



Data Science for the Masses Mission Impossible?

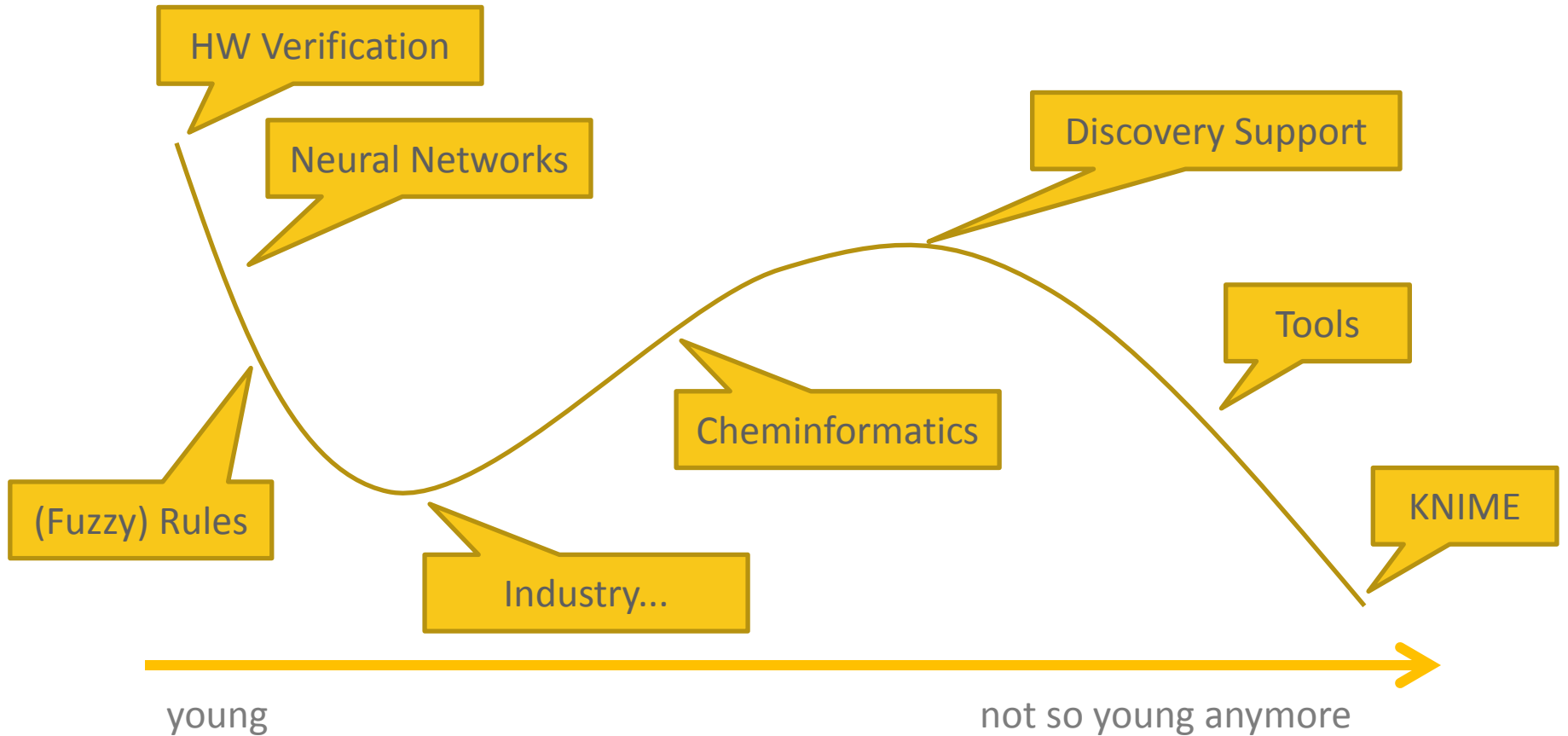
Michael Berthold

Uni Konstanz and KNIME

Menu of the Day

- The Past and a Peak into the Future
- Processes
- Makers and Users

Disclaimer: This is me.

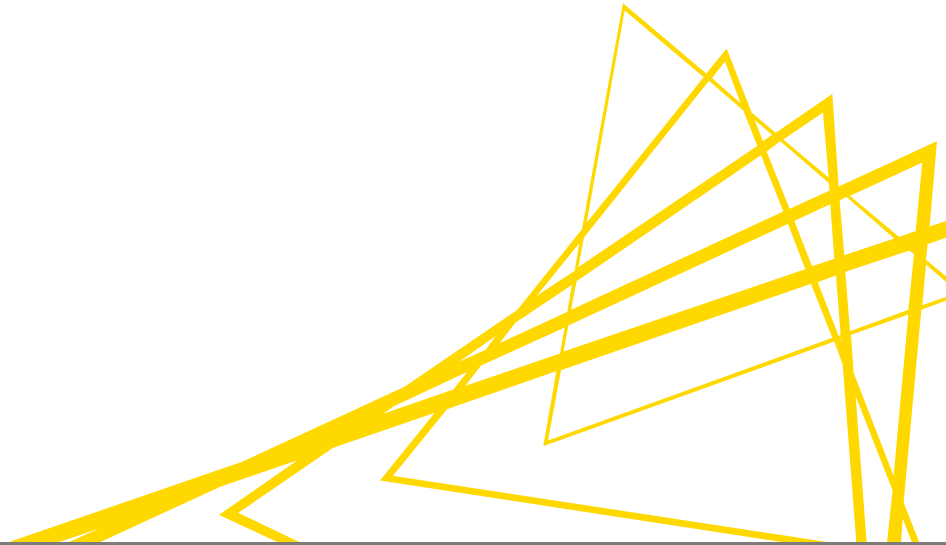


The Data Science Landscape

It's all about

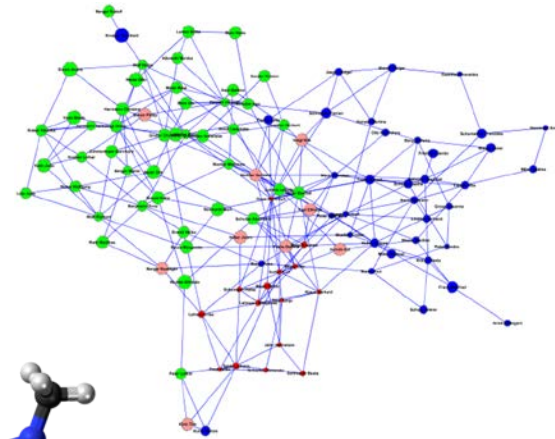
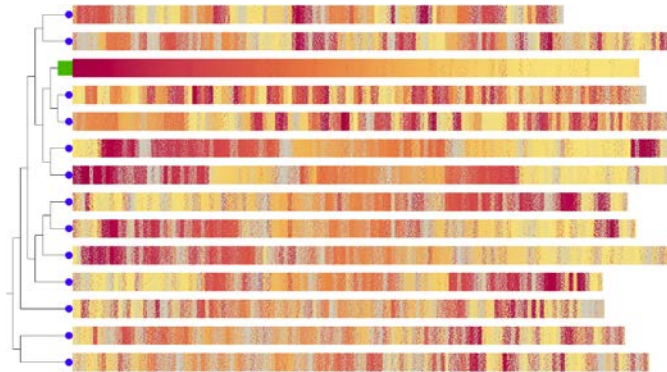
- Data,
- Methods,
- and People.

Data.



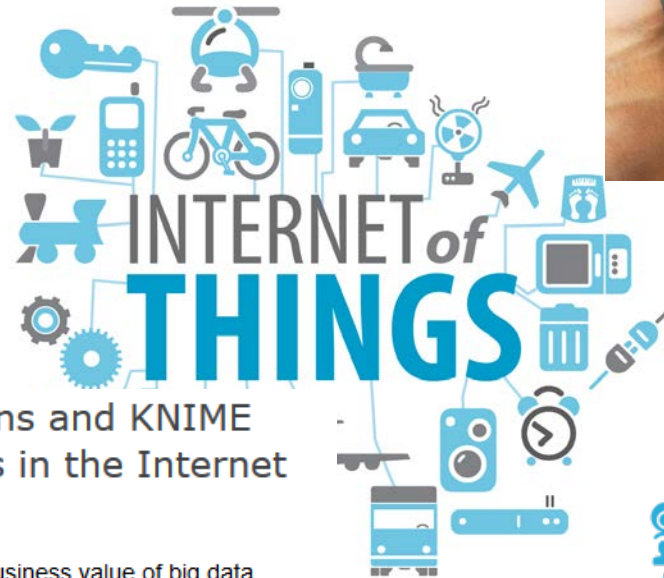
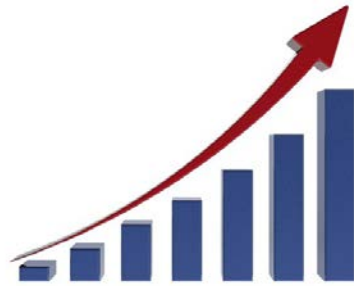
Data. The Past

Sepal length ↕	Sepal width ↕	Petal length ↕	Petal width ↕	Species ↕
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>



deleted cara apps
 internet future style talk
 kid gonna tinder-moe network single ser
 looks hahaha vocé news boys hookup hype pics
 makes leave çekebilir boy using não swiped equity bilton photos
 instagram board fuck day sur game women talking trying omg nyt
 fun messages business night check profile haha swiping nick selam
 shit benchmark 20-something meet pra age-old
 getting jamy fast-growing time social lines
 disruptions found nights girls matched
 matter photo taps true love woman's video
 account gabrielsaporta para tcohookweacy
 butn story girl dates app dating vida
 seen real comes friends
 bio met lol guys people quest guy mas
 takes date matches truth swipe por times phone
 called hope tem likes life left los angeles-based stake
 facebook erichalvorsen quotes millennialmonsters idea
 con cute awkward twitter hot look looking picture friend gerilla
 akma ook message una person ilgini funny como gente research
 god free getir ama hahaha pic factory pictures ridiculous
 matt hooking midnight moment pazzarlama miydi pick

Data. Current Trends



Bosch Software Innovations and KNIME enable advanced analytics in the Internet of Things

- Combined software technology leverages business value of big data

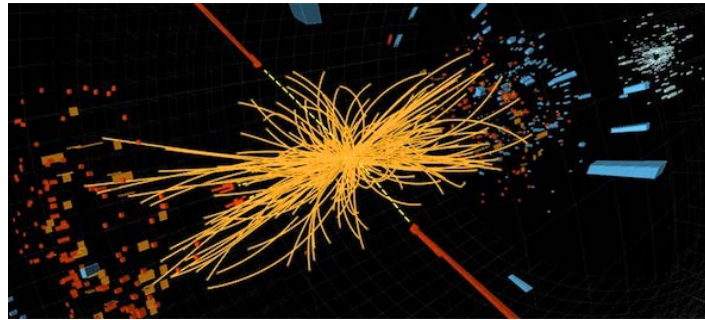
Berlin, 12/16/2014 - Bosch Software Innovations, the Bosch Group's software and systems house, and KNIME.com, provider of the only open platform for data-driven innovation, today announced that they will work together to allow data mining and data analytics for Internet of Things (IoT) applications.

To leverage the business value of big data, Bosch Software Innovations will



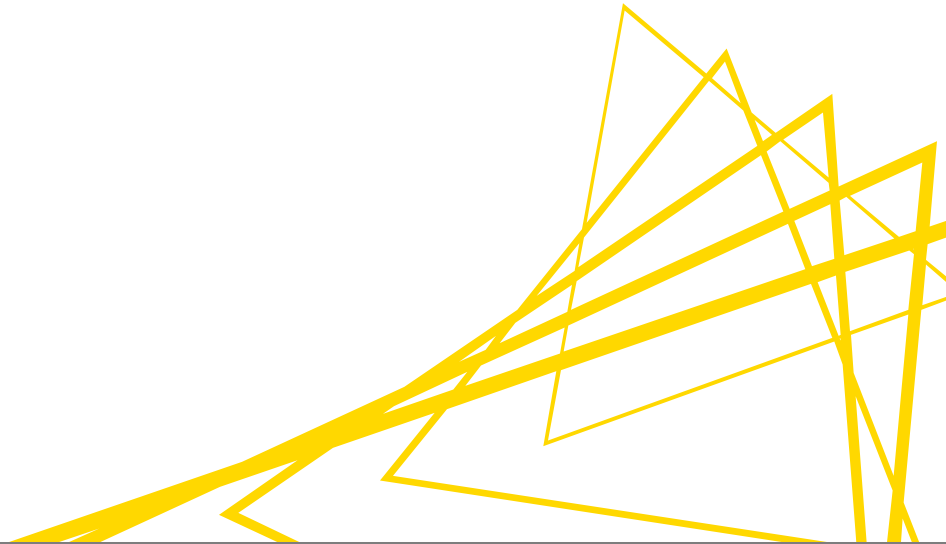
Data. What's up next?

- Even Bigger Data: too much to store.

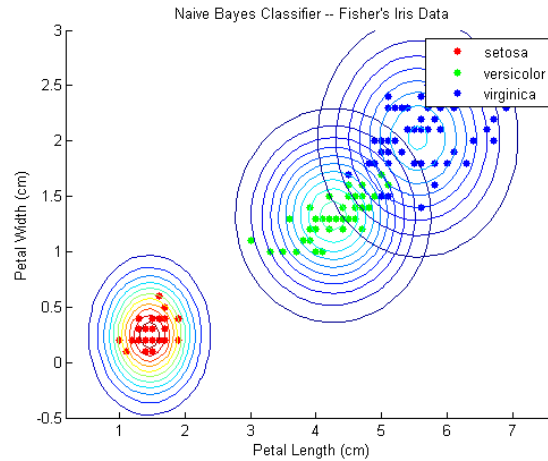
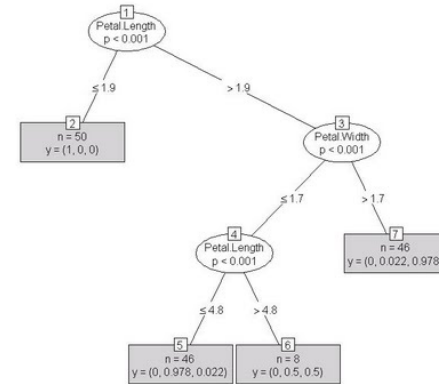
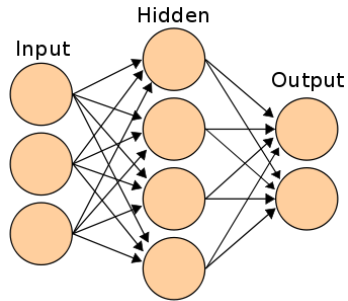


- Even Faster Data: too fast to truly digest.
- Seriously heterogeneous data: crossing domains.
- ???

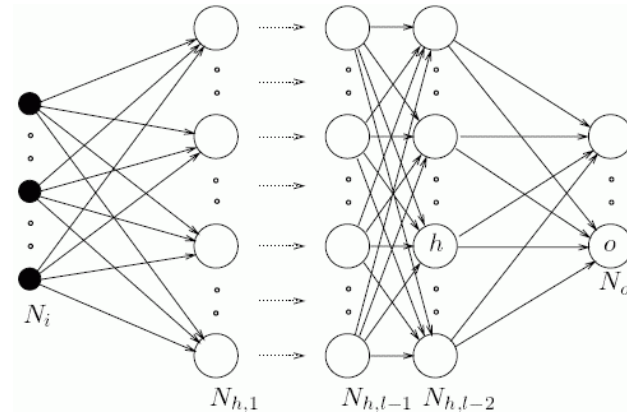
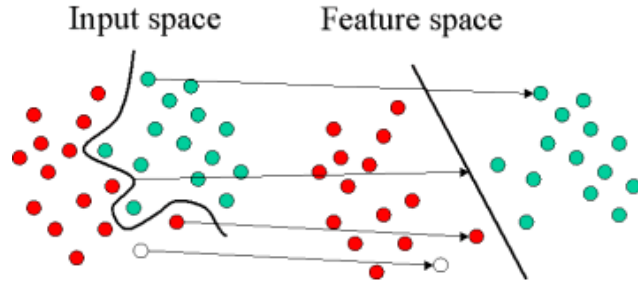
Methods.



Methods. The Past

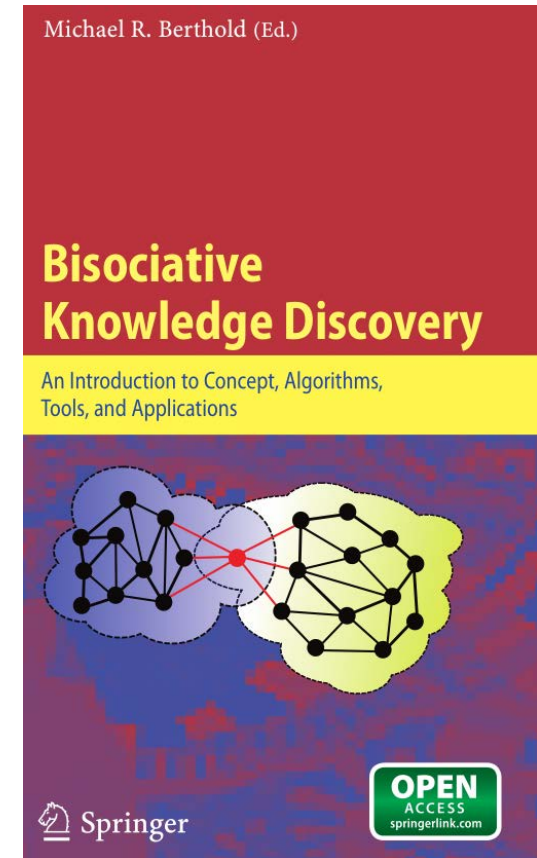


Methods. Current Trends

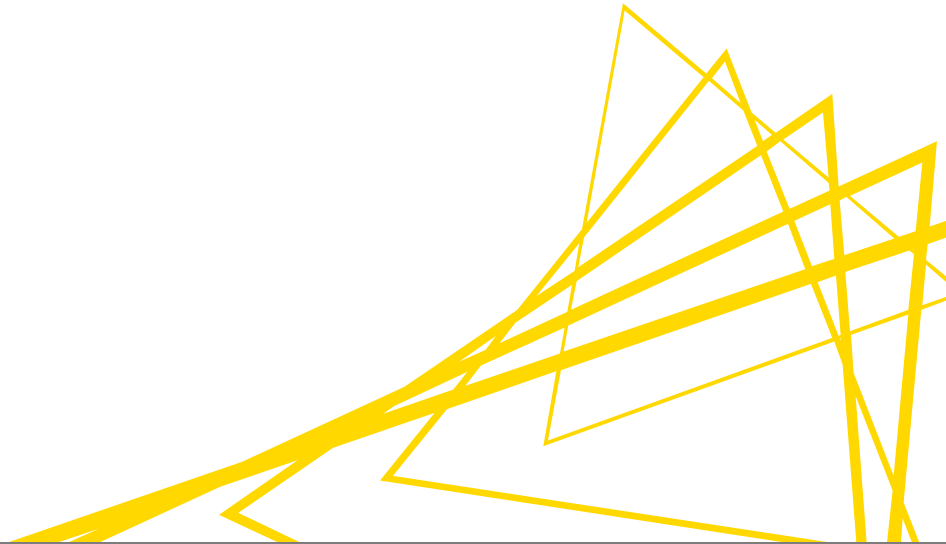


Methods. What's up Next?

- Truly Parallel Algorithms for Data Mining (“Widening”)
- Supporting Discoveries (Bisociative Knowledge Discovery)
- Adaptive & Interactive Methods
- ???



People.



People. The Past

Insights generated
by Experts.

- Statistics
- Machine Learning



People. Current Trends

Insights generated
by more Experts.

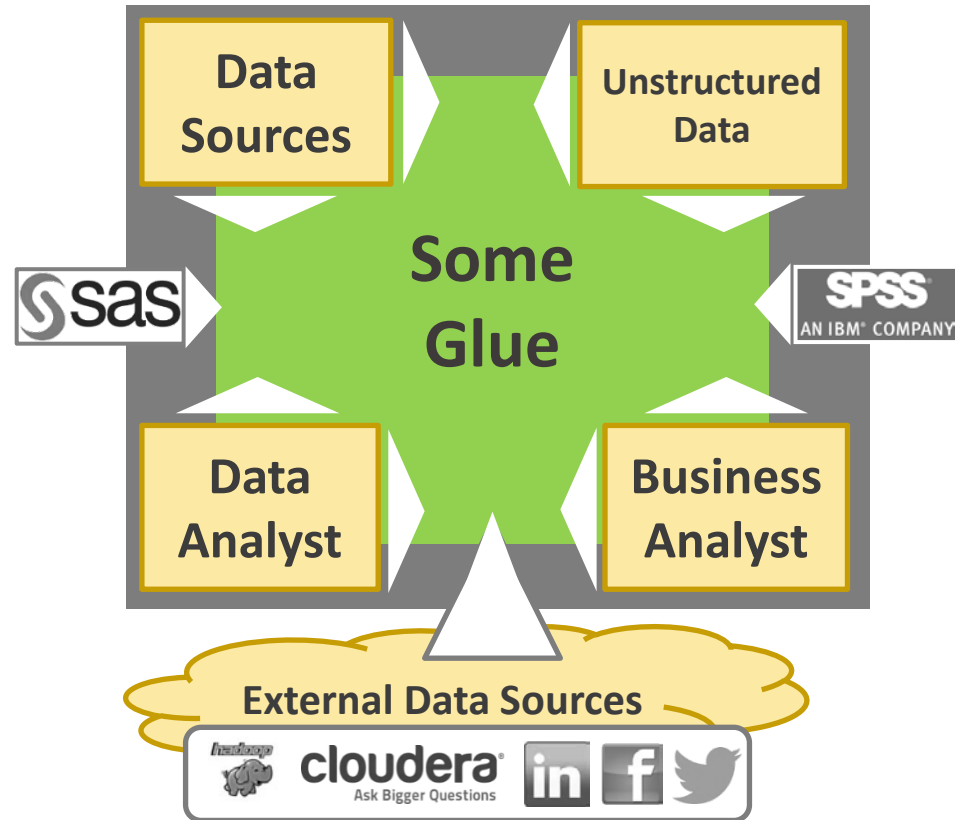
- Data Mining
- Visual Exploration



People. What's up Next?

Insights for
Everybody,
Anytime.

- Casual Analysts
 - Business Users
- “Analytics for the Masses”...



Predicting the Future of Analytics is Hard

Tomorrow	fluent, too much to store ??	interactive, adaptive ??	...Analytics for Everybody, ??
Today	BIG, machine generated, heterogeneous	Data Mining, Deep Learning	Data Scientist, ... Business Analyst
Yesterday	small, structured, static	Statistics, Machine Learning	Data Analyst, Statistician
	Data	Tools	Users

Analytics is about

- discovering insights and
- predicting futures

with

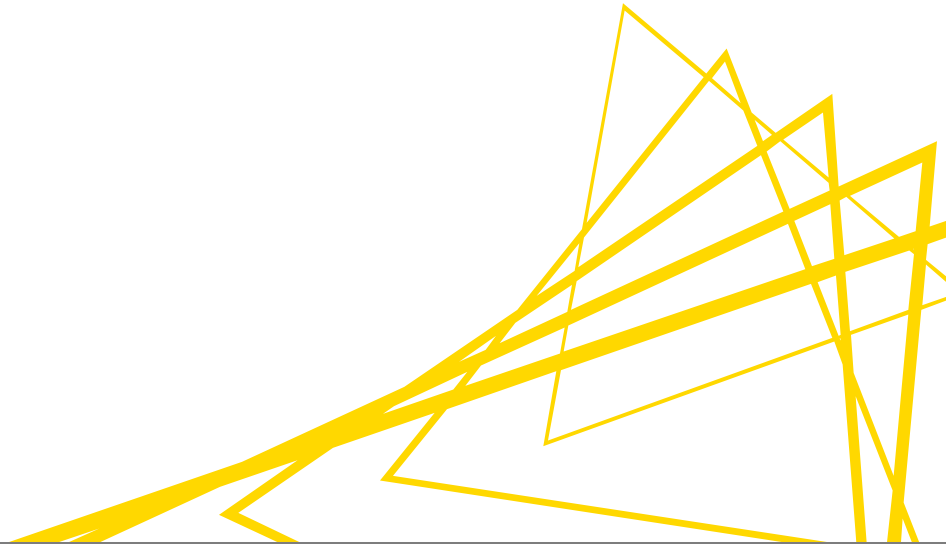
- **yet-to-be-invented tools**

and

- **yet-to-be-collected data**

**...giving Business Analysts
access to Data Science.**

Data Science Processes.



CRISP-DM

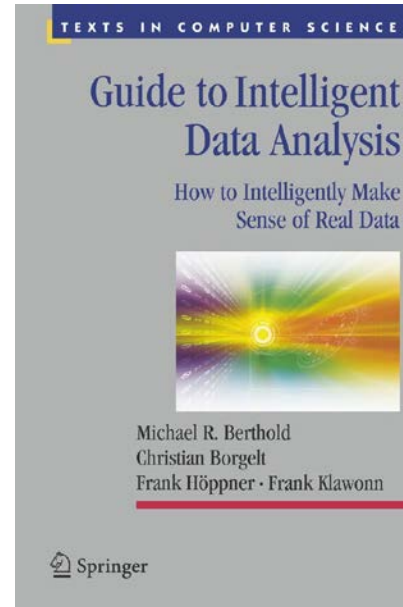
Cross Industry Standard Process for Data Mining



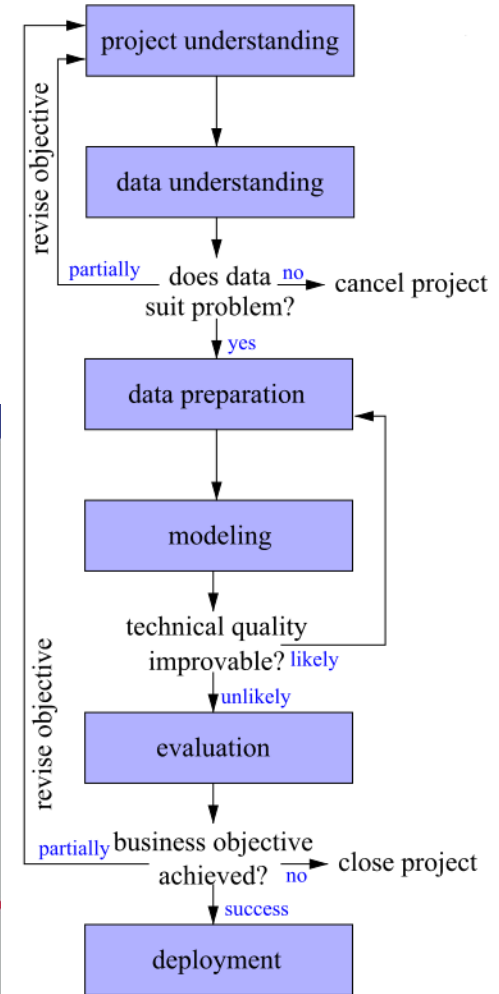
(source Wikipedia)

SEMMA

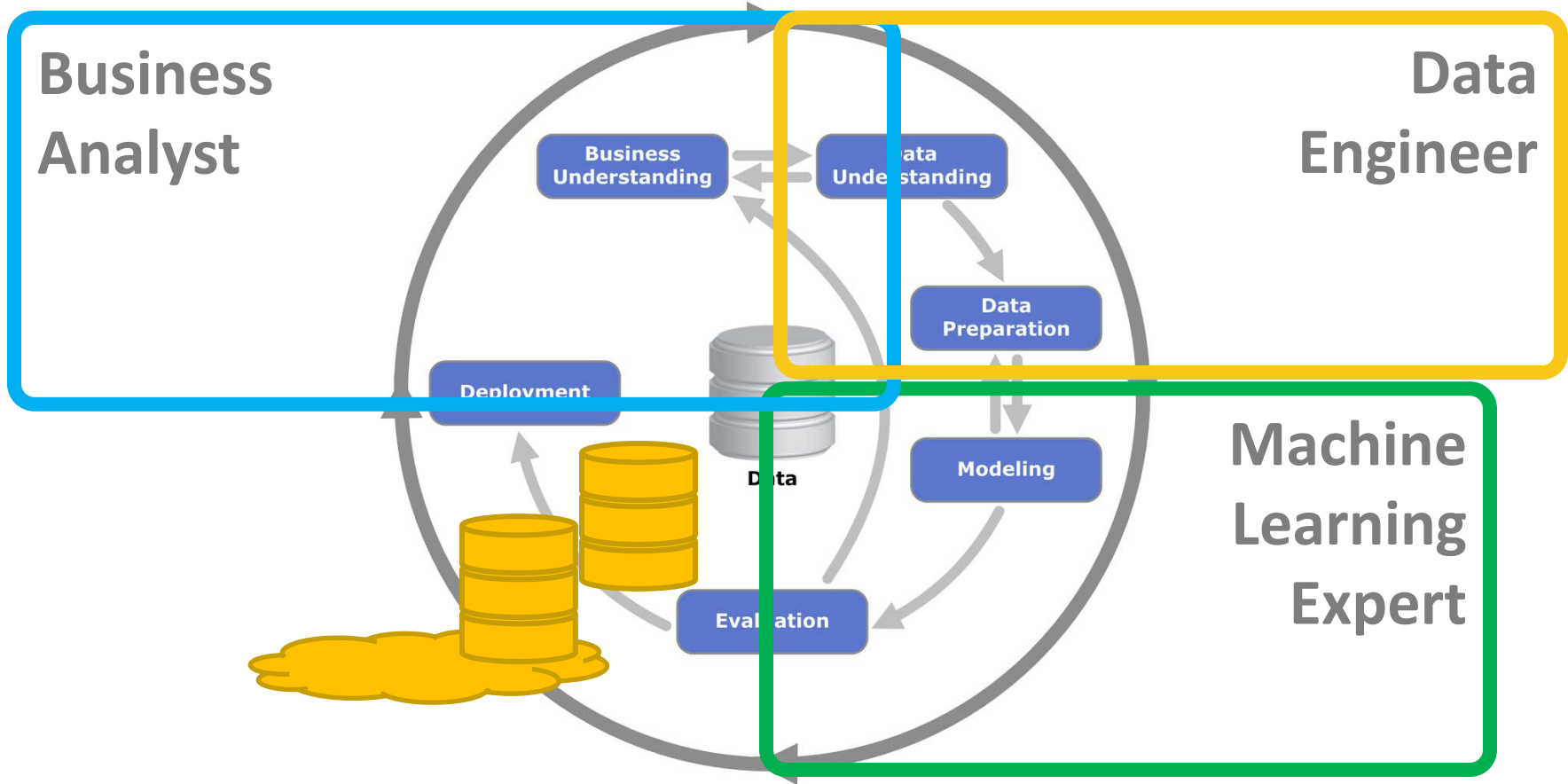
- Sample
- Explore
- Modify
- Model
- Assess



1.2 The Data Analysis Process



CRISP-DM: one for all?



CRISP-DM...

...and the famous magic black box

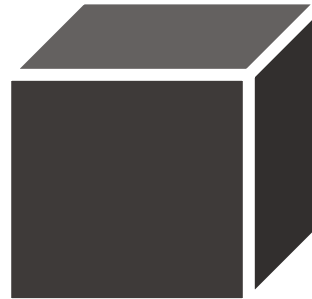
Step 1:

Data Scientist

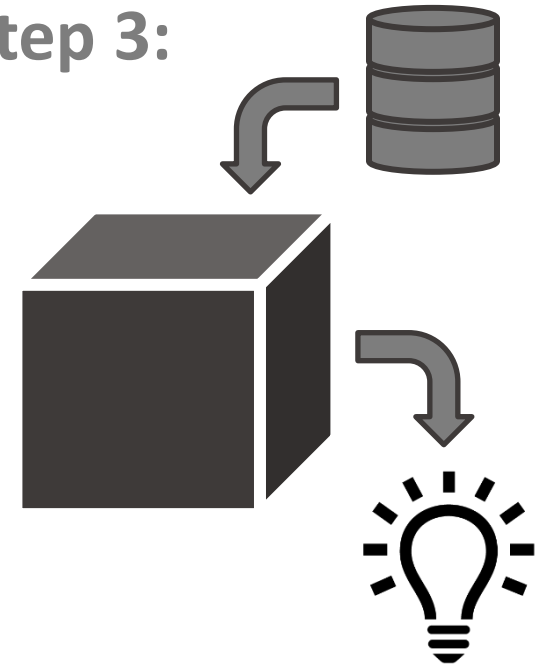
+



Step 2:



Step 3:

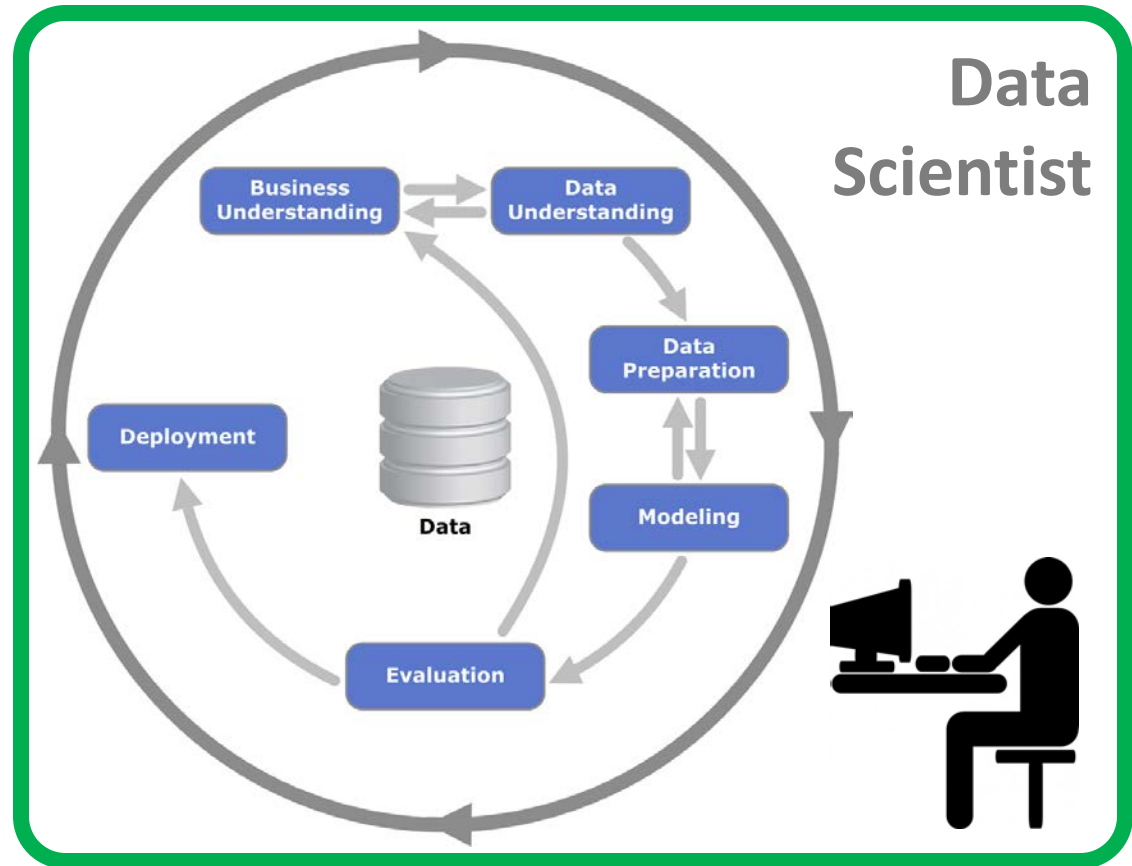


CRISP-DM and the Black Box

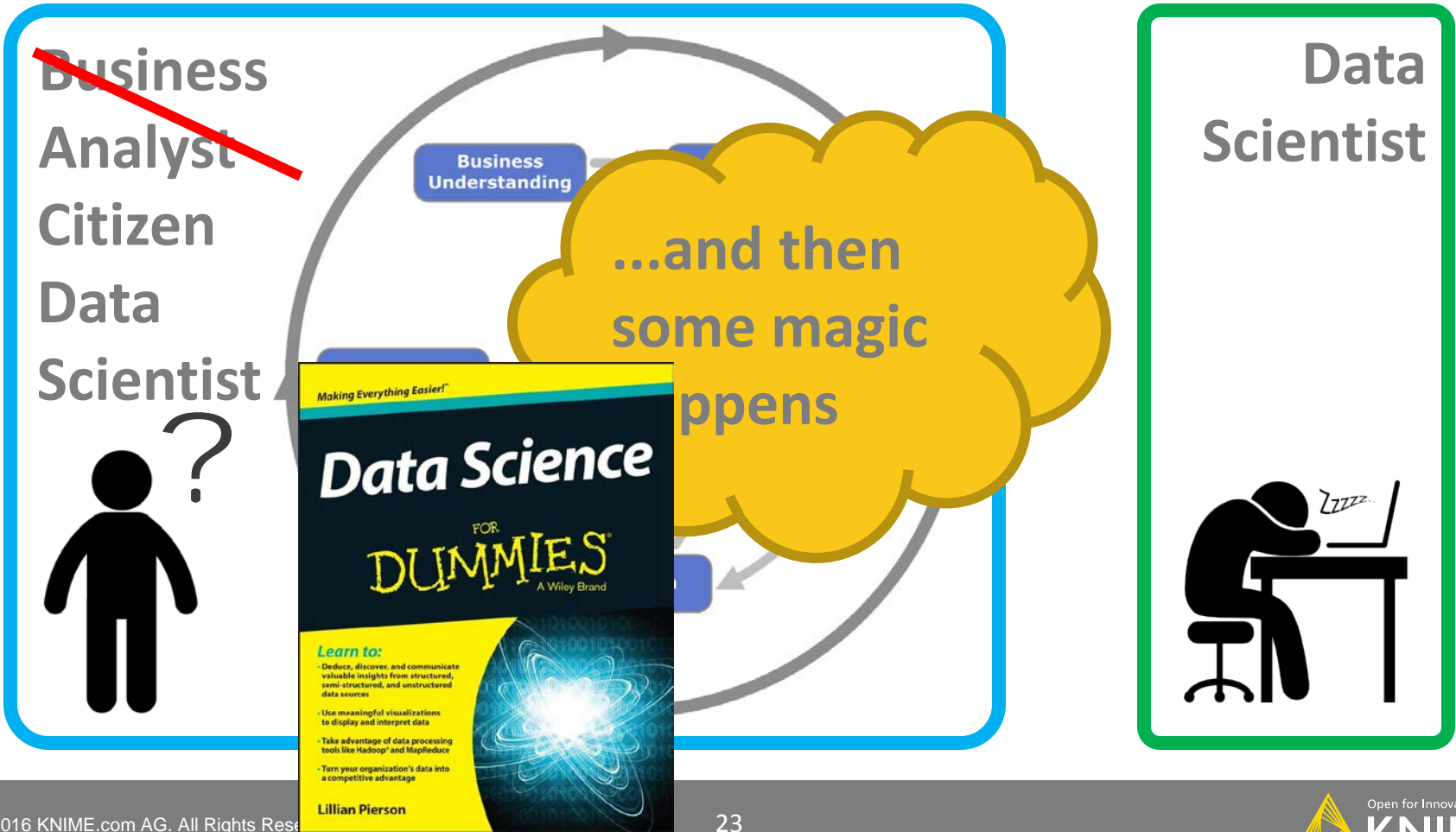
Business Analyst



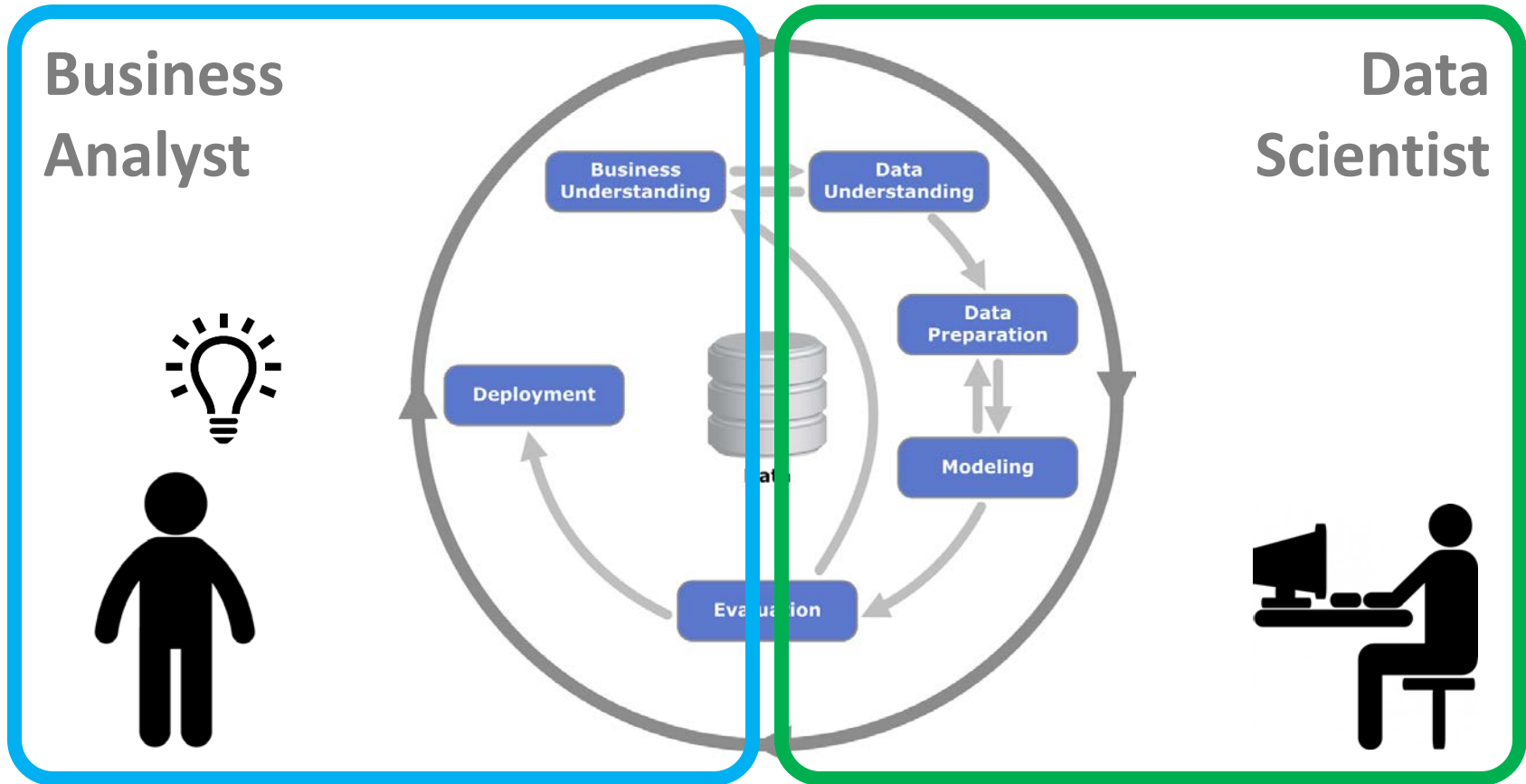
Data Scientist



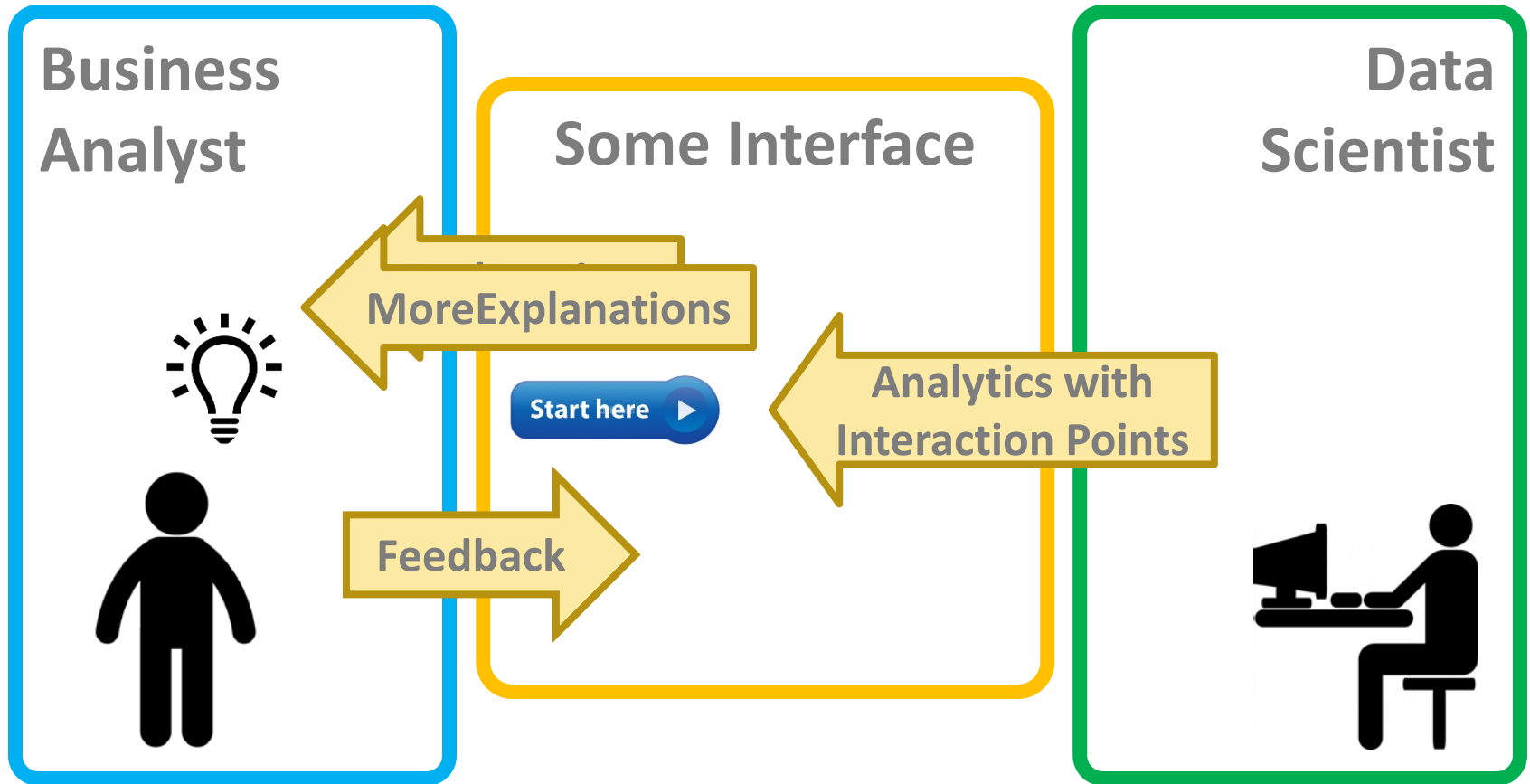
CRISP-DM and the Gartner View



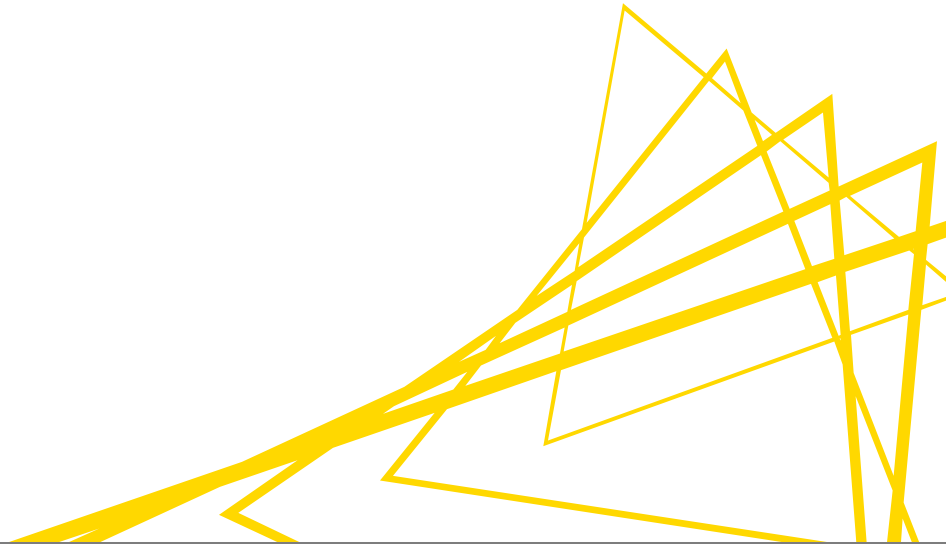
CRISP-DM in an Ideal World



CRISP-DM in the Real World?



Data Science from/for Everybody (the slightly biased KNIME view)



Researchers

New Algorithms come from scientists who write code:

- R, Python, Java, C, ...
- provide access as stand alone tool or as libraries/packages

Input: training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, a differentiable loss function $L(\mathbf{y}, \mathbf{F}(\mathbf{x}))$, number of iterations M .

Algorithm:

1. Initialize model with a constant value:

$$F_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^n L(\mathbf{y}_i, \gamma).$$

2. For $m = 1$ to M :

1. Compute so-called *pseudo-residuals*:

$$r_{im} = - \left[\frac{\partial L(\mathbf{y}_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})} \quad \text{for } i = 1, \dots, n.$$

2. Fit a base learner $h_m(\mathbf{x})$ to pseudo-residuals, i.e. train it using the training set $\{(\mathbf{x}_i, r_{im})\}_{i=1}^n$.
3. Compute multiplier γ_m by solving the following [one-dimensional optimization](#) problem:

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(\mathbf{y}_i, F_{m-1}(\mathbf{x}_i) + \gamma h_m(\mathbf{x}_i)).$$

4. Update the model:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}).$$

3. Output $F_M(\mathbf{x})$.

Data Analysts

Expert Data Analysts use their favorite environment

- Python, R, C, Java, or ...
- don't worry about others.

```
#include <iostream>
#include <seqan/file.h>
#include <seqan/sequence.h>

using namespace seqan;

int computeLocalScore(String<char> subText, String<char> pattern)
{
    int localScore = 0;
    for (unsigned i = 0; i < length(pattern); ++i)
        if (subText[i] == pattern[i])
            ++localScore;

    return localScore;
}

String<int> computeScore(String<char> text, String<char> pattern)
{
    String<int> score;
    resize(score, length(text) - length(pattern) + 1, 0);

    for (unsigned i = 0; i < length(text) - length(pattern) + 1; ++i)
        score[i] = computeLocalScore(infix(text, i, i + length(pattern)), pattern);

    return score;
}

int main()
{
    String<char> text = "This is an awesome tutorial to get to know SeqAn!";
    String<char> pattern = "tutorial";
    String<int> score = computeScore(text, pattern);

    for (unsigned i = 0; i < length(score); ++i)
```

```
library(foreach)
length_divisor<-6
iterations<-5000
predictions<-foreach(m=1:iterations,.combine=cbind)
%do% {
    training_positions <- sample(nrow(training),
    size=floor((nrow(training)/length_divisor))
    train_pos<-1:nrow(training) %in% training_positions
    lm_fit<-lm(y~x1+x2+x3,data=training[train_pos,])
    predict(lm_fit,newdata=testing)
}
predictions<-rowMeans(predictions)
error<-sqrt((sum((testing$y-
predictions)^2))/nrow(testing))
```

```
: public MSDataWritingConsumer
MSDataWritingConsumer allows to change the
disk (to *filename*) using the process()
functions.

set TIC to zero
ng filename) : MSDataWritingConsumer(fi
ra = 0;}
```

```
ng step for spectra before they are writ
MSDataWritingConsumer::SpectrumType & n)

for (Size i = 0; i < n.size(); ++i) { TIC += s[i].getIntens
nr.spectra++;
}
// Empty chromatogram data processing
void processChromatogram(MSDataWritingConsumer::ChromatogramT
);

int main(int argc, const char** argv)
{
    if (argc < 2) return 1;
    // the path to the data should be given on the command line
    String tutorial_data_path(argv[1]);

    // Create the consumer, set output file name, transform
    TICWritingConsumer * consumer = new TICWritingConsumer("Tutori
M2MzFile().transform(tutorial_data_path + "/data/Tutorial.FileI

std::cout << "There are " << consumer->nr.spectra << " spectra
std::cout << "The total ion current is " << consumer->TIC << std::endl;
delete consumer;

return 0;
} //end of main
```

```
*** contribution from Andrew Dalke ***
import sys
from rdkit import Chem
from rdkit.Chem import AllChem

# Download this from http://pypi.python.org/pypi/futures
from concurrent import futures

# Download this from http://pypi.python.org/pypi/progressbar
import progressbar

## On my machine, it takes 39 seconds with 1 worker and 10 seconds with 4.
## 29.855s 0.102x 0:28.68 101.6X 0+0s 0+31s 0pf+0w
#max_workers=1

## With 4 threads it takes 11 seconds.
## 34.923s 0.188s 0:10.89 322.4X 0+0s 125+11s 0pf+0w
#max_workers=4

# (The "user" time includes time spend in the children processes.
# The wall-clock time is 28.68 and 10.89 seconds, respectively.)

# This function is called in the subprocess.
# The parameters (molecule and number of conformers) are passed via a Python
def generateConformations(m, n):
    m = Chem.AddHs(m)
    ids = AllChem.EmbedMultipleConfs(m, numConfs=n)
    for id in ids:
        AllChem.SFFontimizeMolecule(m, confId=id)
    # EmbedMultipleConfs returns a Boost-wrapped type which
    # cannot be pickled. Convert it to a Python list, which can.
    return m, list(ids)

swi_input_file, sdf_output_file = sys.argv[1:3]
n = int(sys.argv[3])

writer = Chem.SDWriter(sdf_output_file)
suppl = Chem.SmilesToSupplier(swi_input_file, titleLine=False)

with futures.ProcessPoolExecutor(max_workers=max_workers) as executor:
    # Submit a set of asynchronous jobs
    jobs = []
    for mol in suppl:
        if mol:
            job = executor.submit(generateConformations, mol, n)
            jobs.append(job)

widgets = ["Generating conformations: ", progressbar.Percentage(), " ",
           progressbar.ETA(), " ", progressbar.Bar()]
pbar = progressbar.ProgressBar(widgets=widgets, maxval=len(jobs))
for job in pbar.futures_as_completed(jobs):
    mol, ids, job_result()
    for id in ids:
        writer.write(mol, confId=id)
writer.close()
```

The Gap.

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M .

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma).$$

2. For $m = 1$ to M :

1. Compute so-called *pseudo-residuals*:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n$$

2. Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. train it using r_{im} .
3. Compute multiplier γ_m by solving the following *one-dimensional* optimization problem:

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$

4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

```
In [3]: modell <- gbm(medv ~ ., data=train, distribution="gaussian", n.trees = 500, interaction.depth = 3, n.minobsinnode = 5, shrinkage = 0.001)
```

```
In [4]: # summarize the model
options(repr.plot.width = 4, repr.plot.height = 4)
summary(modell)
```

Out [4]:

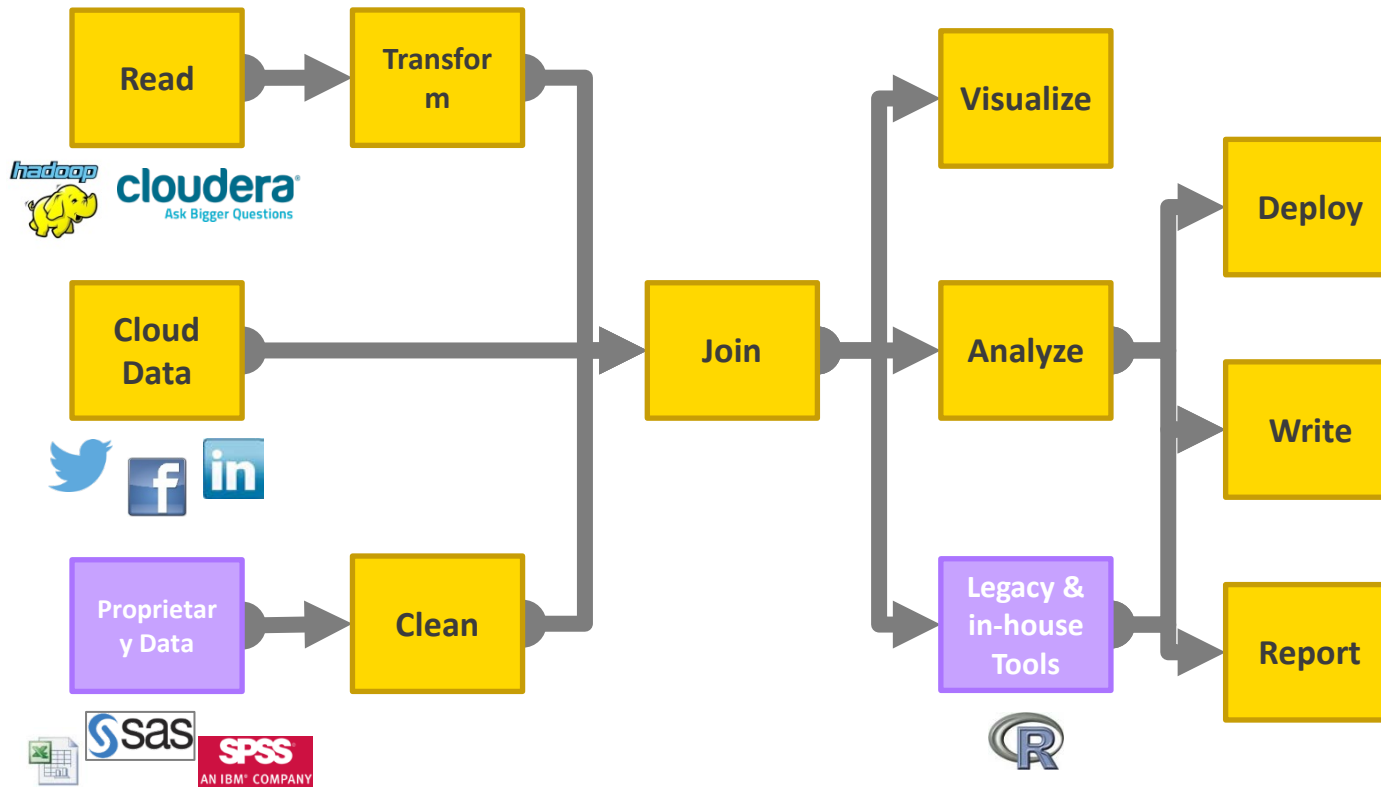
	var	rel.inf
Istat	Istat	42.40815

Methods & Algorithms
(and even libraries)

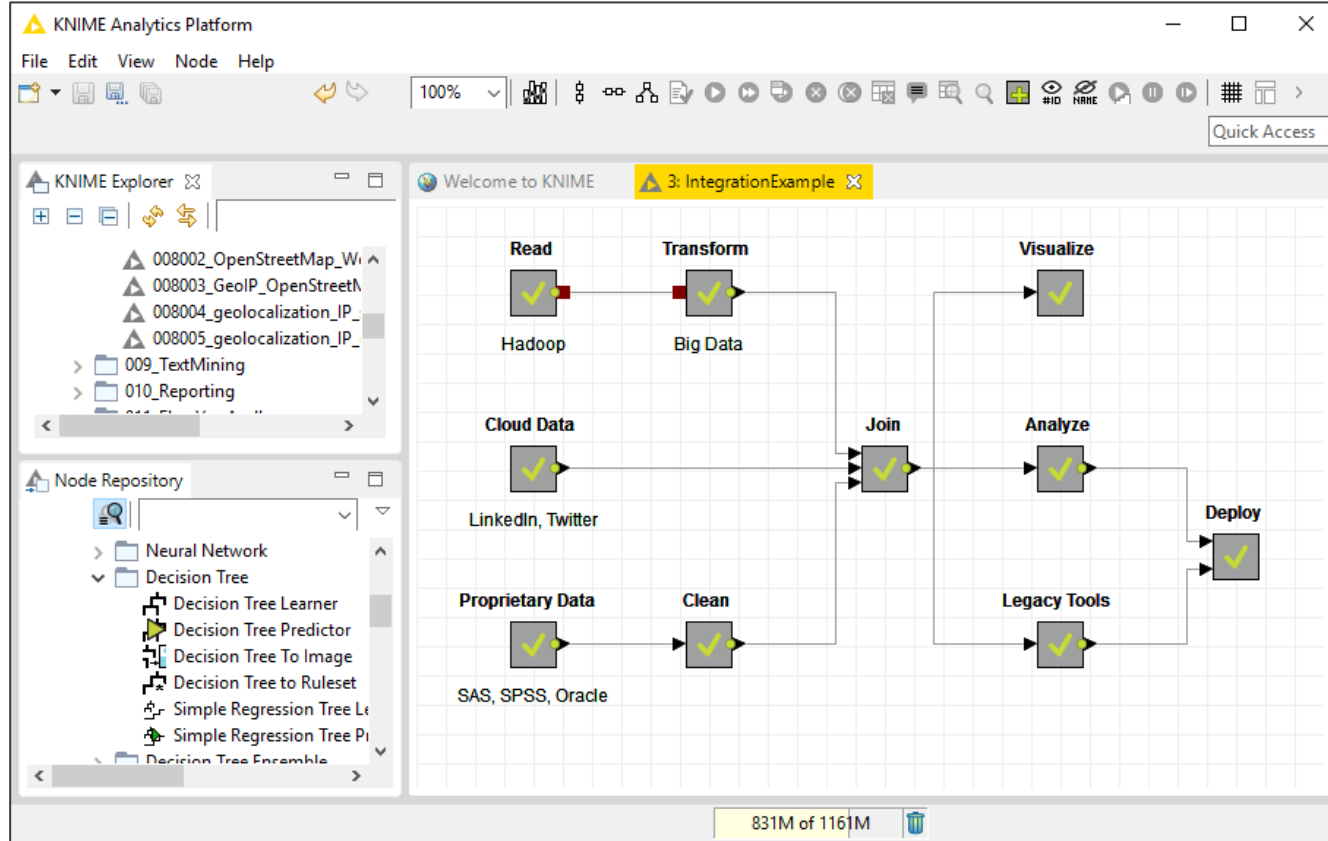
≠
Tools

DATA SCIENCE
DO IT. 

Visual Data Wrangling

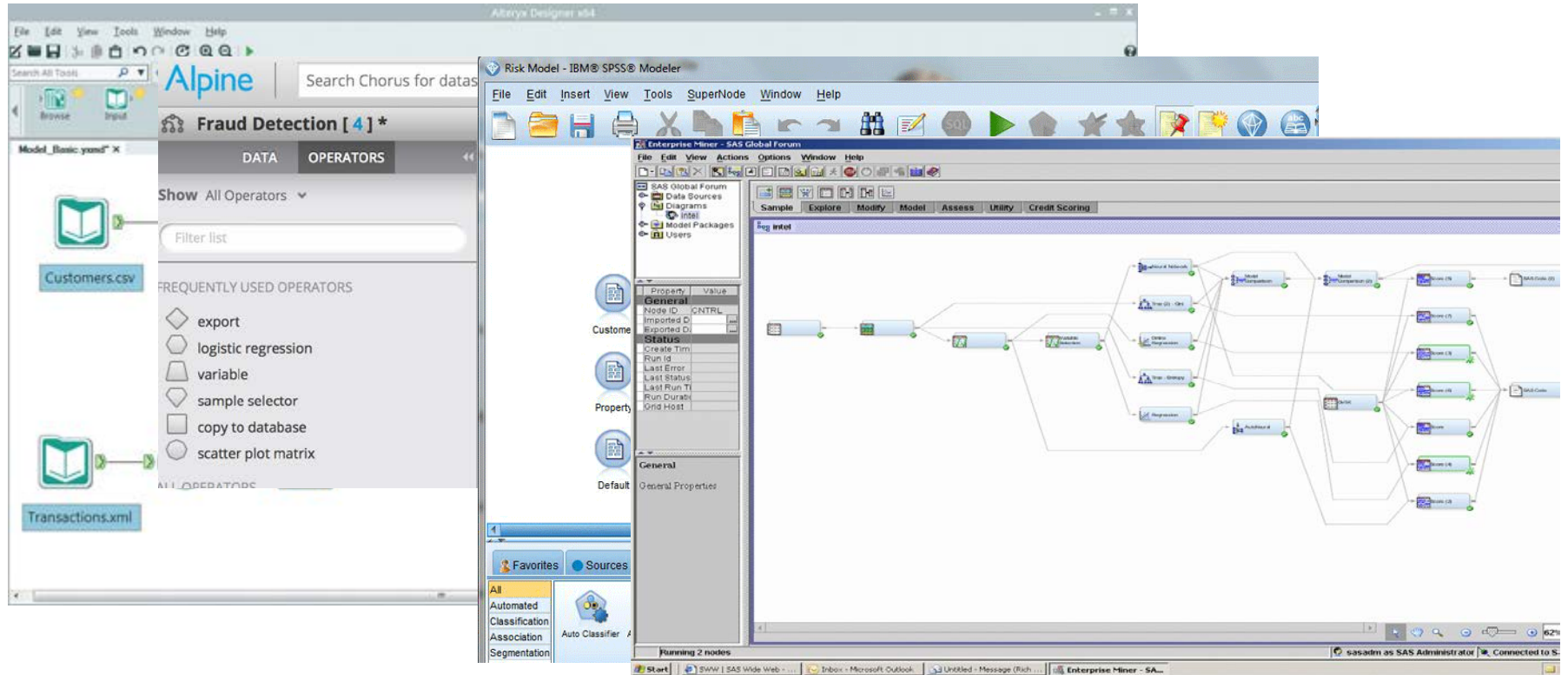


Visual Data Wrangling



Visual Workflows

KNIME is not the only one...



Data Scientists

Algorithms

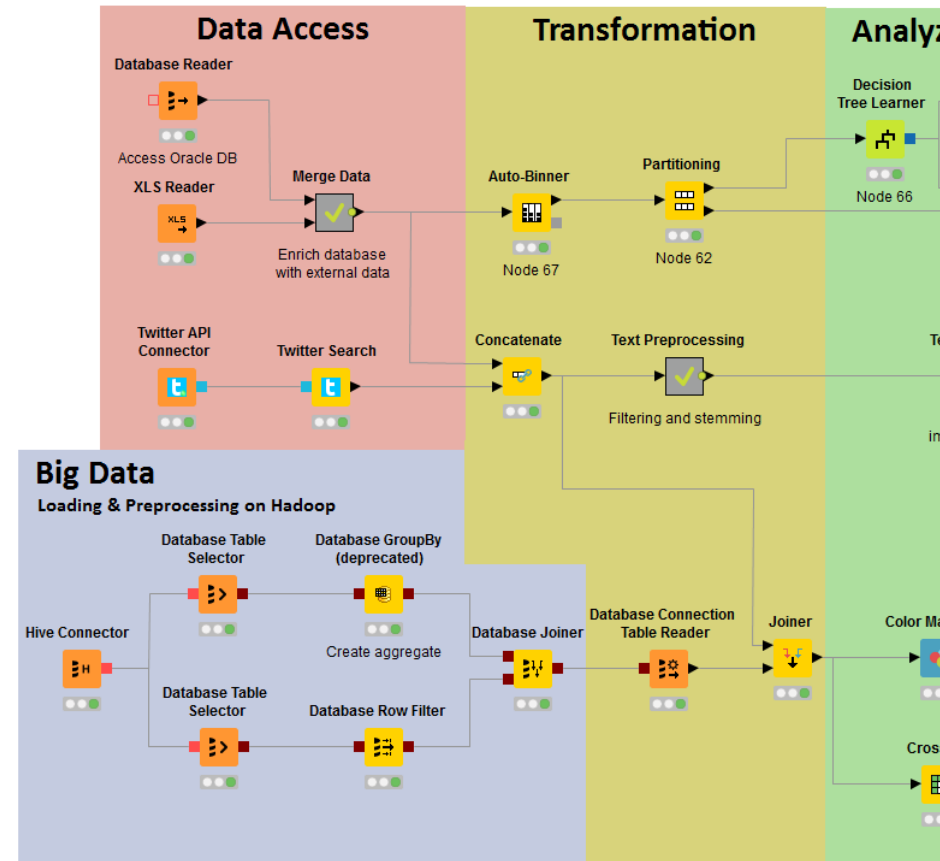
- may know
- mostly

The screenshot shows the Alteryx website's 'Solutions' page. The header includes the Alteryx logo and navigation links for Products, Solutions, Partners, Customers, Events, Resources, and Contact. The main content area features a blue background with the text 'ALTERYX ANALYTIC SOLUTIONS' and 'Data Blending'. Below this, there are two featured articles: 'Enabling Self-Service Data Analytics: 5 Ways Alteryx Improves Excel Processes' and '4 Key Attributes of Data Blending Solutions for Midsize Companies'. Each article has a 'GET THE REPORT' button. The bottom of the page has the headline 'Blend all Relevant Data for More Accurate Analysis' and a sub-headline 'As greater volumes and types of data become necessary for analysts to'.

Visual Data Blending

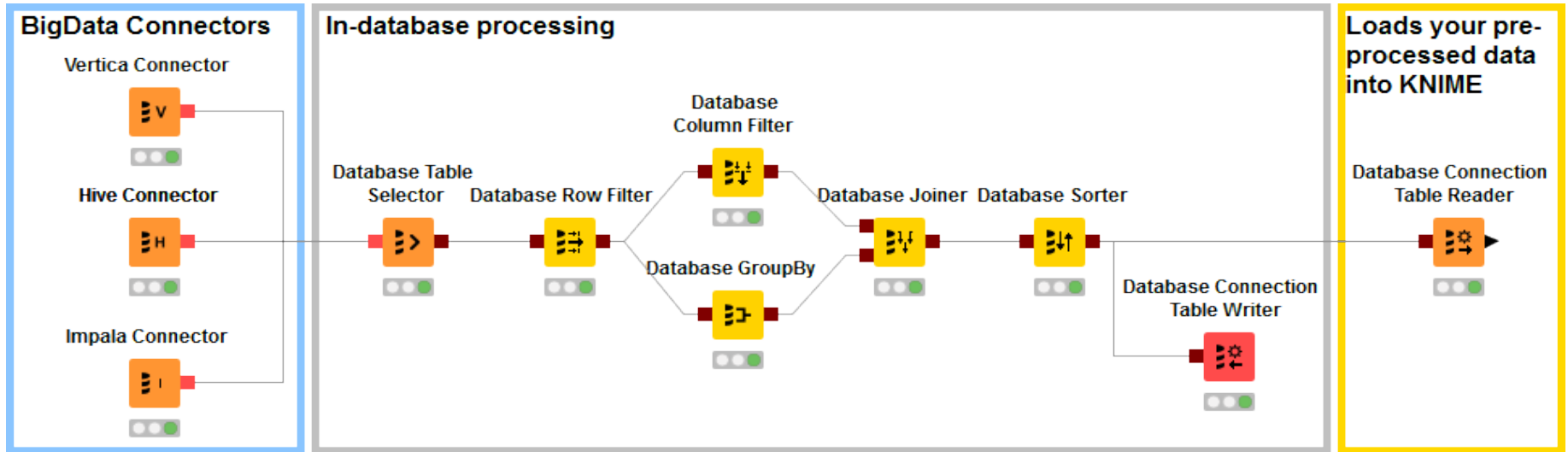
Data Blending:

- merge and transform all relevant sources
- self service (not an IT function!)



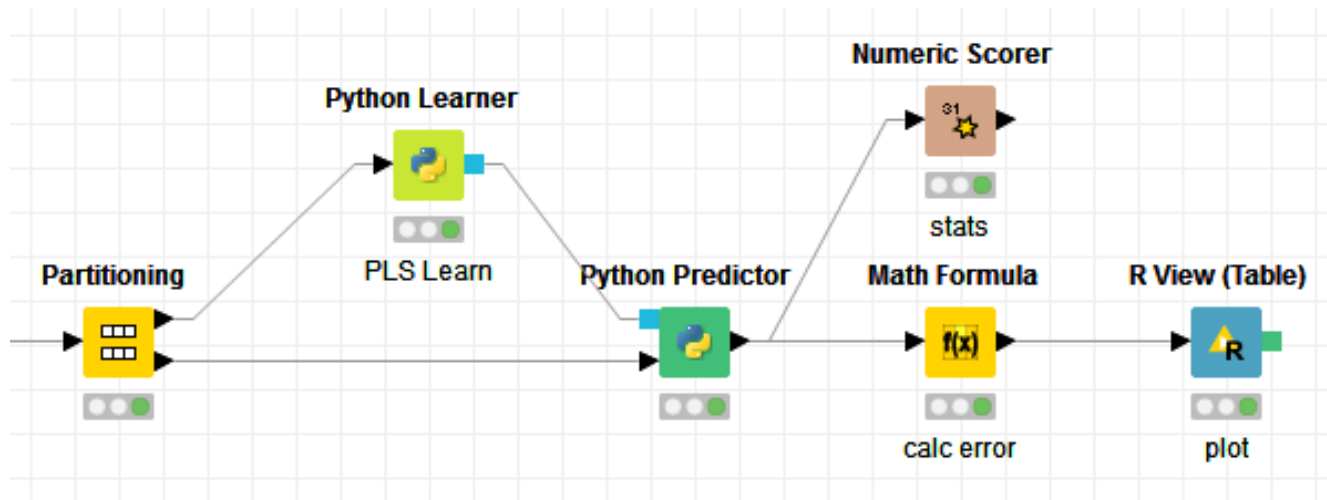
Transforming Data

Visual SQL...



Reusability = Collaboration

Combining and Reusing Expertise without having to care about the (to them: unimportant) details...



Collaboration...

Dialog - 0:27:27 - Python Learner (PLS Learn)

File

Script Options Templates Flow Variables Job Manager Selection Memory Policy

Columns

- 900 nm
- 902 nm
- 904 nm
- 906 nm
- 908 nm
- 910 nm
- 912 nm
- 914 nm
- 916 nm
- 918 nm
- 920 nm
- 922 nm

Flow variables

- knime.workspace

```

1 from sklearn.pls import PLSRegression
2
3 # Assign target variable & components
4 target_name = 'Octane'
5 n = 2
6
7 # Create subsets for fitting
8 y = input_table[target_name].copy()
9 X = input_table.copy().drop(target_name,1)
10
11 # Create and fit PLS model
12 pls = PLSRegression(n_components=n)
13 pls.fit(X, y)
14
15 # Export results
16 output_model = pls
    
```

Name	Type	Value
PLSRegres...	ABCMeta	<class 's...
X	DataFrame	900...
flow_vari...	dict	{u'knime...
input_table	DataFrame	900...
n	int	2
output_model	PLSRegres...	PLSRegres...
pls	PLSRegres...	PLSRegres...
target_name	str	Octane
y	Series	1 -0.1...

Execute script Execute selection Reset workspace

Execution successful

OK Apply Cancel ?

Dialog - 0:27:30 - R View (Table) (plot)

File

R Snippet PNG Settings Templates Flow Variables Job Manager Selection Memory Policy

Create Template...

Column List

- 900 nm
- 902 nm
- 904 nm
- 906 nm
- 908 nm
- 910 nm
- 912 nm
- 914 nm
- 916 nm
- 918 nm
- 920 nm
- 922 nm

R Script

```

15 color = Deviation
16
17
18
19 model = lm(knime.in$"Octane"~knime.in$"Prediction")
20
21
22 intercept = model$coefficients[1]
23 slope = model$coefficients[2]
24
25 line = geom_abline(intercept = intercept, slope = sl
26
27 note = annotate("text", x = 2:3, y = 20:21, label = .
28
29 qplot(x,y,color=color) + labels + line
    
```

Workspace

Name	Type
color	numeric
intercept	numeric
knime.flow.in	pairlist
knime.in	data.frame
labels	labels
line	proto
model	lm
note	proto
slope	numeric
x	numeric
y	numeric

Flow Variable List

- knime.workspace

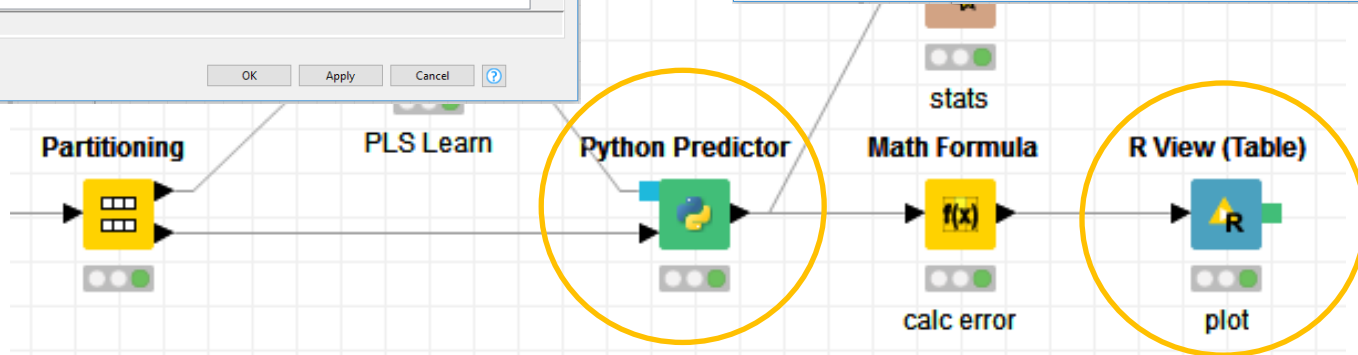
Eval Script Eval Selection Reset Workspace Show Plot

Console

```

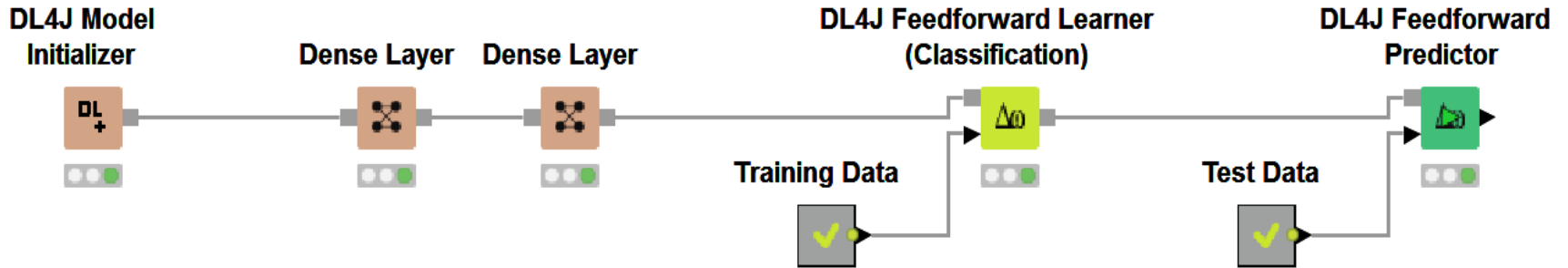
> # This example relies on ggplot2 and grid, if these packages are not part
+ # your R installation, please add them!
+ require(ggplot2)
+ require(grid)
+ library(ggrr)
    
```

OK Apply Cancel ?



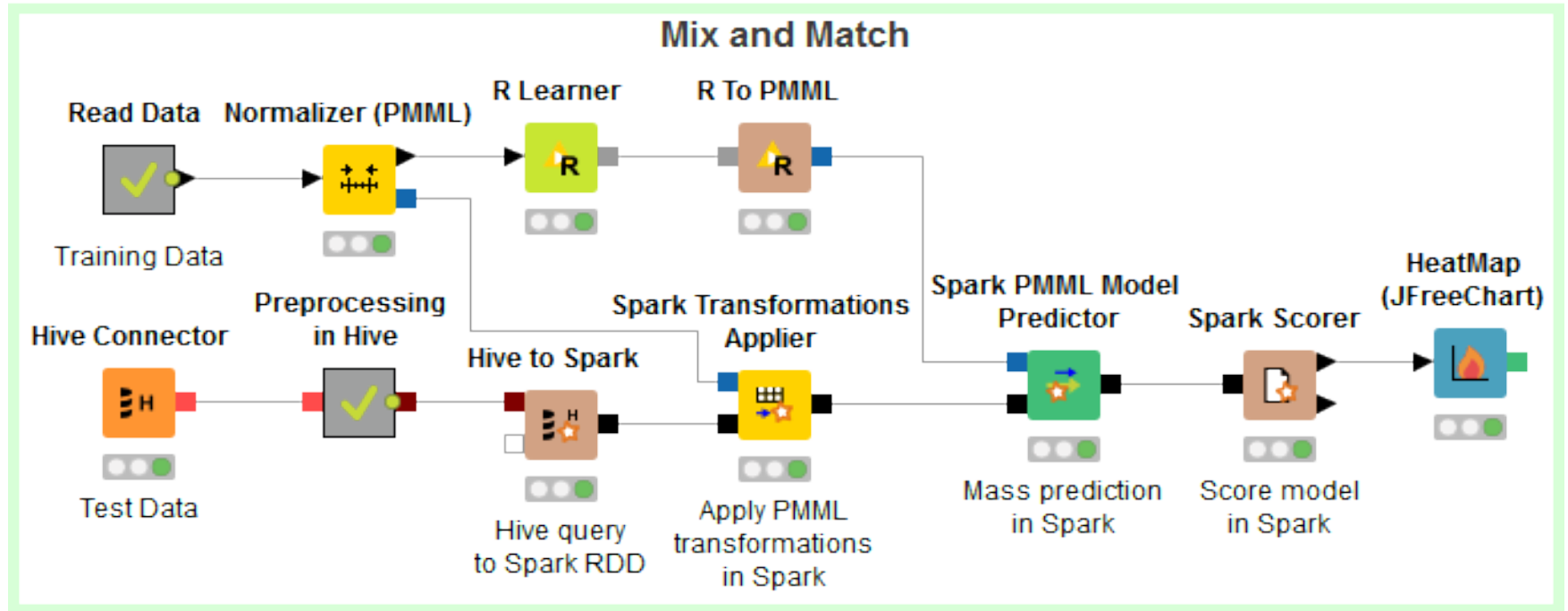
...Integration...

...getting quick access to new developments:



...Mixing & Matching.

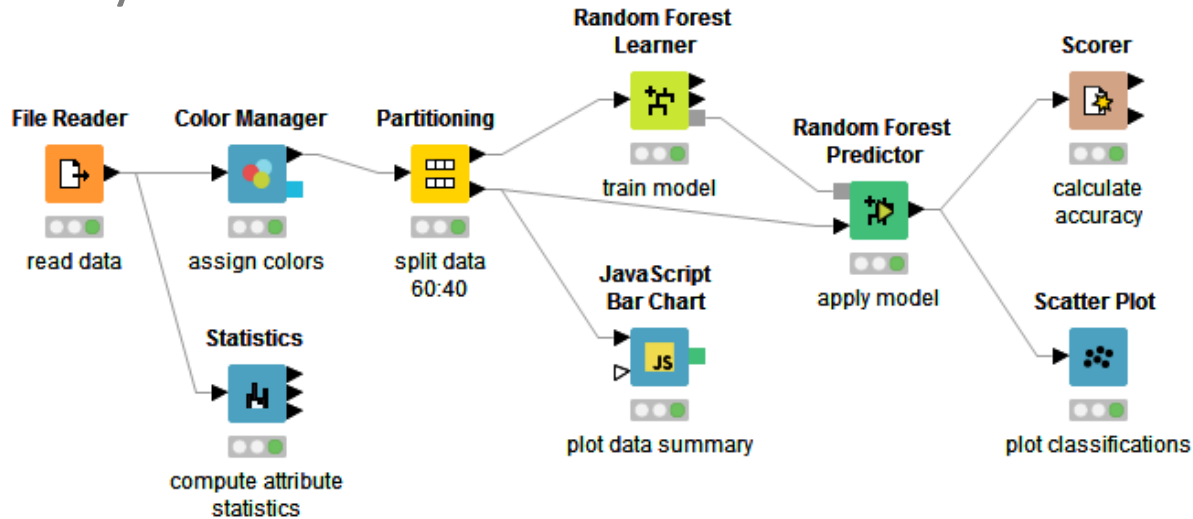
...and using data/tools where it makes sense!



Citizen Data Scientists

Analytical “Best Practices” are re-used by others:

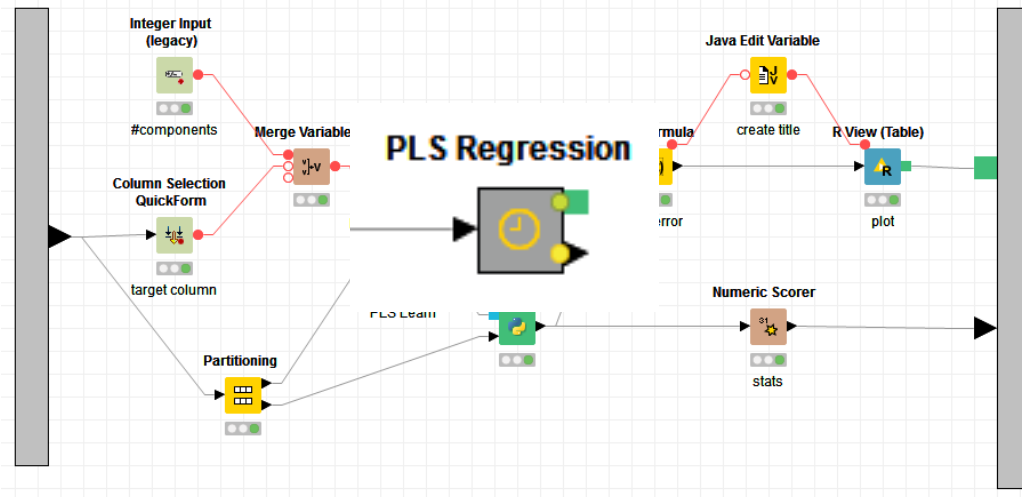
- maybe knows Java, Python, R, C, SQL or ...
- struggles with backwards compatibility, reproducibility, reusability...



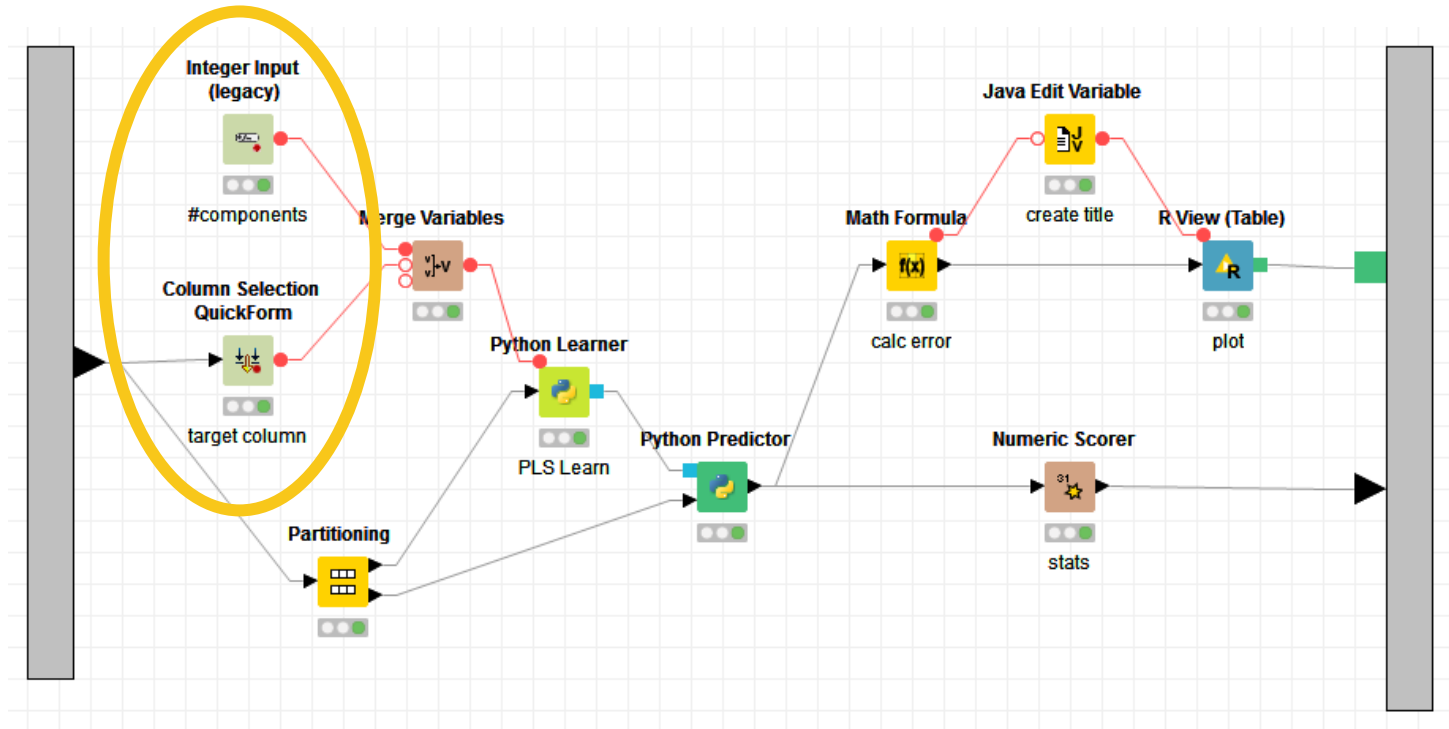
Citizen Data Scientists

Analytical “Best Practices” are re-used by others:

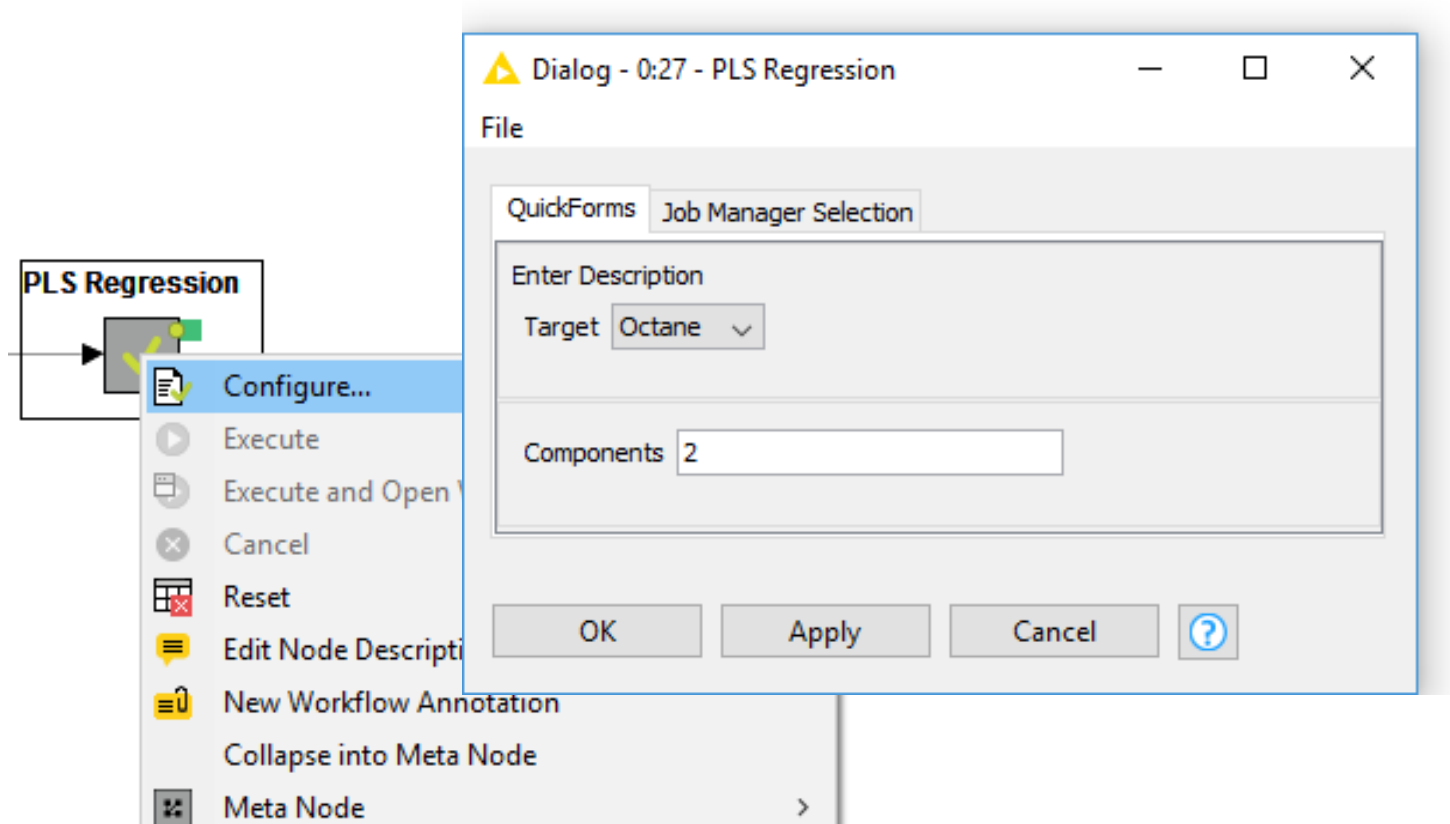
- maybe knows Java, Python, R, C, SQL or ...
- struggles with backwards compatibility, reproducibility, reusability...



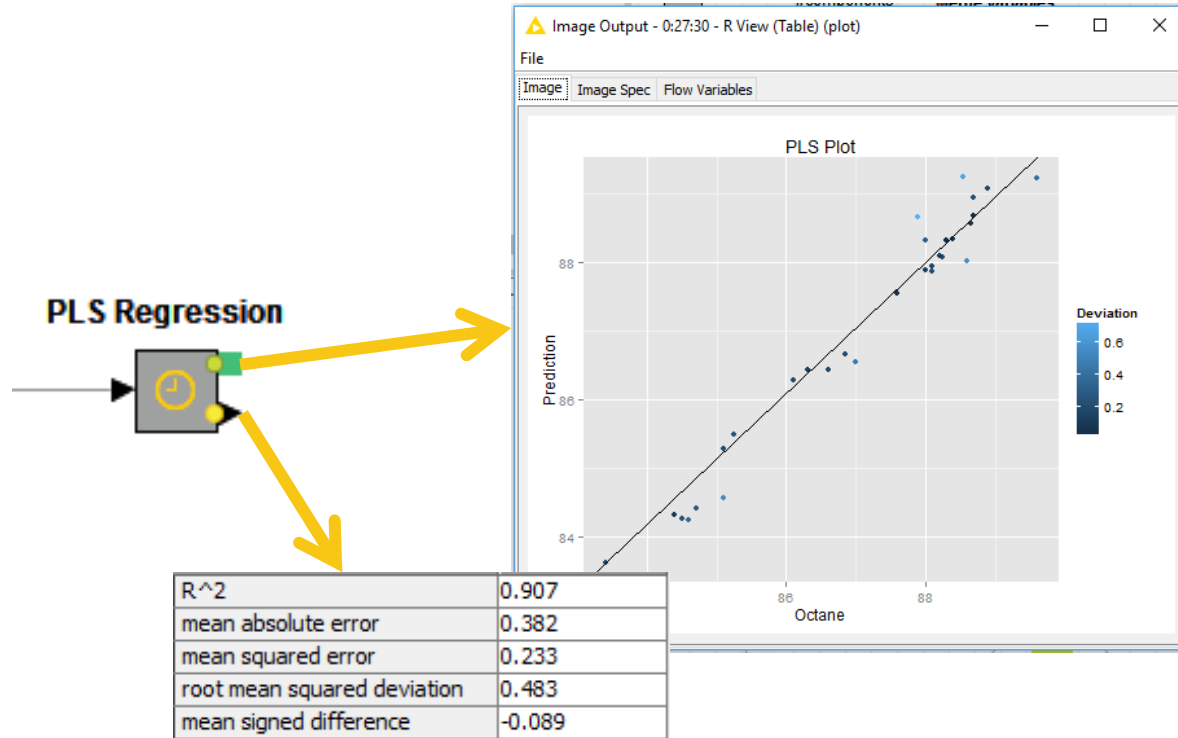
KNIME Metanodes: Controlling Parameters



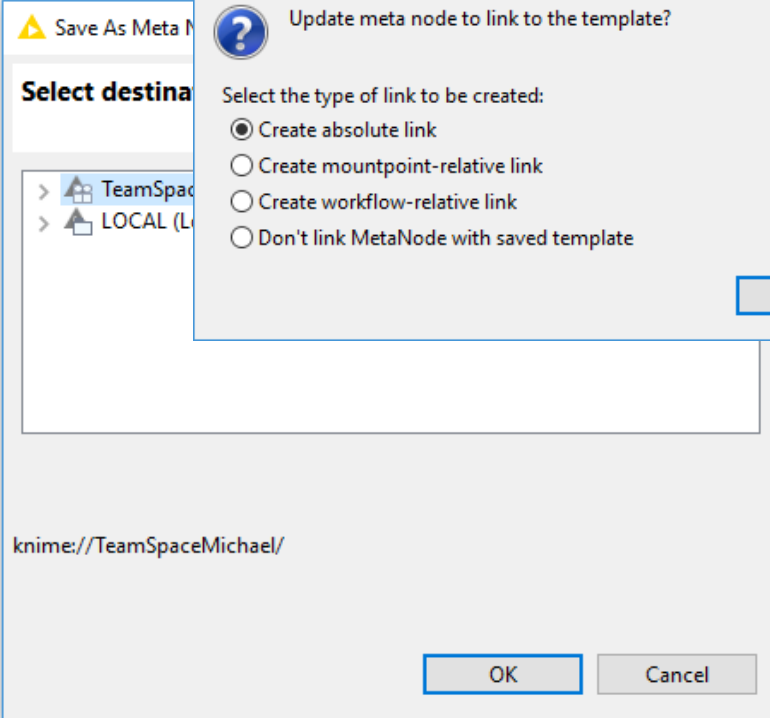
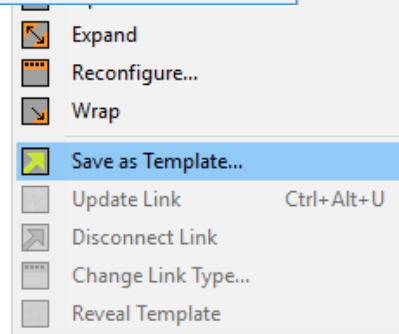
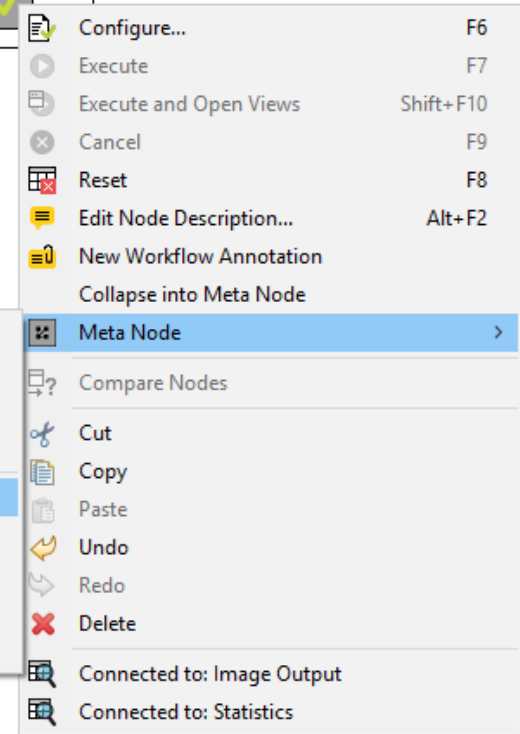
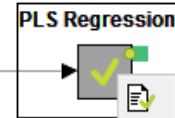
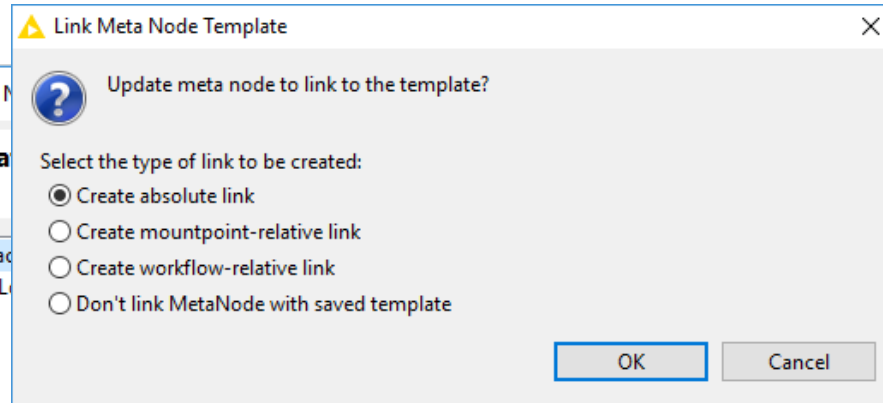
KNIME Metanodes: Exposing select parameters



KNIME Metanodes: Encapsulating Data Science



Sharing Metanodes



Workflow Templates

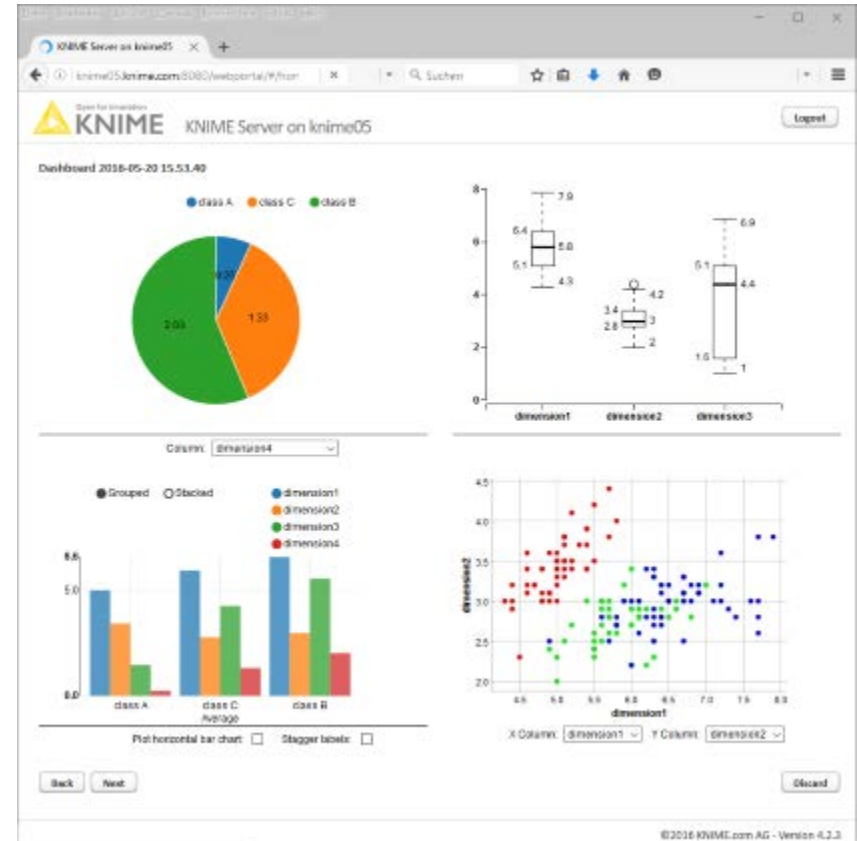
Visual Workflows...

- provide built-in Documentation
- allow easy Abstraction / Encapsulation
- enable Governance
- (KNIME workflows also guarantee Reproducibility)

Business Analyst

Ultimately the business analysts want to just use it:

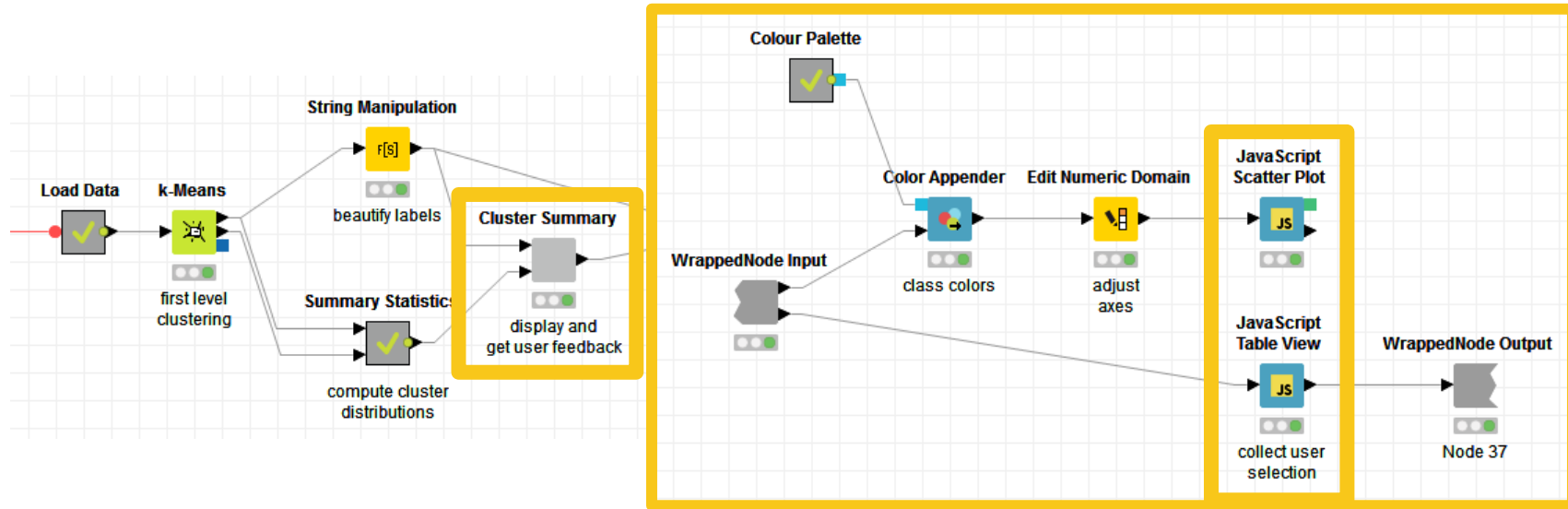
- they have domain expertise...
- ...but don't know (and should not know) C, R, Python, Java, or ...



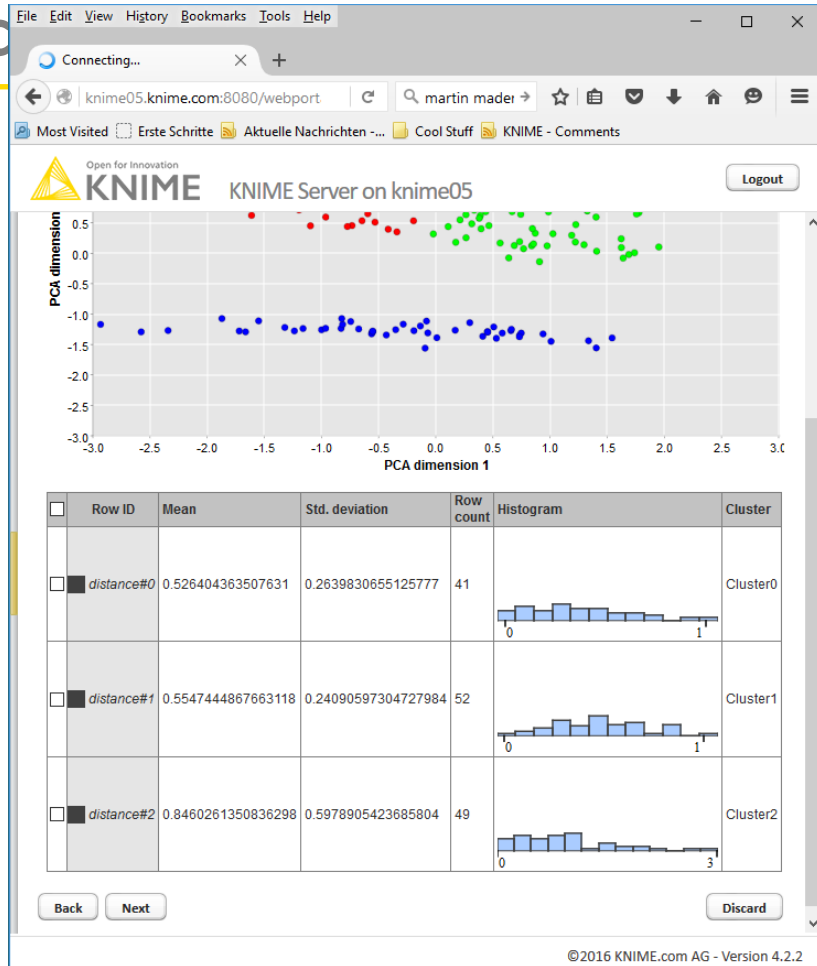
Guided Analytics

The image shows two overlapping browser windows from the KNIME Server web portal. The left window displays the login page with the KNIME logo and the text "Open for Innovation KNIME KNIME Server on knime05 Version 4.2.2". It includes input fields for "Username" (containing "michel.berthold") and "Password" (masked with dots), and a "Login" button. The right window shows the configuration page for the "Universal Clusterer Webportal 2016-01-12 15.01.23". It features a left sidebar with a tree view of workflows, including "Universal Clusterer Webportal" and "Universal Clusterer Webportal". The main content area contains configuration options: "Column Delimiter" (a text input field with a comma), "File to use" (a "Change File" button and "Uploaded file 'cars-85.csv' (28 KB)"), "Comment Character" (a text input field with "#"), "Quote Character" (a text input field with " "), "Has Row Header" (an unchecked checkbox), and "Has Column Header" (a checked checkbox). "Back", "Next", and "Discard" buttons are at the bottom. The footer of the right window reads "©2016 KNIME.com AG - Version 4.2.2".

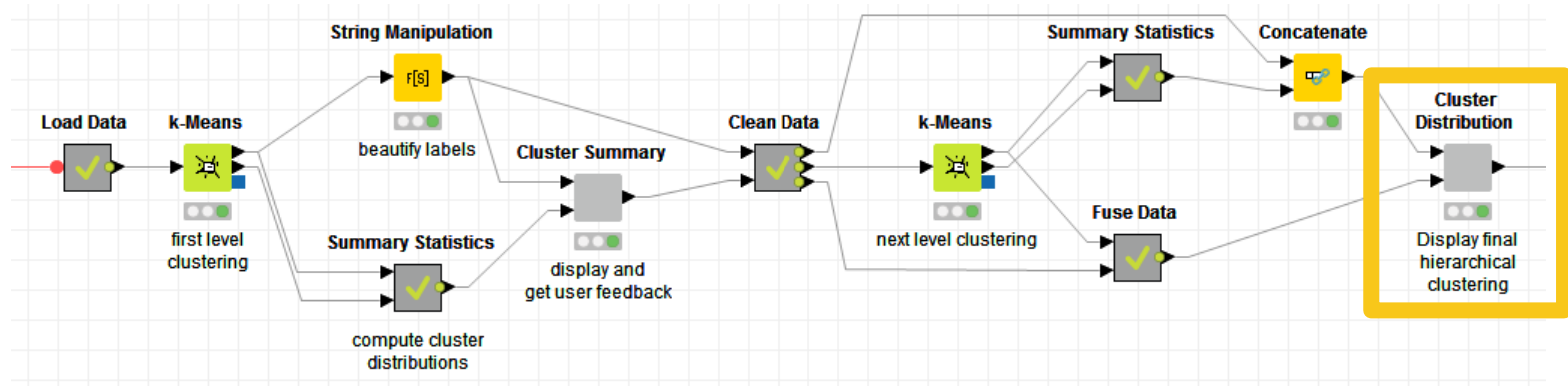
Guided Workflow Execution



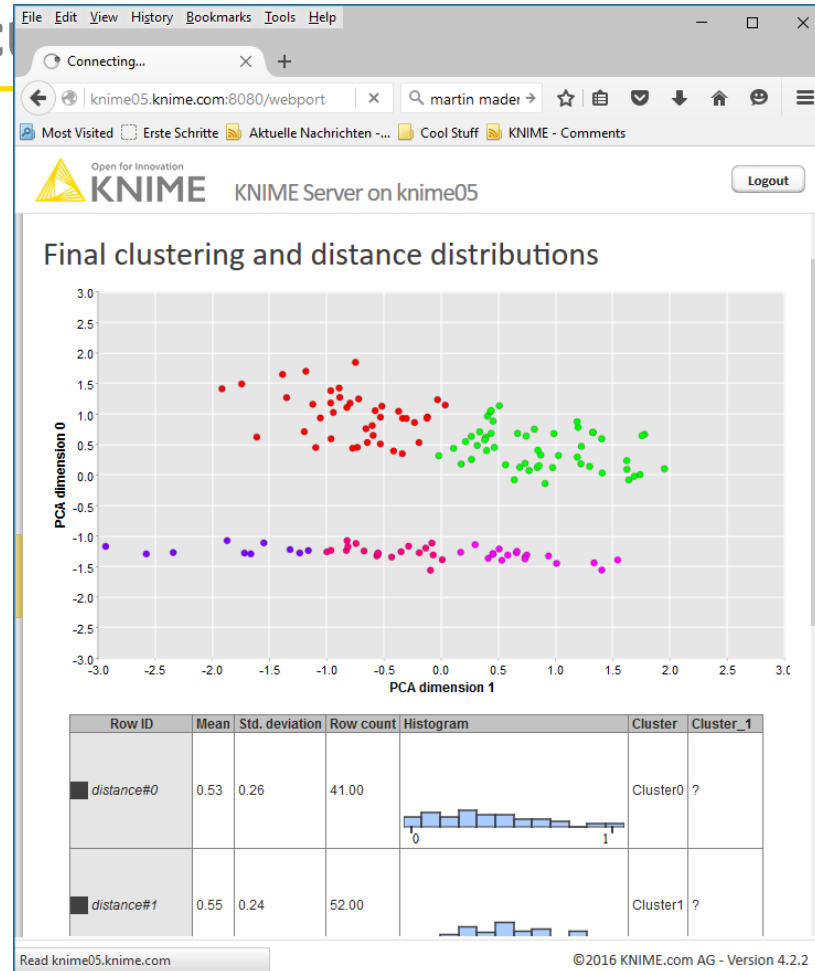
Guided Workflow Exec



Guided Workflow Execution



Guided Workflow Exec



Guided Analytics and KNIME

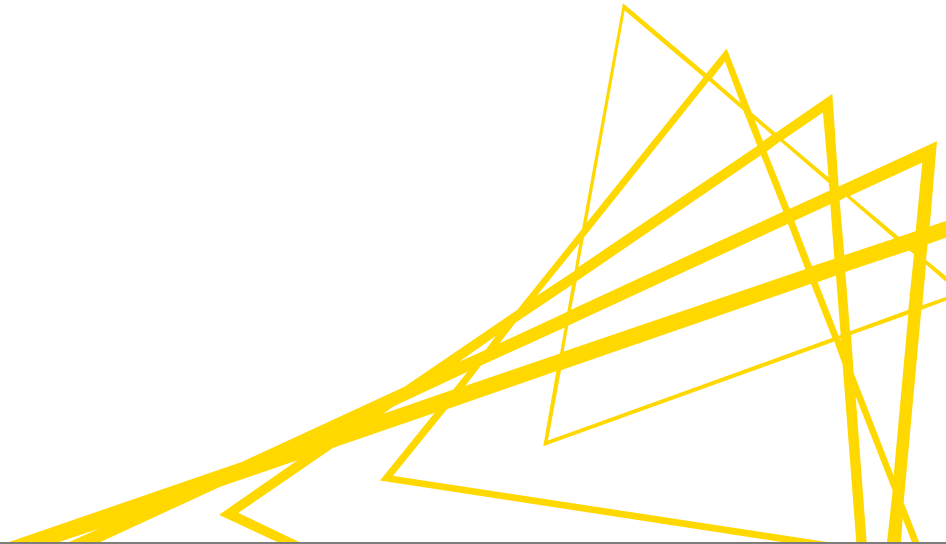
Data Scientist

- designs workflow with interaction points...
- ...which provide sufficient flexibility to incorporate user focus.

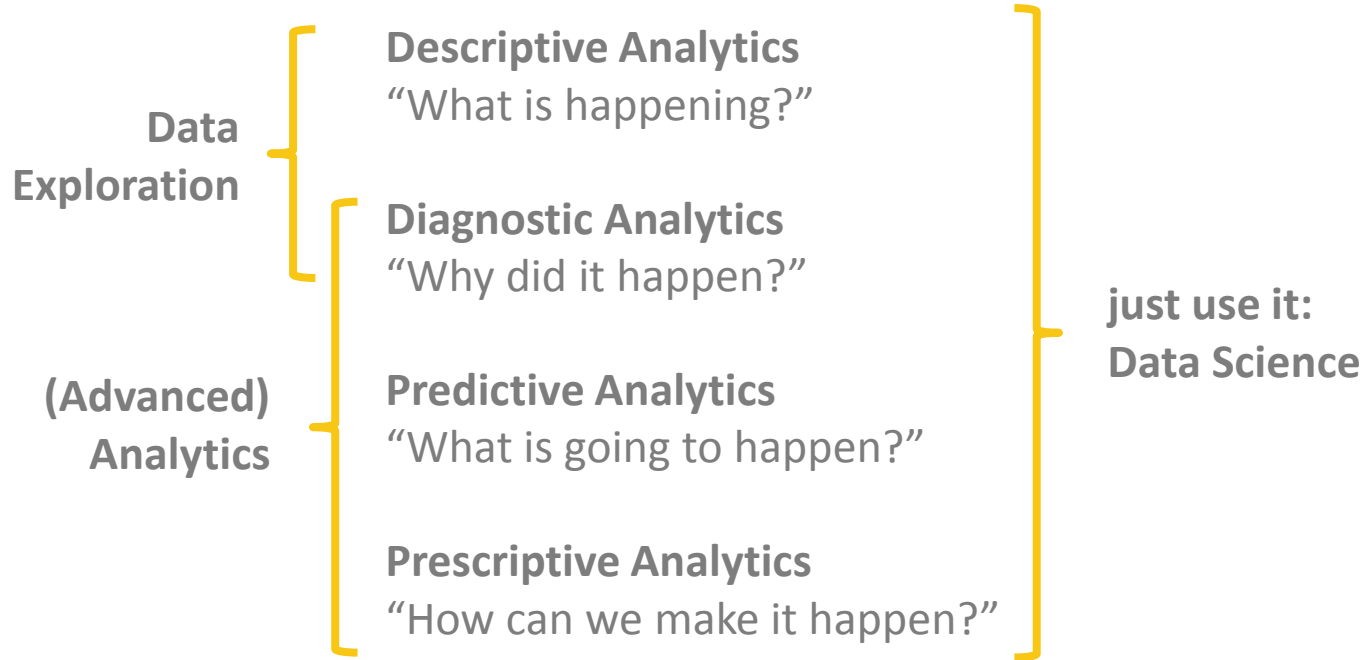
Deployed to *Analytics Consumer* as an easy UI with

- appropriate abstraction level for his/her level of expertise...
- ...and hiding unnecessary complexity.

Take Aways



1: Data Science and Flavors of Analytics



2: Data Science is not (a new) Science

Data Science is about:

- ease of use (focus on the analytical part)
- collaboration (combine expertise)
- reusability (recycle previous/others' work), and
- deployment (give others easy access).

→ Data Scientists need Tools that (let them) work.

3: Data Science for the Masses

- Black Box:
Limited Applicability (narrow, well-defined tasks)
- Analytical Templates:
Limited Audience (req. Data Science!)
- Guided Analytics - Analytical Applications:
Business Analyst in Driver's Seat (with Safety Belt!)

Thank You!



The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME.com AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.