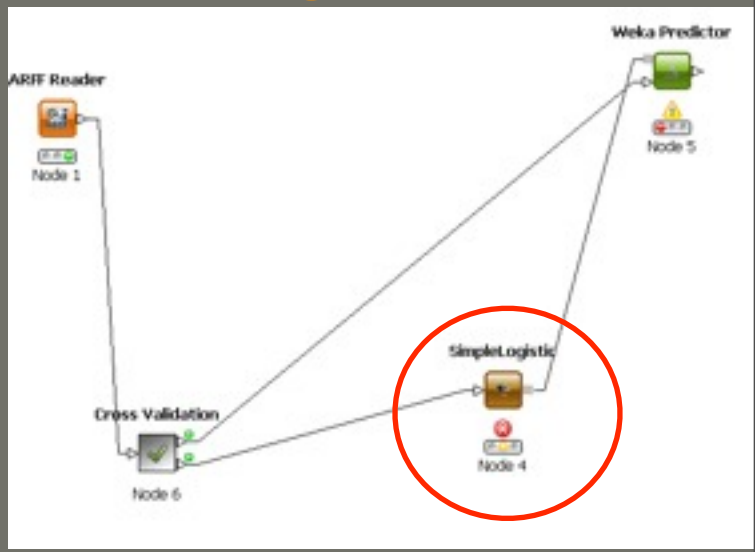


Train data for supervised learning

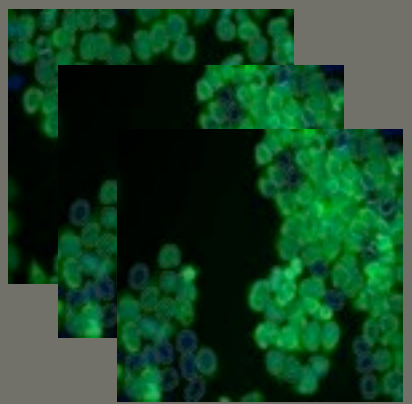


WEKA/R-Project nodes (KNIME)

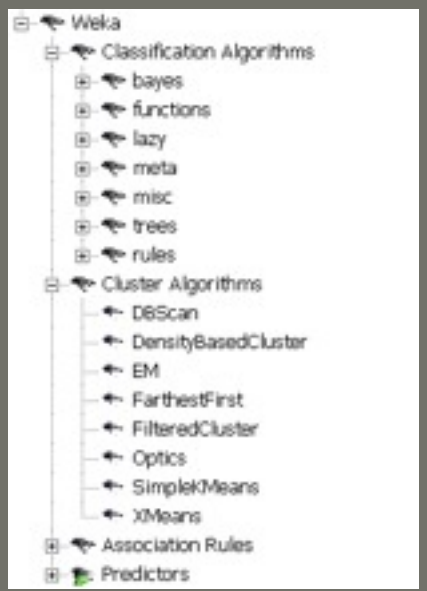


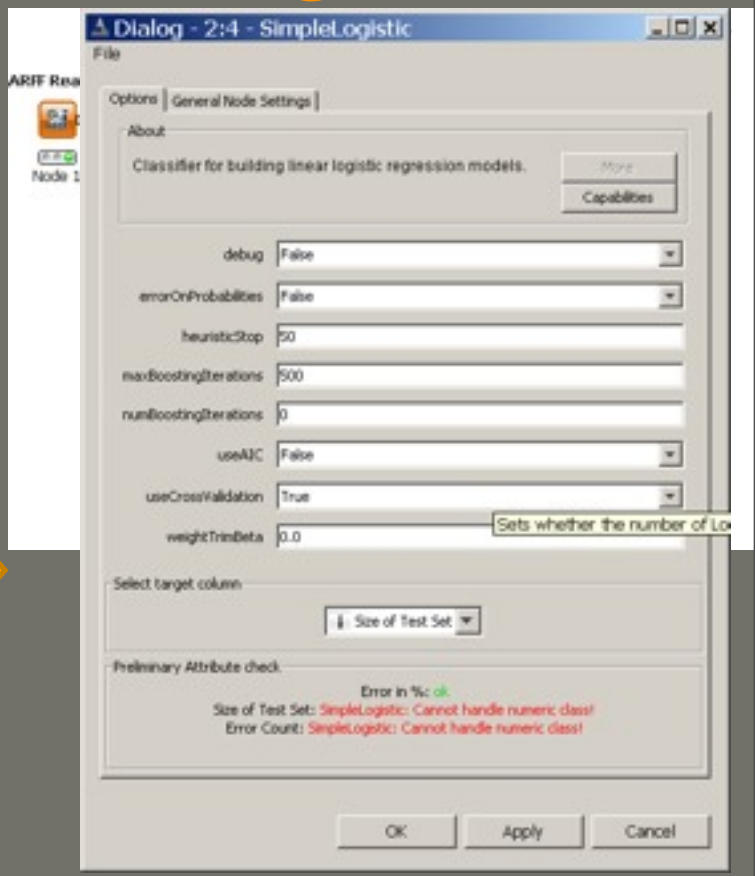


Train data for supervised learning



WEKA/R-Project nodes (KNIME)

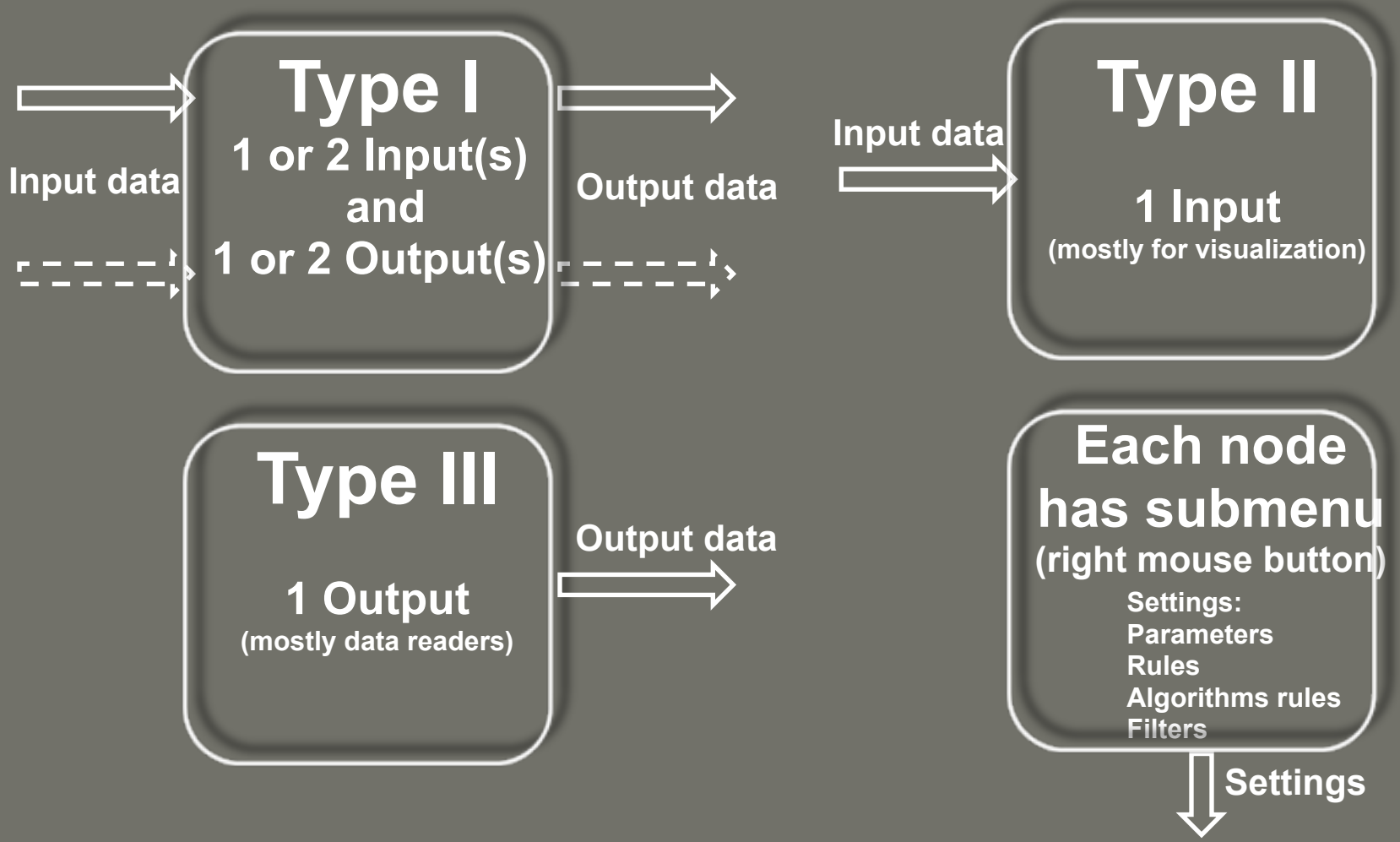




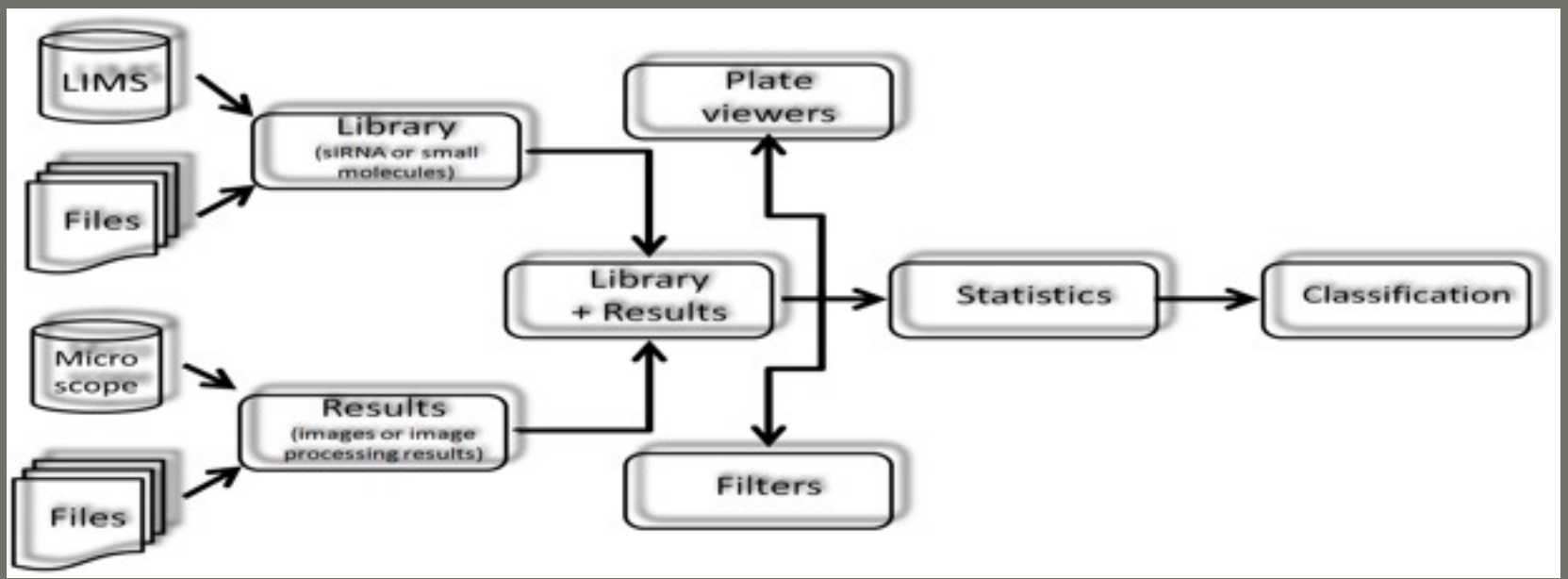
Niels Landwehr, Mark Hall, Eibe Frank (2005). Logistic Model Trees.

92% Accuracy

HC/DC

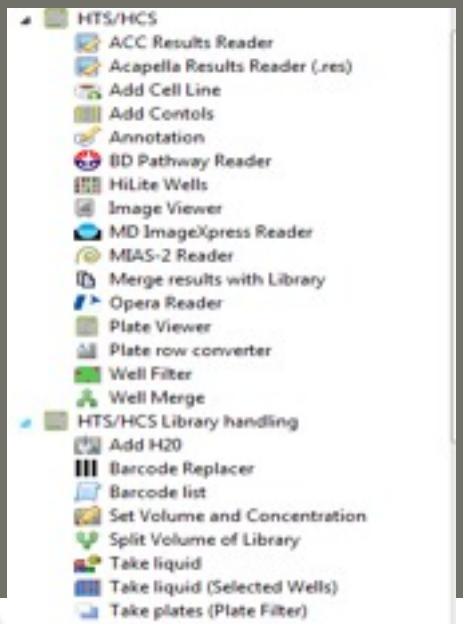

Type I
1 or 2 Input(s)
and
1 or 2 Output(s)
Input data
Output data
Type II
1 Input
(mostly for visualization)
Input data
Type III
1 Output
(mostly data readers)
Output data
Each node
has submenu
(right mouse button)
Settings:
Parameters
Rules
Algorithms rules
Filters
Settings

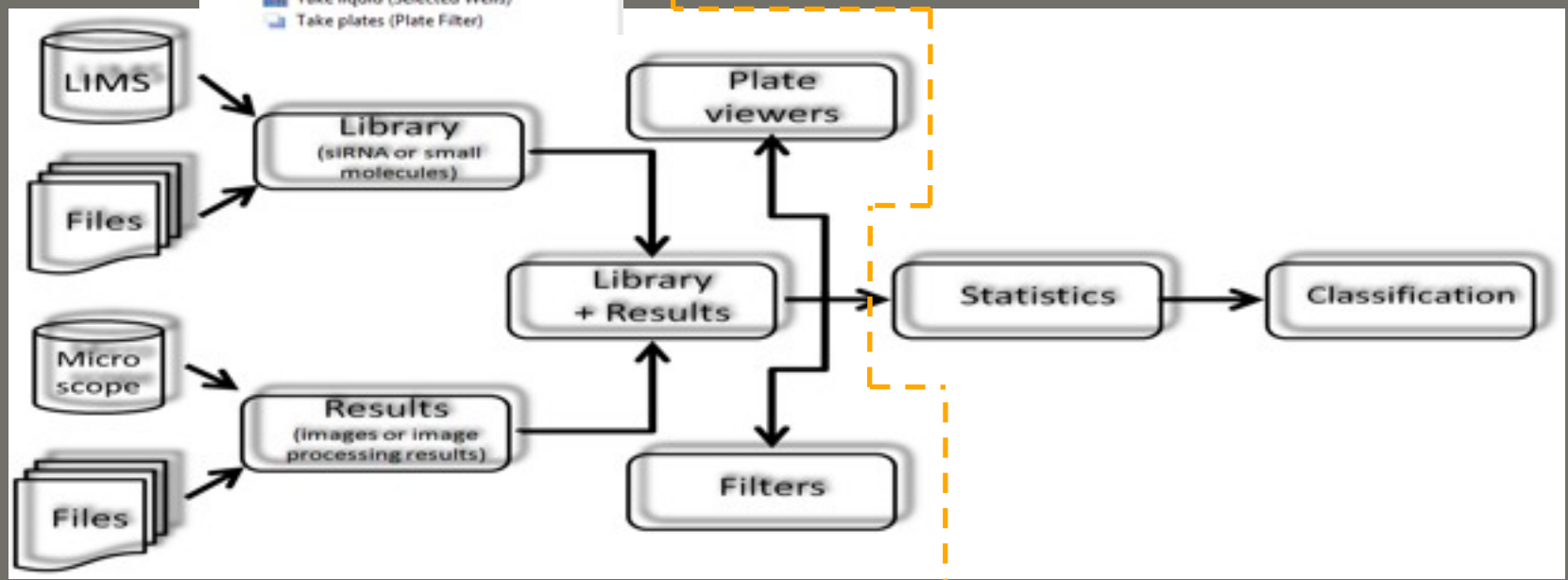
HC/DC


LIMS
Files
Micro scope
Files
Library
(siRNA or small molecules)
Results
(images or image processing results)
Plate viewers
Library + Results
Filters
Statistics
Classification

HC/DC

HC/DC nodes





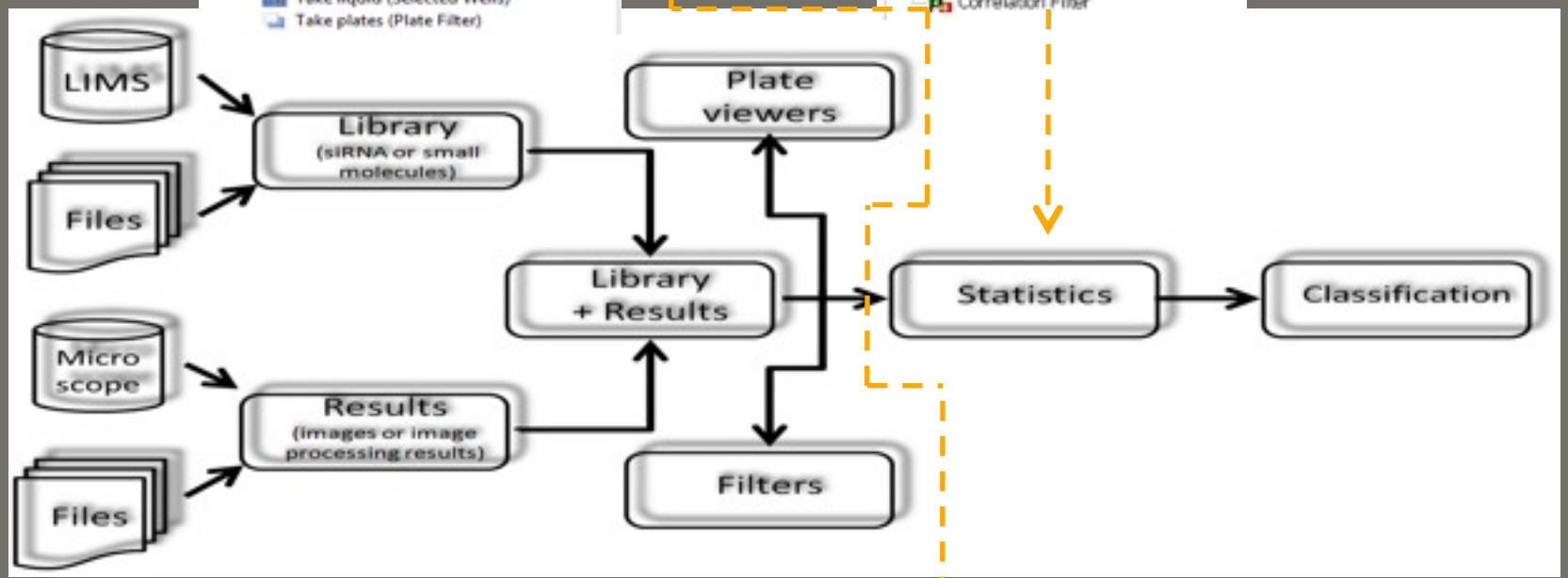
HC/DC

HC/DC nodes

- HTS/HCS
  - ACC Results Reader
  - Acapella Results Reader (.res)
  - Add Cell Line
  - Add Controls
  - Annotation
  - BD Pathway Reader
  - HiLite Wells
  - Image Viewer
  - MD ImageXpress Reader
  - MIAS-2 Reader
  - Merge results with Library
  - Opera Reader
  - Plate Viewer
  - Plate row converter
  - Well Filter
  - Well Merge
- HTS/HCS Library handling
  - Add H20
  - Barcode Replacer
  - Barcode list
  - Set Volume and Concentration
  - Split Volume of Library
  - Take liquid
  - Take liquid (Selected Wells)
  - Take plates (Plate Filter)

KNIME nodes

- Property
- JFreeChart
  - Box Plot
  - Conditional Box Plot
  - Histogram
  - Histogram (interactive)
  - Interactive Table
  - Lift Chart
  - Line Plot
  - Parallel Coordinates
  - Pie chart
  - Pie chart (interactive)
  - Scatter Matrix
  - Scatter Plot
- Statistics
  - Regression
    - Linear Regression (Learner)
    - Polynomial Regression Learner
    - Regression (Predictor)
  - Linear Correlation
  - Statistics View
  - Value Counter
  - Correlation Filter



HC/DC

HC/DC nodes

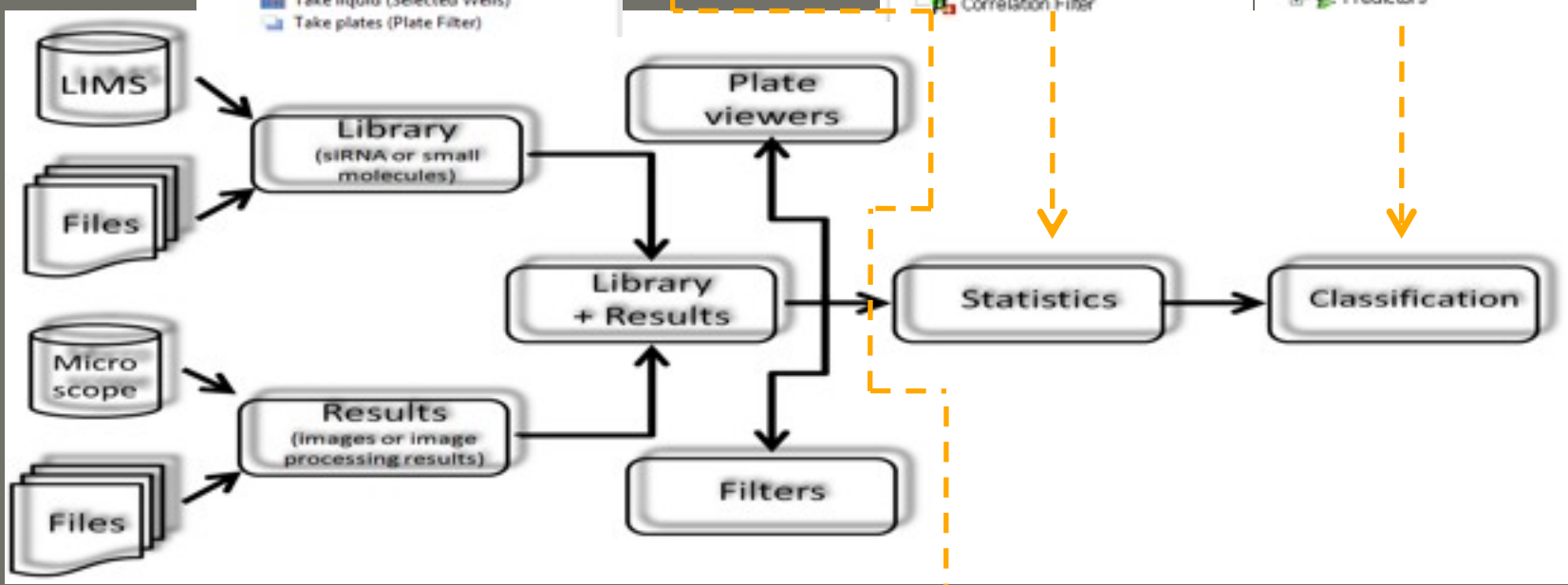
- HTS/HCS
  - ACC Results Reader
  - Acapella Results Reader (.res)
  - Add Cell Line
  - Add Controls
  - Annotation
  - BD Pathway Reader
  - HiLite Wells
  - Image Viewer
  - MD ImageXpress Reader
  - MIAS-2 Reader
  - Merge results with Library
  - Opera Reader
  - Plate Viewer
  - Plate row converter
  - Well Filter
  - Well Merge
- HTS/HCS Library handling
  - Add H20
  - Barcode Replacer
  - Barcode list
  - Set Volume and Concentration
  - Split Volume of Library
  - Take liquid
  - Take liquid (Selected Wells)
  - Take plates (Plate Filter)

KNIME nodes

- Property
- JFreeChart
- Box Plot
- Conditional Box Plot
- Histogram
- Histogram (interactive)
- Interactive Table
- Lift Chart
- Line Plot
- Parallel Coordinates
- Pie chart
- Pie chart (interactive)
- Scatter Matrix
- Scatter Plot
- Statistics
- Regression
  - Linear Regression (Learner)
  - Polynomial Regression Learner
  - Regression (Predictor)
- Linear Correlation
- Statistics View
- Value Counter
- Correlation Filter

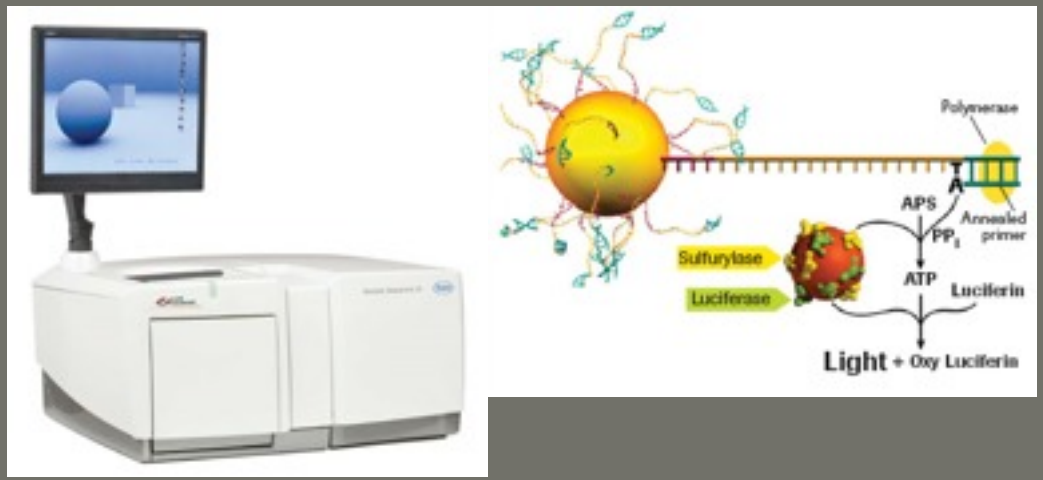
WEKA/R-Project nodes (KNIME)

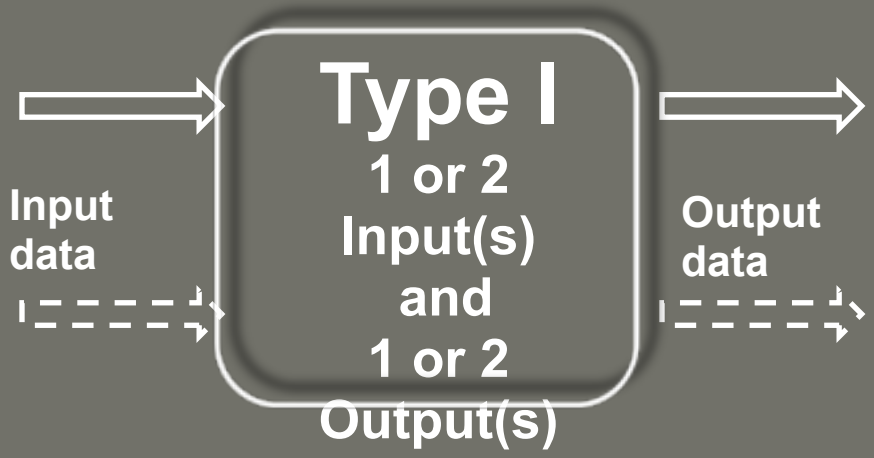
- Weka
  - Classification Algorithms
    - bayes
    - functions
    - lazy
    - meta
    - misc
    - trees
    - rules
  - Cluster Algorithms
    - DBScan
    - DensityBasedCluster
    - EM
    - FarthestFirst
    - FilteredCluster
    - Optics
    - SimpleKMeans
    - XMeans
  - Association Rules
  - Predictors



Next-Generation sequencing nodes

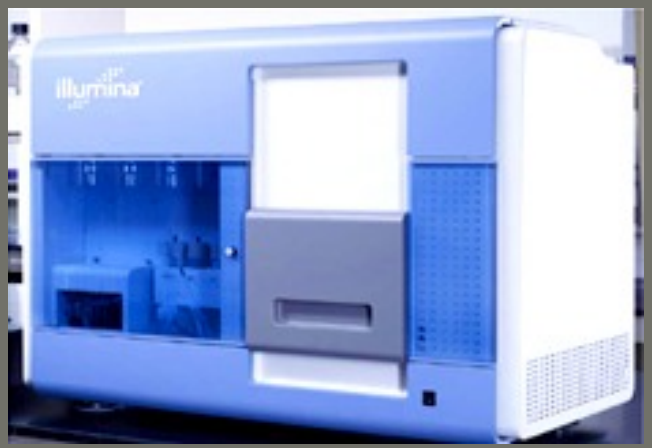
Roche 454





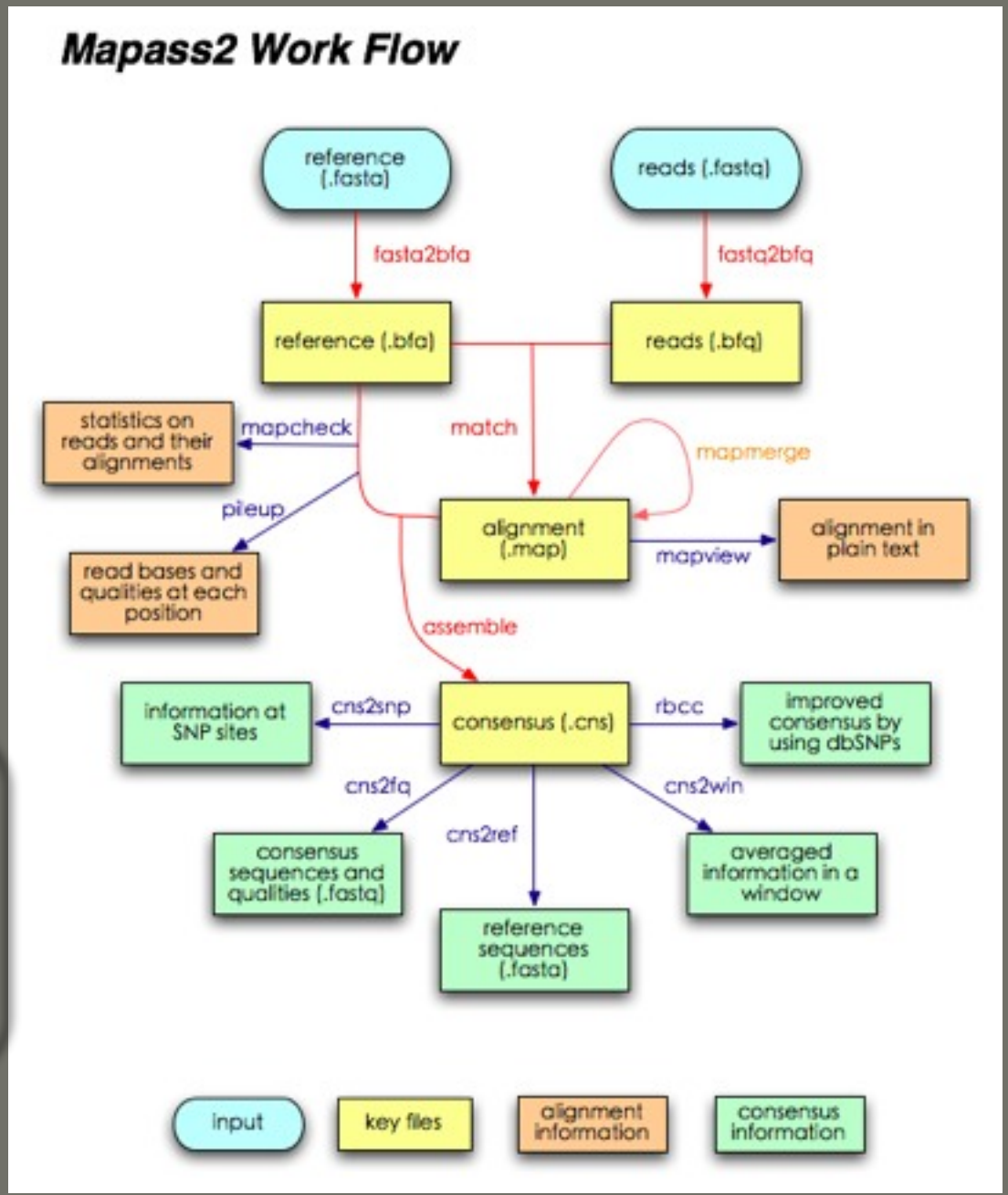
Next-Generation sequencing nodes

SOLEXA
Illumina

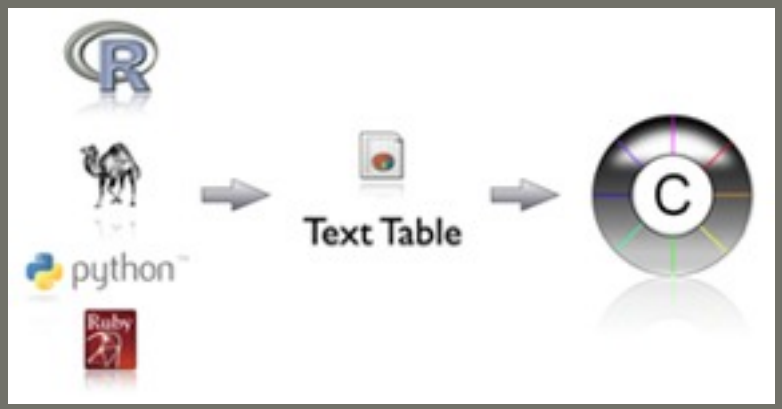


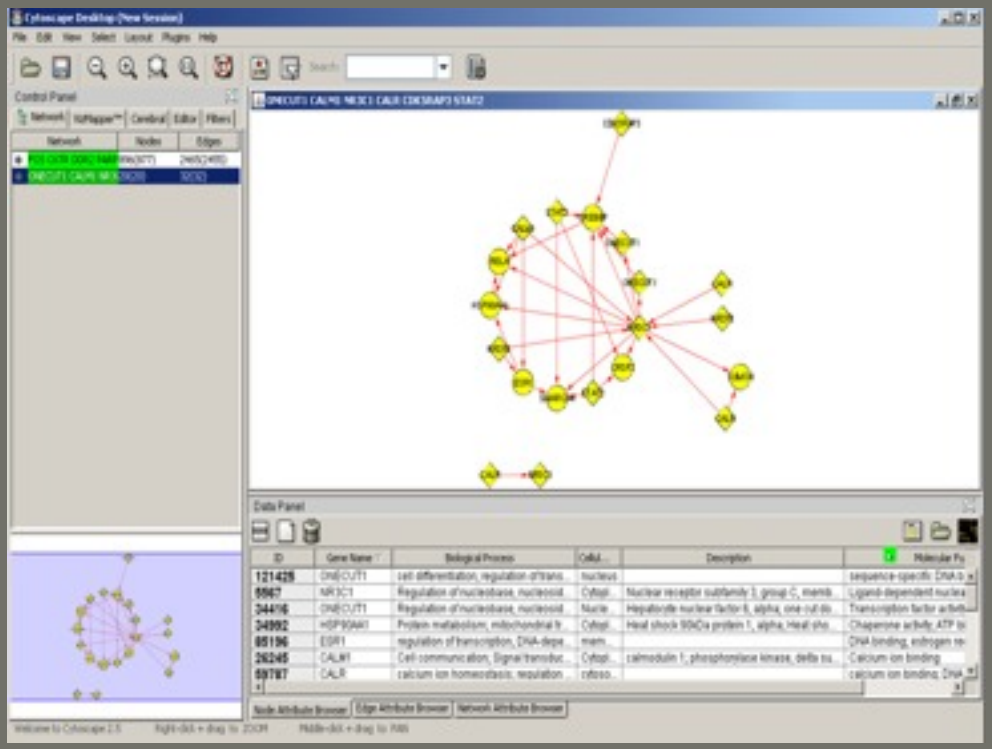
Input data

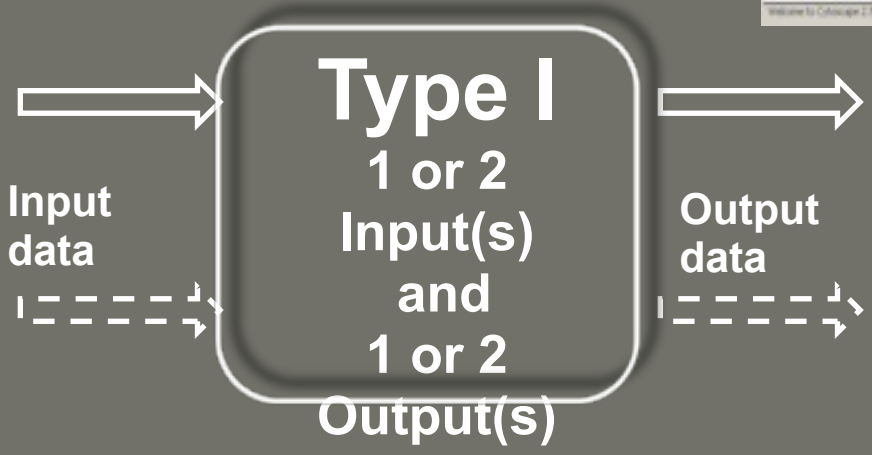
Type I
1 or 2 Input(s) and 1 or 2 Output(s)



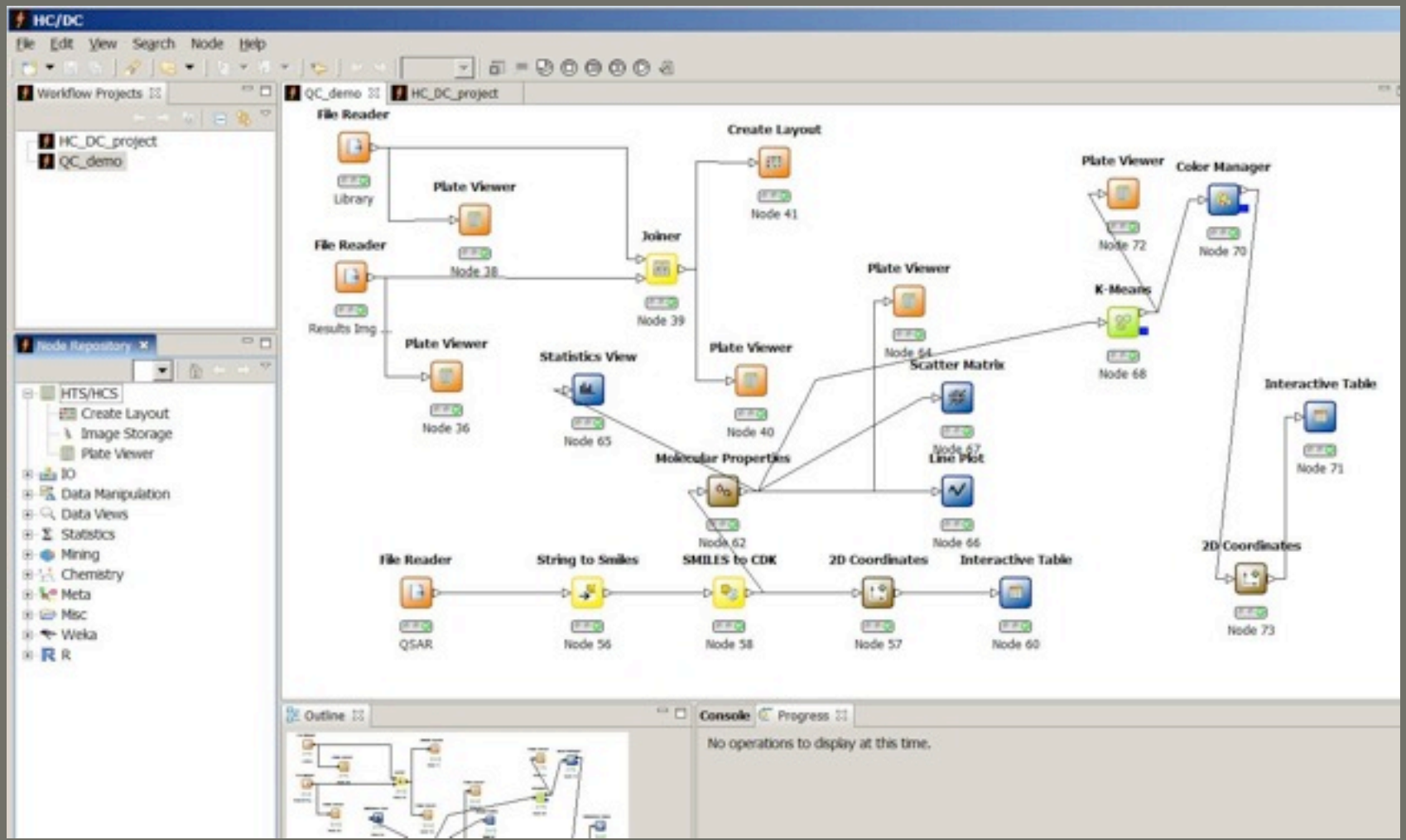
Proteomics


Text Table
python

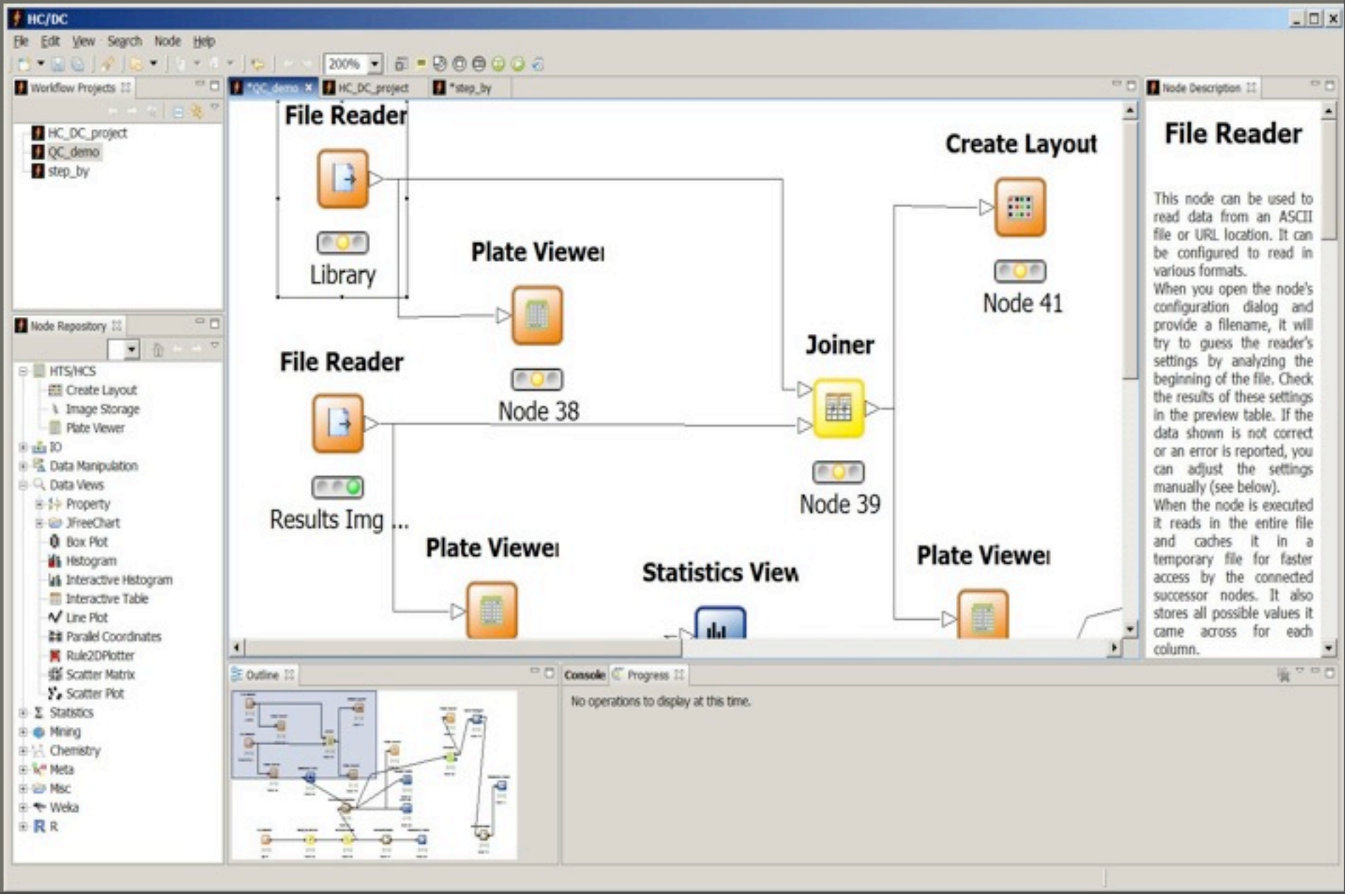



Type I
1 or 2
Input(s)
and
1 or 2
Output(s)
Input
data
Output
data

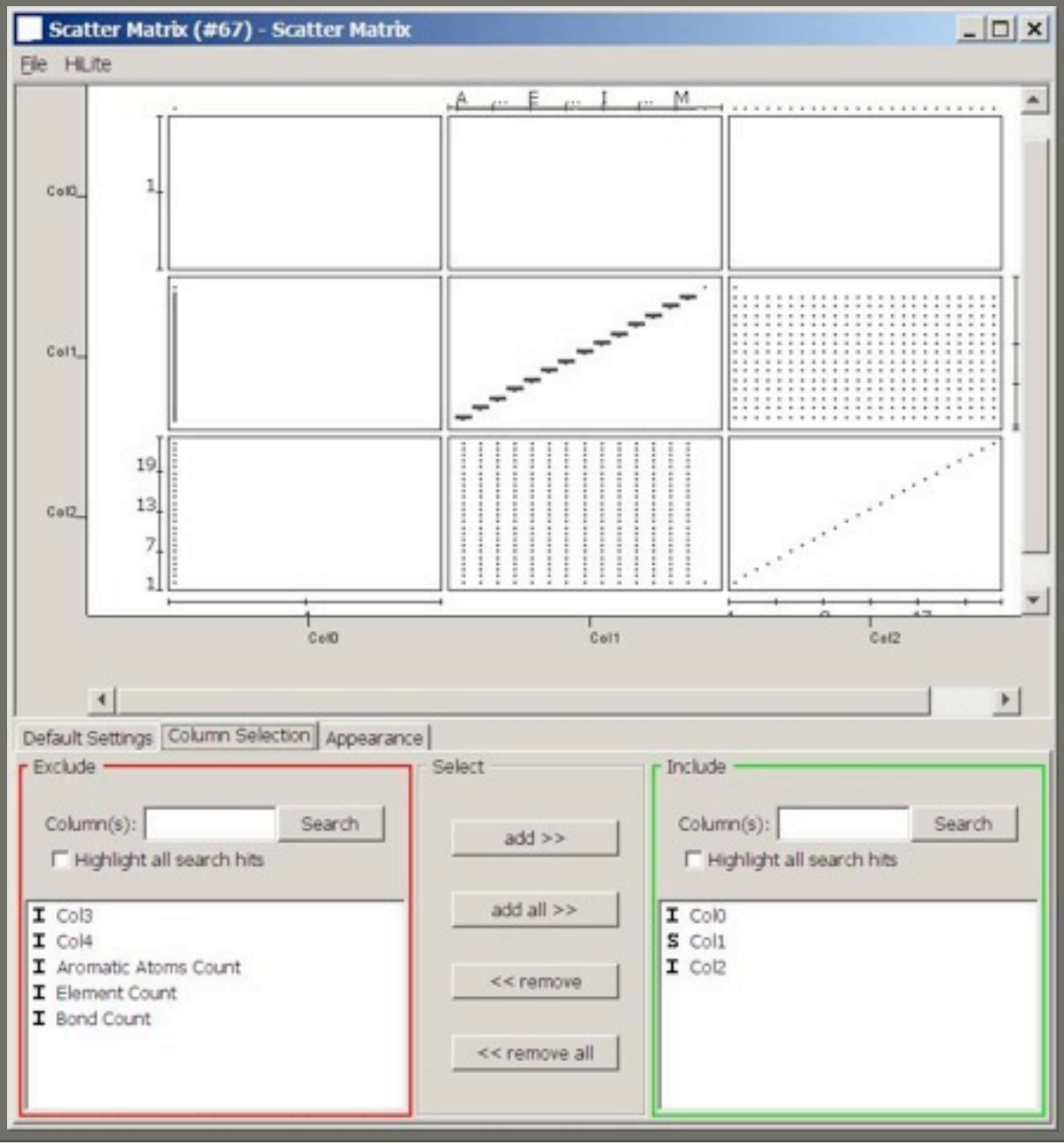
Drag-and-drop



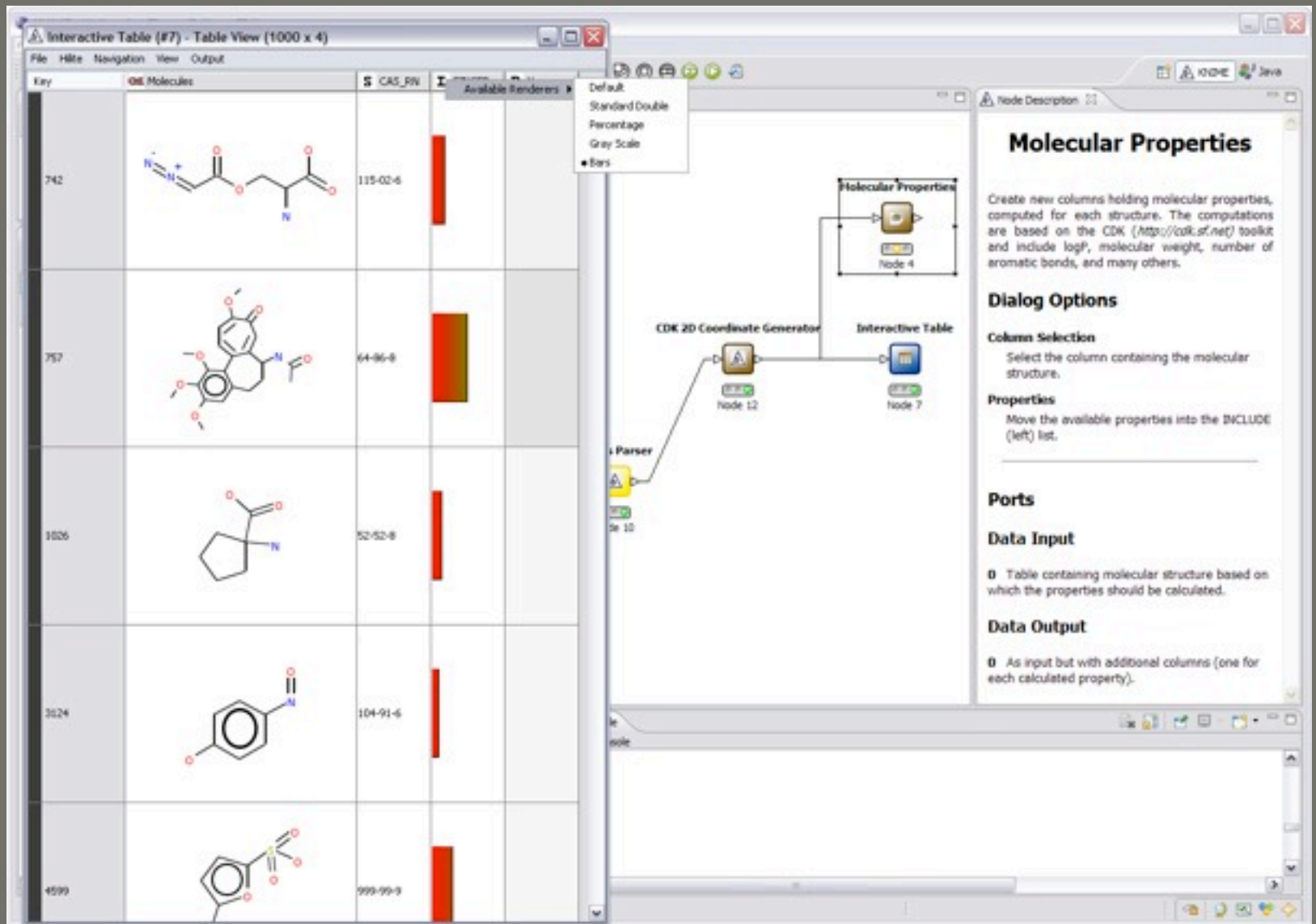
Pioneering study



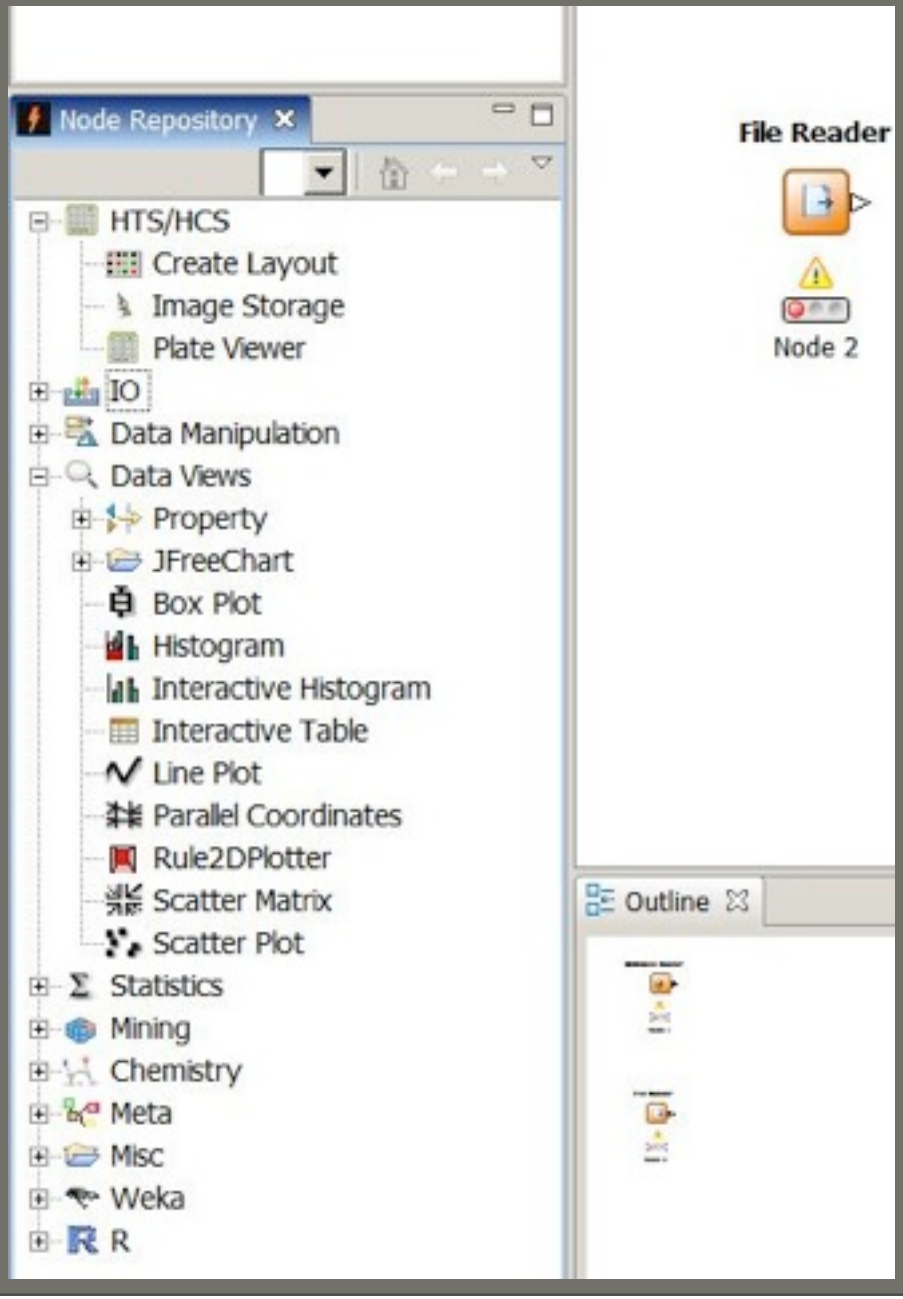
Plot image parameters/descriptors



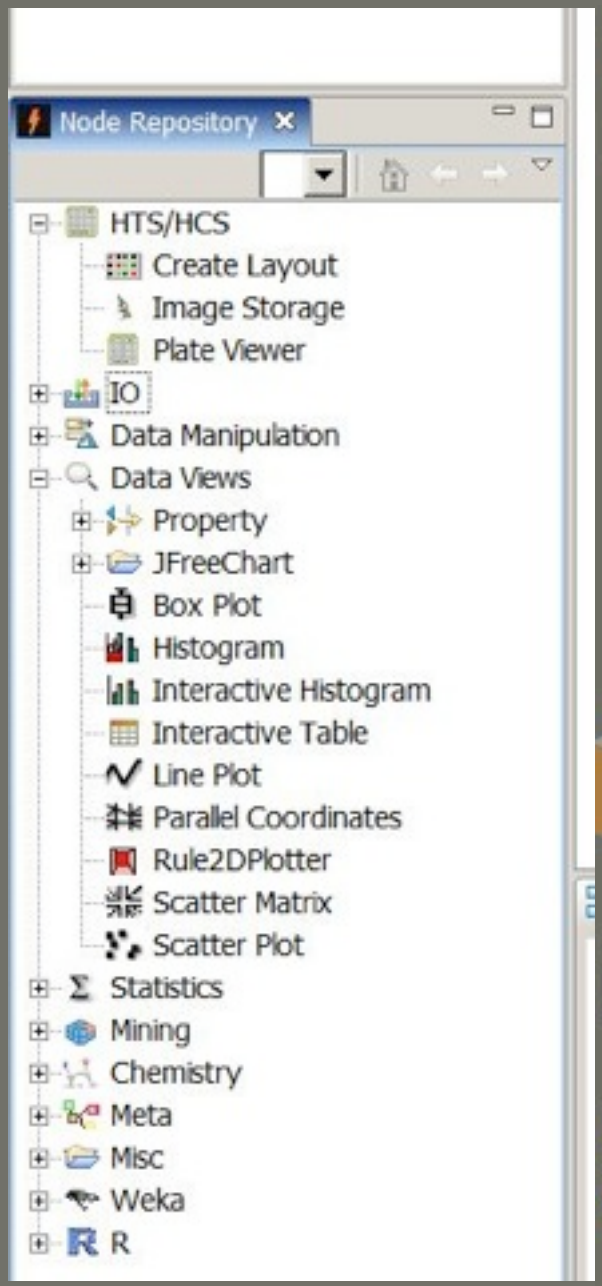
Compound screening

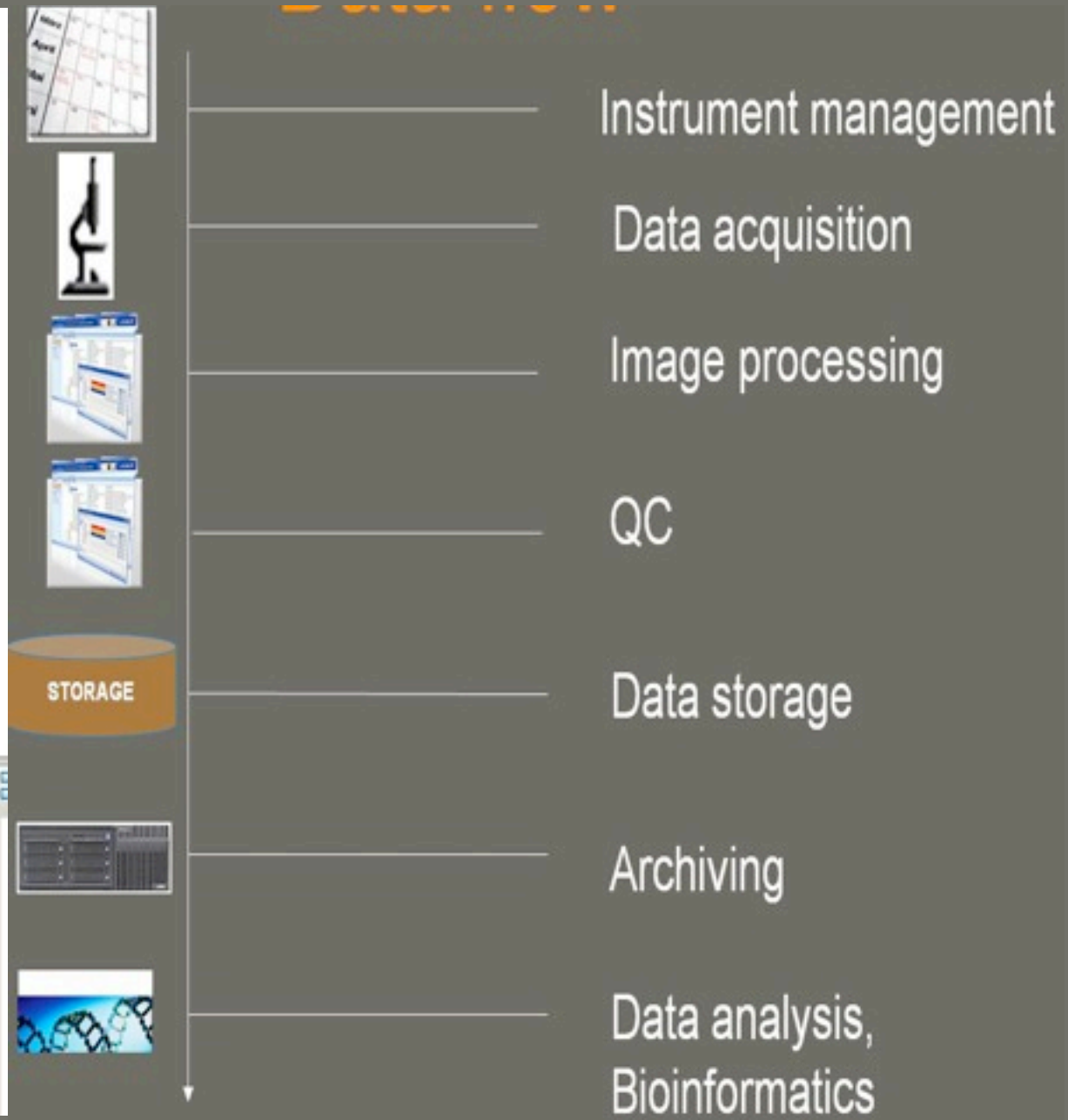


Menu HC/DC

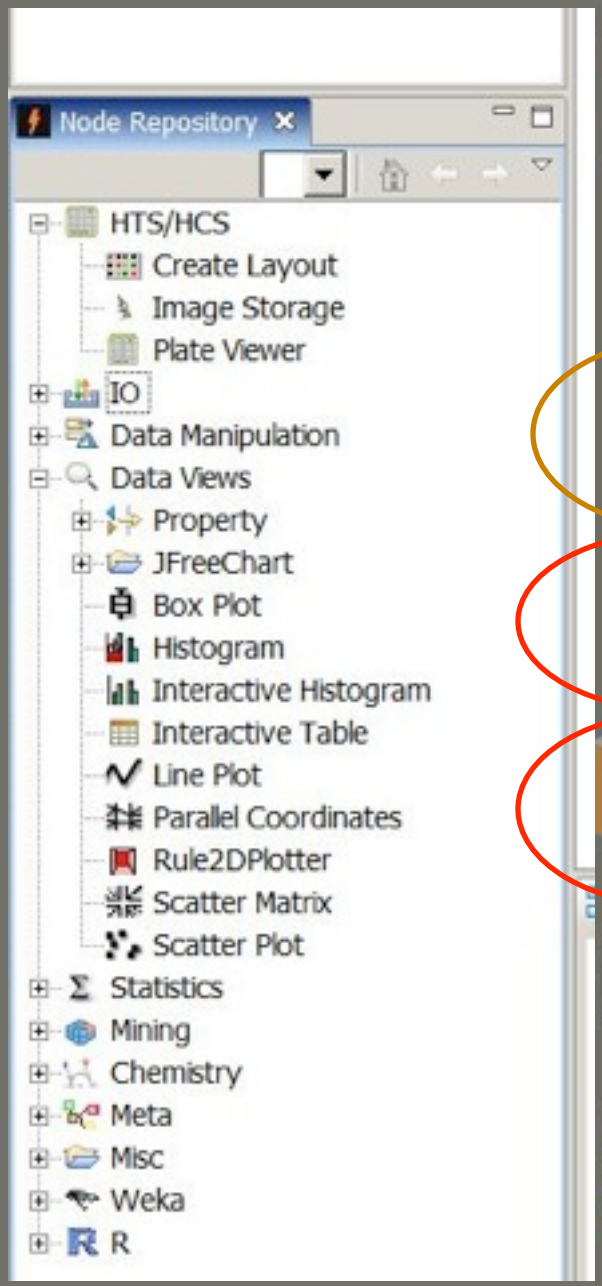

Node Repository
HTS/HCS
Create Layout
Image Storage
Plate Viewer
IO
Data Manipulation
Data Views
Property
JFreeChart
Box Plot
Histogram
Interactive Histogram
Interactive Table
Line Plot
Parallel Coordinates
Rule2DPlotter
Scatter Matrix
Scatter Plot
Statistics
Mining
Chemistry
Meta
Misc
Weka
R
File Reader
Node 2
Outline

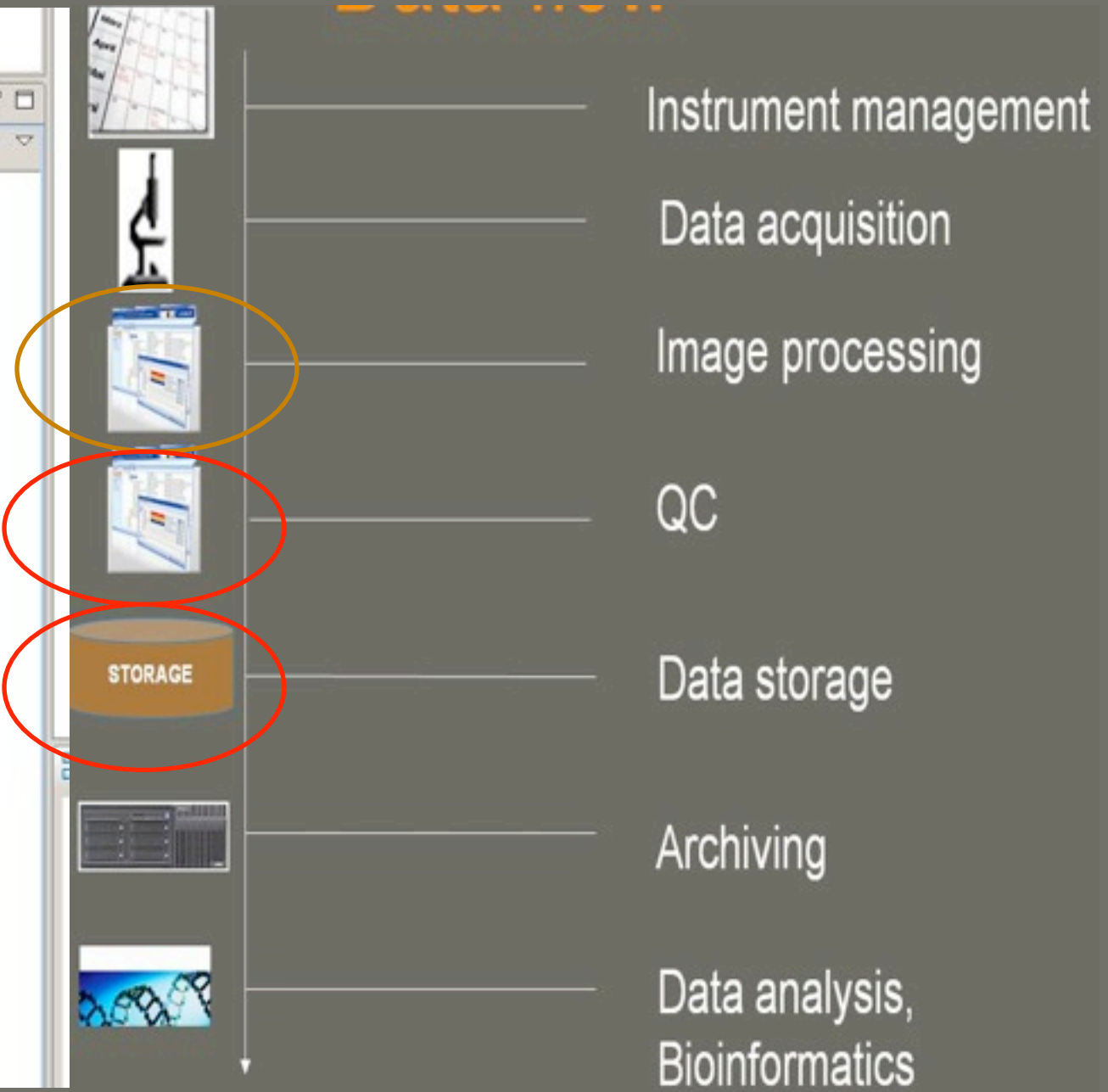
Menu HC/DC


Node Repository
HTS/HCS
Create Layout
Image Storage
Plate Viewer
IO
Data Manipulation
Data Views
Property
JFreeChart
Box Plot
Histogram
Interactive Histogram
Interactive Table
Line Plot
Parallel Coordinates
Rule2DPlotter
Scatter Matrix
Scatter Plot
Statistics
Mining
Chemistry
Meta
Misc
Weka
R

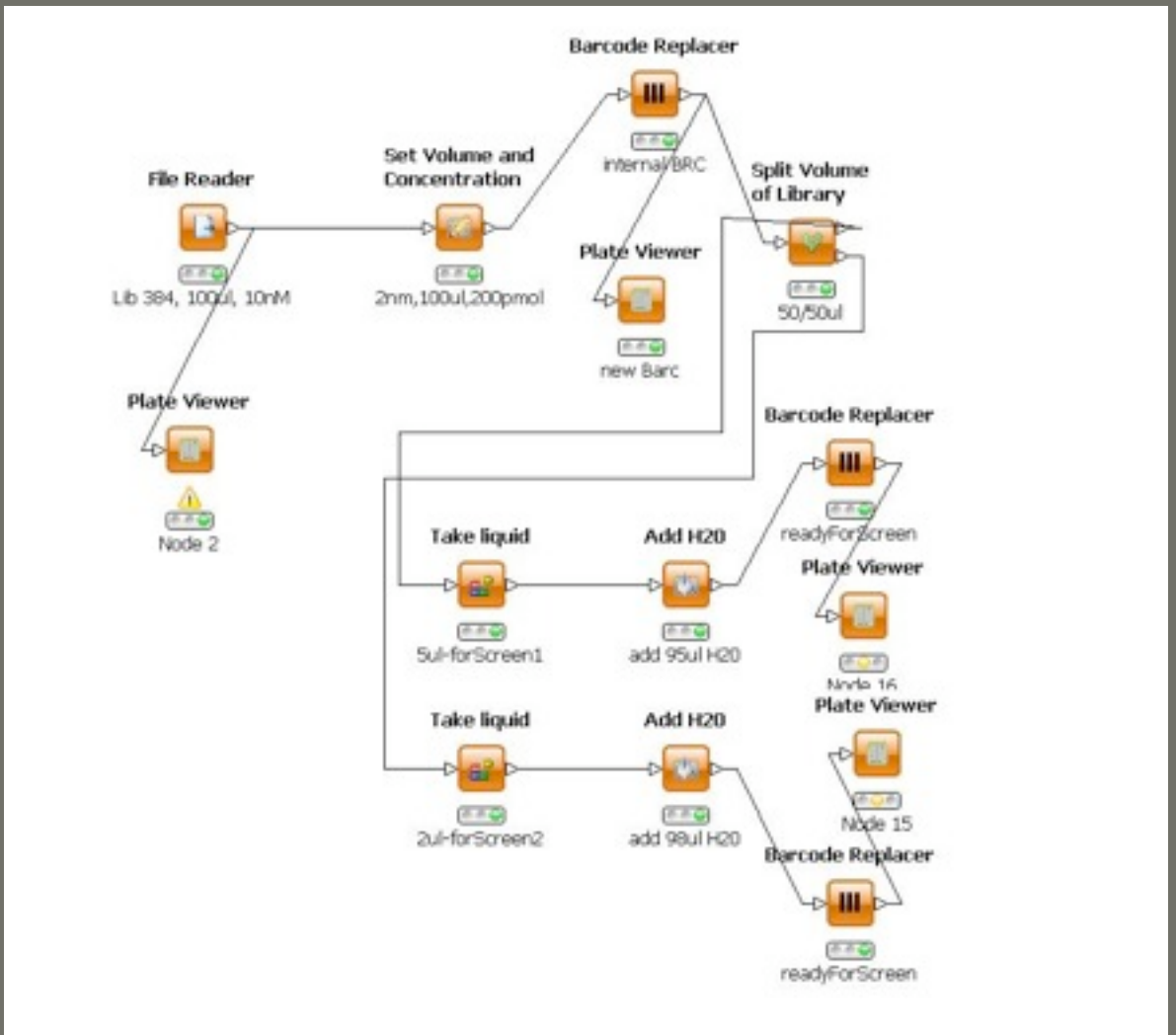

STORAGE
Instrument management
Data acquisition
Image processing
QC
Data storage
Archiving
Data analysis,
Bioinformatics

Menu HC/DC


Node Repository
HTS/HCS
Create Layout
Image Storage
Plate Viewer
IO
Data Manipulation
Data Views
Property
JFreeChart
Box Plot
Histogram
Interactive Histogram
Interactive Table
Line Plot
Parallel Coordinates
Rule2DPlotter
Scatter Matrix
Scatter Plot
Statistics
Mining
Chemistry
Meta
Misc
Weka
R

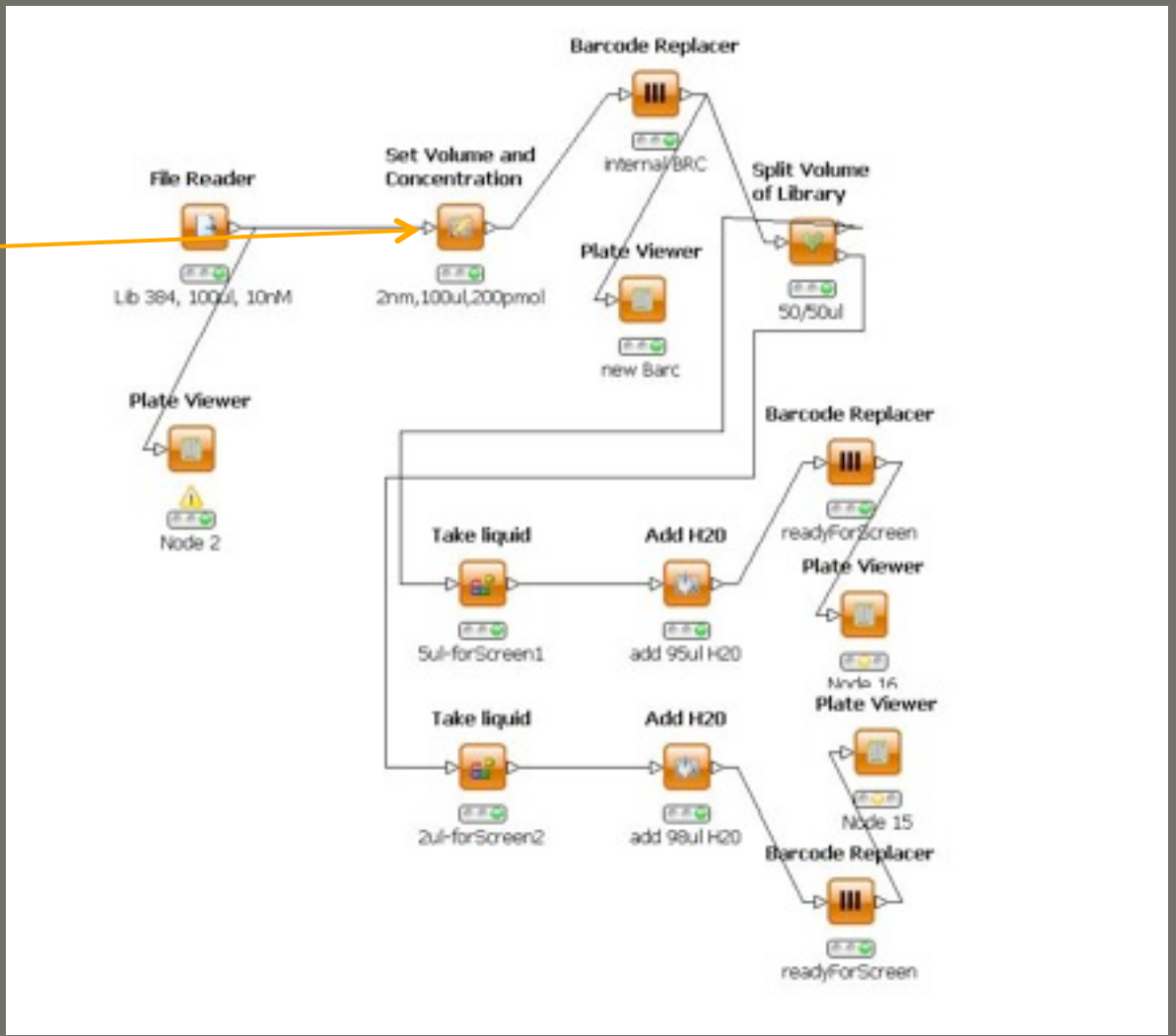

STORAGE
Instrument management
Data acquisition
Image processing
QC
Data storage
Archiving
Data analysis,
Bioinformatics

Library Handling



Library Handling

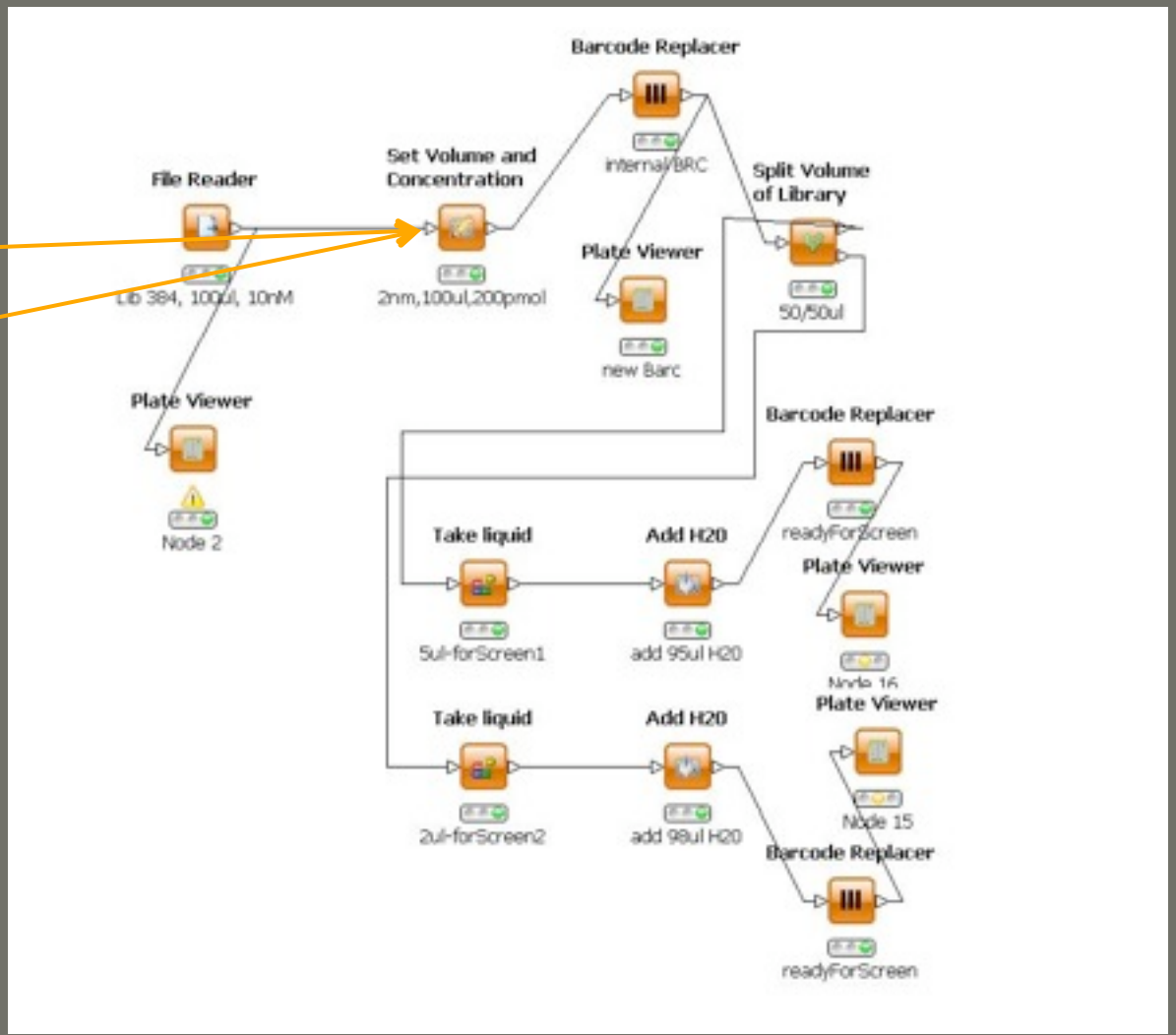
Data Automation:

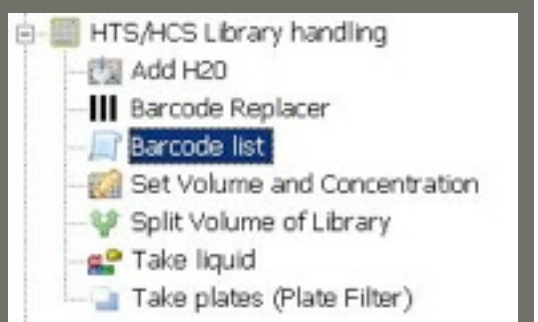


Library Handling

Data Automation:

- Volume

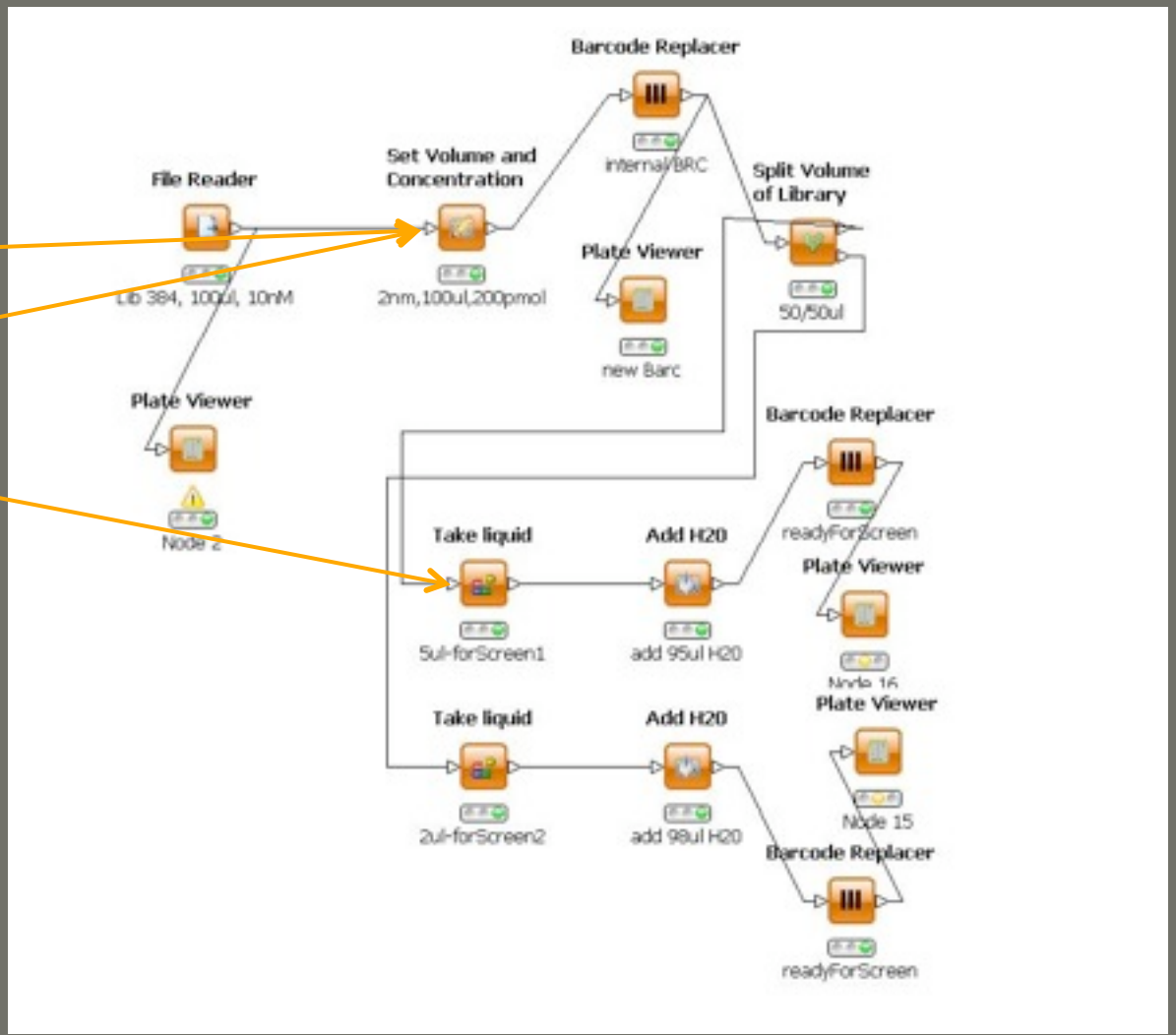


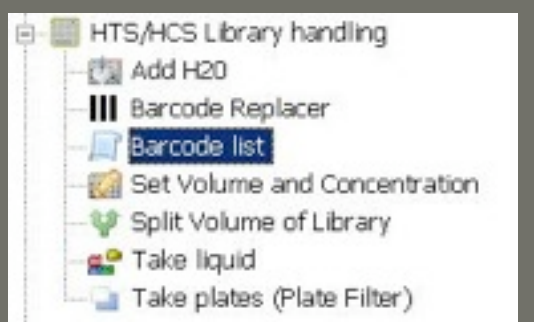


Library Handling

Data Automation:

- Volume
- Concentration

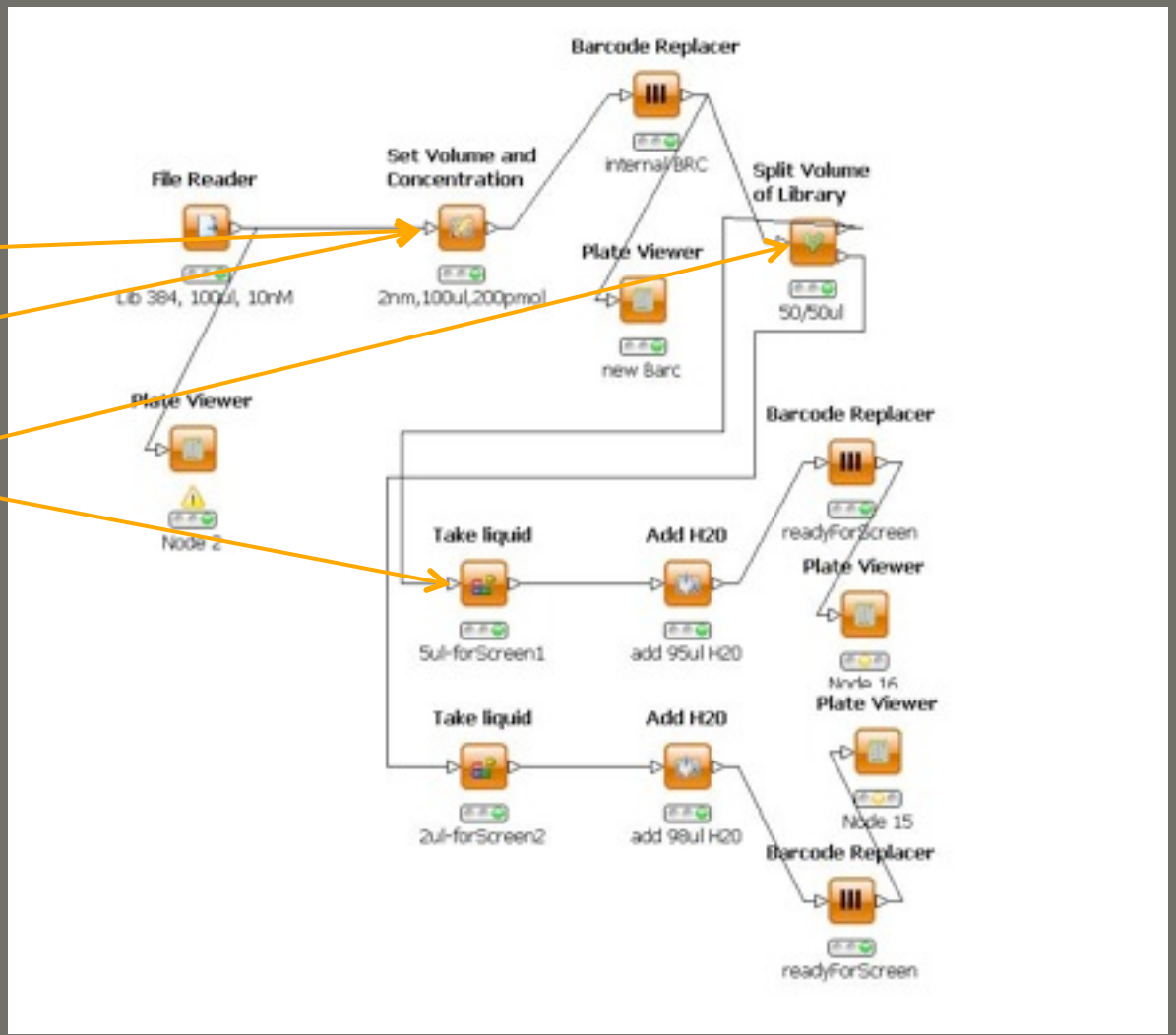




Library Handling

Data Automation:

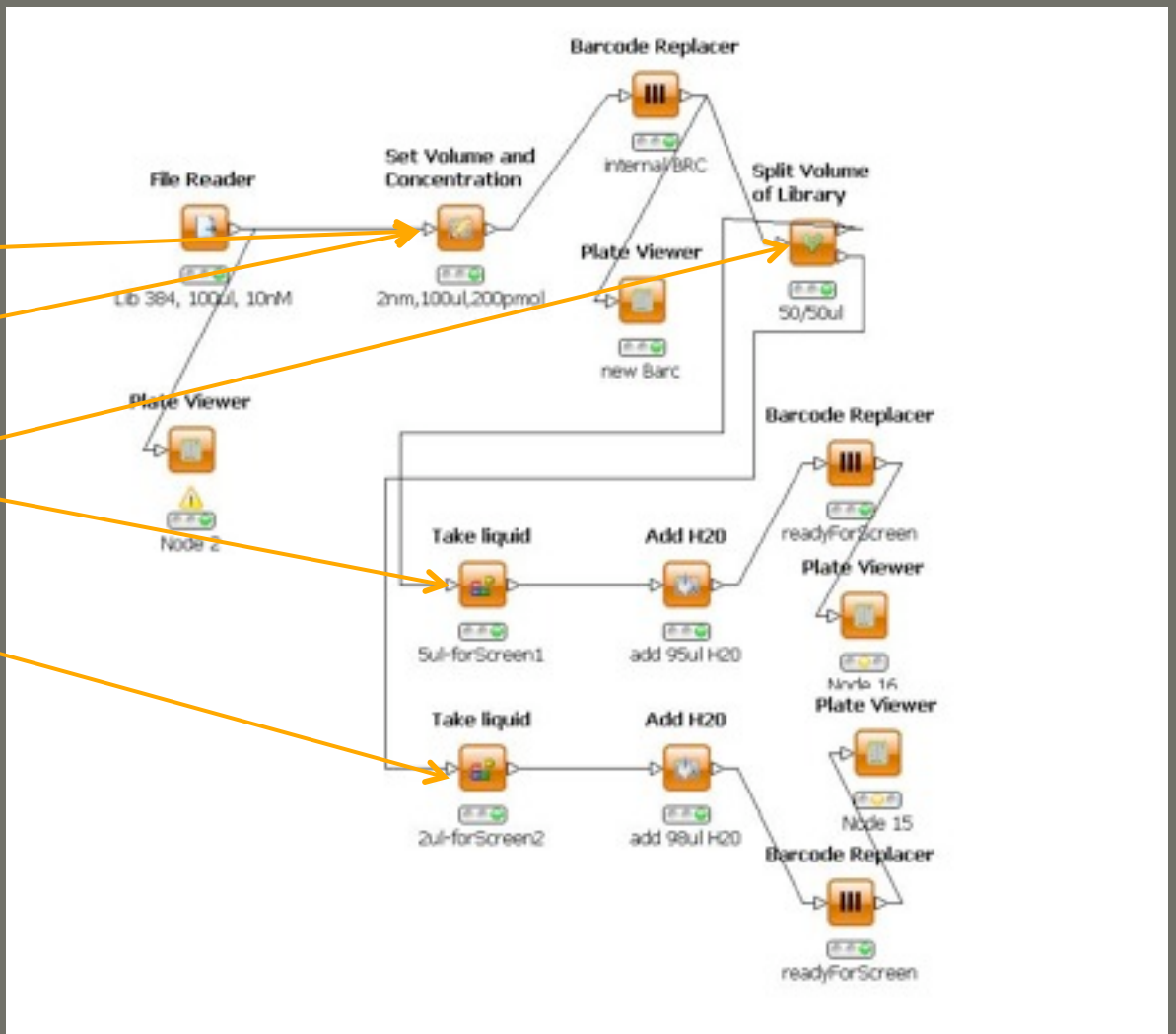
- Volume
- Concentration
- Dilutio



Library Handling

Data Automation:

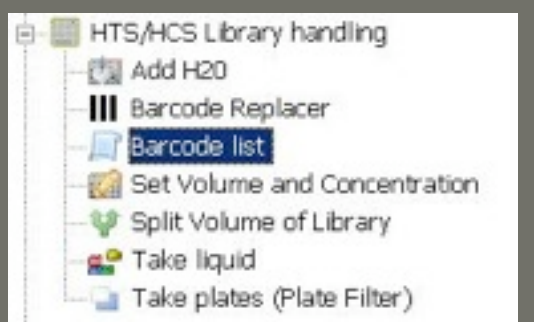
- Volume
- Concentration
- Dilutio
- Splitting

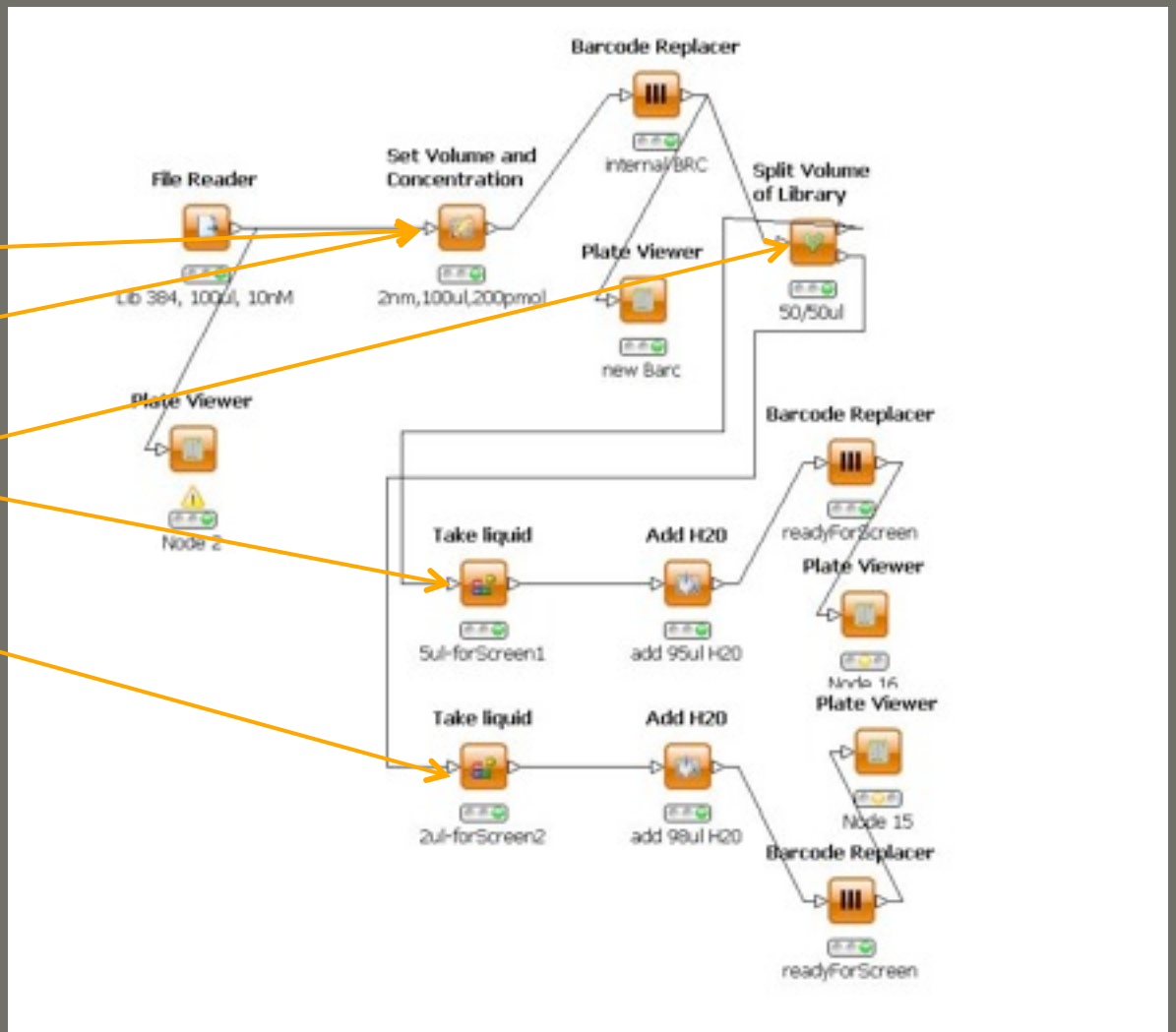


Library Handling

Data Automation:

- Volume
- Concentration
- Dilutio
- Splitting
- Take liquid

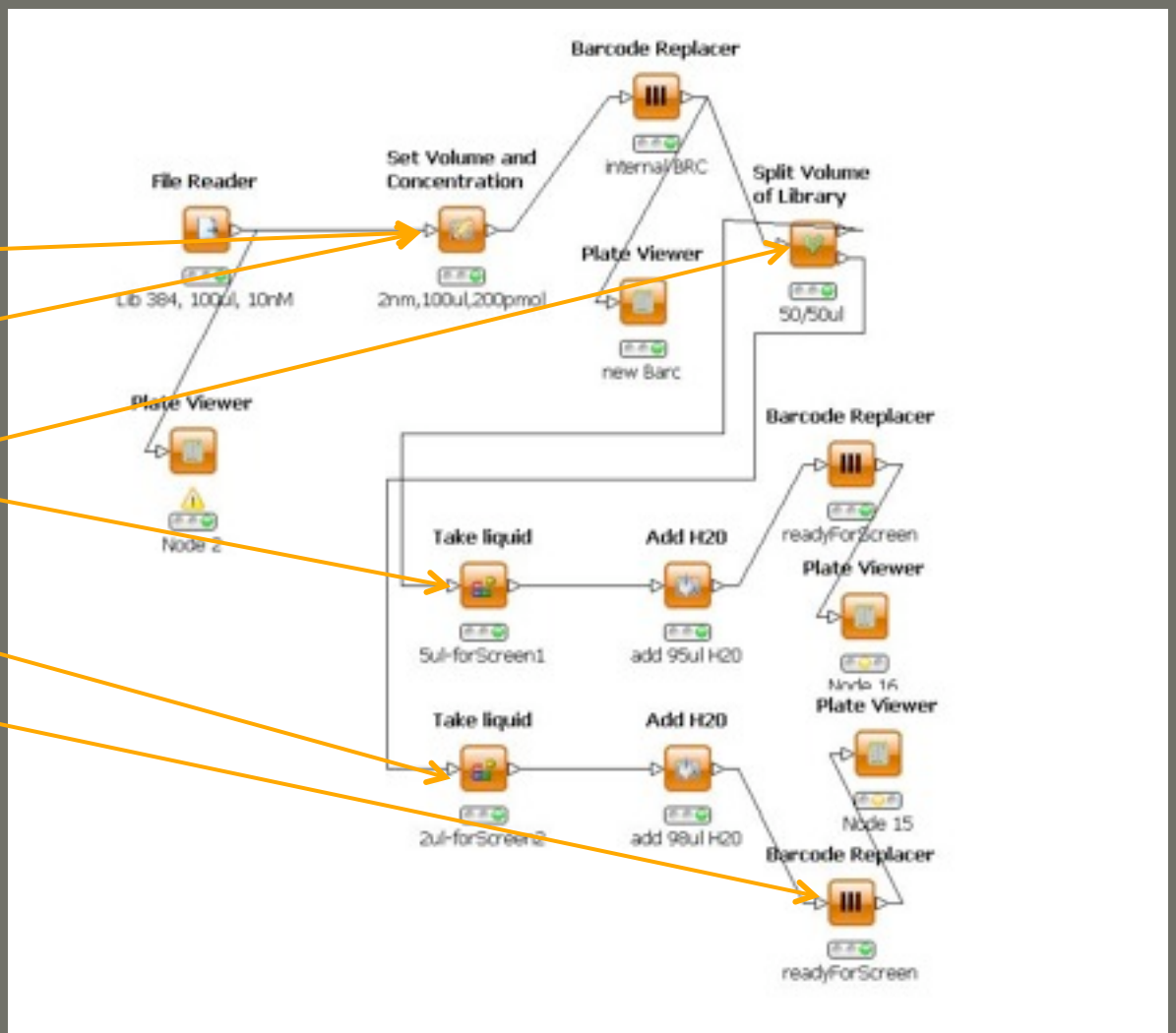




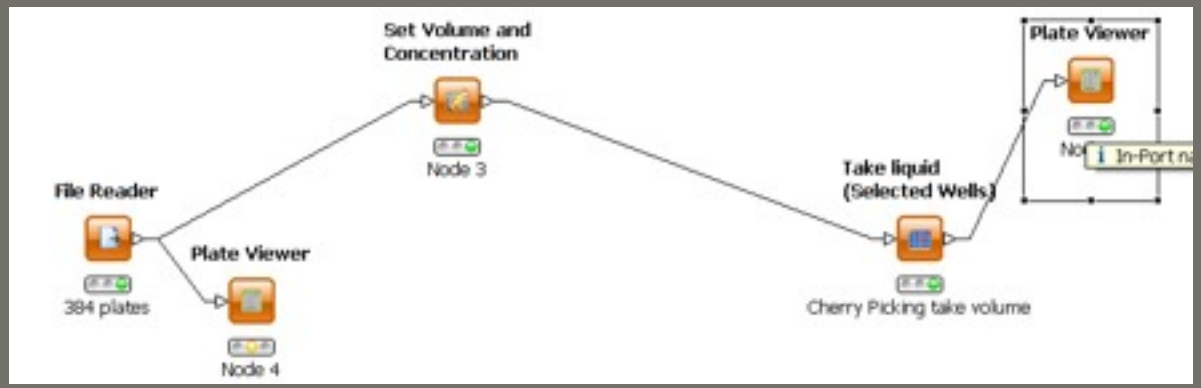
Library Handling

Data Automation:

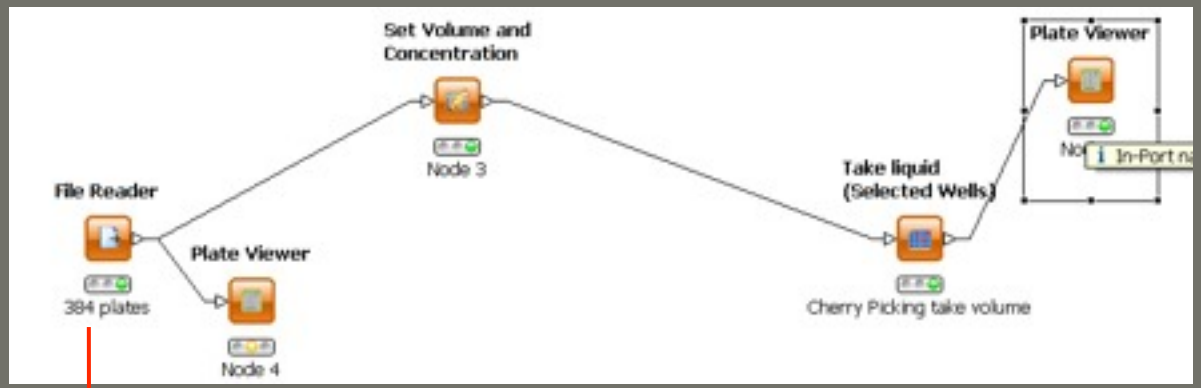
- Volume
- Concentration
- Dilutio
- Splitting
- Take liquid
- Barcode

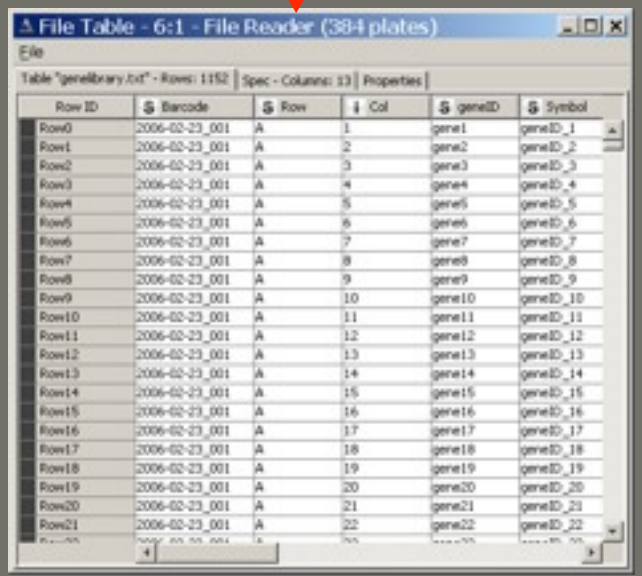


Workflows for HCS

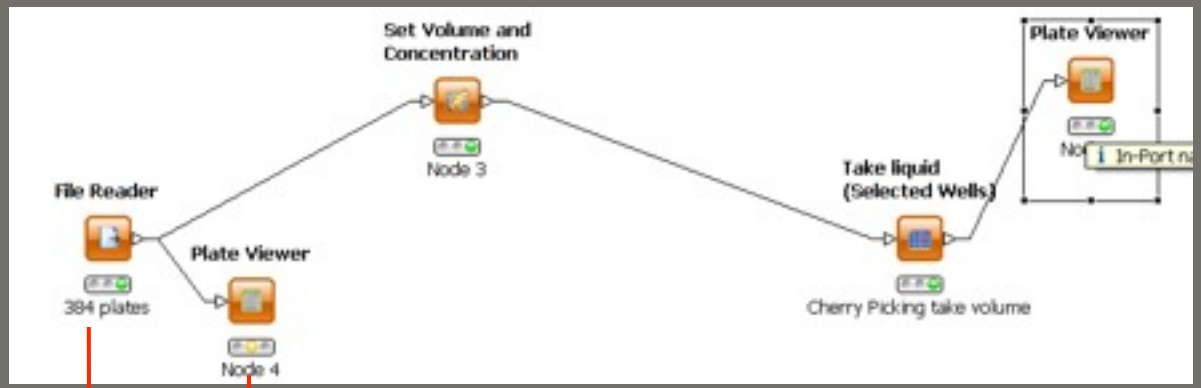


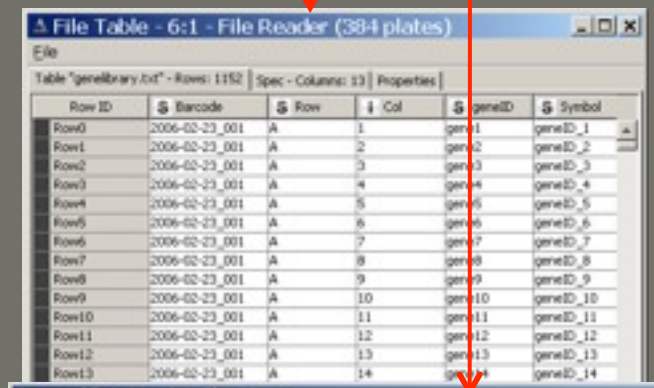
Workflows for HCS



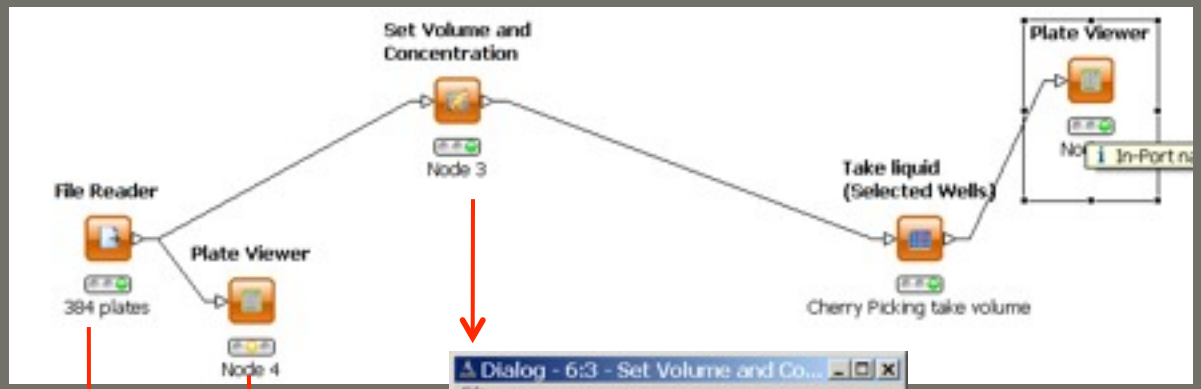


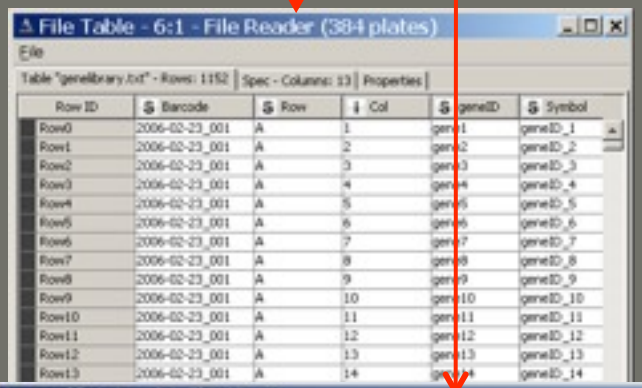
Workflows for HCS



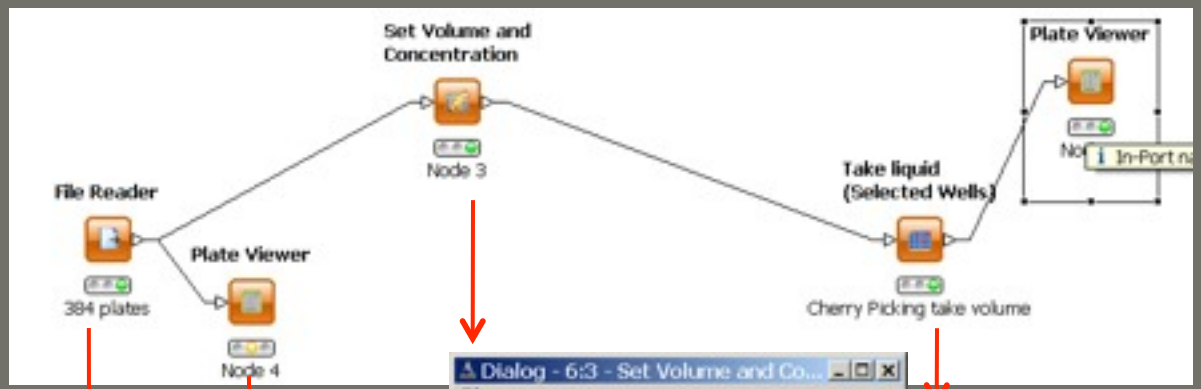


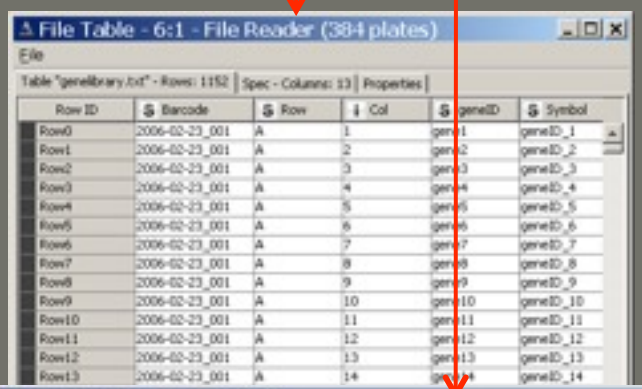
Workflows for HCS

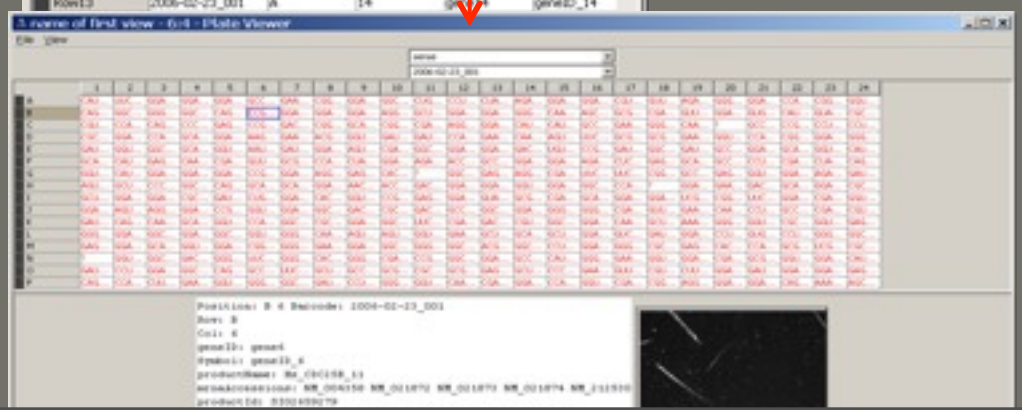


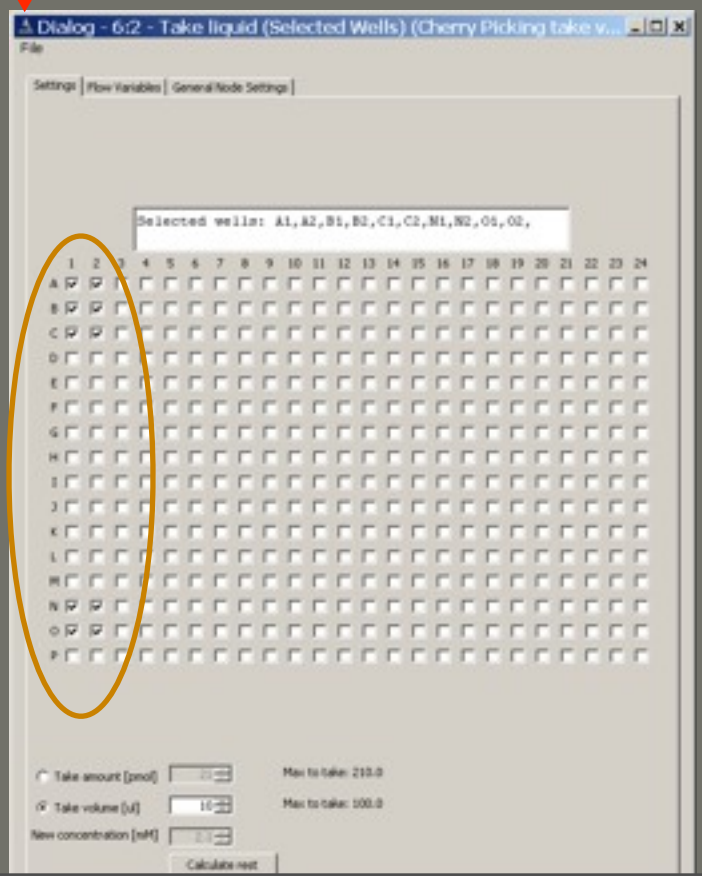


Workflows for HCS

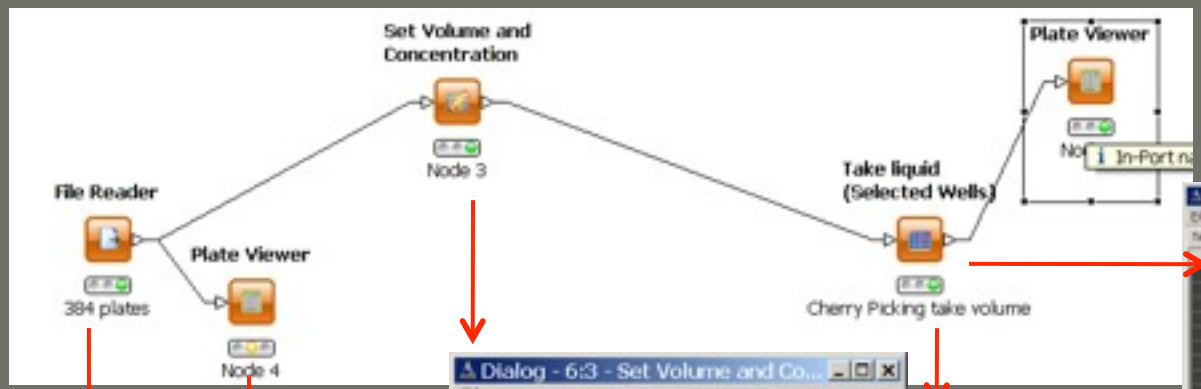


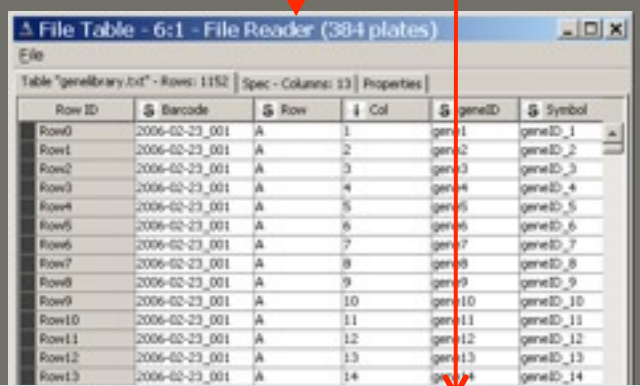


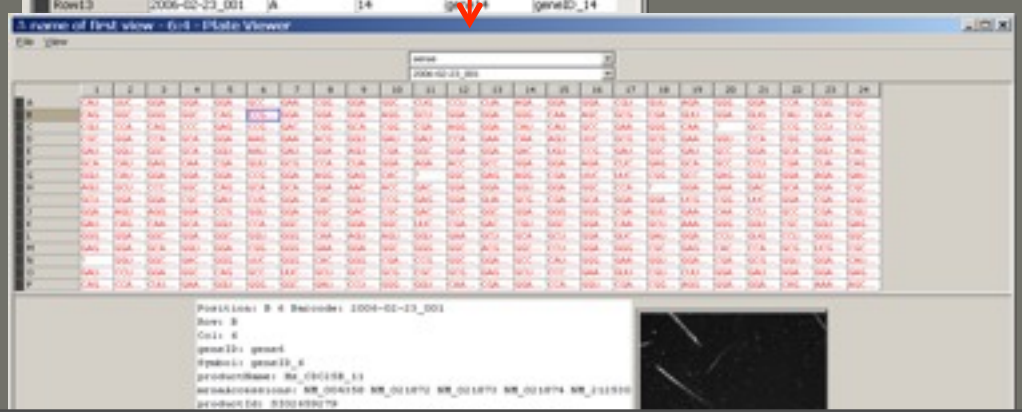


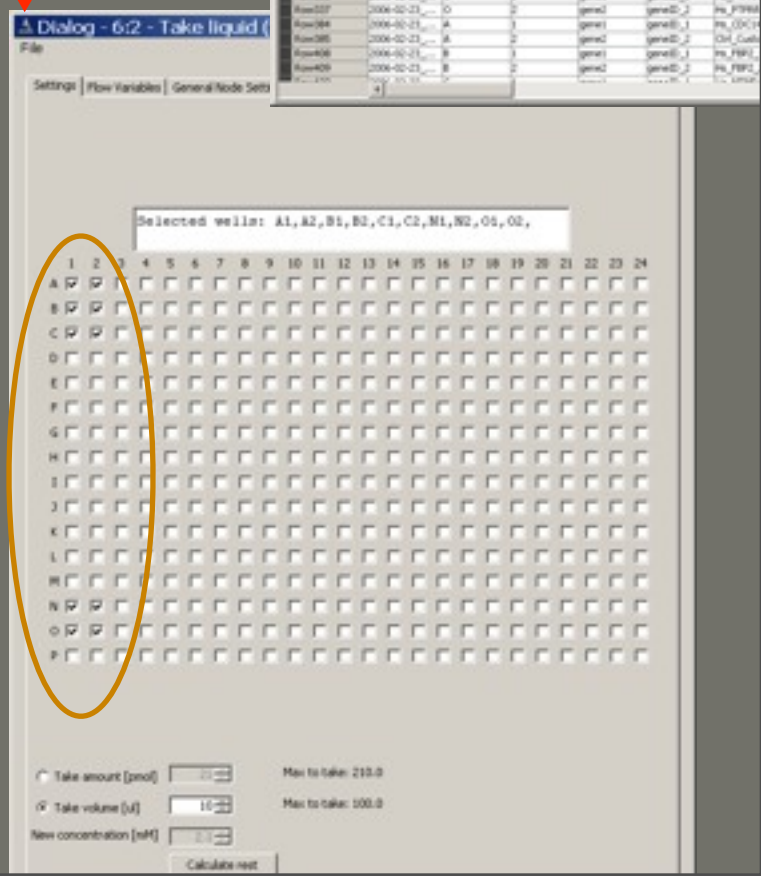


Workflows for HCS

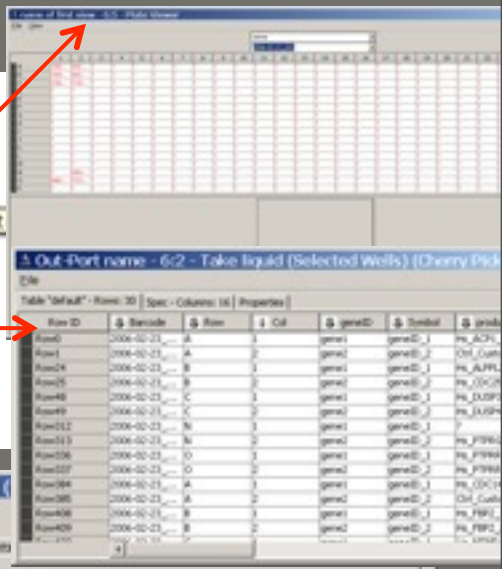


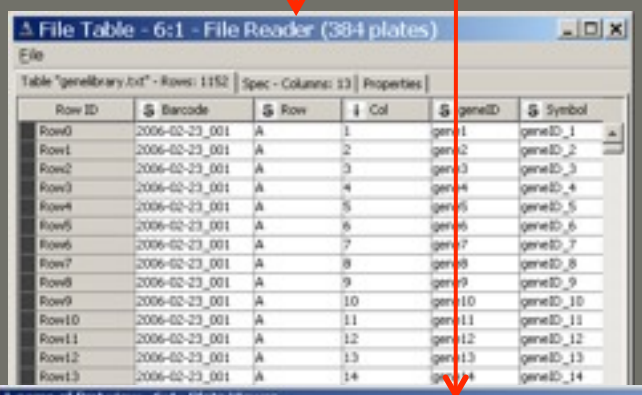


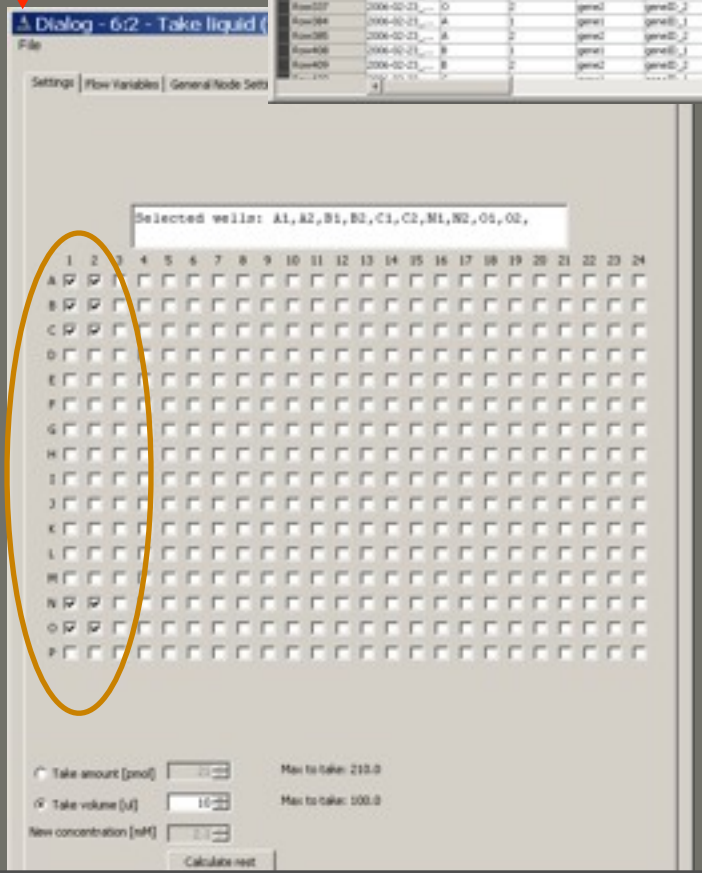


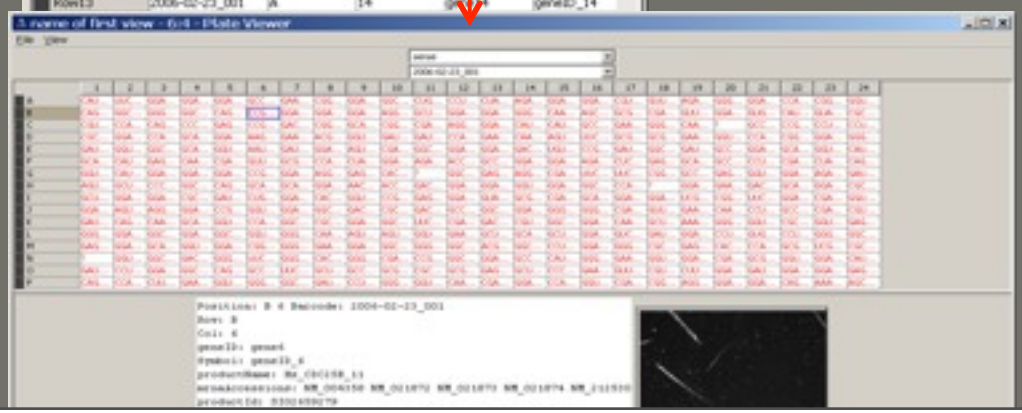


Workflows for HCS

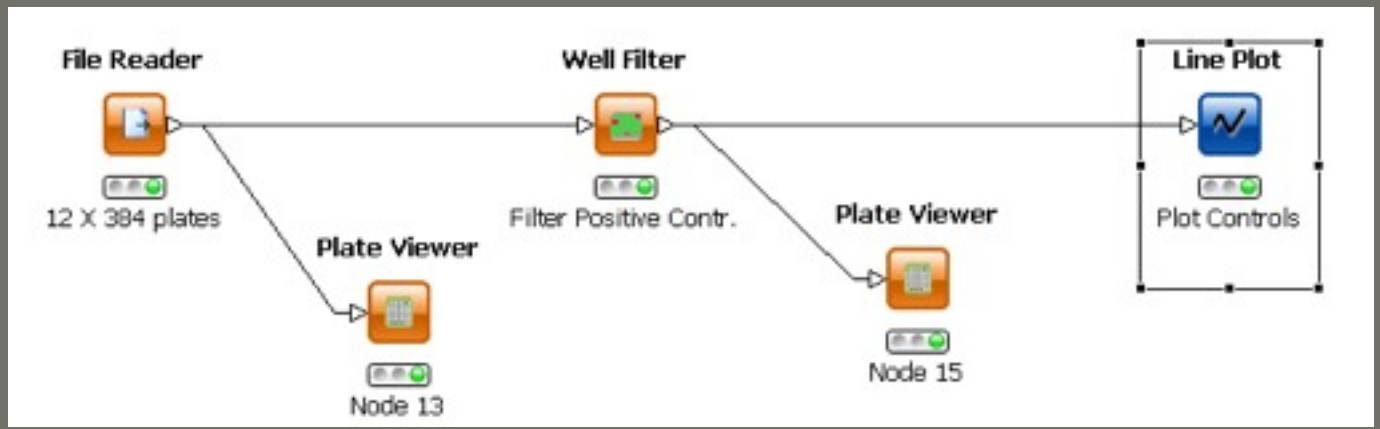




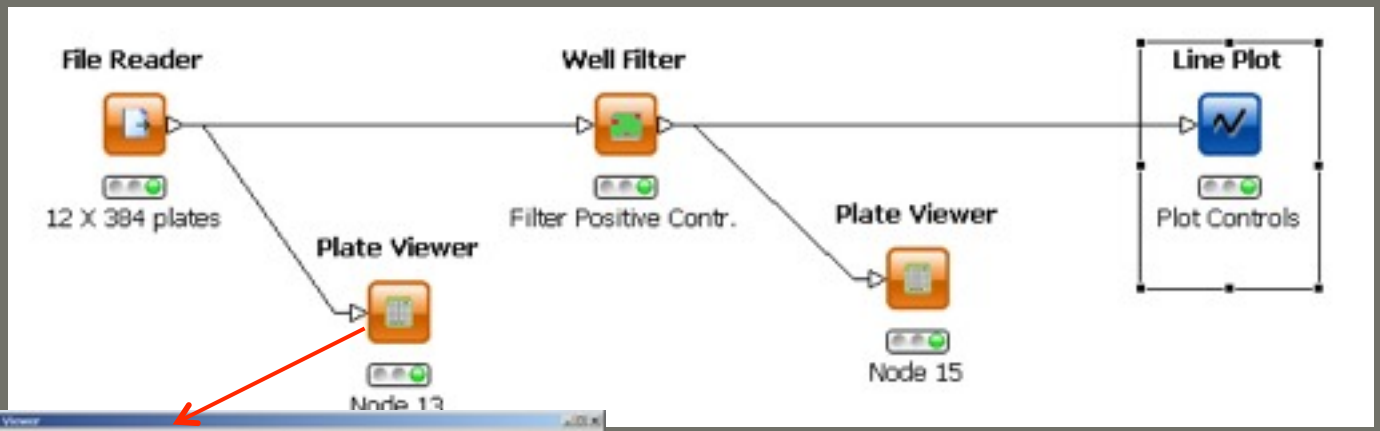




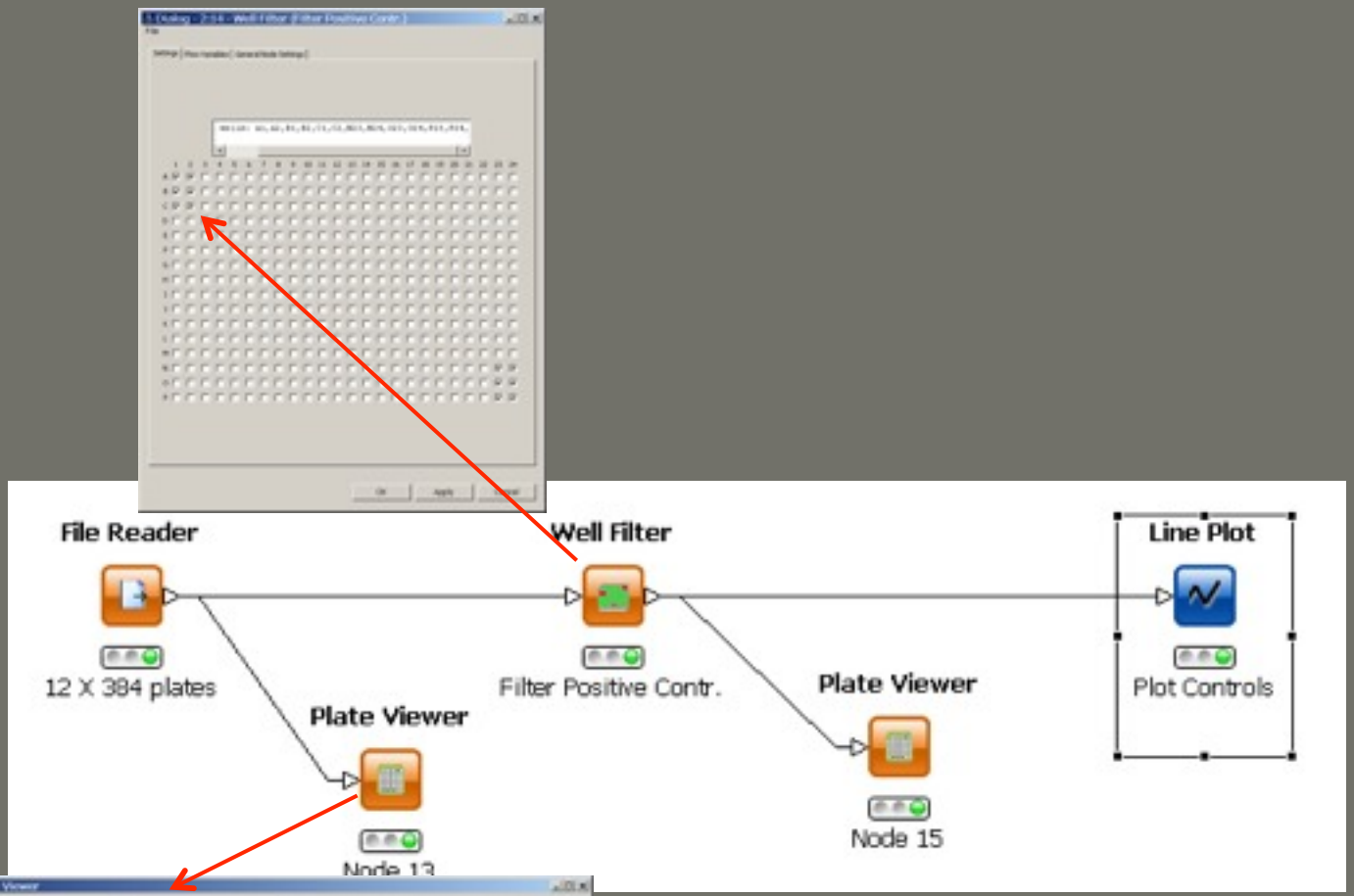
Workflows for HCS



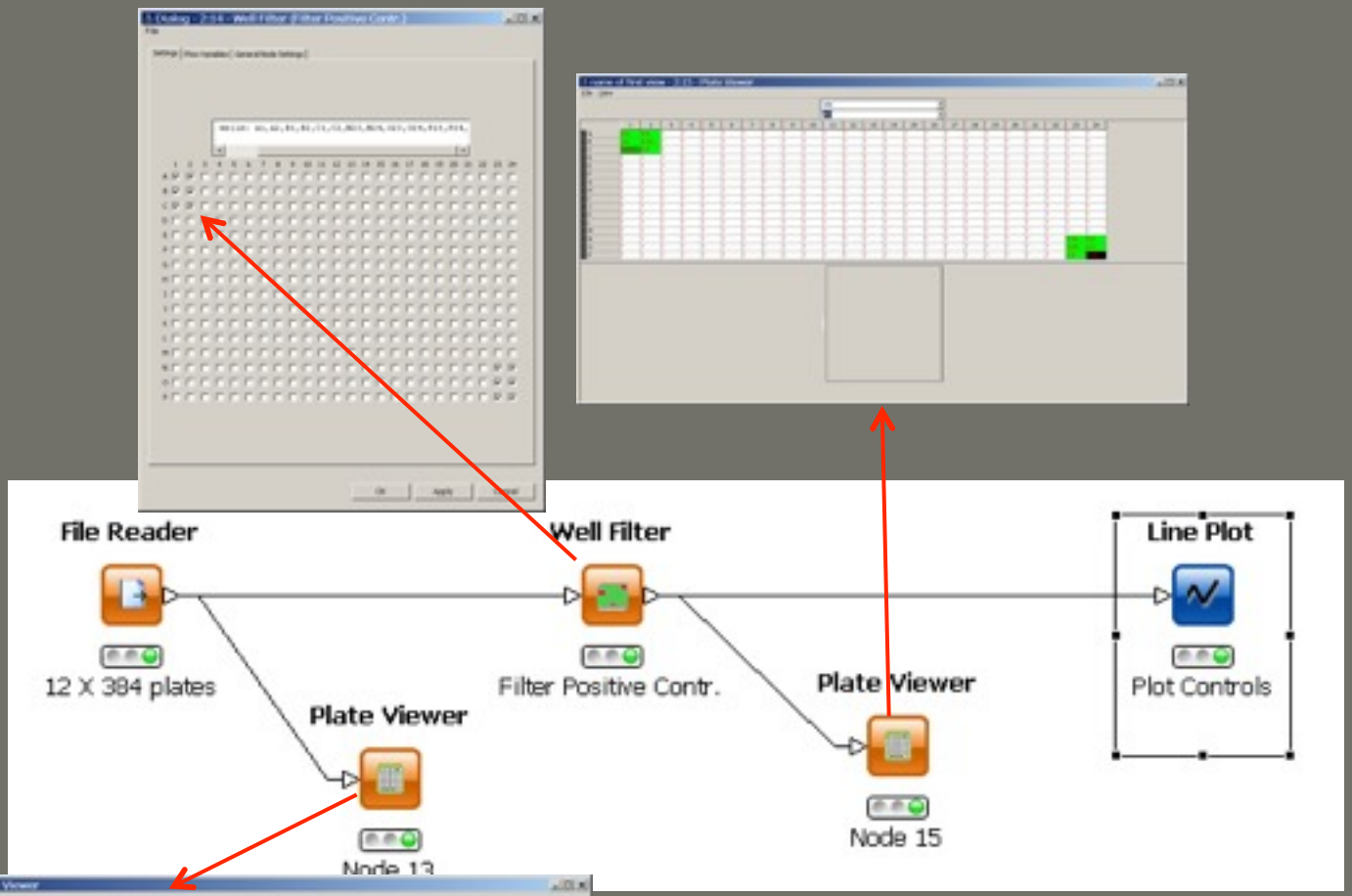
Workflows for HCS



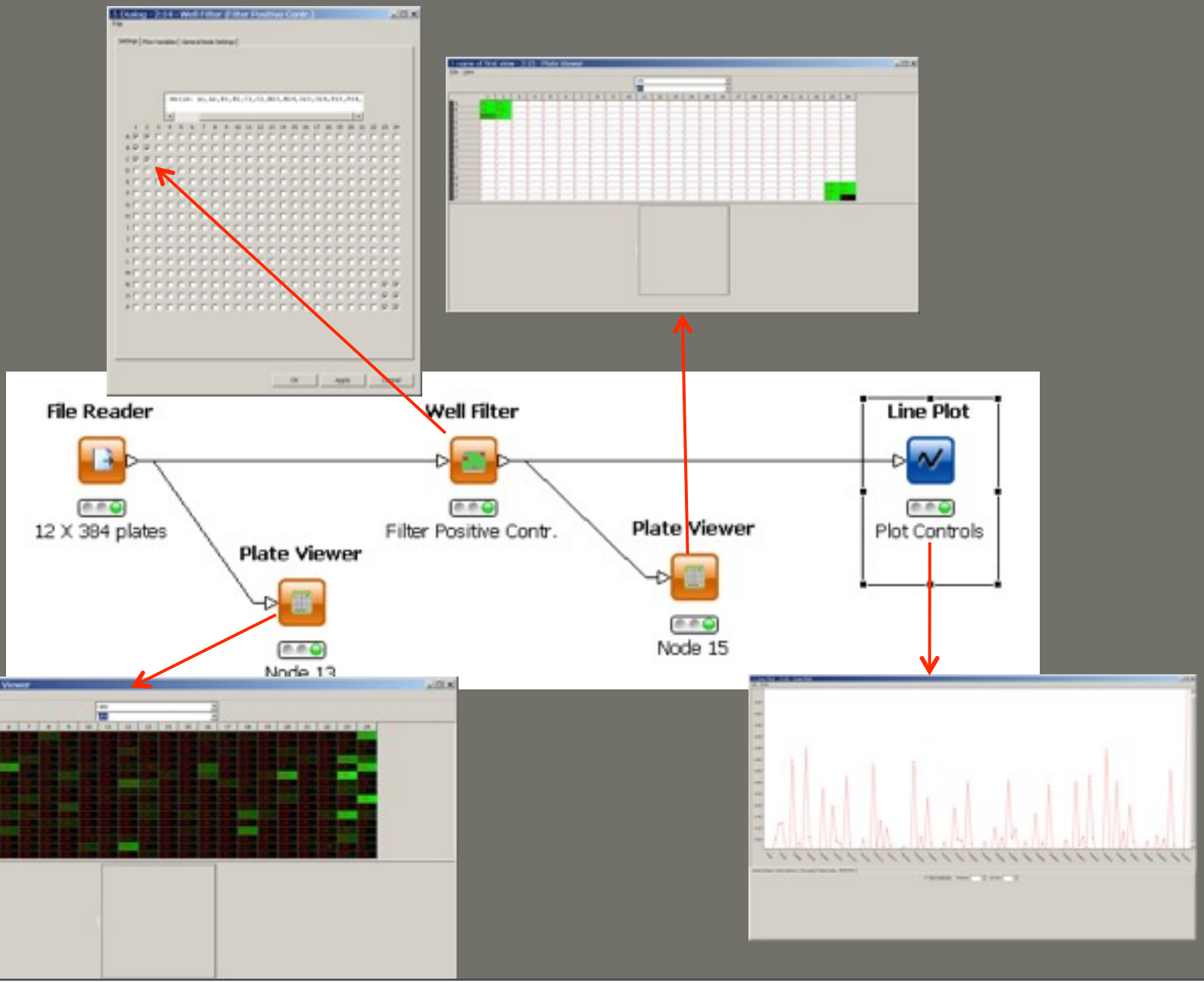
Workflows for HCS



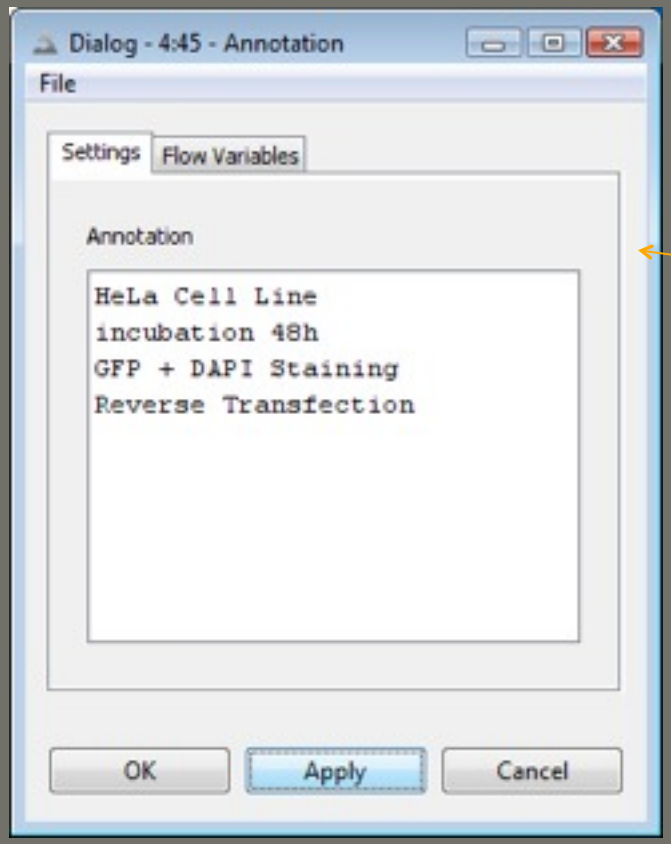
Workflows for HCS

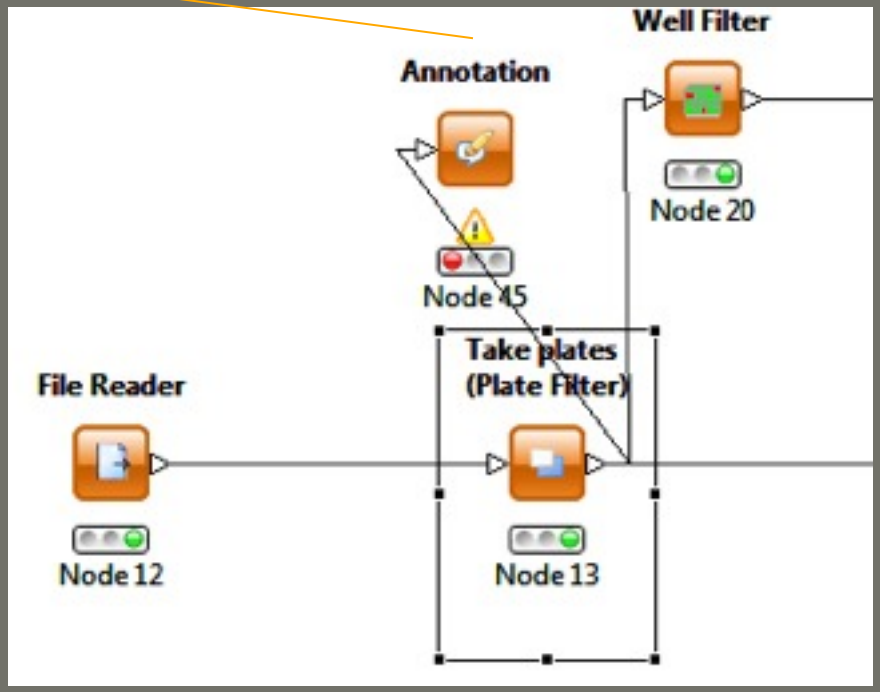


Workflows for HCS

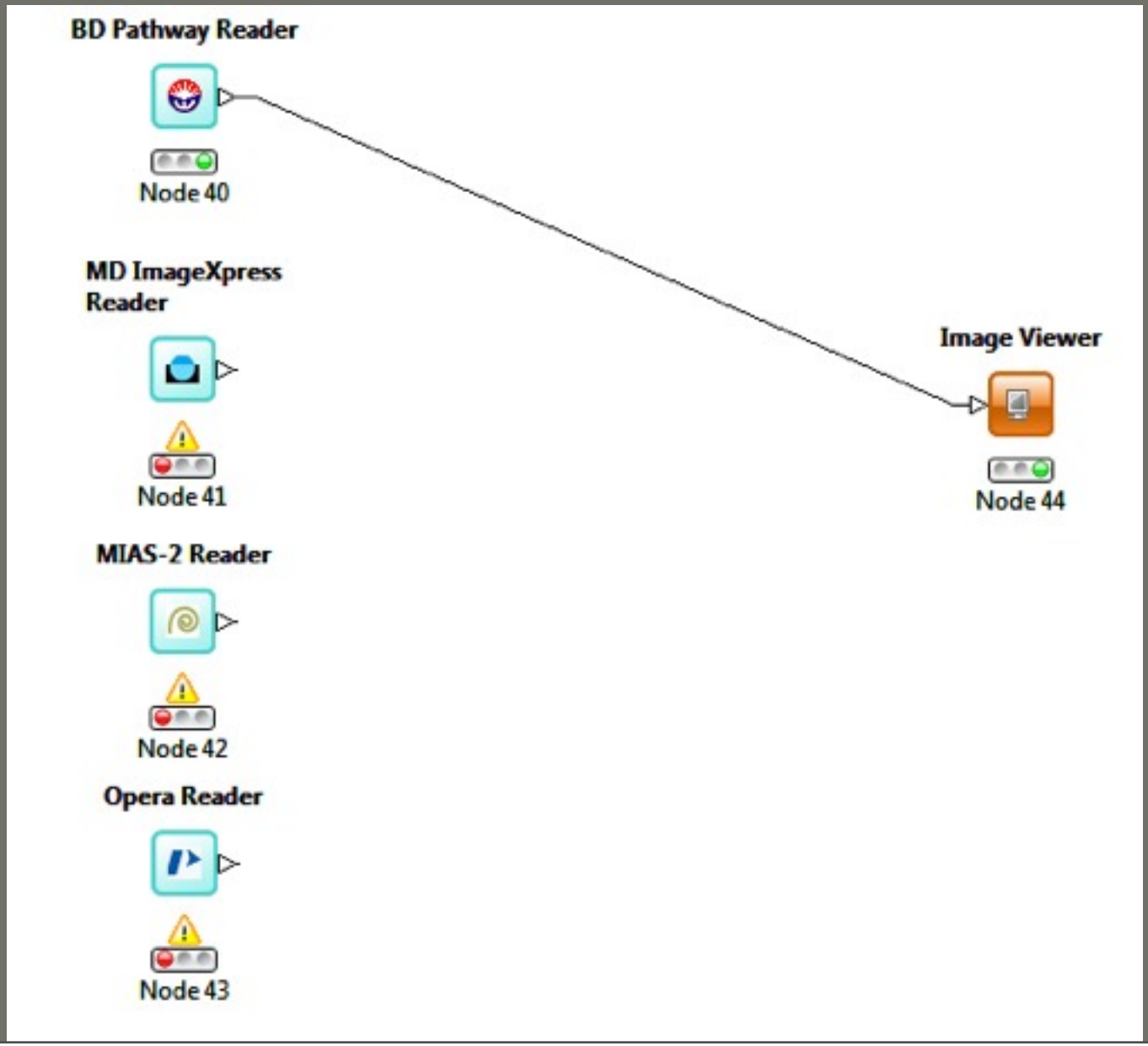


Workflows for HCS

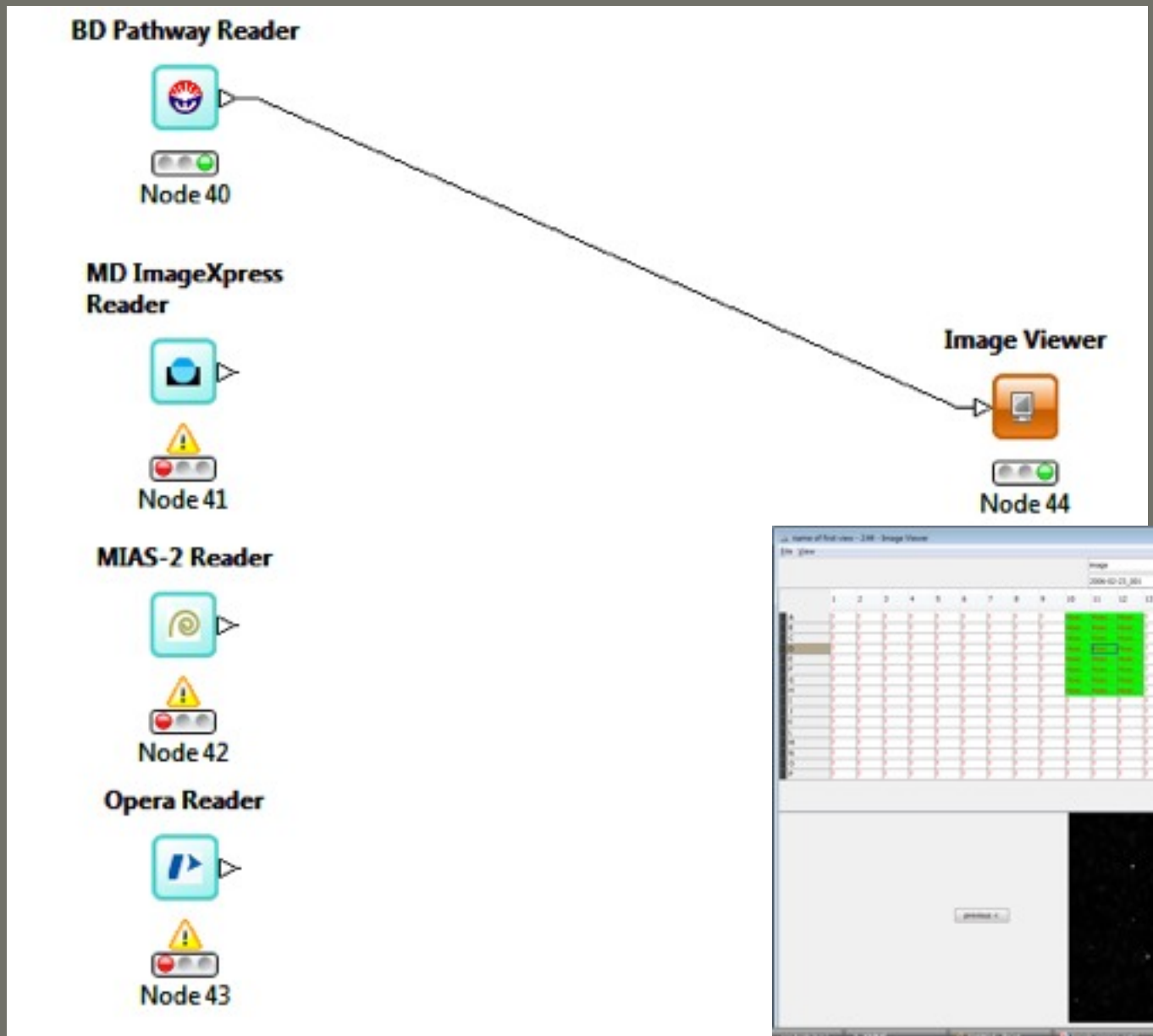


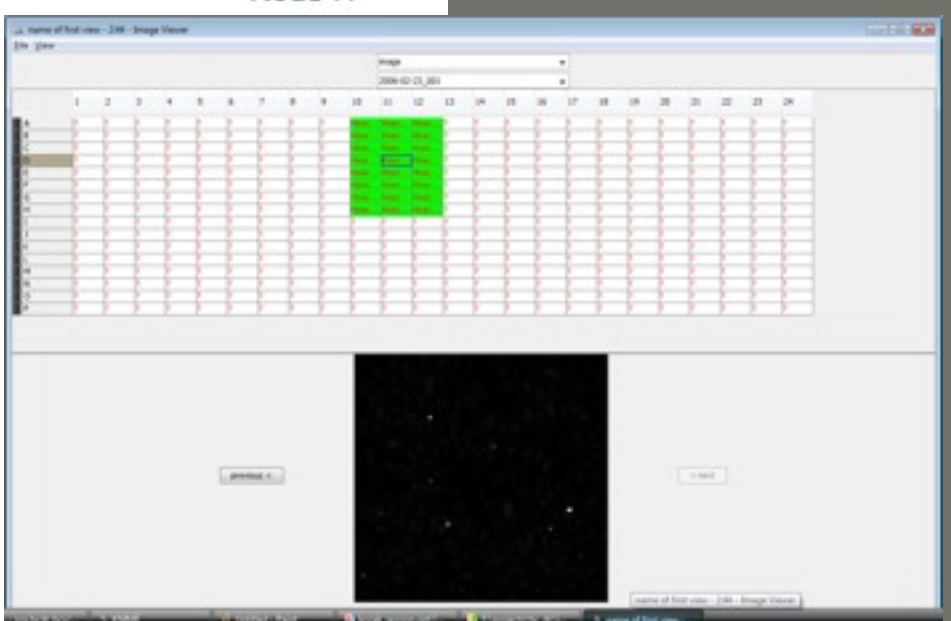


Workflows for HCS

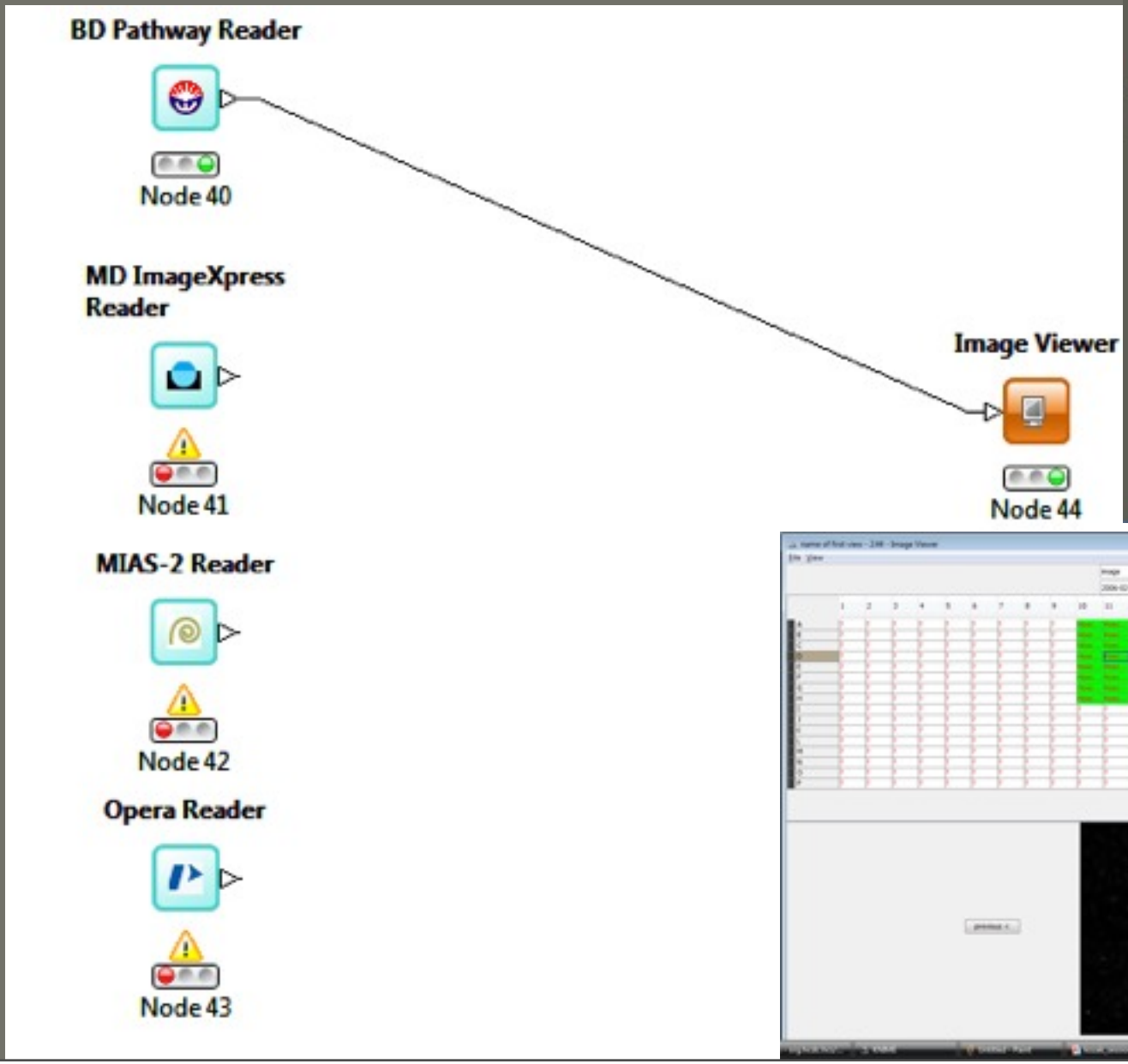


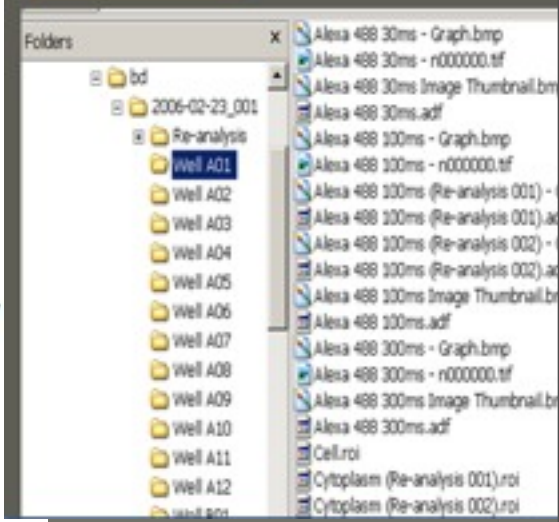
Workflows for HCS

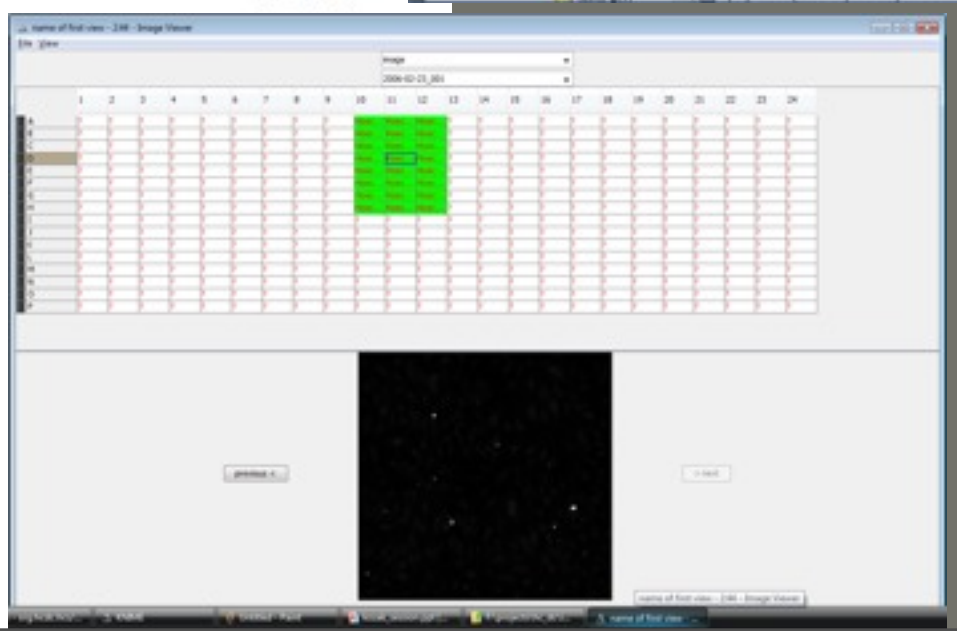




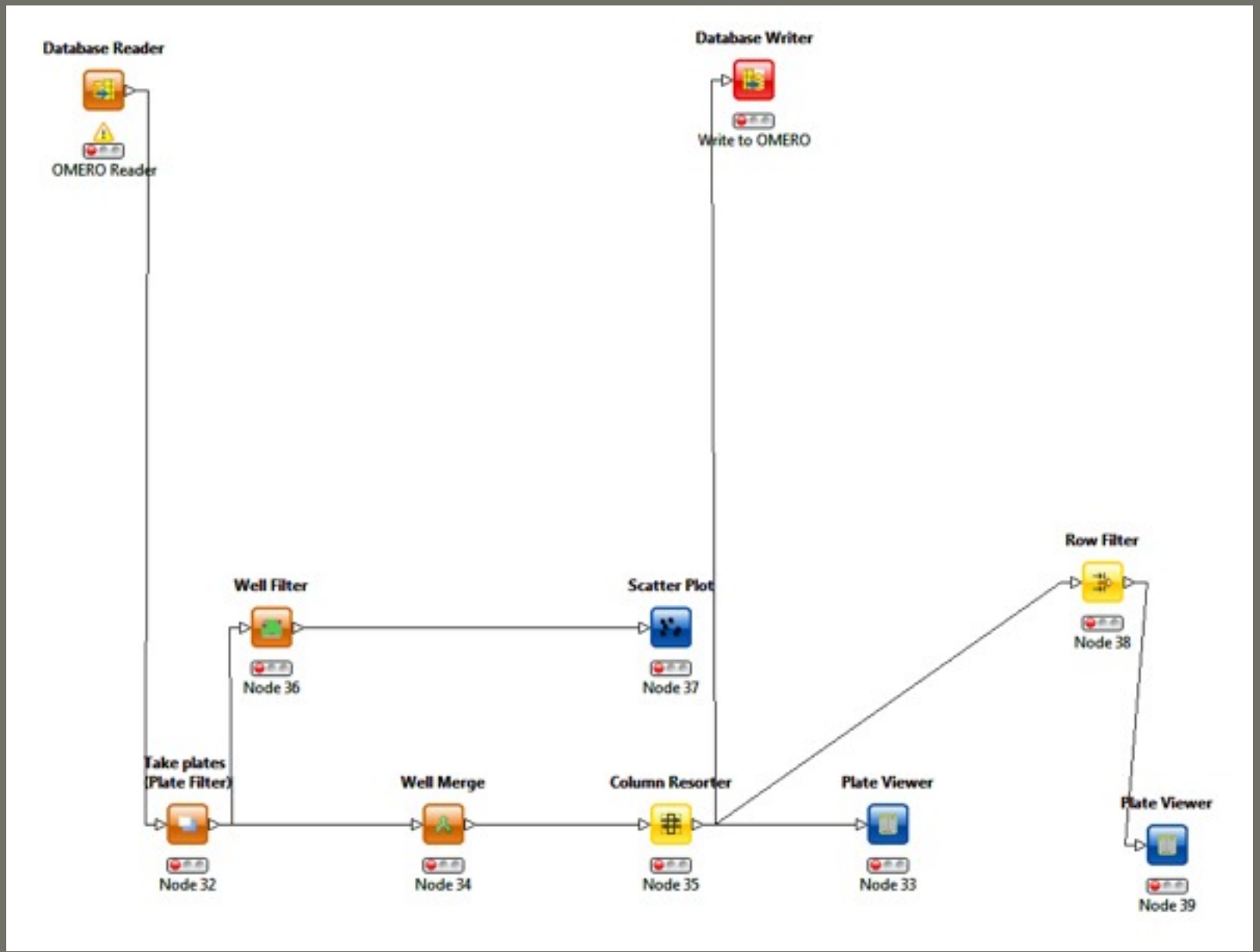
Workflows for HCS







Link to OME/OMERO



Acknowledgements & Partners:

Acknowledgements & Partners:

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

KNIME
Konstanz Information Miner

Contribution in development:

MPI-CBG, Dresden
Eberhard Krausz (former) & team
Eugenio Fava & team
Marc Bickle & team

ETH Zurich
LMC-RISC: Gabor Csucs & team
SystemX Adrian Honegger & team

MPI-IB, Berlin
Nikolaus Machuy & team

WebPage

HCDC: http://hcdc.ethz.ch

Workshops HCDC+ KNIME:

16-17.10.2009

2 days ETH Zurich