COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

0. Introduction and Overview

LU 0B – OPENMS AND KNIME

- Workflows - definition
- Conceptual ideas behind OpenMS and TOPP
- Installation of KNIME and OpenMS extensions
- Overview of KNIME
- Simple workflows in KNIME
  - Loading tabular data, manipulating rows, columns
  - Visualization of data
  - Preparing simple reports
  - Embedding R scripts
  - Simple OpenMS ID workflow: finding all proteins in a sample

High-Throughput Proteomics

- Analyzing one sample is usually not a big deal
- Analyzing 20 can be tiresome
- Analyzing 100 is a really big deal

- ***High-throughput experiments require high-throughput analysis***

- ***Compute power scales much better than manpower***

Pipelines and Workflows

pipeline |ˈpīpˌlīn| *noun*

1. a long pipe, typically underground, for conveying oil, gas, etc., over long distances. […]

2. *Computing* a linear sequence of specialized modules used for pipelining.

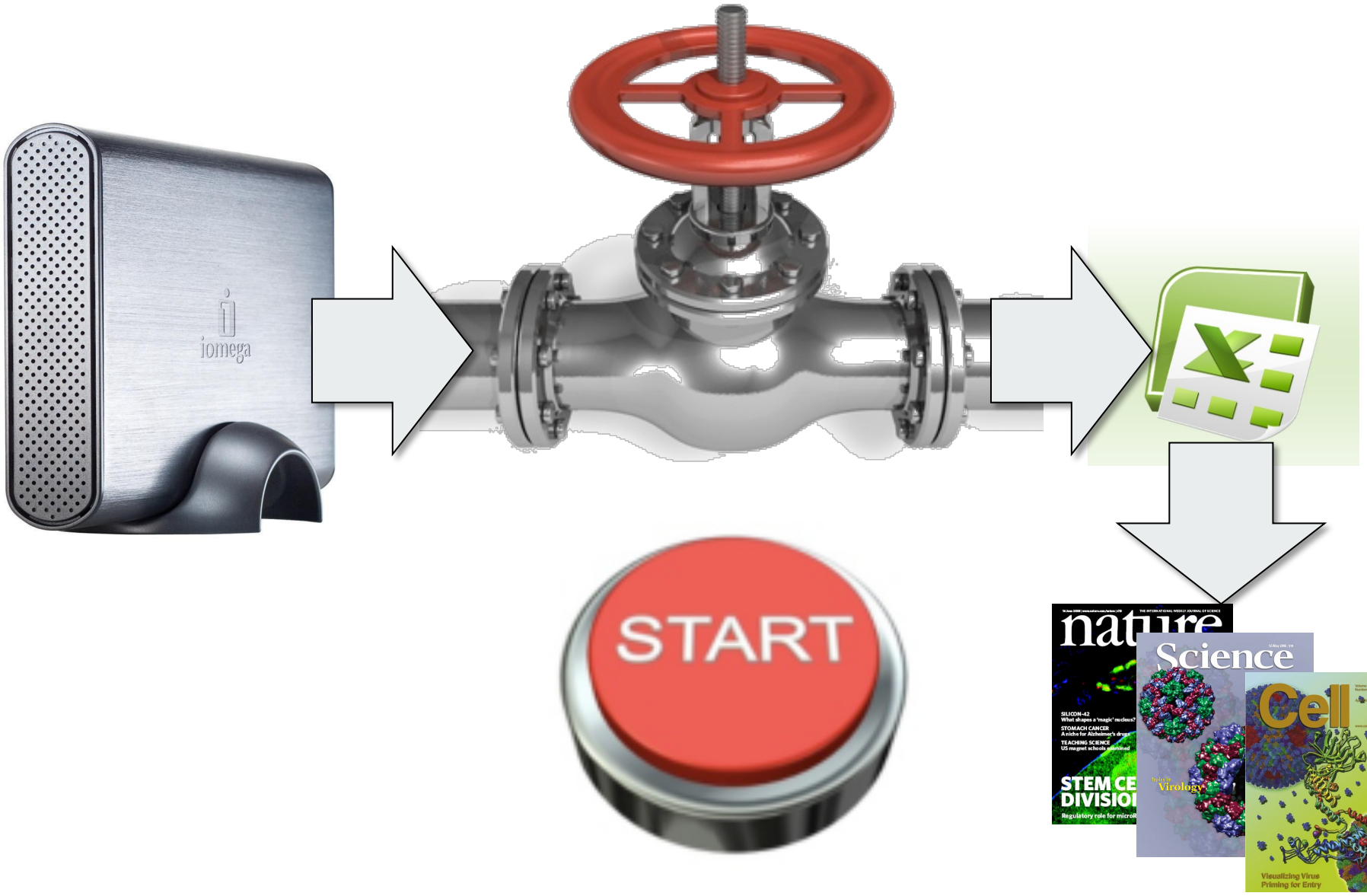
3. *(in surfing)* the hollow formed by the breaking of a large wave.

workflow |ˈwərkˌflō| *noun*

- the sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.

http://oxforddictionaries.com/definition/american_english/pipeline
http://oxforddictionaries.com/definition/american_english/workflow

Bioinformatics – The Holy Grail



KNIME and OpenMS

- Constructing workflows requires
  - **Tools** – making up the nodes of the workflows
  - A **workflow engine** – executing the nodes in a predefined order
- In the context of this course, we will use **OpenMS** tools to analyze mass spectrometric data
- We will design the workflow engine and data mining tool **KNIME** to construct and execute these workflows in a convenient manner
- We will briefly intro both tools – they are open-source software and freely available on all major platforms

OpenMS/TOPP

- **OpenMS** – an open-source C++ framework for computational mass spectrometry
- Jointly developed at ETH Zürich, FU Berlin, University of Tübingen
- **Open source**: BSD 3-clause license
- **Portable**: available on Windows, OSX, Linux
- **Vendor-independent**: supports all standard formats and vendor-formats through proteowizard
- **TOPP – The OpenMS Proteomics Pipeline**
  - Building blocks: One application for each analysis step
  - All applications share **identical user interfaces**
  - Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats
- **TOPP tools** can be integrated in various **workflow systems**
  - TOPPAS – TOPP Pipeline Assistant
  - Galaxy
  - WS-PGRADE/gUSE
  - **KNIME**

Kohlbacher et al., Bioinformatics (2007), 23:e191

TOPP – Concepts

- **TOPP – The OpenMS Proteomics Pipeline**
- No programming skills required
- **Graphical User Interface**: TOPPView and TOPPAS
- Building blocks: One application for each analysis step
- All applications share **identical user interfaces**
- Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats

Kohlbacher et al., Bioinformatics (2007), 23:e191

TOPP Tools – Implementation

- Very easy to implement thanks to the OpenMS framework
- Usually short (200 lines of code on average, mostly concerned with parameter handling)
- Make use of the OpenMS framework functionality

IDMapper.C:

```
[...]
vector<ProteinIdentification> protein_ids;
vector<PeptideIdentification> peptide_ids;
String document_id;
IdXMLFile().load(getStringOption_
   ("id"), protein_ids,peptide_ids, document_id);
IDMapper mapper;
[...]
ConsensusXMLFile file;
ConsensusMap map;
file.load(in, map);
mapper.annotate(map, peptide_ids, protein_ids, false);
file.store(out, map);
```

Interoperability

- Pipeline components (tools) have to be compatible
- Data formats have to be compatible
- Alternatives
  - **Glue code** to convert parameters, adapt settings
  - **Converters** translating one data format into another
- Issues
  - Portability
  - Loss of information

PSI Standard Formats

Numerous open and standardized **XML formats** have been proposed by the **HUPO Proteomics Standards Initiative (HUPO PSI)**:

- **mzML** (successor of mzData) for storing mass spectrometry data
- **mzIdentML** for storing peptide/protein identifications
- **traML** for storing transition and inclusion lists (Deutsch et al., MCP, 2012)
- **mzQuantML** for storing quantitation results (Walzer et al., MCP, 2013)
- **mzTab** for summary information of quantitative and qualitative results, Excel-compatible TSV format (Griss et al., MCP, 2014)
- **qcML** for storing and mining quality control information (Walzer et al., MCP, 2014)

Advantages

- Open, documented, no closed-source libraries required
- Will still be readable in 10 years from now
- Interoperable with different software packages

Disadvantages

- Initial raw data conversion required (and often awkward)
- File size
- Poor support by instrument software

Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.OpenMS.de)

FeatureFinder

The feature detection application for quantitation.

pot. predecessor tools		pot. successor tools
PeakPicker	→ FeatureFinder →	FeatureLinker
MapAligner		SeedListGenerator

This module identifies "features" in a LC/MS map. By feature, we understand a peptide in a MS sample that reveals a characteristic isotope distribution. The algorithm computes positions in rt and m/z dimension and a charge estimate of each peptide.

The algorithm identifies pronounced regions of the data around so-called seeds. In the next step, we iteratively fit a model of the isotope profile and the retention time to these data points. Data points with a low probability under this model are removed from the feature region. The intensity of the feature is then given by the sum of the data points included in its regions.

How to find suitable parameters and details of the different algorithms implemented are described in the **TOPP tutorial**.

Note:

that the wavelet transform is very slow on high-resolution spectra (i.e. FT, Orbitrap). We recommend to use a noise or intensity filter to remove spurious points first and to speed-up the feature detection process.

Specialized tools are available for some experimental techniques: **SILACAnalyzer**, **ITRAQAnalyzer**.

The command line parameters of this tool are:

```
FeatureFinder -- Detects two-dimensional features in LC-MS data.
Version: 1.7.0 Sep  3 2010, 15:13:04, Revision: 7349

Usage:
  FeatureFinder <options>
```

Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.openms.de)

```
Common TOPP options:
  -ini <file>        Use the given TOPP INI file
  -threads <n>       Sets the number of threads allowed to be used by the TOPP tool (default: "1")
  -write_ini <file>  Writes the default configuration file
  --help             Shows options
  --helphelp         Shows all options (including advanced)

The following configuration subsections are valid:
 - algorithm   Algorithm section

You can write an example INI file using the '-write_ini' option.
Documentation of subsection parameters can be found in the
doxygen documentation or the INIFileEditor.
Have a look at OpenMS/doc/index.html for more information.
```

For the parameters of the algorithm section see the algorithms documentation:
centroided
isotope_wavelet
mrm

In the following table you can find example values of the most important parameters for different instrument types.
These parameters are not valid for all instruments of that type, but can be used as a starting point for finding suitable parameters.

'centroided' algorithm:

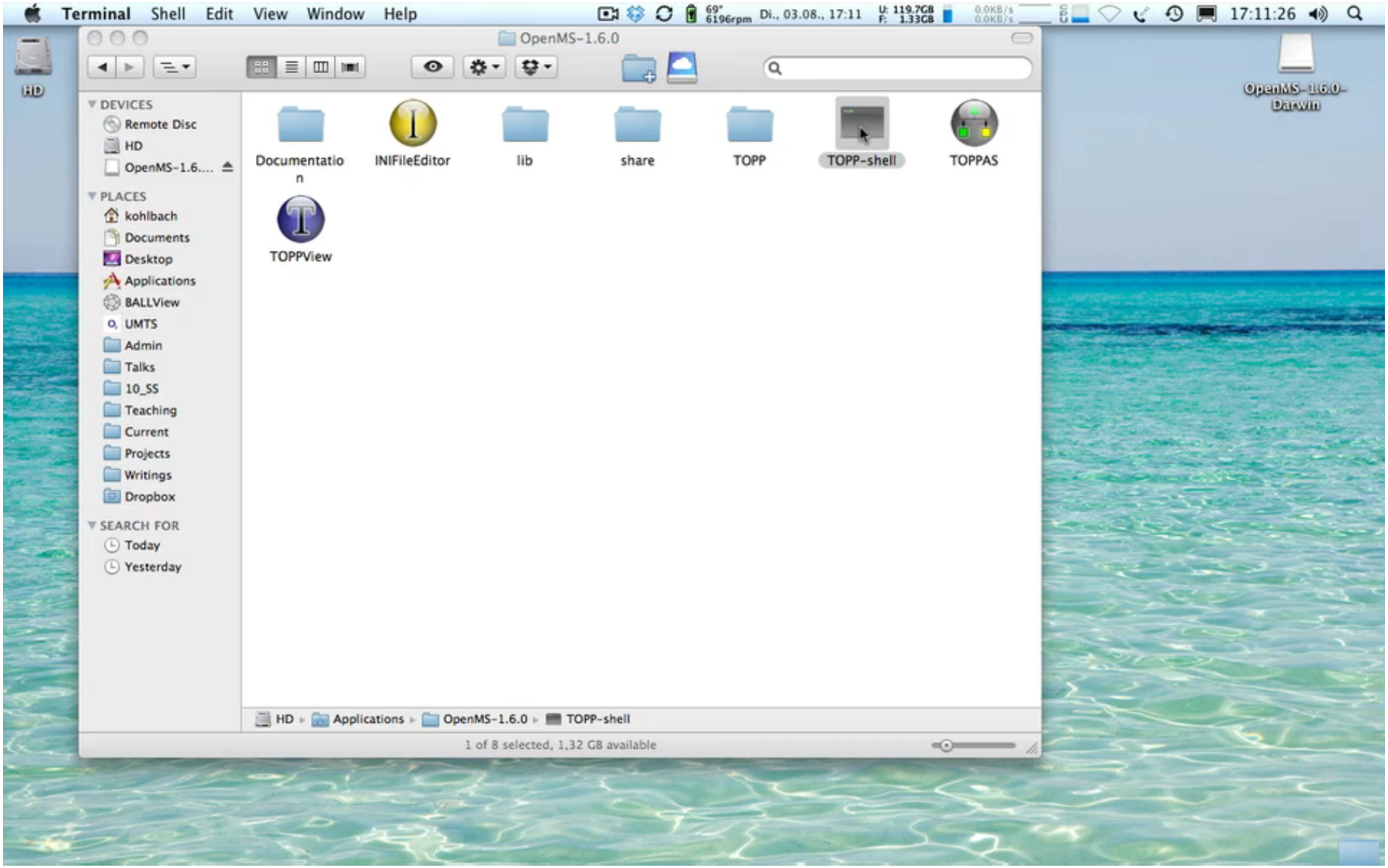
	Q-TOF	LTQ Orbitrap
intensity:bins	10	10
mass_trace:mz_tolerance	0.02	0.004
isotopic_pattern:mz_tolerance	0.04	0.005

For the *centroided* algorithm centroided data is needed. In order to create centroided data from profile data use the **PeakPicker**.

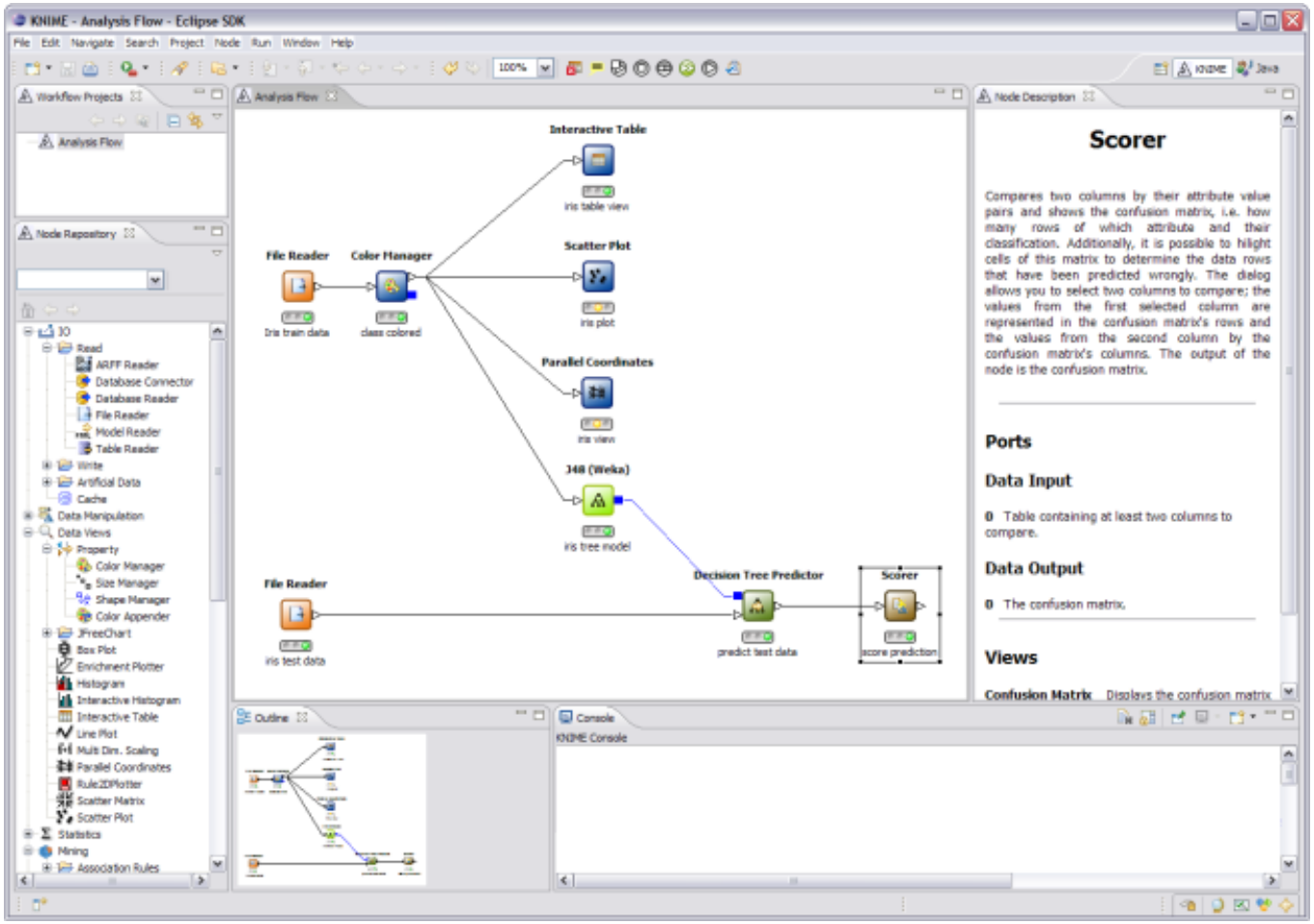
Installation of OpenMS

- Standalone version for command line and cluster environments
- Pre-built installers for Windows and Mac OS X
- Installer and installation instructions: http://open-ms.sourceforge.net/downloads/
- Bleeding edge development versions: http://ftp.mi.fu-berlin.de/OpenMS/nightly_binaries/
- Linux? Build your own OpenMS from git: https://github.com/OpenMS/OpenMS

Use on the Command Line



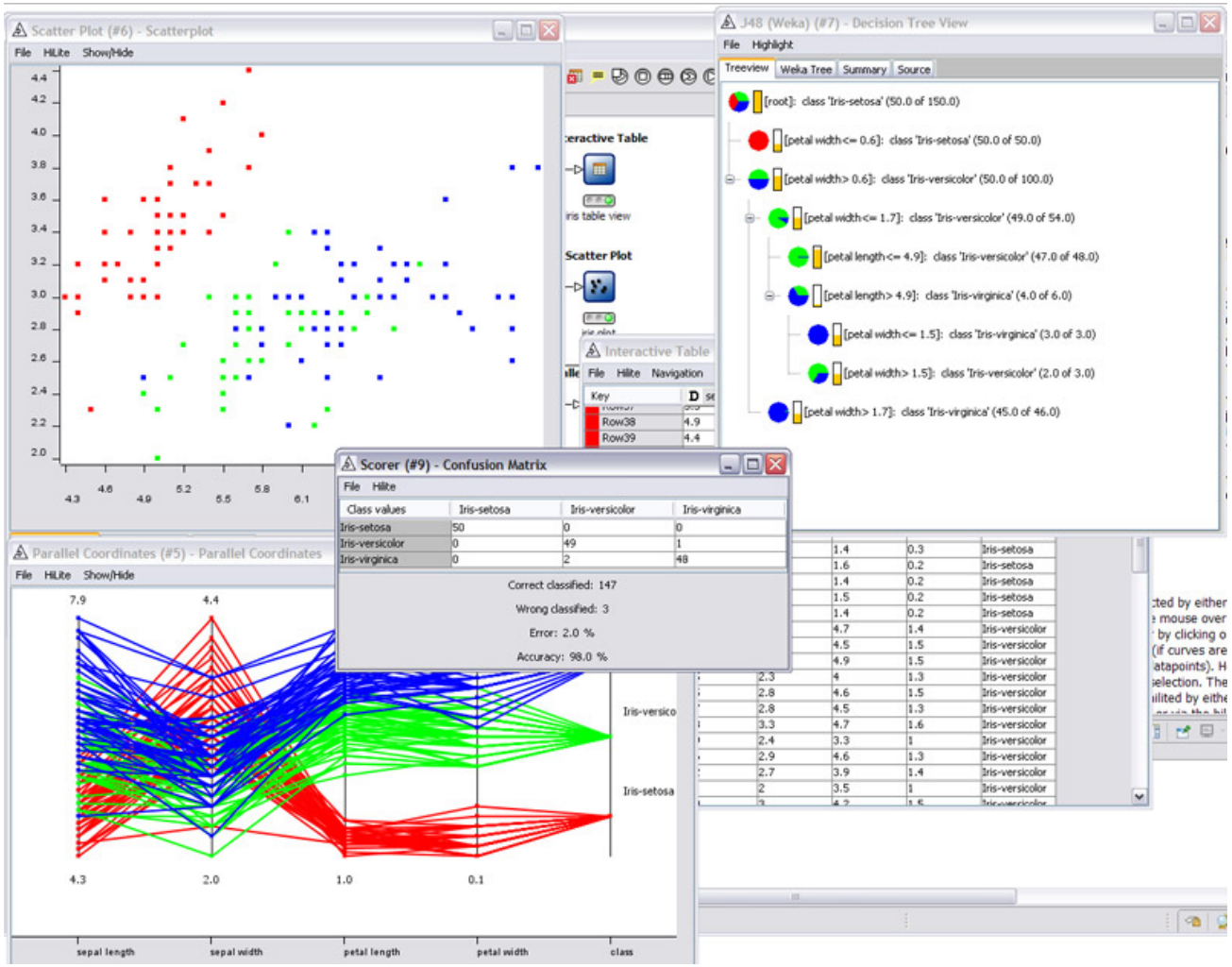
KNIME – KoNstanz Information MinEr



- Industrial-strength general-purpose workflow system
- Convenient and easy-to-use graphical user interface
- Available for Windows, OSX, Linux at http://KNIME.org

http://knime.org

KNIME – KoNstanz Information MinEr

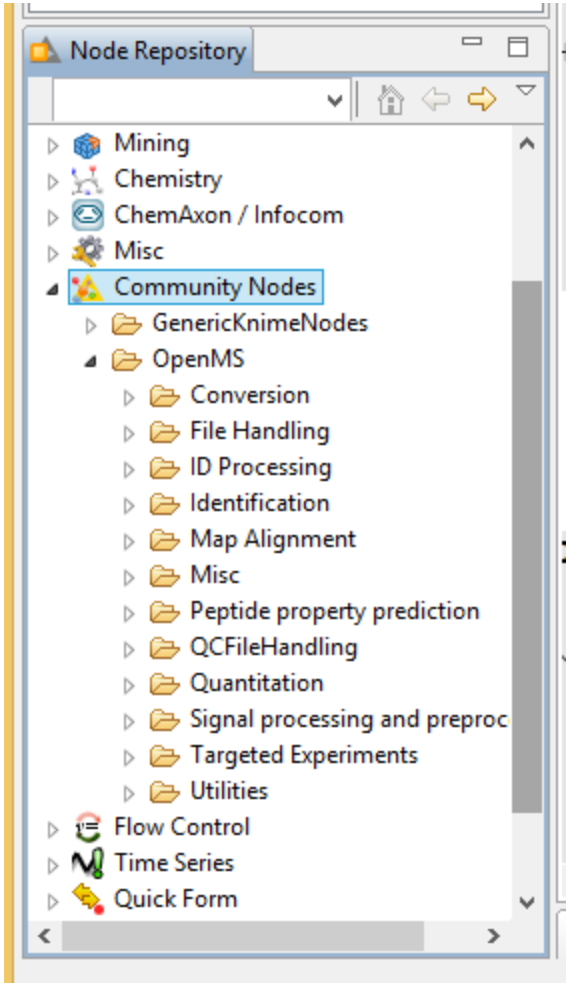


- Visualization capabilities
- Data mining & advanced statistical methods

http://knime.org

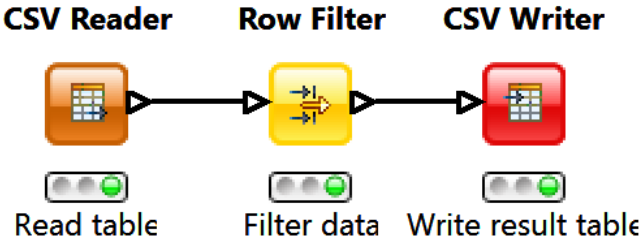
Installation of KNIME and OpenMS

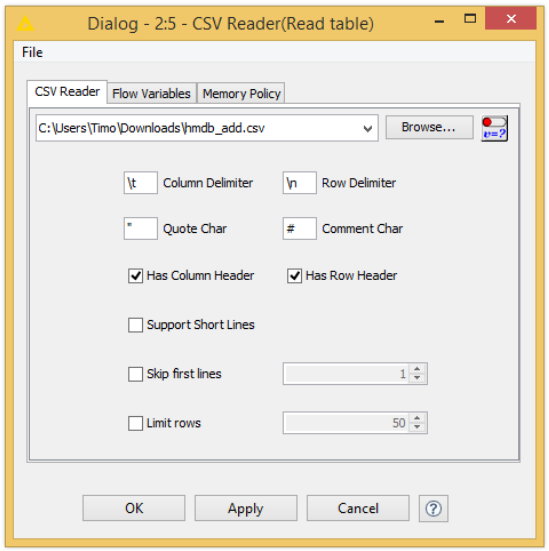
- KNIME installers available from: www.knime.org
- KNIME provides a sophisticated plugin system:
  - Many additional nodes can be installed as KNIME extensions
  - OpenMS installation in KNIME provides all TOPP tools as separate nodes
  - Nodes can be found in the folder ‘Community Nodes’
  - Detailed instructions on how to install OpenMS nodes in the additional materials

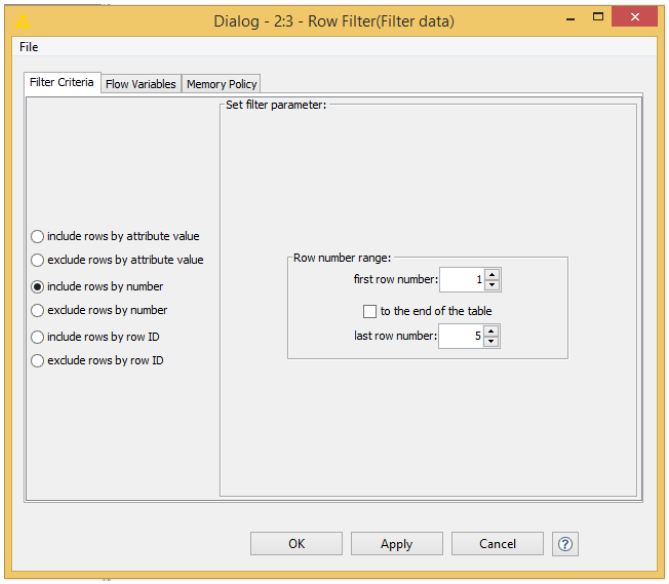


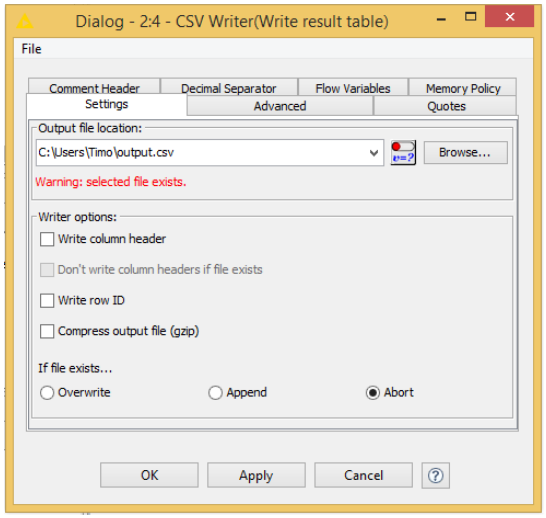
Simple Workflows in KNIME

- KNIME workflows consist of distinct nodes that can be assembled into workflows
- Workflow construction via drag and drop
- Either **tables** or **files** are exchanged between nodes along the edges of the workflow
- Files are marked by square ports, tables by triangular ports
- **Configuration dialogs** exist for all nodes

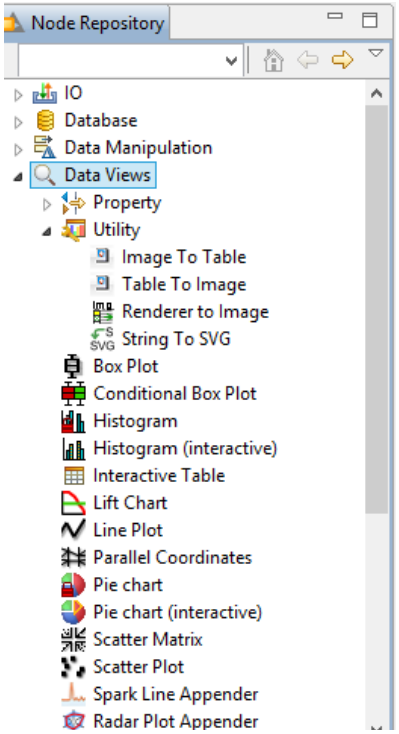








Simple Workflows in KNIME

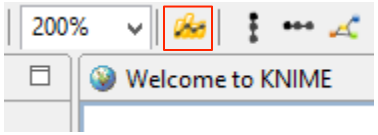


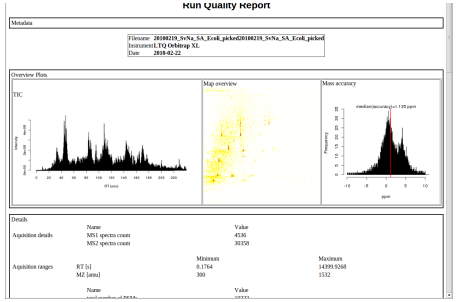
Plotting

- Data View nodes offer interactive visualization of tables
- Data can be explored interactively

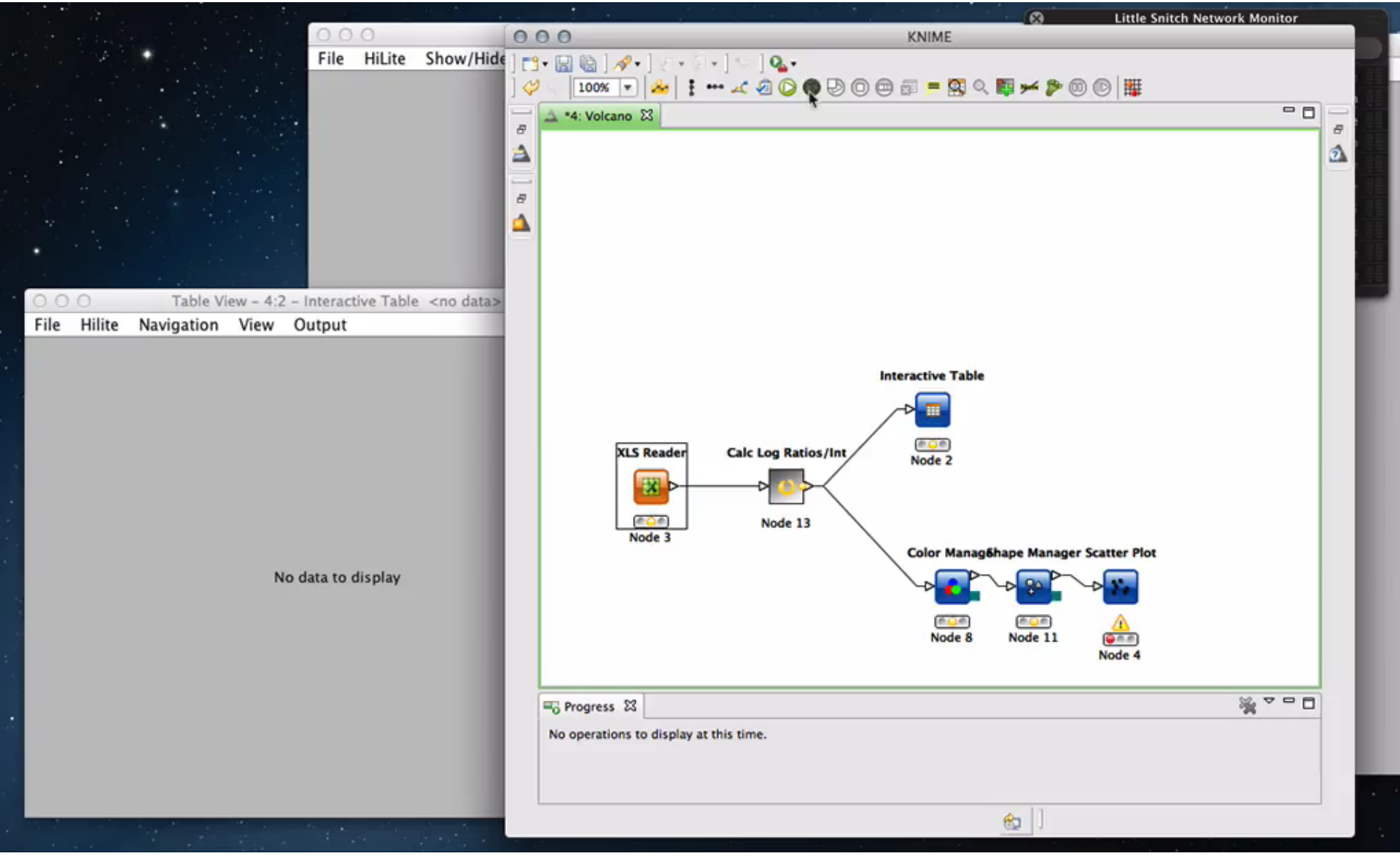
Generating reports

- Single file (e.g., pdf) from workflows
- "Data To Report" and "Image To Report" nodes specify what will be reported
- Visual construction and layout in **report perspective**



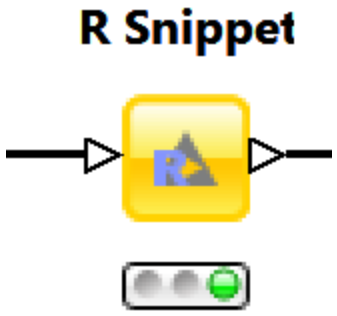


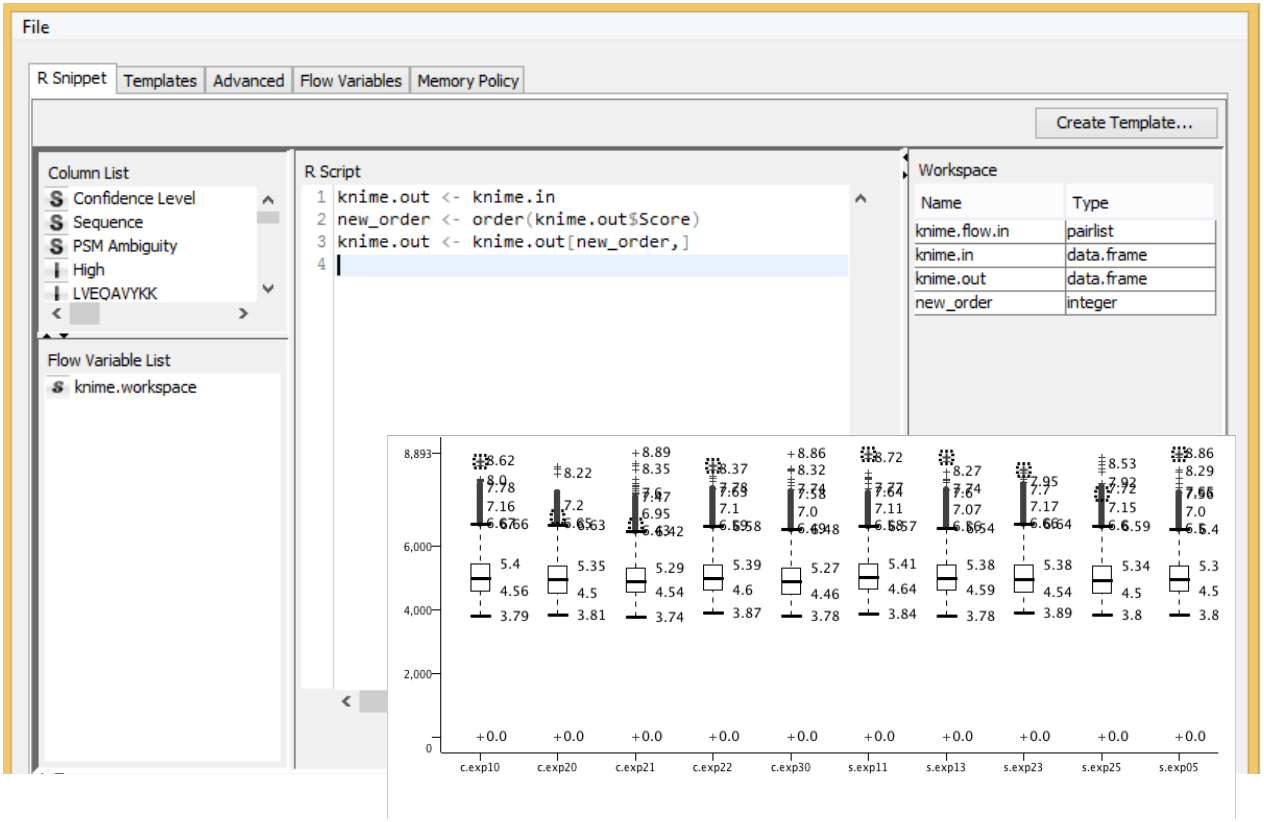
KNIME Interactive Analysis



Simple Workflows in KNIME

- KNIME permits the embedding of R code for advanced statistics
- Embedding of R scripts using the R Snippet node
- All plotting capabilities of R can be used as well





Protein Identification Workflow

- Finding all proteins in multiple samples
- Mass spectra enter workflow on the left
- Loop nodes permit execution of parts of the workflow
- Identified proteins end up in result files (right side)