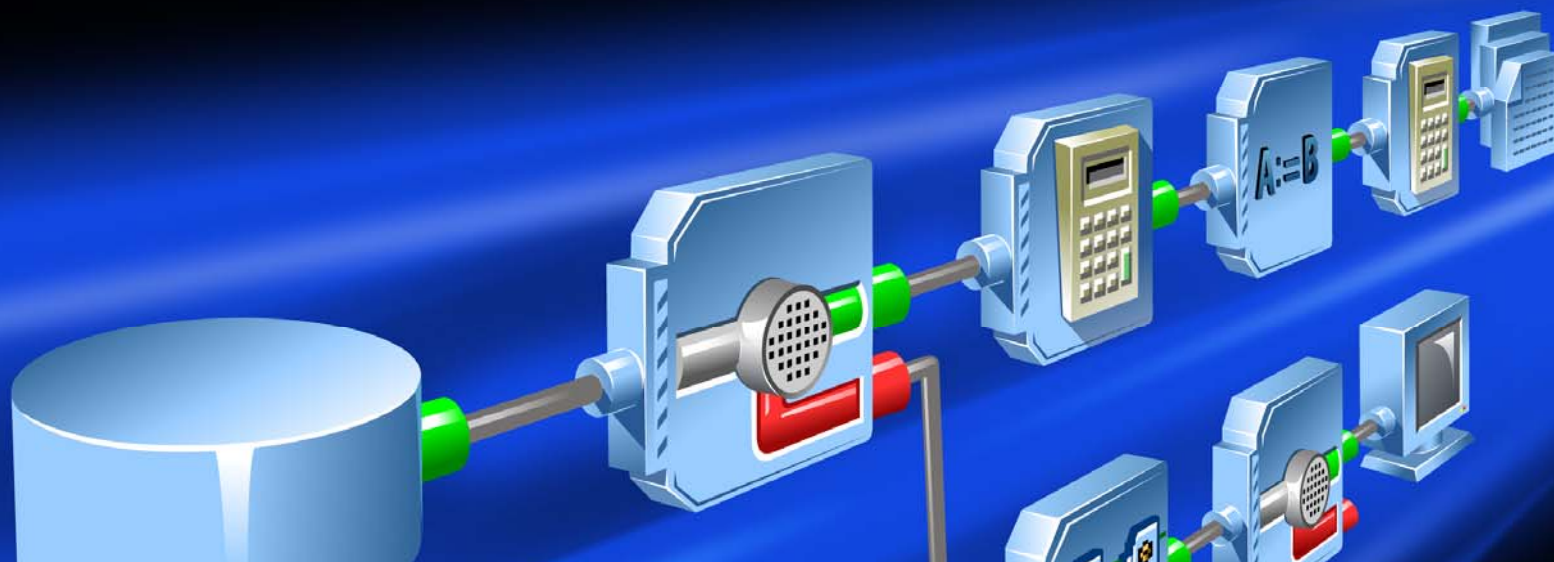




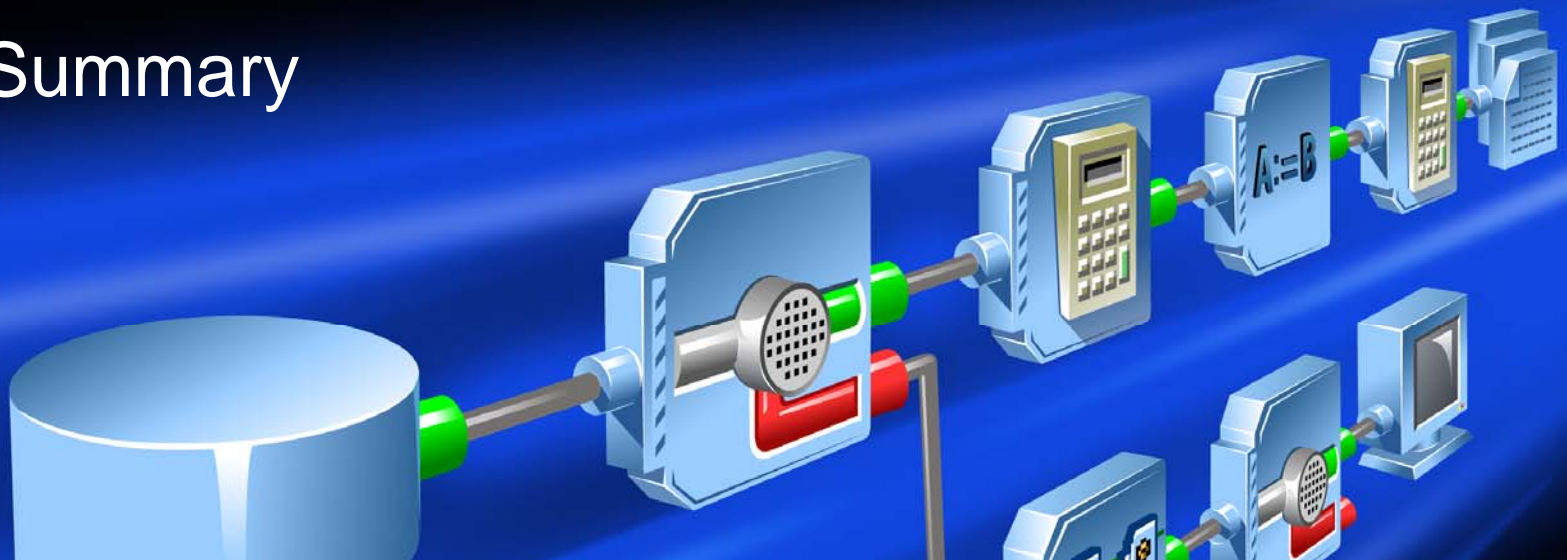
ask *more* of your *data*

# Automation of Cheminformatics Tasks using Pipeline Pilot



# Outline

- Challenges in Cheminformatics
- An Introduction to Data Pipelining
- Examples
  1. Data integration, mining and reporting
  2. Deployment through end-user clients
- Summary

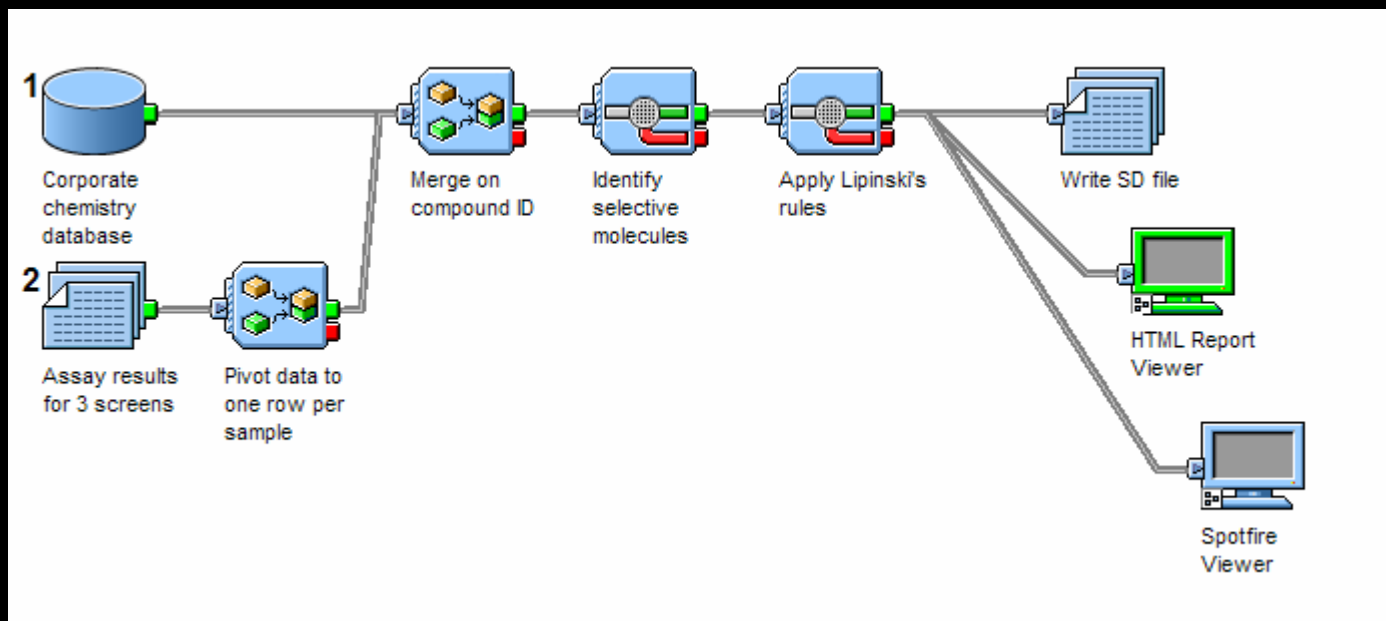


# Challenges in Cheminformatics

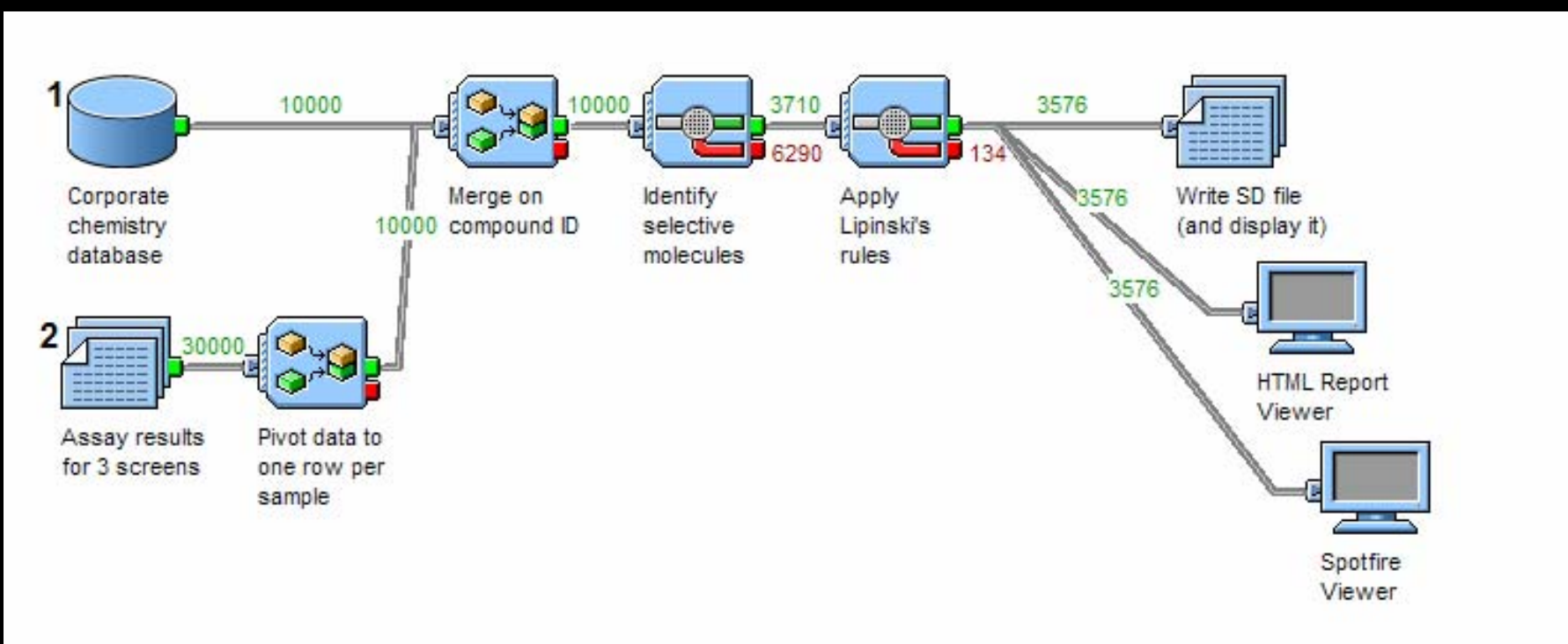
- Integrating diverse data & applications
- Automating data processing & analysis
- Capturing & deploying best-practice processes

# Data Pipelining

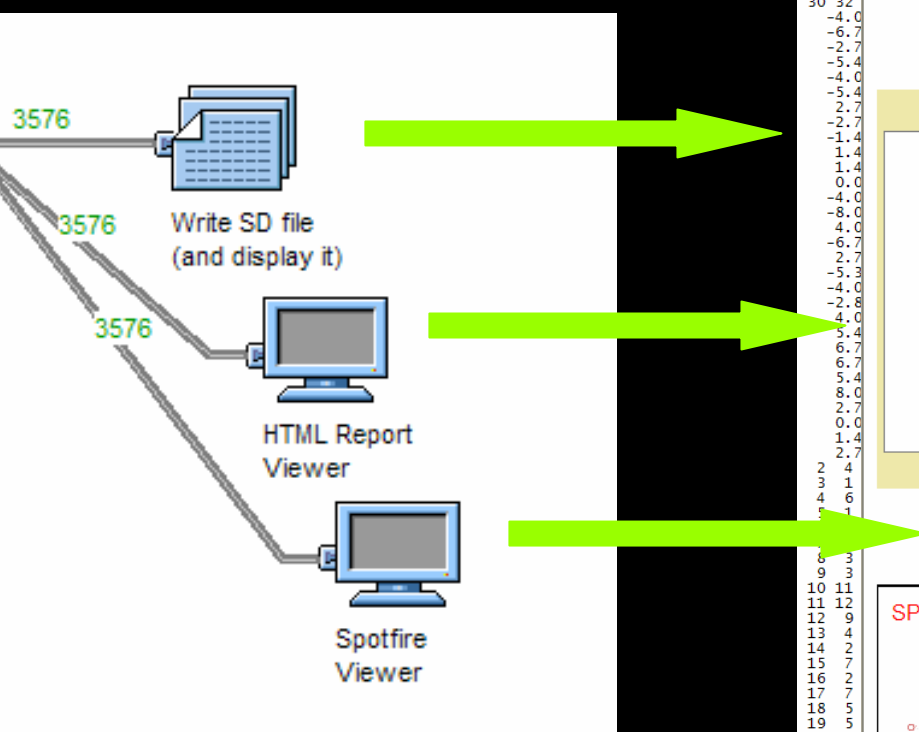
- A powerful paradigm for data processing
- Pipelines guide the flow of data through a series of modular computational components



# Execution...



# Results...



## Project ABC-123 Activity Report

Selectivity

egfr

Spotfire DecisionSite 6.3 - 163.tmp - [Scatter Plot]

Query Devices

- CODE (ALL)
- cdk2 (ALL)
- egfr (ALL)
- pkca (ALL)
- selectivity

Details-on-Demand

Click on a record or mark several to see details here.

SPB

pkca:  
egfr:  
cdk2:

SPB 07585    SPB 07587    SPB 07588    SPB 07589

3576 out of 3576 records visible (100.00 %), 0 marked

SPB 075

Scite

30 32  
-4.0  
-6.7  
-2.7  
-5.4  
-4.0  
-5.4  
2.7  
-2.7  
-1.4  
1.4  
1.4  
0.0  
-4.0  
-8.0  
-6.7  
-5.3  
-4.0  
-2.8  
4.0  
-6.7  
2.7  
-5.3  
-4.0  
2.7  
0.0  
1.4  
2.7

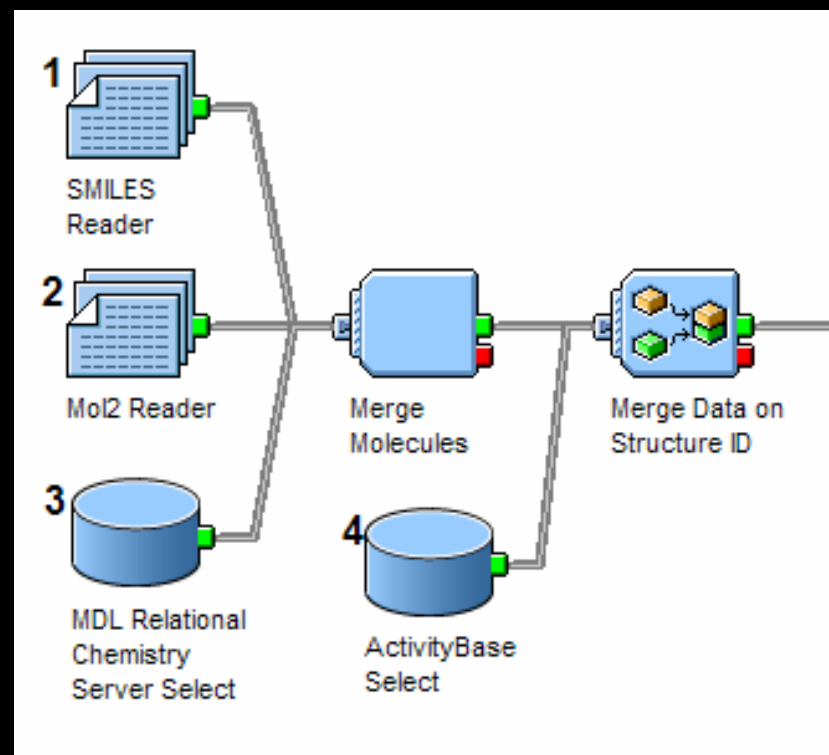
2 4  
3 1  
4 6  
5 1

6 3  
7 6  
8 3  
10 11  
11 12  
12 9  
13 4  
14 2  
15 7  
17 7  
16 2  
18 5  
19 5  
20 5  
21 15  
22 15  
23 24  
24 22  
25 21  
26 23  
27 11  
28 12  
29 28  
30 29  
8 13  
27 30  
25 23  
M CHG  
M END  
> <Name  
SPB 075

Chemical structures: c1ccccc1, c1ccc(F)cc1, c1ccc(N)cc1, O=C(O)N

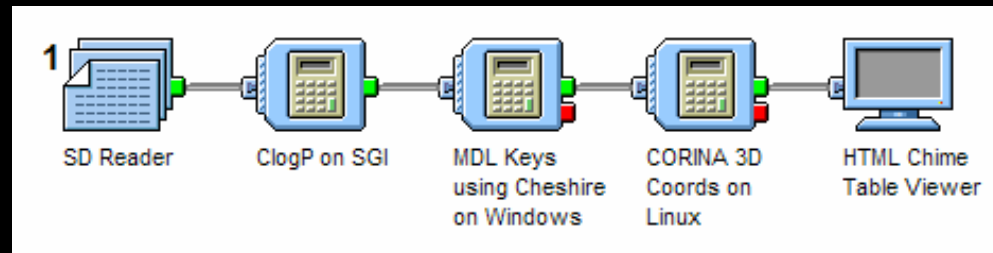
# Data Pipelining Enables

- Integration of data from multiple disparate data sources
- Integration of disparate applications
- Automated execution of routine processes
- Capture and deployment of best practice



# Data Pipelining Enables

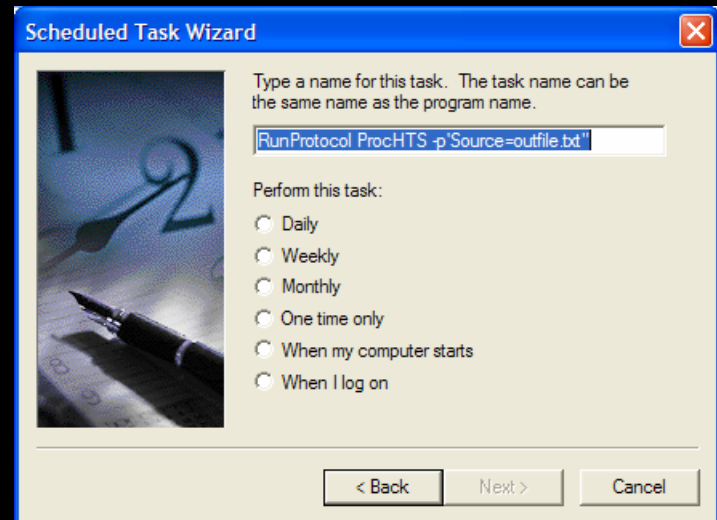
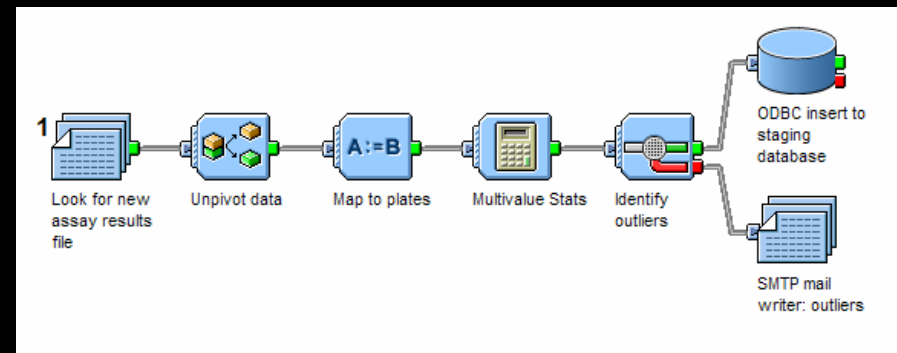
- Integration of data from multiple disparate data sources
- Integration of disparate applications
- Automated execution of routine processes
- Capture and deployment of best practice



The Structure	CLogP	CLogP_MR	MDL_SS_Keys_166_N
	2.002	5.160	32
	1.372	3.591	28
	0.721	3.651	32

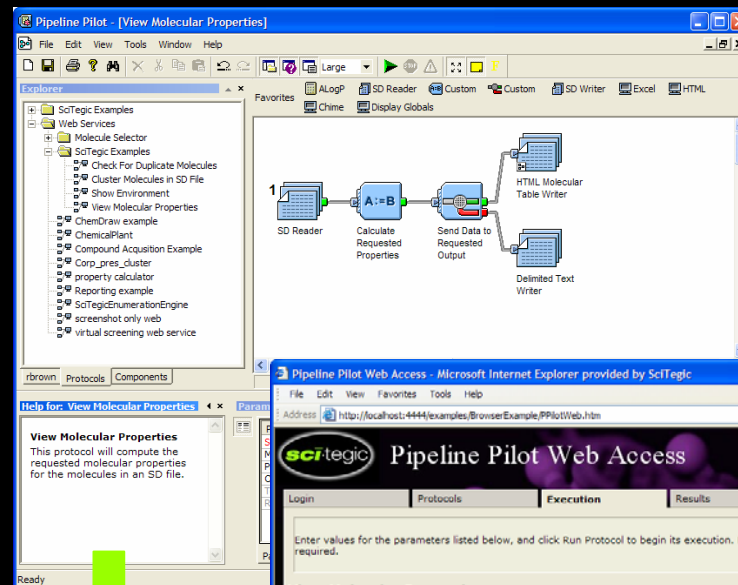
# Data Pipelining Enables

- Integration of data from multiple disparate data sources
- Integration of disparate applications
- Automated execution of routine processes
- Capture and deployment of best practice



# Data Pipelining Enables

- Integration of data from multiple disparate data sources
- Integration of disparate applications
- Automated execution of routine processes
- Capture and deployment of best practice



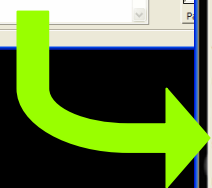
The screenshot shows the Pipeline Pilot Web Access interface in a Microsoft Internet Explorer browser. The address bar shows 'http://localhost:4444/examples/BrowserExample/PPloWeb.htm'. The page has a header with the SciTeGic logo and 'Pipeline Pilot Web Access'. Below the header are tabs for 'Login', 'Protocols', 'Execution', 'Results', and 'Help'. The main content area is titled 'View Molecular Properties' and contains the following information:

Run an external program on the server

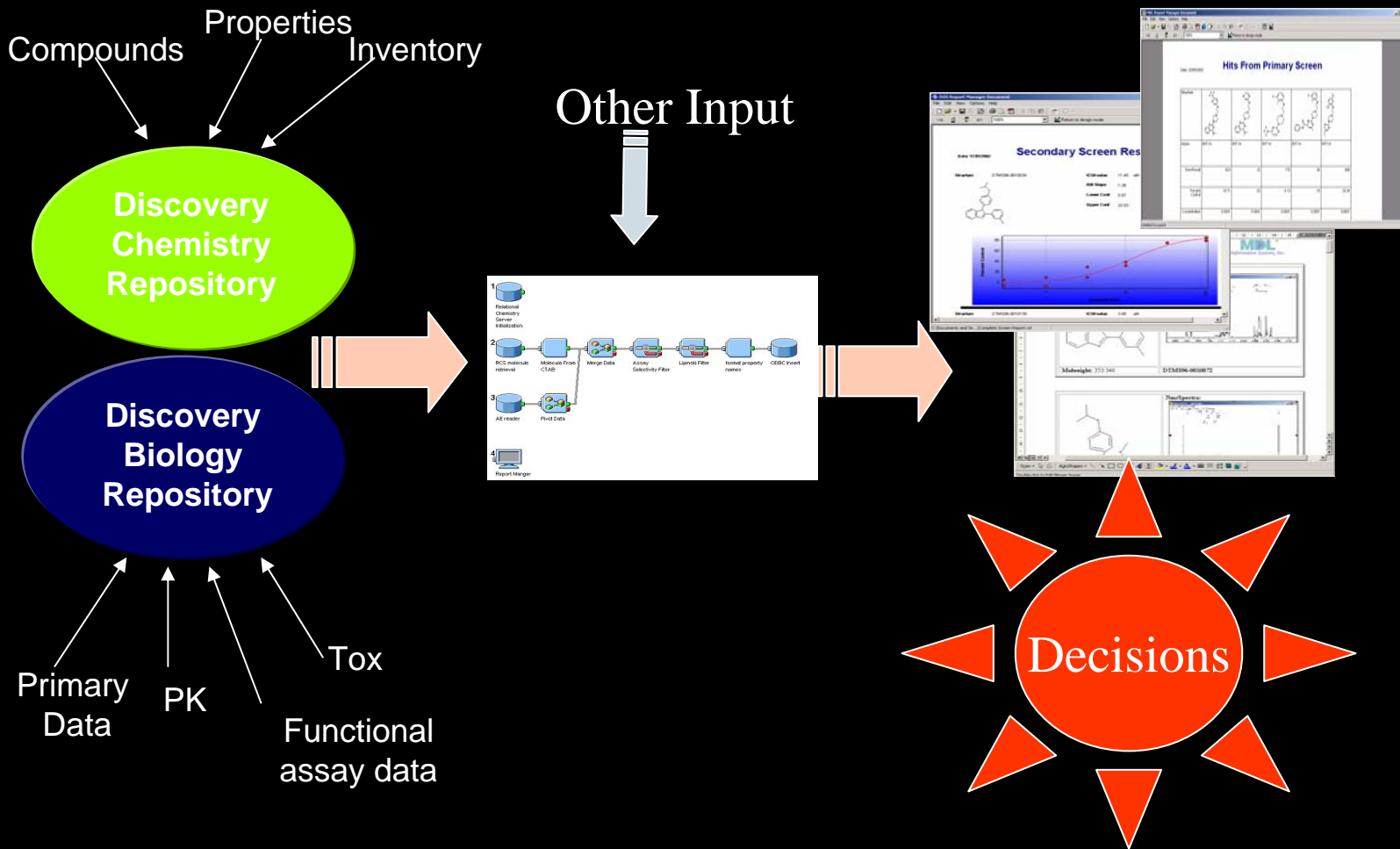
**Parameters:**

- Source:** users/rbrown/nci\_drugs.sd [Remember this file] [Add a file](#)
- Maximum:** 5000
- Properties To View:** Molecular Weight, Molecular Formula, AlogP, Num\_RotatableBonds
- Output Results To:** HTML Report

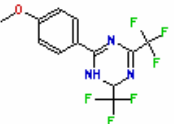
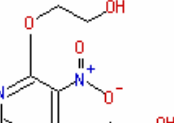
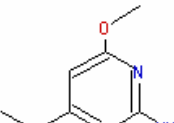

At the bottom of the form is a 'Run protocol' button.



# Data Integration Example



# Data Integration Example

Molecule	CDBREGNO	COMPOUND_ID
	1	DTSM96-00001
	2	DTSM96-00002
	3	DTSM96-00003
	4	DTSM96-00004

Chemistry in ISIS

Biology in Assay Explorer

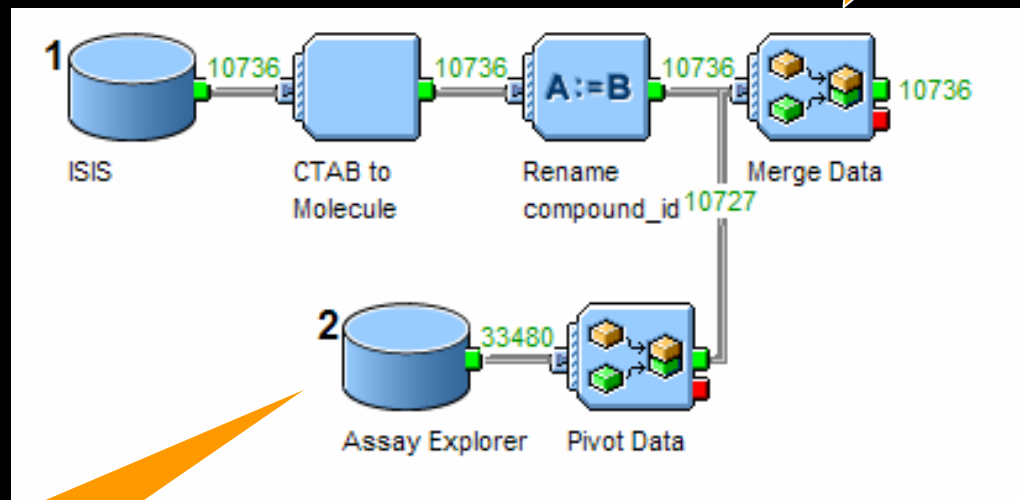
SAMPLE_ID	PLATE_ID	WELL_ROW	WELL_COLUMN	ASSAY	PROTOCOLNAME	PCT_RESULT	HIT
DTSM96-07726	Plate.023	L	12	EGFR	HTS for EGFR	96.099	N
DTSM96-07822	Plate.023	L	13	EGFR	HTS for EGFR	112.234	N
DTSM96-07734	Plate.023	L	14	EGFR	HTS for EGFR	106.887	N
DTSM96-07742	Plate.023	L	16	EGFR	HTS for EGFR	95.344	N
DTSM96-07750	Plate.023	L	18	EGFR	HTS for EGFR	101.861	N
DTSM96-07846	Plate.023	L	19	EGFR	HTS for EGFR	100.594	N
DTSM96-07758	Plate.023	L	20	EGFR	HTS for EGFR	90.941	N
DTSM96-07854	Plate.023	L	21	EGFR	HTS for EGFR	103.228	N
DTSM96-07766	Plate.023	L	22	EGFR	HTS for EGFR	99.443	N
DTSM96-07766	Plate.023	L	22	EGFR	HTS for EGFR	99.443	N
DTSM96-07951	Plate.023	M	1	EGFR	HTS for EGFR	104.842	N
DTSM96-07863	Plate.023	M	2	EGFR	HTS for EGFR	99.335	N
DTSM96-07959	Plate.023	M	3	EGFR	HTS for EGFR	110.526	N
DTSM96-07871	Plate.023	M	4	EGFR	HTS for EGFR	66.501	N
DTSM96-07967	Plate.023	M	5	EGFR	HTS for EGFR	108.139	N

# Step 1: Retrieve and Merge

Retrieve chemistry data

- One row per molecule
- Convert text string CTAB to Pipeline Pilot molecule

Merge on common key (sample id)

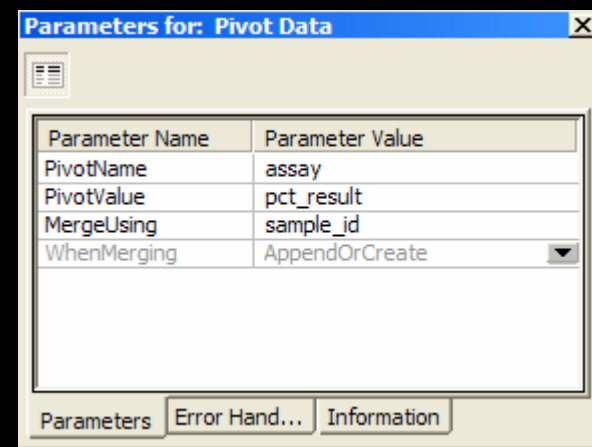
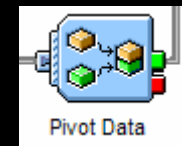


Retrieve biology data

- One row per assay
- Pivot to get one row per molecule

# Data pivoting

- Data is typically organized for storage
  - Long-skinny format
- Pivot converts to a short-wide format
  - View single compound multiple assay results
  - Make the data more compact
  - Suitable for joining to chemistry



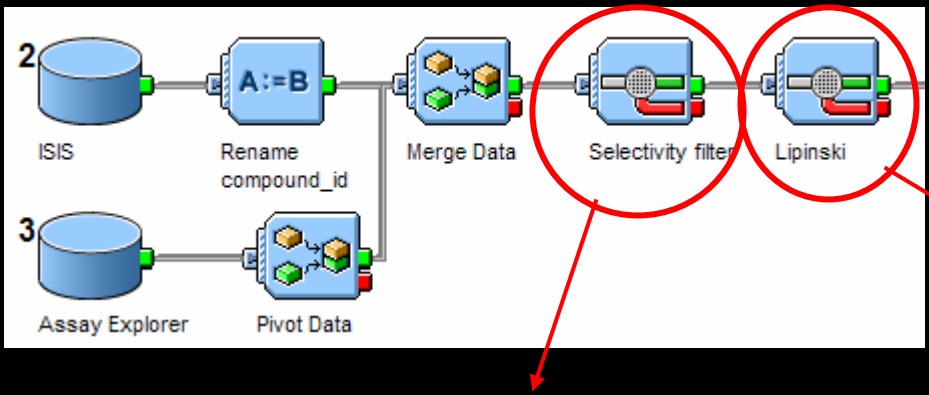
	A	B	C
1	<b>SAMPLE_ID</b>	<b>ASSAY</b>	<b>PCT_RESULT</b>
2	DTSM96-00009	CDK2	107.043
3	DTSM96-00010	CDK2	107.682
4	DTSM96-00009	EGFR	94.924
5	DTSM96-00010	EGFR	96.980
6	DTSM96-00009	PKCa	90.132
7	DTSM96-00010	PKCa	84.059

	A	B	C	D
1	<b>SAMPLE_ID</b>	<b>CDK2</b>	<b>EGFR</b>	<b>PKCa</b>
2	DTSM96-00009	107.043	94.924	90.132
3	DTSM96-00010	107.682	96.980	84.059





# Workflow Example



**PilotScript**

Enter PilotScript.  
Press F4 for Calculable Properties. Press F5 for Keywords and Functions.  
To update the Calculable Properties Press F7

```
@failed :=0;
IF ((N_Count + O_Count) > 10) THEN
@failed++;
END IF;
IF (Molecular_Weight > 500) THEN
@failed++;
END IF;
IF (Num_H_Donors > 5) THEN
@failed++;
END IF;
IF (AlogP > 5) THEN
@failed++;
END IF;
@failed <= 1;
```

Initial Expression    Expression    Final Expression

Buttons: OK, Cancel, Find..., Replace..., Go To Line..., Settings..., Print, Check Syntax, Help

**PilotScript**

Enter PilotScript.  
Press F4 for Calculable Properties. Press F5 for Keywords and Functions.  
To update the Calculable Properties Press F7

```
if (pkca IS DEFINED AND egfr IS DEFINED AND cdk2 IS DEFINED and
pkca<=50 AND egfr>50 AND cdk2>50) then
selectivity := 'pkca';
elseif (pkca IS DEFINED AND egfr IS DEFINED AND cdk2 IS DEFINED and
pkca>50 AND egfr<=50 AND cdk2>50) then
selectivity := 'egfr';
elseif (pkca IS DEFINED AND egfr IS DEFINED AND cdk2 IS DEFINED AND
pkca>50 AND egfr>50 AND cdk2<=50) THEN
selectivity := 'cdk2';
end if;

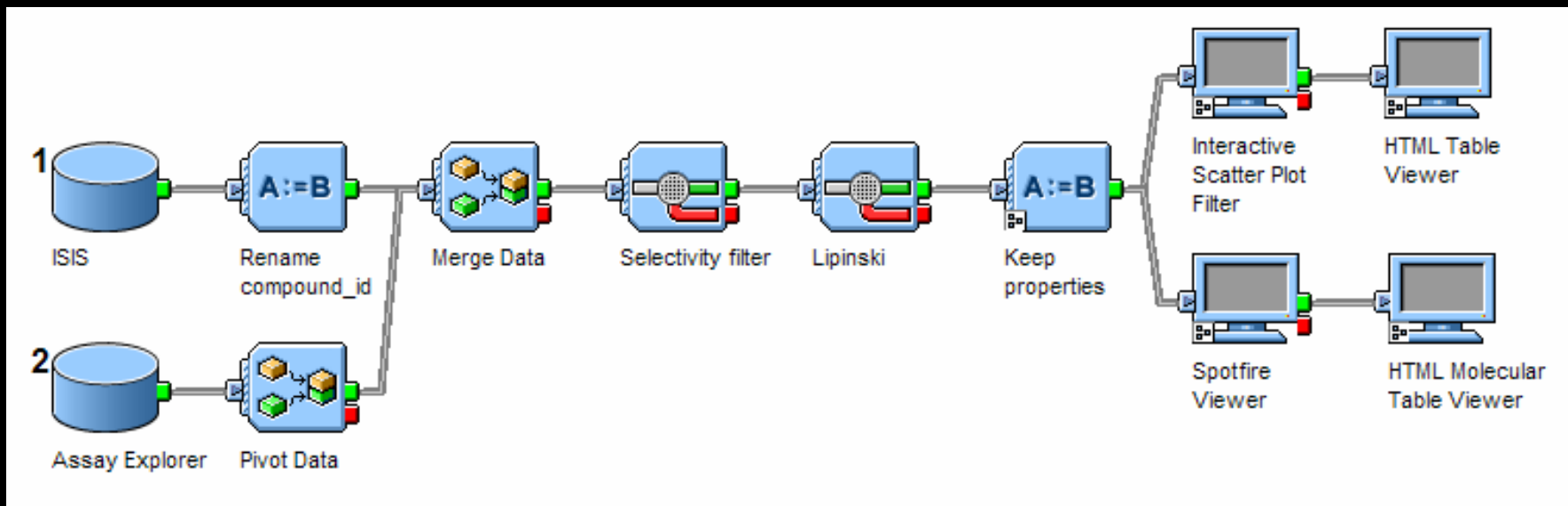
selectivity is defined
```

Initial Expression    Expression    Final Expression

Buttons: Cancel, Find..., Replace..., Go To Line..., Settings..., Print, Check Syntax, Help

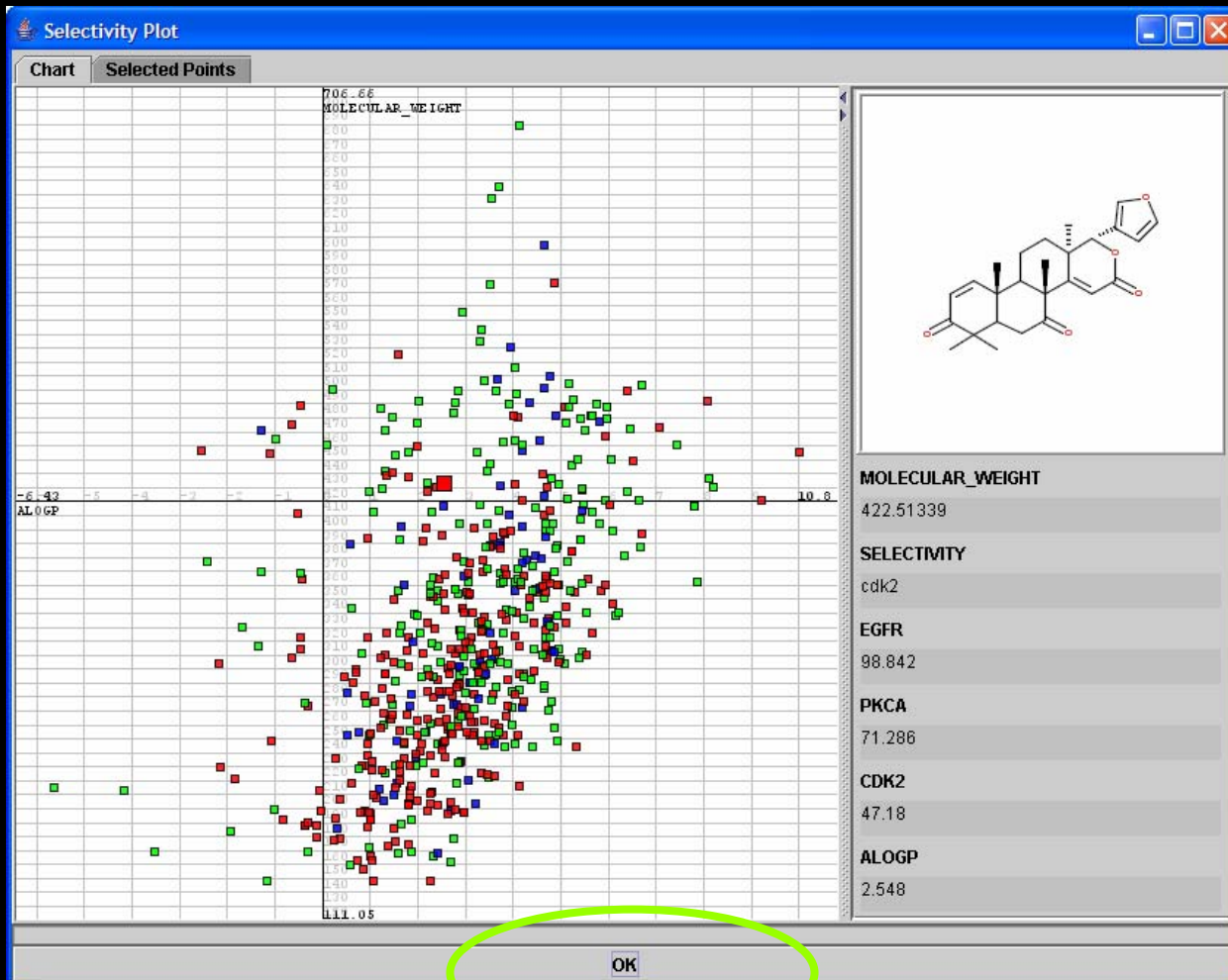


# Interactive Visualization

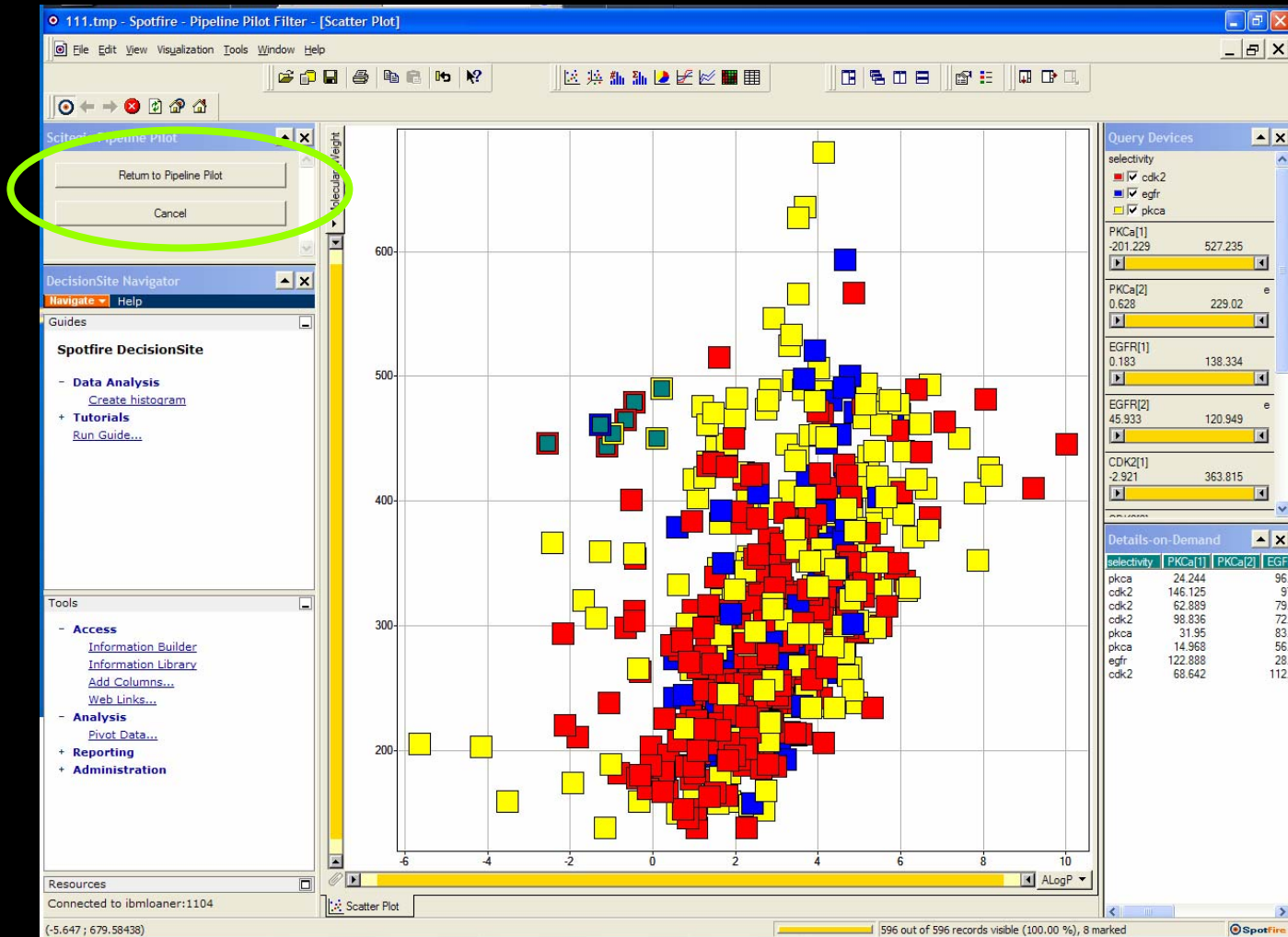


[Run Demo](#)

# Interactive Scatter Plot



# Using Spotfire as a Filter



### Query Devices

selectivity  
 cdk2  
 egfr  
 pkca

PKCa[1]  
-201.229      527.235

PKCa[2]  
0.628      229.02

EGFR[1]  
0.183      138.334

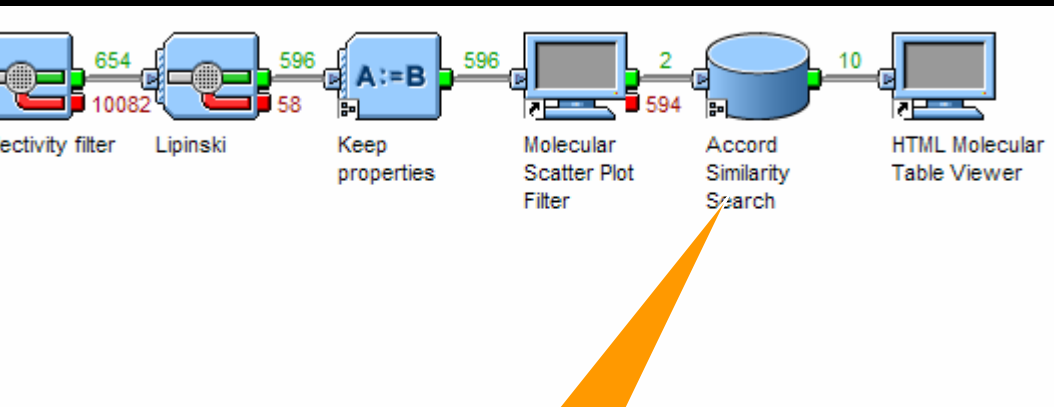
EGFR[2]  
45.933      120.949

CDK2[1]  
-2.921      363.815

### Details-on-Demand

selectivity	PKCa[1]	PKCa[2]	EGFR
pkca	24.244		96.
cdk2	146.125		97.
cdk2	62.889		79.
cdk2	98.836		72.
pkca	31.95		83.
pkca	14.968		56.
egfr	122.808		28.
cdk2	68.642		112.

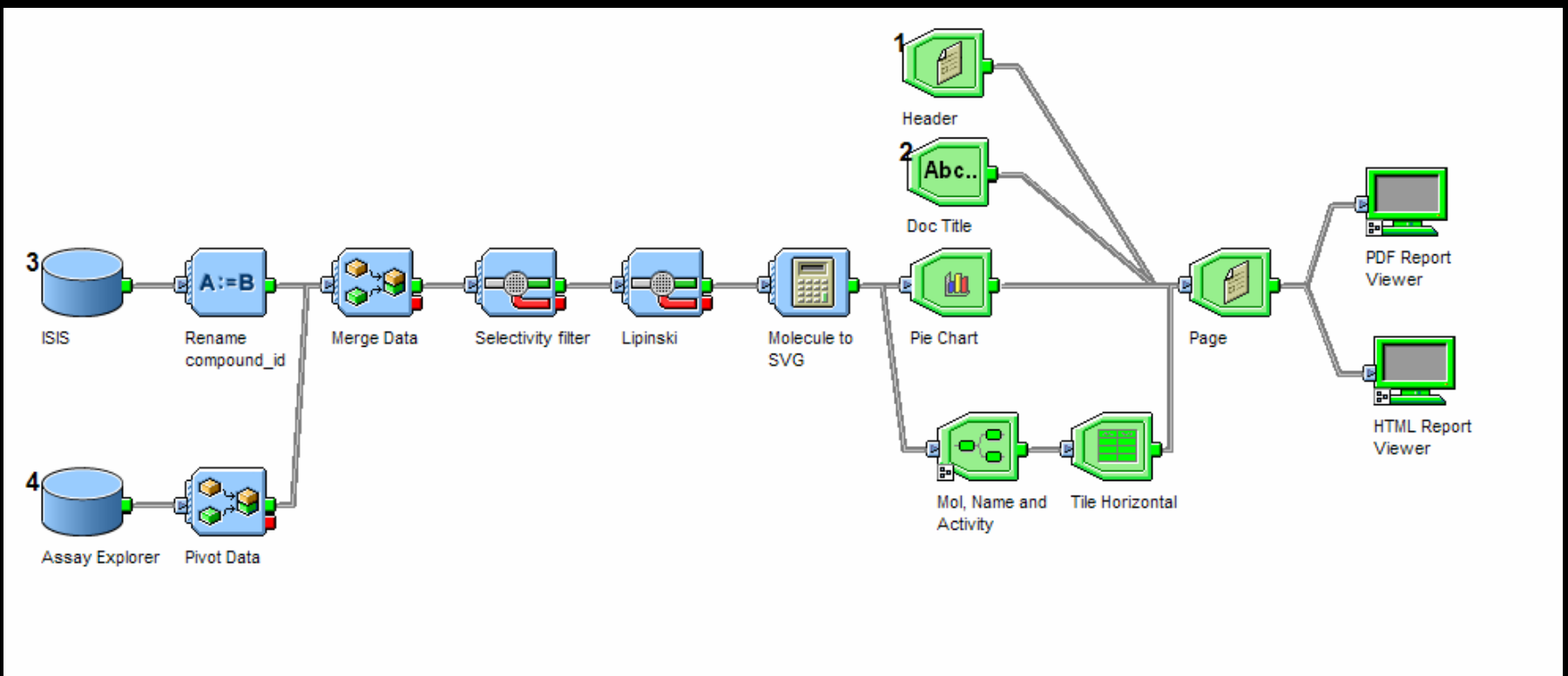
# Hit elaboration



Select similar molecules from Maybridge using Accord Cartridge similarity search

Molecule	REGNO	IUPAC_NAME	SIMILARITY
	50549	2-methylbicyclo[2.2.2]oct-2-ene	81.48
	40319	17-(1,5-dimethylhex-4-enyl)-4,4,10,14-tetramethyl-2,3,4,5,6,7,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[a]phenanthren-3-ol	75
	50538	5-methylbicyclo[2.2.1]hept-2-ene	75
	28696	13-methyl-2,3,6,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[ a]phenanthren-17-ol	72.22
	28704	13-methyl-2,3,6,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[ a]phenanthren-17-one	72.22

# Reporting...



# Report Output

PDF

HTML

Confidential: Compound Selectivity Report

## Project ABC-123 Activity Report

'Distribution of Selectivity'

<b>cdk2</b>	<b>pkca</b>	<b>egfr</b>
-------------	-------------	-------------

**DTSM96-00024**  
 pkca: 59.526  
 egfr: 88.318  
 cdk2: 40.568

**DTSM96-00029**  
 pkca: 40.458  
 egfr: 88.646  
 cdk2: 90.241

**DTSM96-00035**  
 pkca: 24.244  
 egfr: 96.158  
 cdk2: 60.82

**DTSM96-00037**  
 pkca: 30.013  
 egfr: 75.527  
 cdk2: 104.232

Confidential: Compound Selectivity Report

## Project ABC-123 Activity Report

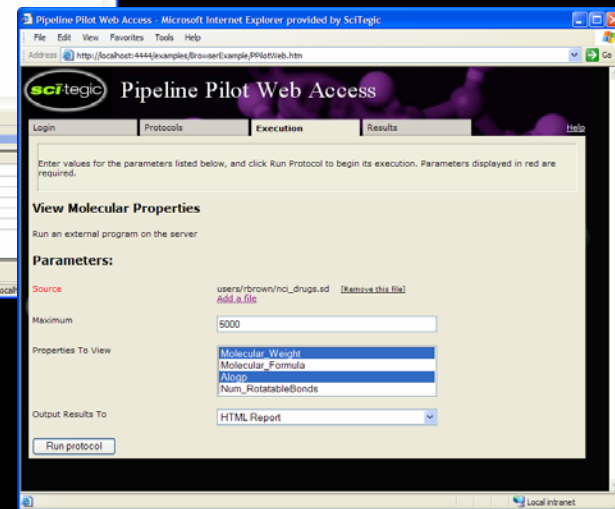
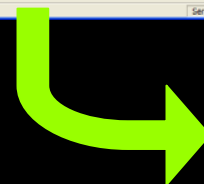
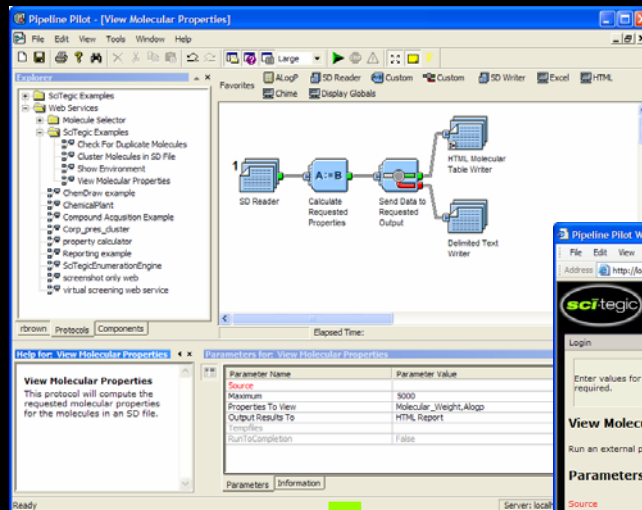
'Distribution of Selectivity'

<b>cdk2</b>	<b>pkca</b>	<b>egfr</b>
-------------	-------------	-------------

<b>DTSM96-00024</b>	<b>DTSM96-00029</b>	<b>DTSM96-00035</b>	<b>DTSM96-00037</b>
pkca: 59.526	pkca: 40.458	pkca: 24.244	pkca: 30.013
egfr: 88.318	egfr: 88.646	egfr: 96.158	egfr: 75.527
cdk2: 40.568	cdk2: 90.241	cdk2: 60.82	cdk2: 104.232
<b>DTSM96-00038</b>	<b>DTSM96-00041</b>	<b>DTSM96-00044</b>	<b>DTSM96-00046</b>
pkca: 23.09	pkca: 25.398	pkca: 47.138	pkca: 37.24
egfr: 95.366	egfr: 84.388	egfr: 96.687	egfr: 94.96
cdk2: 226.896	cdk2: 179.619	cdk2: 100.559	cdk2: 65.356

# Deployment to End Users

- Deployment of best-practice processes through an application that end-users are already familiar with speeds uptake and cuts training, for example
  - Accord for Excel
  - SpotFire
  - WebBrowser



# Accord for Excel – Reagent Searching

The screenshot shows the Microsoft Excel interface with the 'Tools' menu open. The menu items are:

- Spelling... (F7)
- Research... (Alt+Click)
- Error Checking...
- Speech
- Shared Workspace...
- Share Workbook...
- Track Changes
- Compare and Merge Workbooks...
- Protection
- Online Collaboration
- Goal Seek...
- Scenarios...
- Formula Auditing
- Macro
- Add-Ins...
- AutoCorrect Options...
- Customize...
- Options...
- PP retrieve as SD
- PP retrieve as XML
- PP retrieve as csv

In the spreadsheet, cell A1 contains a chemical structure of a bicyclic compound (a benzene ring fused to a five-membered ring). Below it, in cell A2, is the text 'Chemistry 0'. The formula bar shows the formula for cell A1 as '=Chemistry 0'.



# Under the hood...

```

Private Function RunProtocol(protocolName As String, _
                            inputArgs() As struct_ParameterStringValue) _
As struct_ParameterStringValue()
' Login to the pp server, run the protocol using the arguments and return an
' array of output args from the protocol
Dim pp As New clsWS_PPSOAPAPI
Dim sessionHandle As String
Dim runHandle As String
Dim running As Boolean
Dim info As struct_RunInfo
Dim outputArgs() As struct_ParameterStringValue

sessionHandle = pp.wsm_Login(gPPUser, gPPPassword)

runHandle = pp.wsm_RunProtocolWithStringValues_A(sessionHandle, _
        protocolName, inputArgs)

' Wait for completion
running = True
While running
    Set info = pp.wsm_GetRunInfo(runHandle)

    Application.Wait (Now + TimeValue("0:00:1"))
    running = Not (info.Status = "Finished" Or info.Status = "Error")
Wend

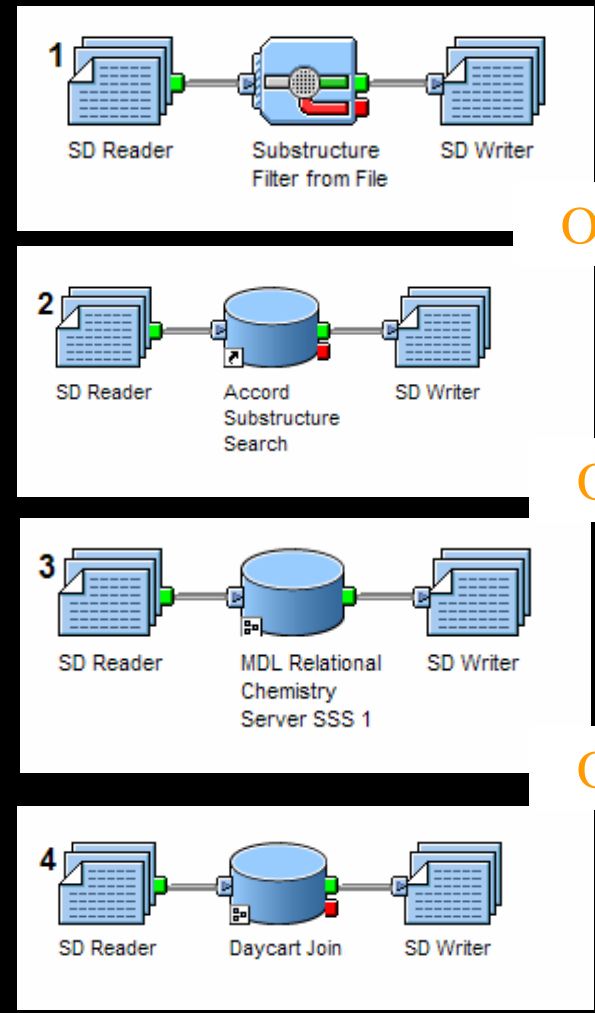
If info.Status = "Error" Then
    Call Err.Raise(100, "Pipeline Pilot returned 'Error'")
End If

outputArgs = pp.wsm_GetRunStringResults(runHandle)

|

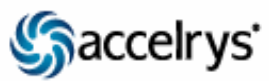
If UBound(outputArgs) = -1 Then
    Call Err.Raise(100, "Output Array from PP Protocol was empty", "Output Array")
End If

RunProtocol = outputArgs
End Function
    
```



# Out-of-the-box integration of 3<sup>rd</sup> parties

## Cheminformatics



## Modeling



## Bioinformatics

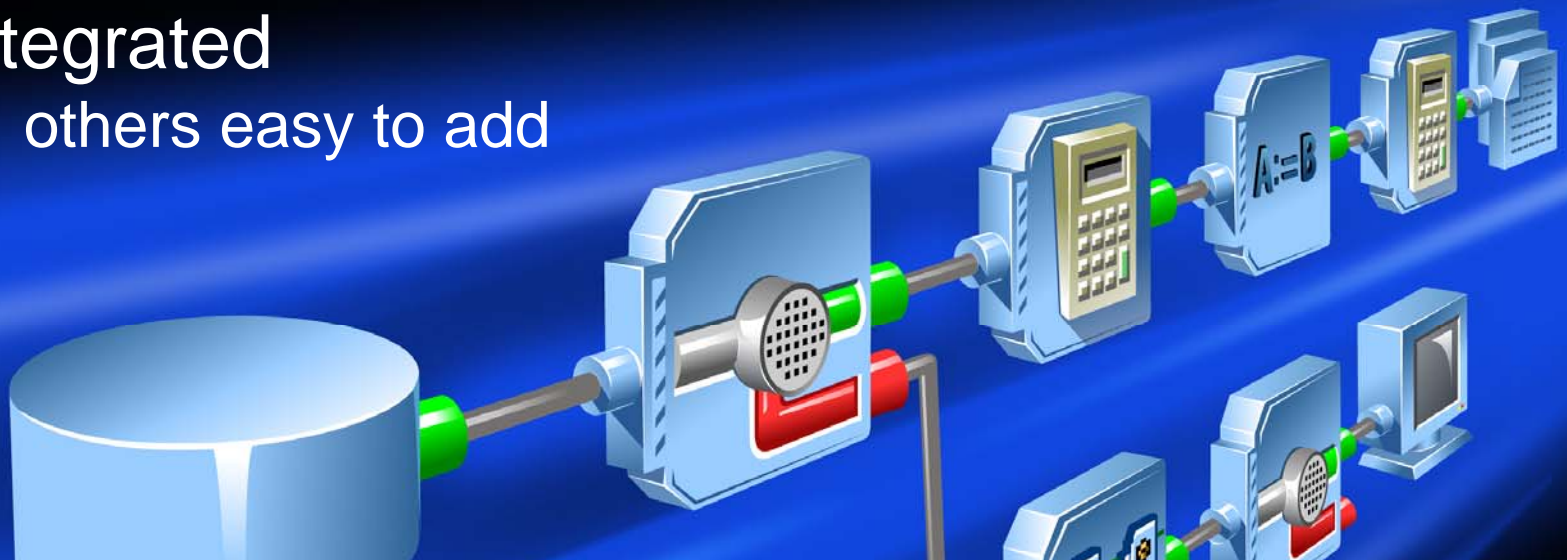


## Other



## Conclusion

- A flexible platform for data and application integration
- Best practice workflows can be deployed to end-users through familiar clients
- Many data sources and applications already integrated
  - others easy to add



## With thanks to

- MDL – for software, data and protocols
  - David Evans
  - Gerd Blanke
- Accelrys
  - Jim Clark and the Accord for Excel team
- Spotfire
- SciTegic
  - Andrei Caracoti – Spotfire integration
  - Phil Cochrane – Scatter plot viewer
  - Rahim Lila