Integrating Data, Tools, and Science

Michael R. Berthold

KNIME.com GmbH, Zurich, Switzerland

and

University of Konstanz, Germany

Agenda

KNIME: 2004 – today

or

How we went out to build a data analysis workbench and ended up building something else...

Why KNIME?

Why?

- University Group – Research Focus:
  - Data Mining / Machine Learning
  - Explorative Methods
  - „...Sparking Ideas“ (...PKDD/ECML 2005)
- Need for platform to develop, test, deploy, and use new methods.

Requirements

- Modular & flexible
- Professional strength (industry collaborators)
- Highly scalable (dito)
- Easy to extend (legacy and third party tools)

„I am working on KNIME“

(PhD Student, 2004/2005)

„I am working with KNIME“

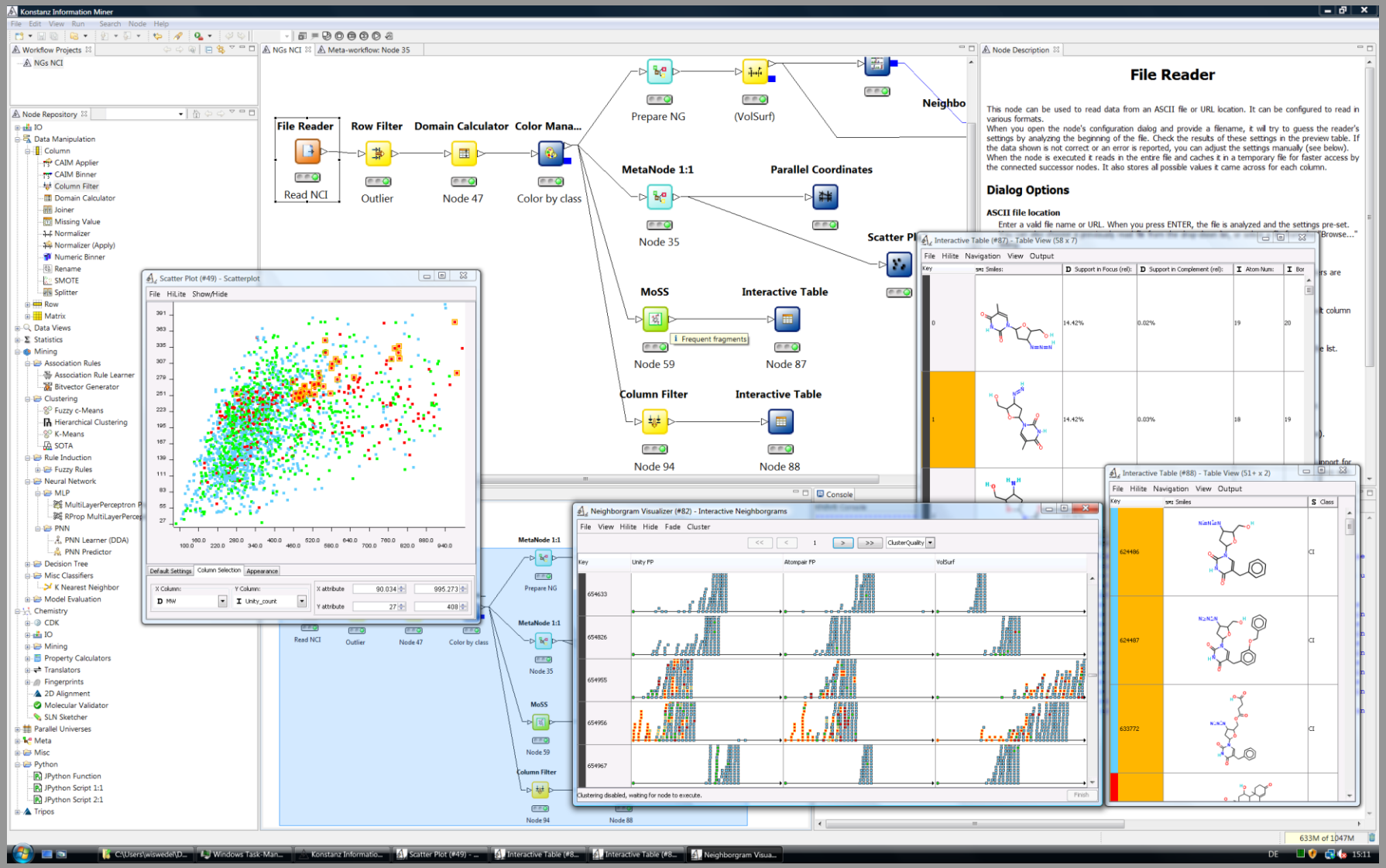
(PhD Students, 2006...)

June 2006:

First Open Source Release (v1.0)

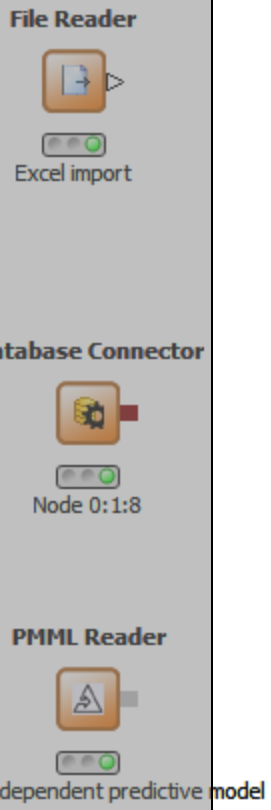
What is KNIME?

The KNIME Platform

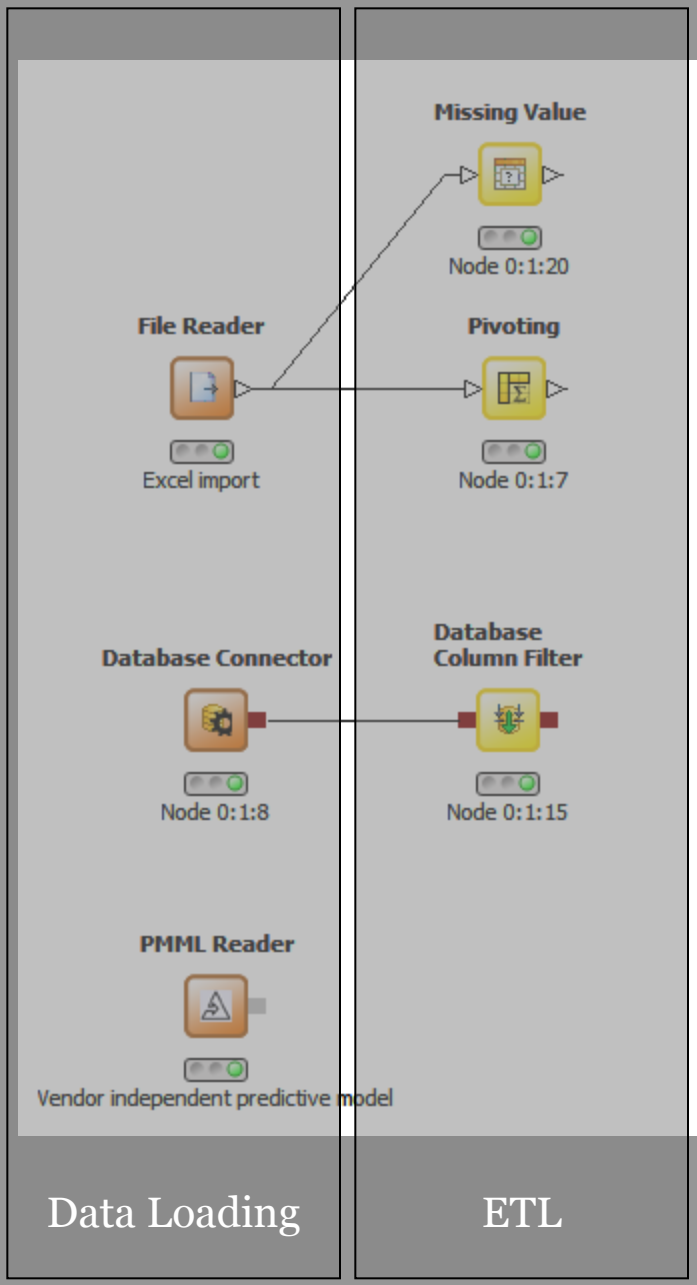


KNIME loads and integrates data from diverse data sources:

- Different data bases
- Various file formats (CSV, XML, SDF, etc.)



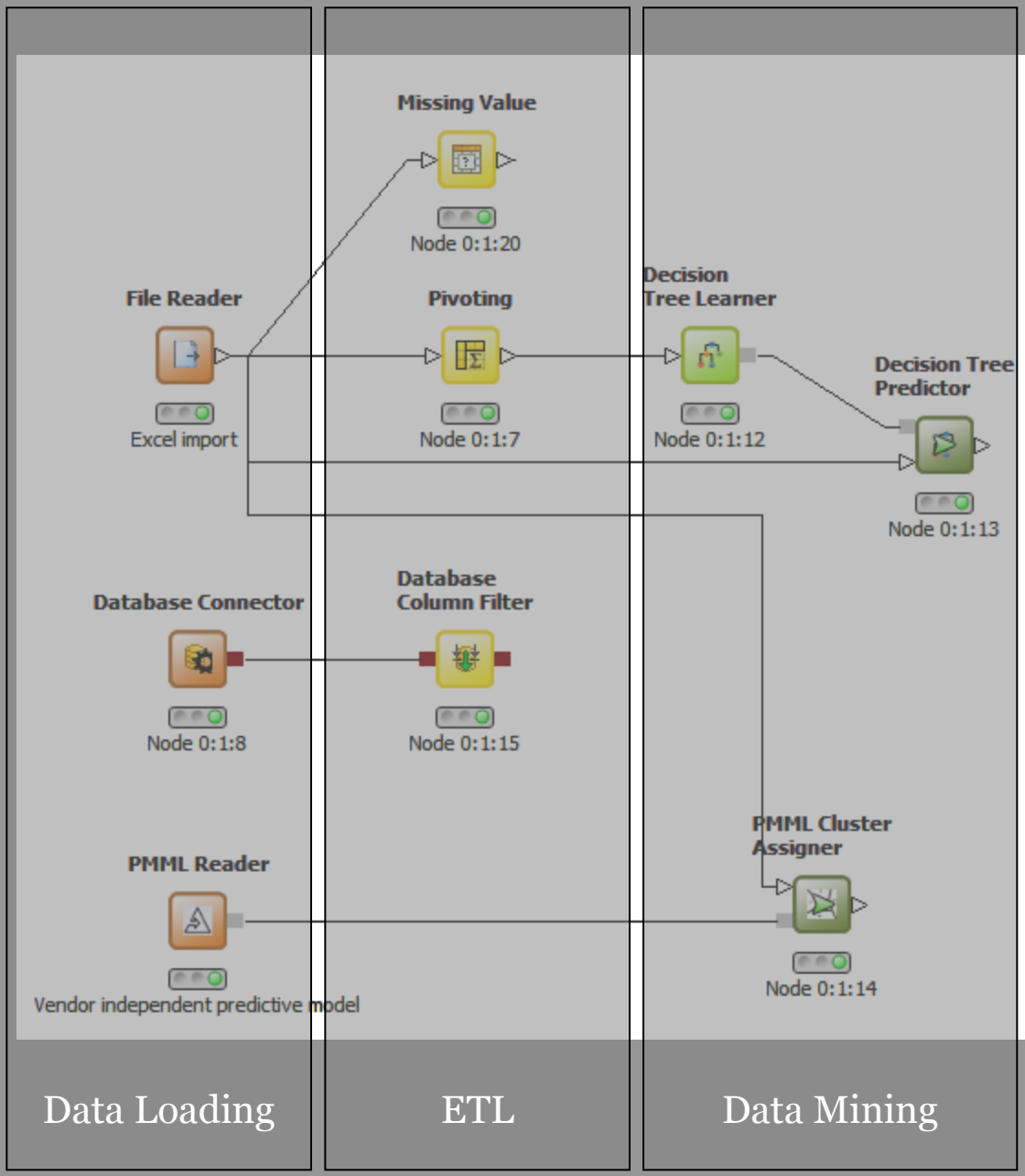
KNIME



KNIME provides huge repository of modules for easy-to-use, modular

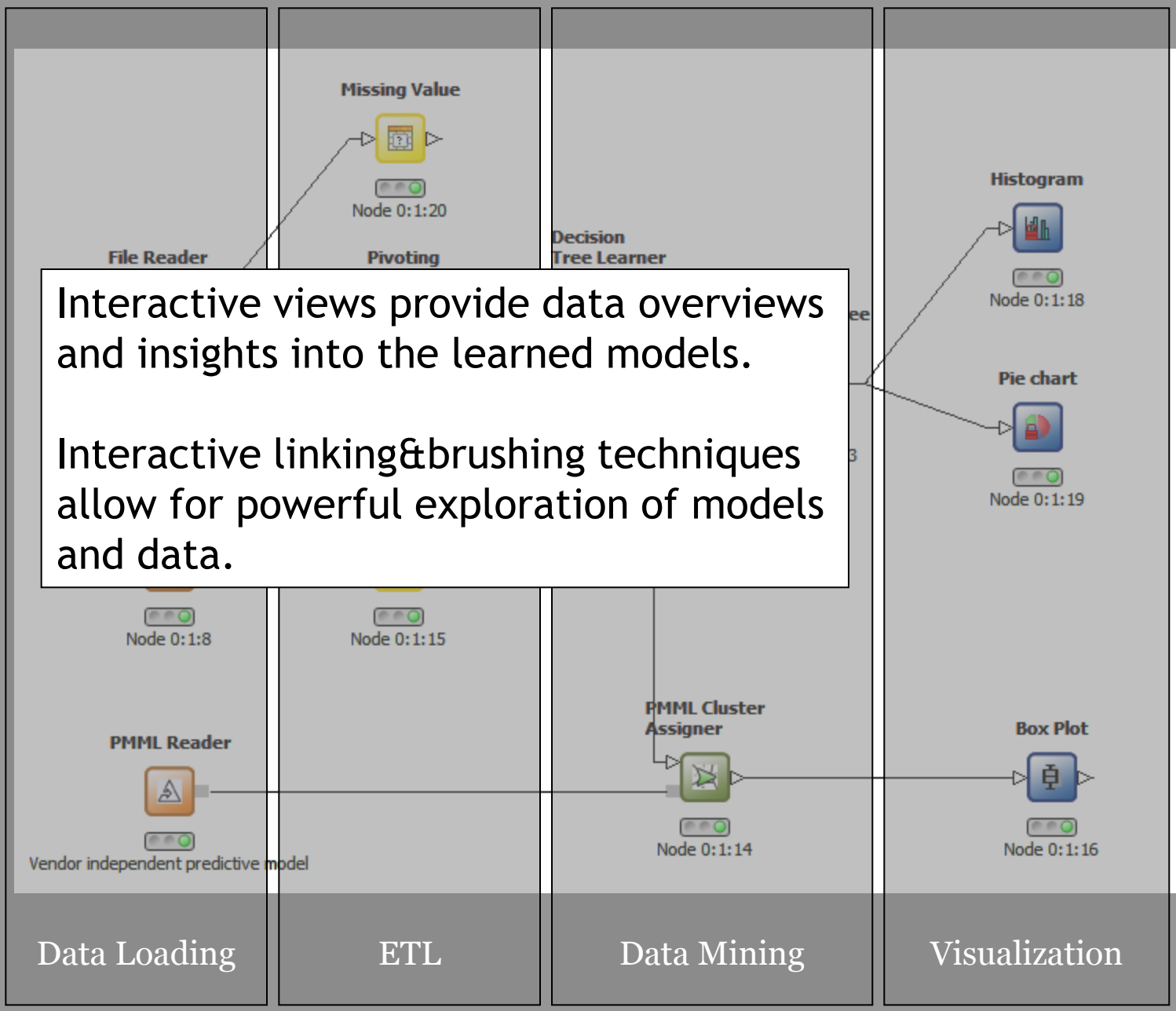
- Data preprocessing
- Data fusion
- Data transformation

KNIME



In addition to standard data mining techniques, KNIME adds cutting edge data analysis algorithms.
(…thanks to its academic roots)

KNIME



Interactive views provide data overviews and insights into the learned models.

Interactive linking&brushing techniques allow for powerful exploration of models and data.

A peak under the hood

- Java
  - Runs on Linux, Windows, Mac
- Eclipse plugin
  - Extension wizard for own nodes
  - Simple installation of additional modules (text, image, time series, ... analysis)
- Out of memory data handling
  - Caches data to HD
  - Scales linearly (e.g. 200 – 20.000.000 images)
- Support for multi threaded execution

- Architecture designed by professional SW engineer

„But I have all those R scripts“

„Where is this algorithm?
(Weka has it!)“

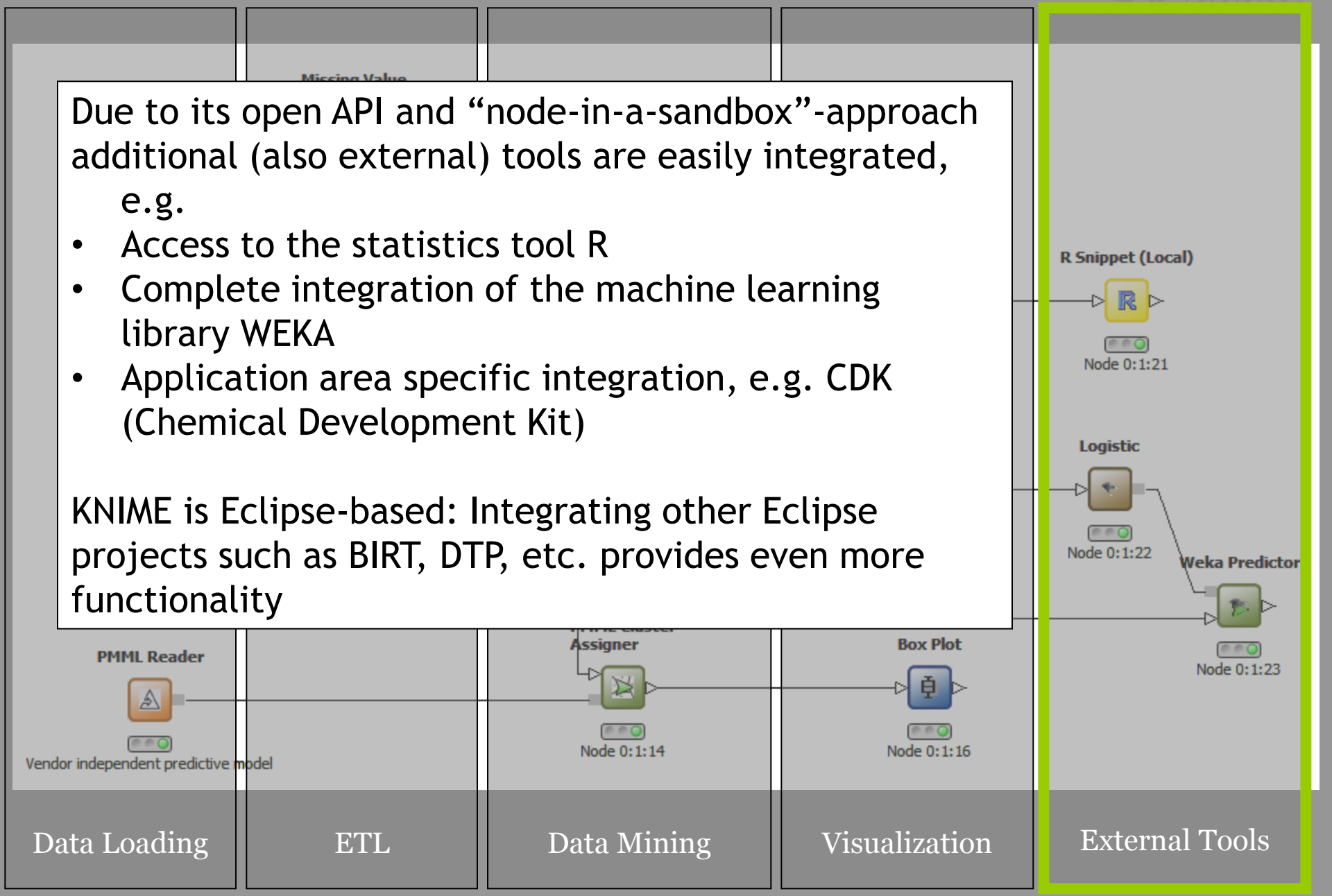
...re-invent the wheel?

KNIME

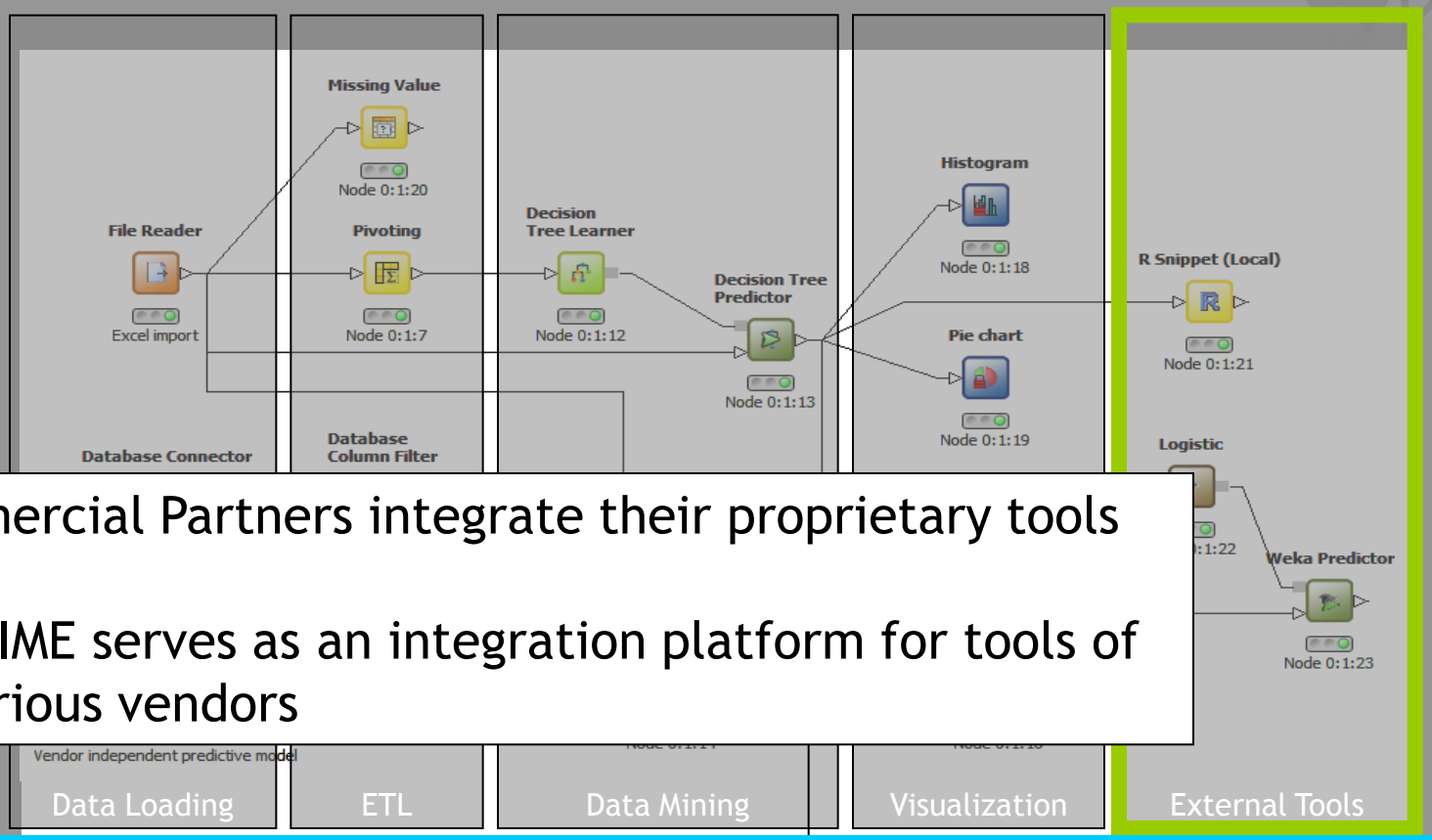
Due to its open API and “node-in-a-sandbox”-approach additional (also external) tools are easily integrated, e.g.

- Access to the statistics tool R
- Complete integration of the machine learning library WEKA
- Application area specific integration, e.g. CDK (Chemical Development Kit)

KNIME is Eclipse-based: Integrating other Eclipse projects such as BIRT, DTP, etc. provides even more functionality

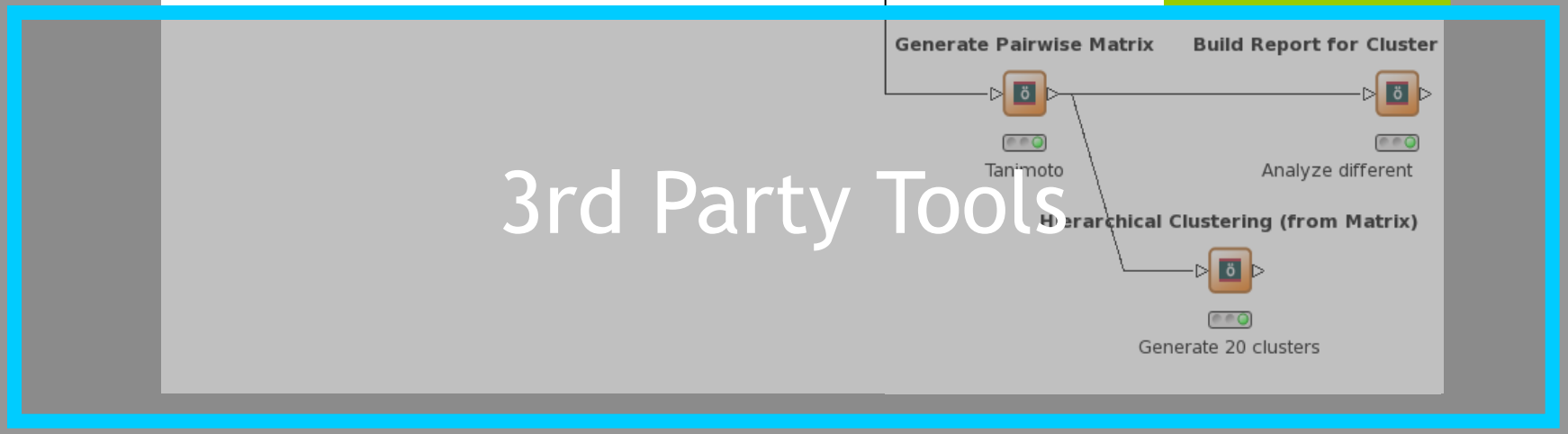


„We bought this wonderful tool“



Commercial Partners integrate their proprietary tools

⇒ KNIME serves as an integration platform for tools of various vendors

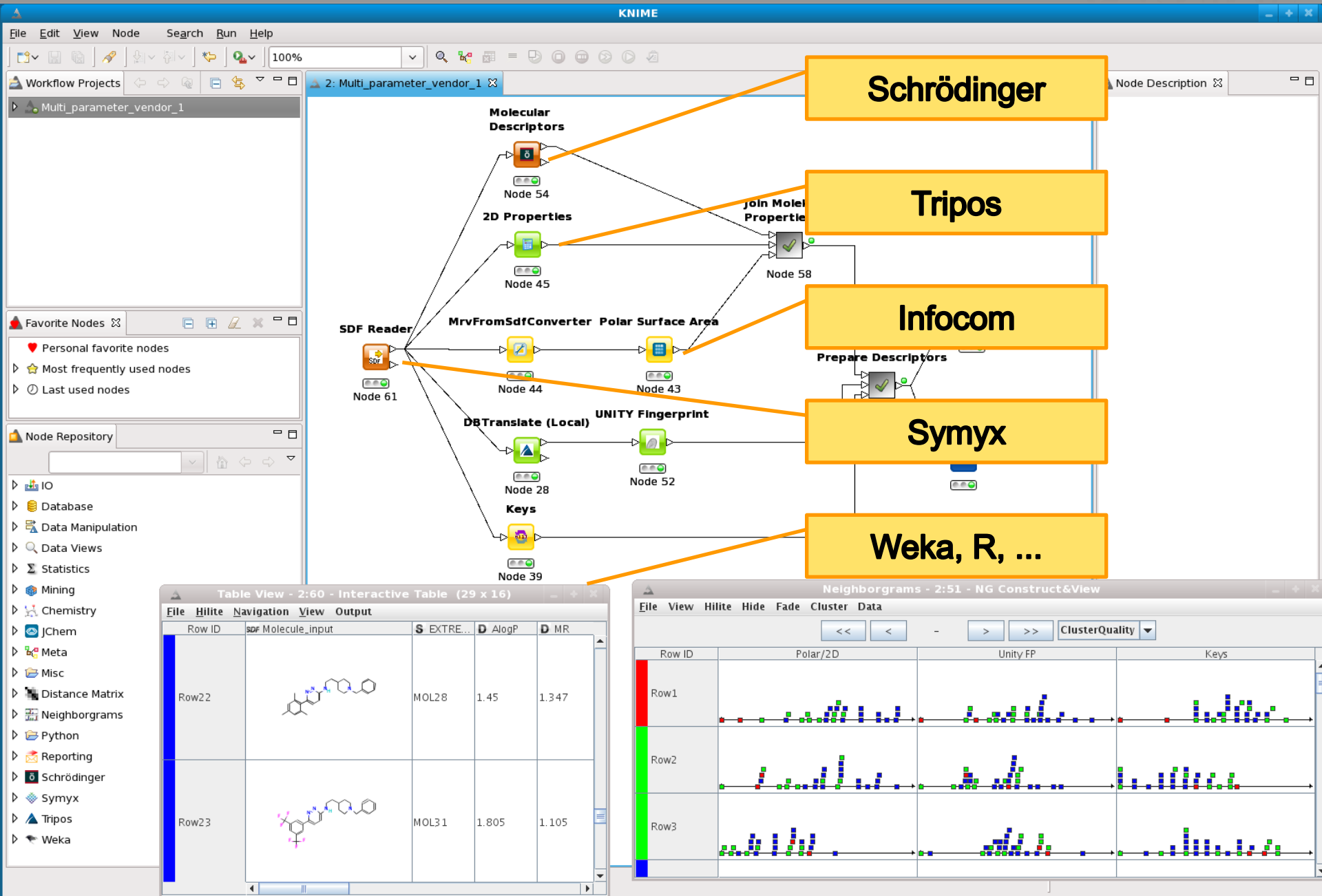


3rd Party Tools

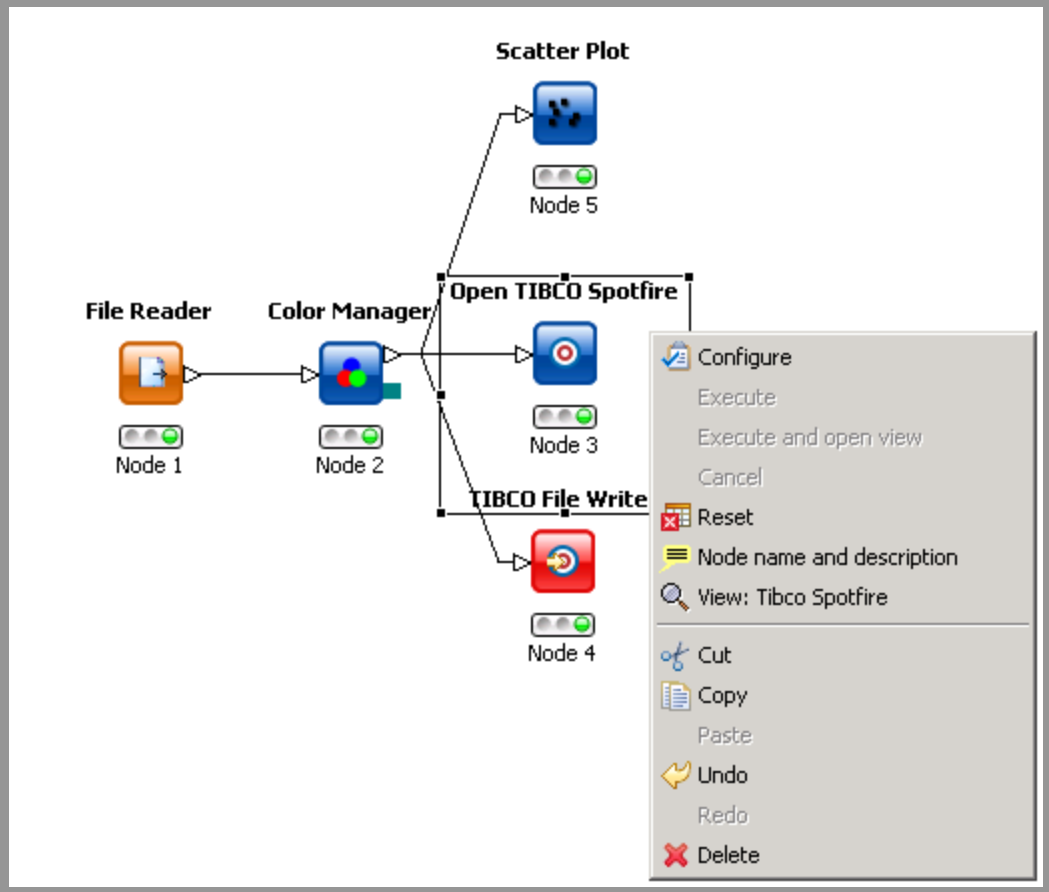
Some KNIME Partners

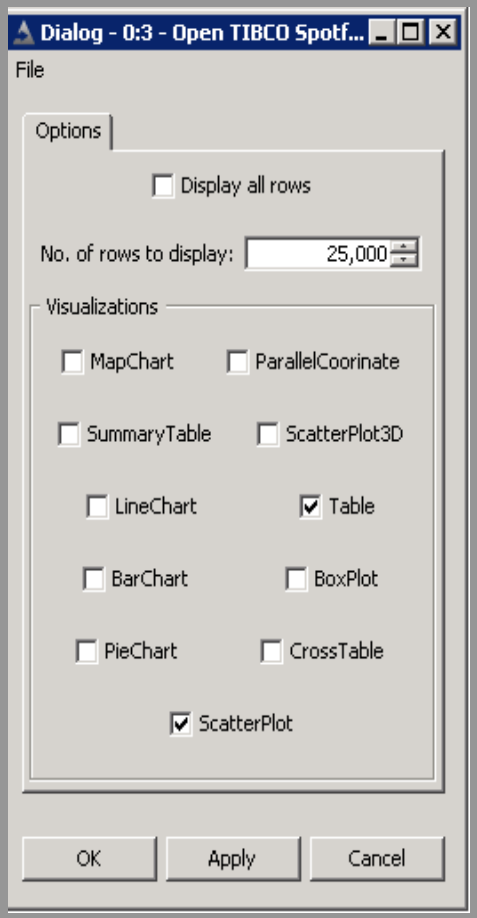
- CambridgeSoft
- Schrödinger
- Symyx Technologies
- Tripos
- ChemAxon (via Infocom)
- CDK
- Molecular Discovery
- Treweren
- Tibco / Spotfire
- …
- And others: BioSolve IT, CCG, …

Partner Extensions at Work

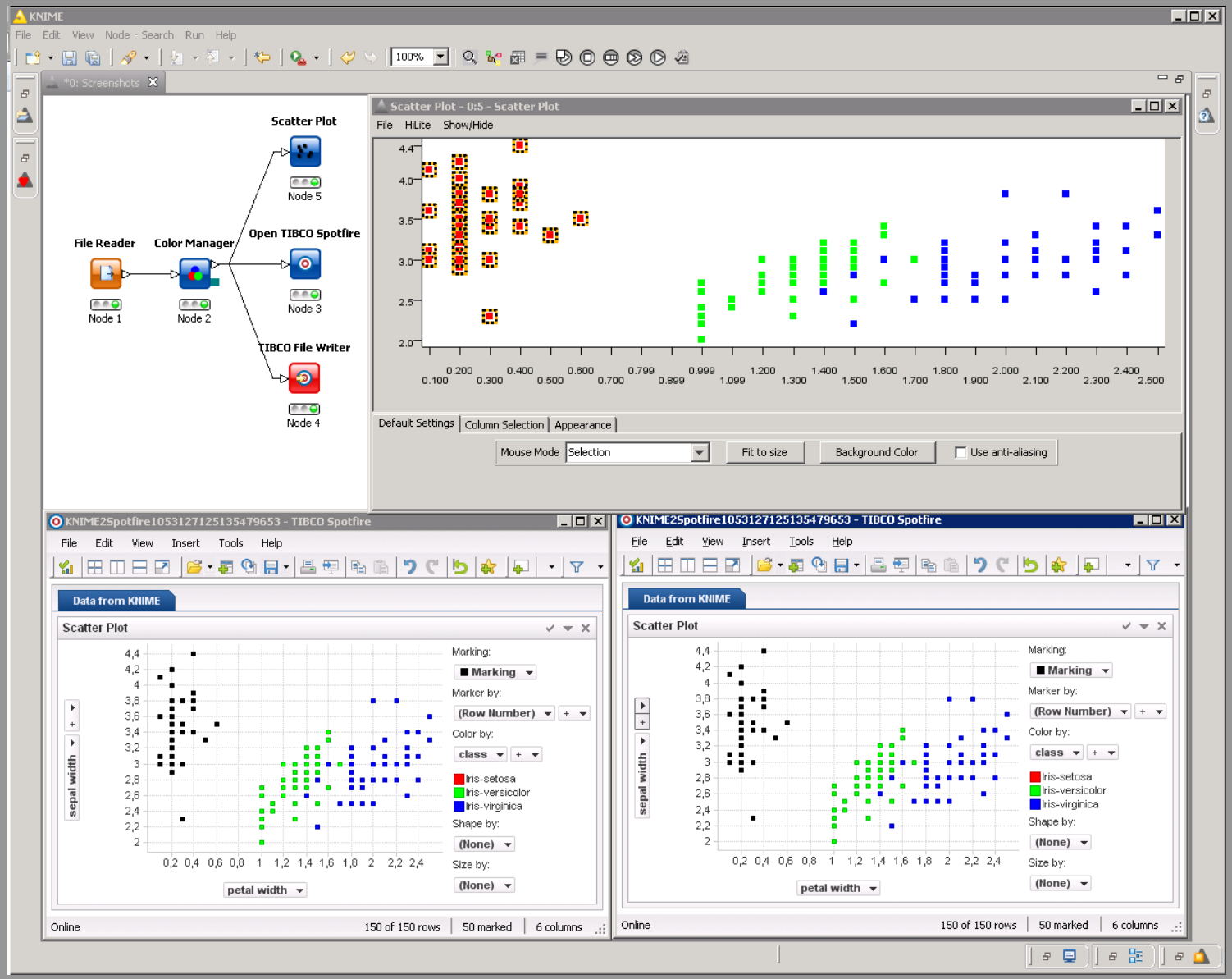


Partner Extensions at Work: Spotfire





Partner Extensions at Work: Spotfire



„I didn‘t mean someone else‘s tool – our own is so much better.

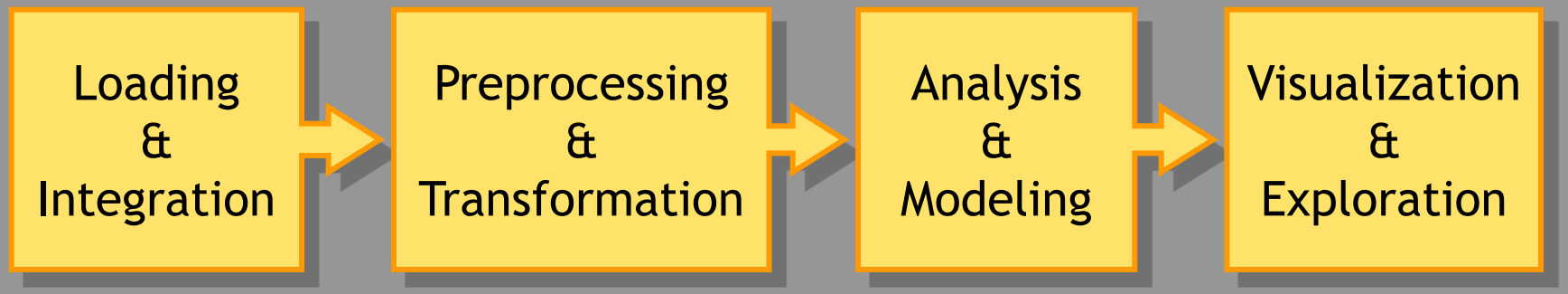
(nobody understands it anymore but everybody uses it)!“

Integration of Tools, cont.

- External Tool Node
  (launches shell commands)
- Web Service node
- New Node Wizard
  - for thin wrappers
  - for your own prototyping / development!

KNIME:

“Integrating Data, Tools, and Science”

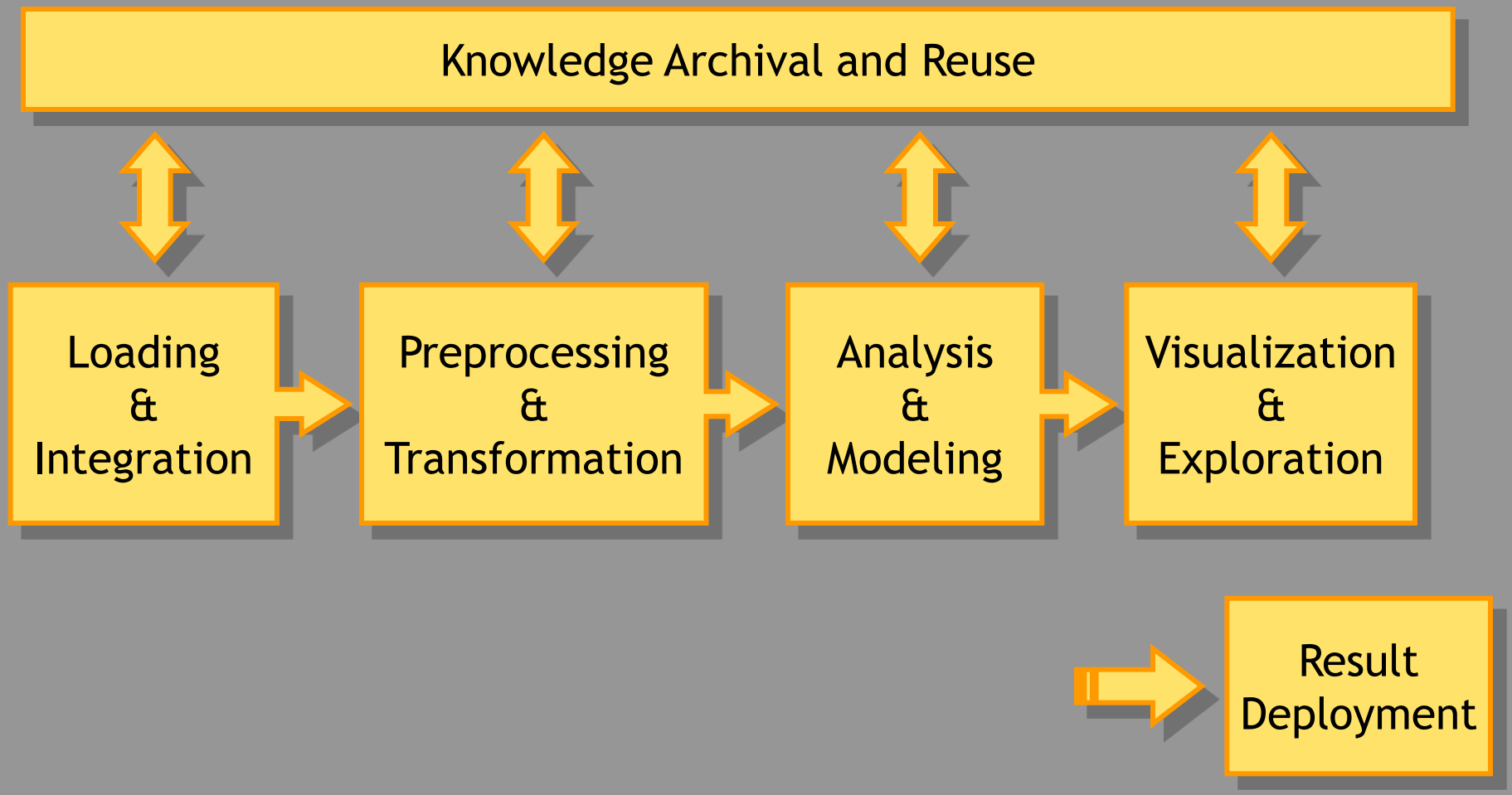


Data: Integrate distributed & heterogeneous data sources
Tools: Integrate Open Source, Third Party and In-house Tools
Science: Standardize Analysis and Modeling Workflows

„Very nice, but...“

- „...I want to share workflows!“
- „...I want to run workflows on a server...“
- „...and our cluster!“
- „...I want to show the results to upper management!“

The Larger Picture



Wait a second – is this still research?

June 2008:

KNIME.com GmbH
founded in Zurich, Switzerland

KNIME.com GmbH

Software & Services:

- Support for Open Source KNIME Platform
- Consulting
- Development:
  - Customizations
  - KNIME Extensions
  - Enterprise Components

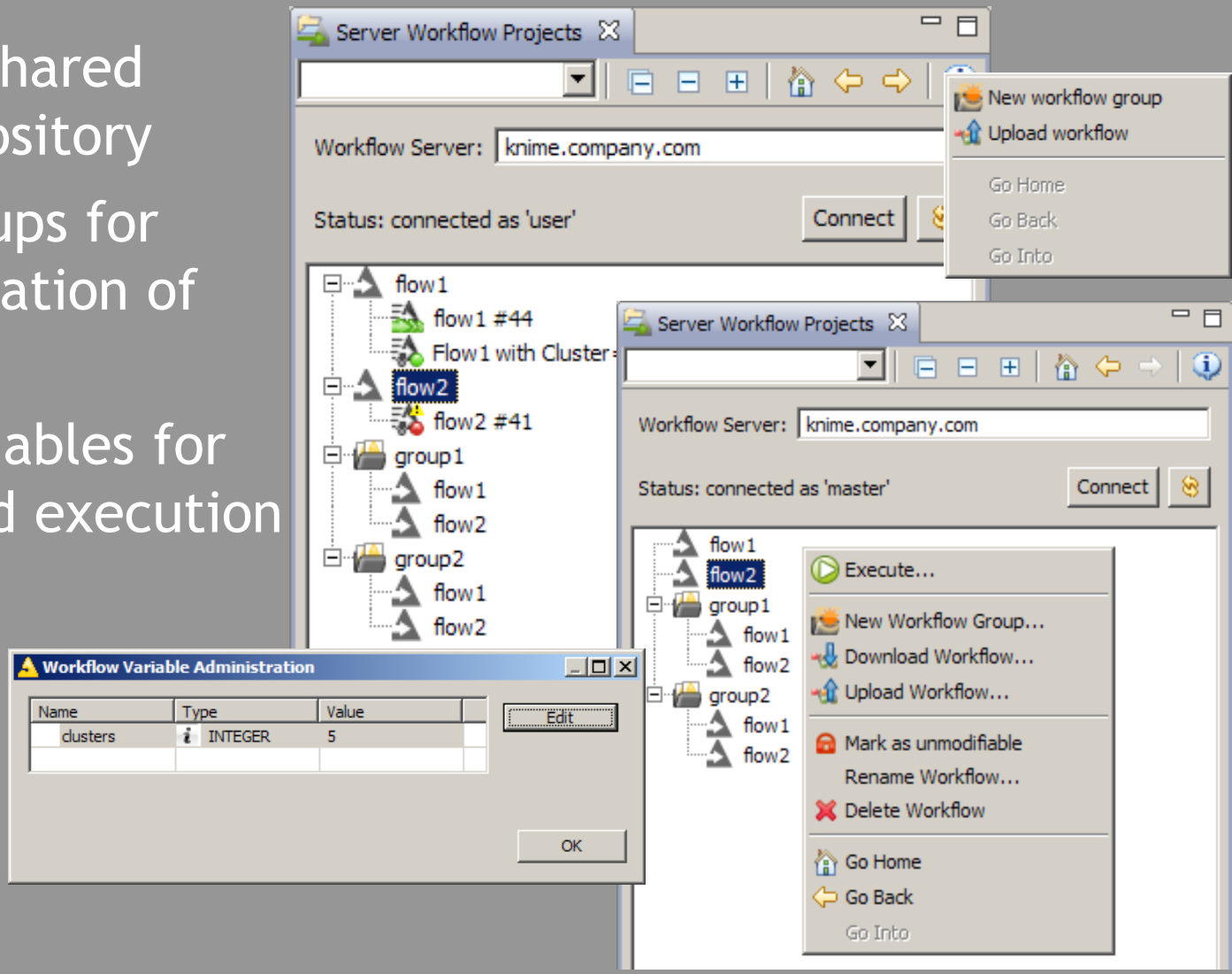
Knowledge Archival and Reuse

- Workflow Repository:
  - predefined flows: standardization
  - Workflow archival: knowledge repository
- Metanode Repository:
  - workflow fragments: expert knowledge archive

⇒ KNIME Enterprise Server
- Central corporate repository for workflows
- User access control
- Scheduled execution
- Backend data processing engine for SOA integration

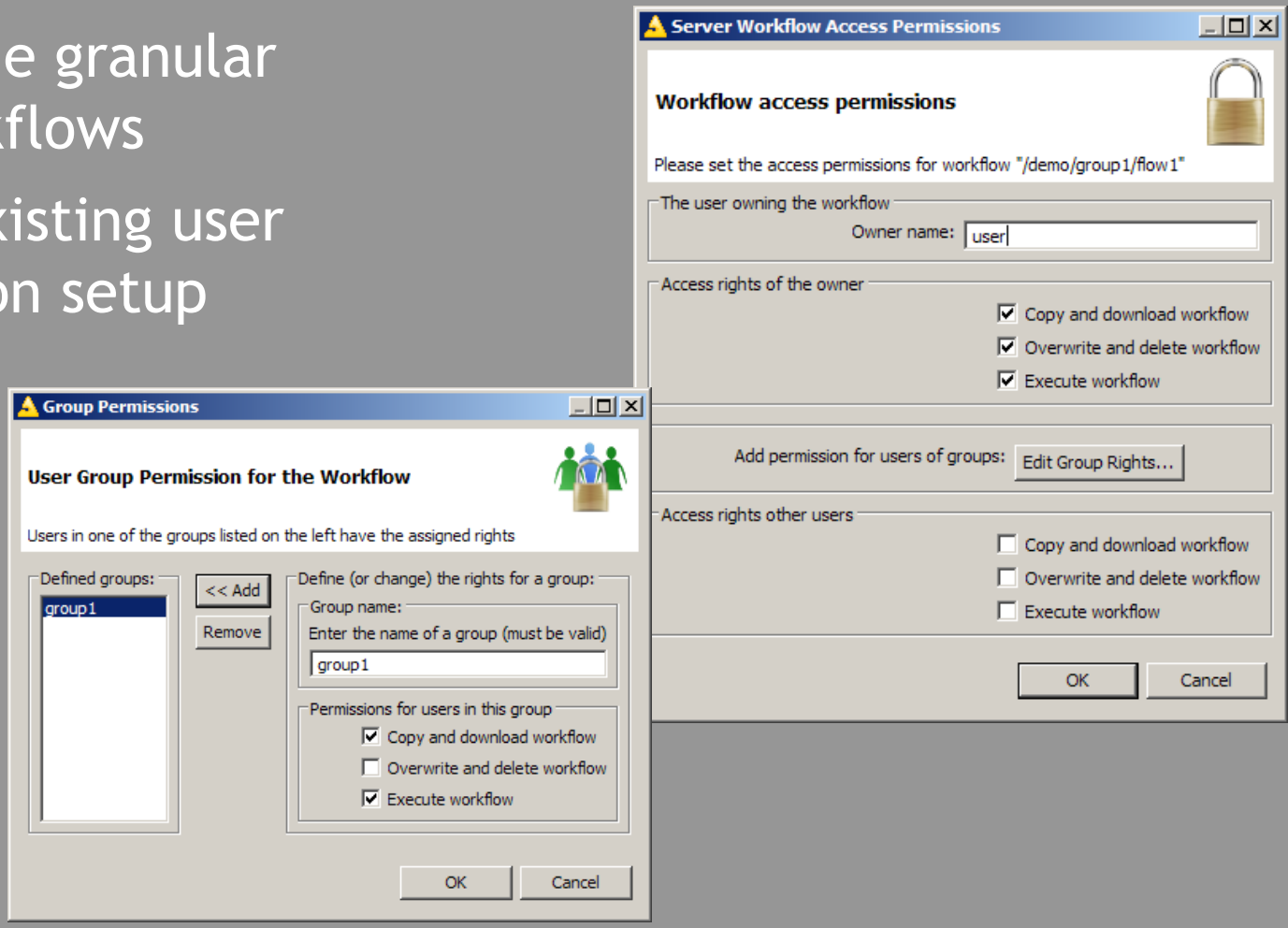
Enterprise Server

- Centralized, shared workflow repository
- Workflow groups for easier organization of repository
- Workflow Variables for parameterized execution



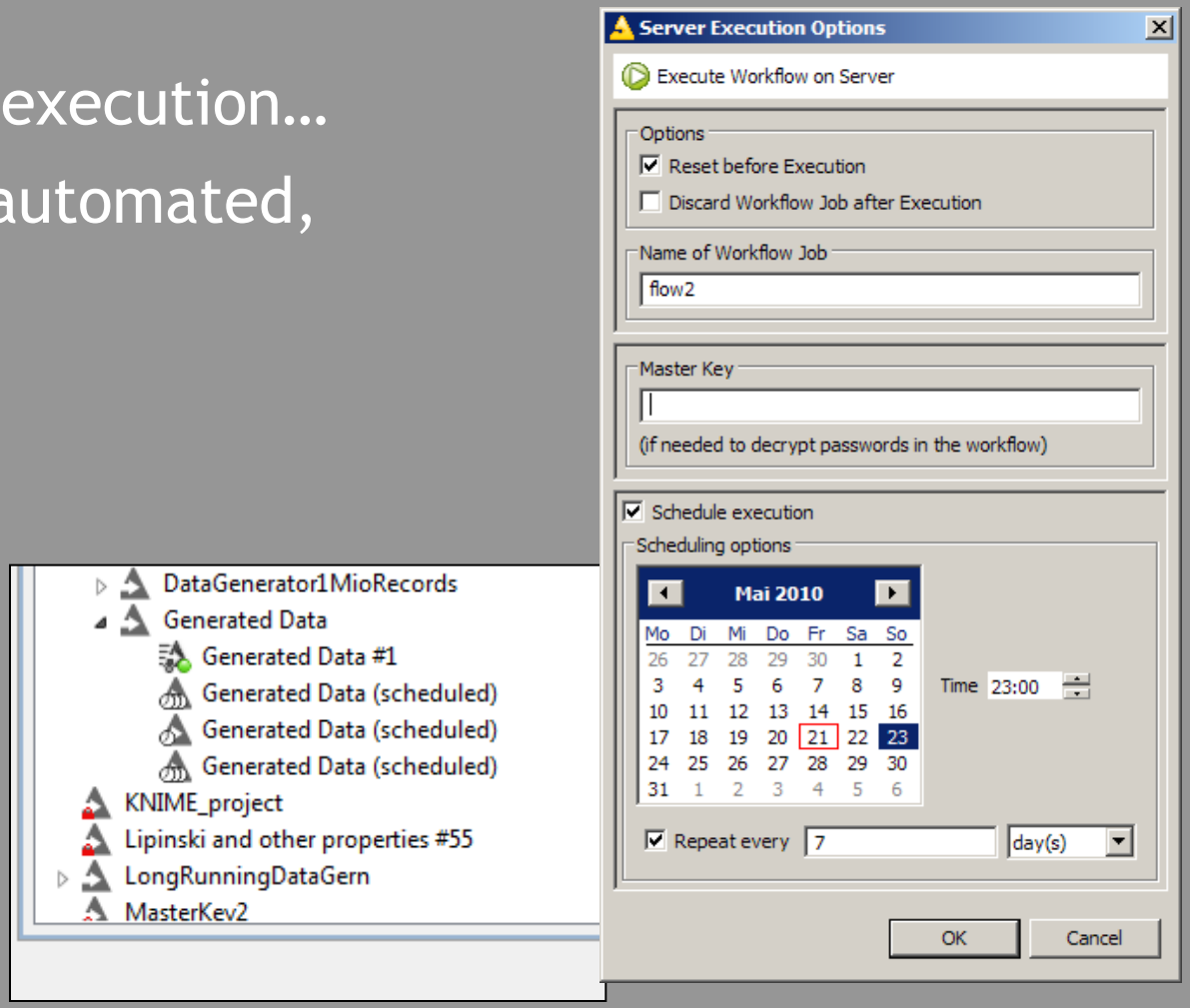
User Rights Management

- Controlled, fine granular access to workflows
- Connects to existing user authentification setup



Scheduled Execution

- Manual server side execution…
- …or scheduled for automated, regular execution.



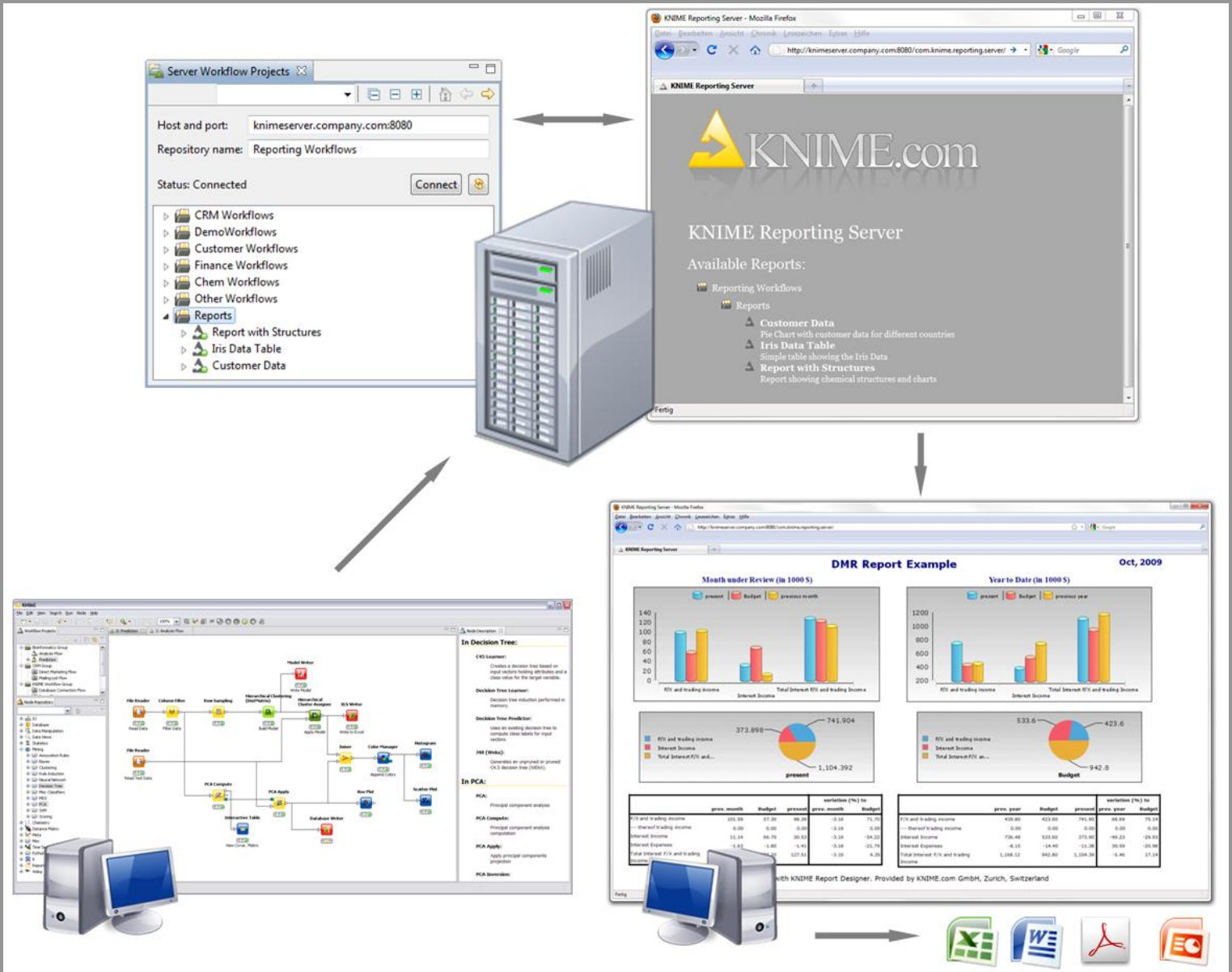
Result Deployment - 1

- Reports:
  - KNIME workflows load, aggregate, analyze data
  - Reports summarize and enrich workflow output (tables, graphs, images) – a BIRT integration, btw…
  - "one click" update for new data or parameters

⇒ KNIME Report Designer and Server

- Provides access to KNIME workflows without exposing the underlying workflows and nodes
- Templates created by expert
- End users execute template and export report to desired format (PPT, DOC, PDF, HTML, …)

KNIME Report Server



Result Deployment - 2

- Models:
  - Export of Trained Models
- Workflows
  - Access to workflows from other applications

⇒ PMML Export of models
⇒ Access to workflows via web service calls

KNIME

Just another

Data Analysis Workbench?

KNIME: the backbone of your data

Specialized Extensions
- Chemo and Bio Informatics
- Image Analysis
- Time Series, Text, …
- … Partner Tools

Analytics / Modeling
- Built-In Data Analytics
- R-Project
- Weka, LibSVM, CDK…
- …Partner Tools

Exploration
- Interactive Views
- KNIME Reports
- …Spotfire

Integration
- KWaaS

Information Backbone:
KNIME Workflows

Legacy SW
- Ext Tool
- WS / PP

Collaboration
- KNIME Enterprise Server

HPC
- Cluster Execution
- KNIME o.t. Cloud

Deployment
- KNIME Report Server
- PMML
- KWaaS

Status Quo

- KNIME 2.2.0 released (July 2010)
- >5,500 active users
  - ~50% Life Science
  - ~50% Business Intelligence, Analytics
- 10+ commercial partners
- Distributors in Japan, Italy, China, Brazil
- Since Aug 2010: Global Support and Distribution Agreement with CambridgeSoft

Integration!

- **Data**: form one view for all activities (transformation, modeling, analytics, reporting)
- **Tools**: open source and commercial (own and others)
- **Science** (&Expertise): archive and reuse workflows

try it out – it's free:

www.knime.org