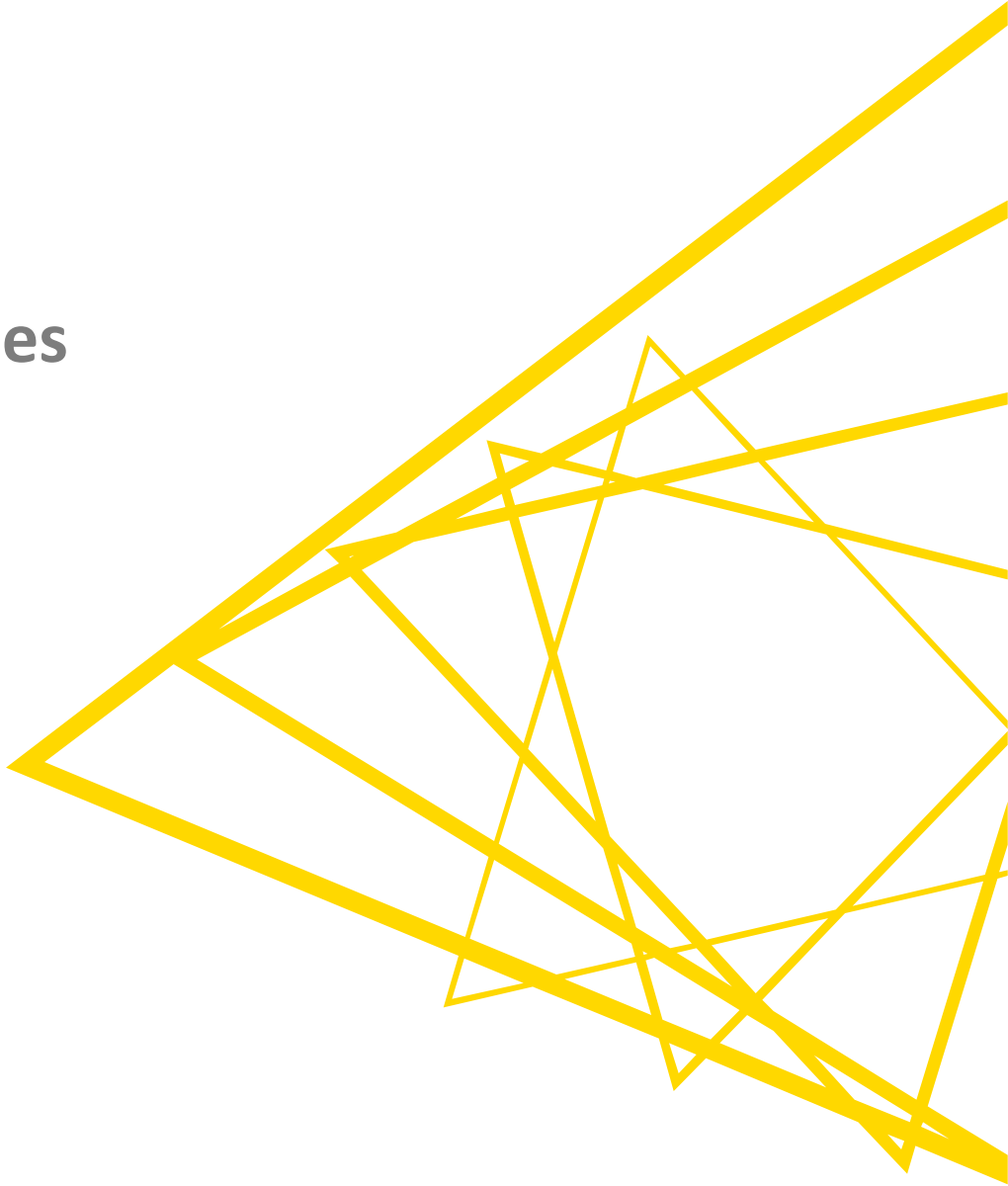
Data Science for the Masses
Mission Impossible?

Michael Berthold

Uni Konstanz and KNIME



© 2016 KNIME.com AG. All Rights Reserved.

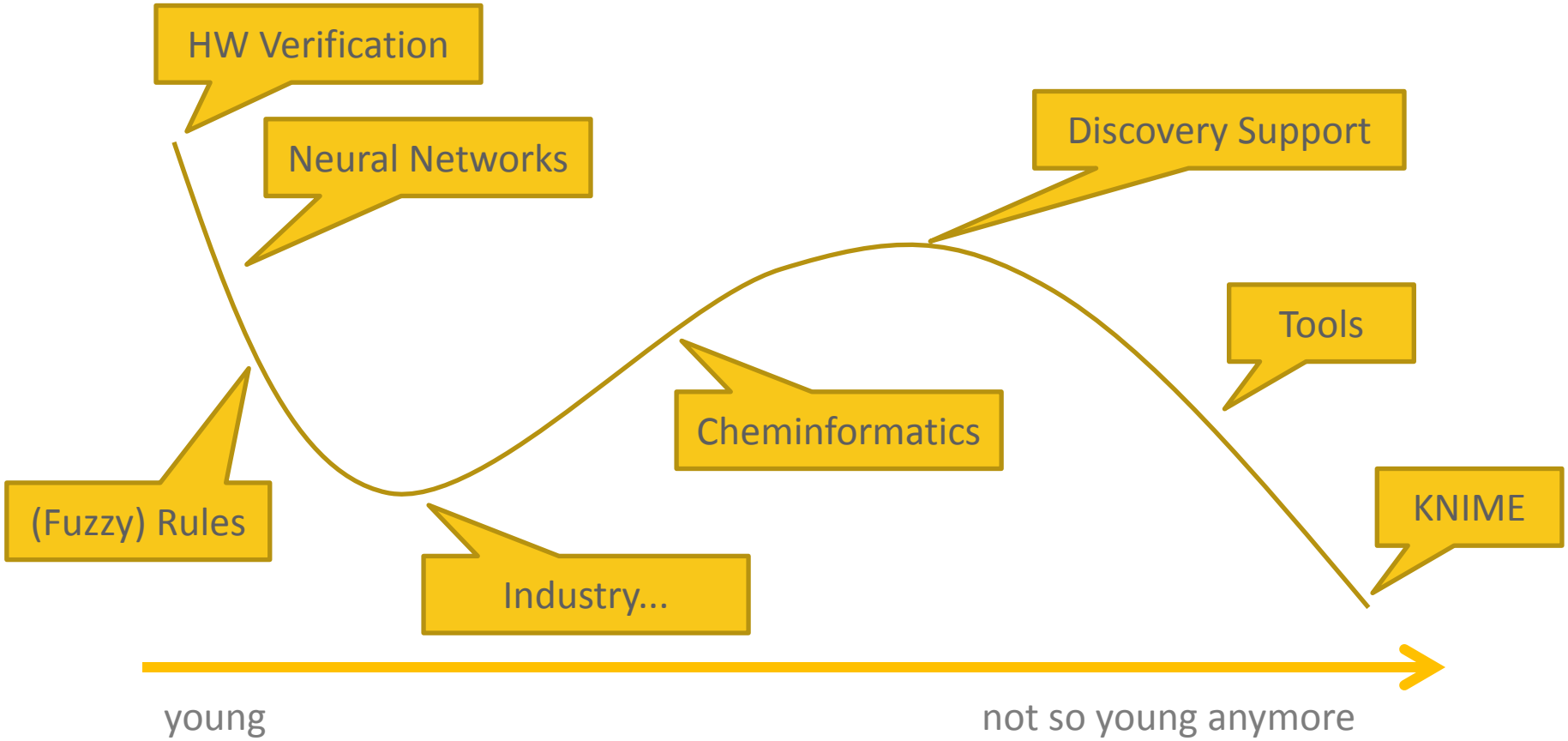
Menu of the Day

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

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Disclaimer: This is me.



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

The Data Science Landscape

It’s all about

- Data,
- Methods,
- and People.

© 2016 KNIME.com AG. All Rights Reserved.

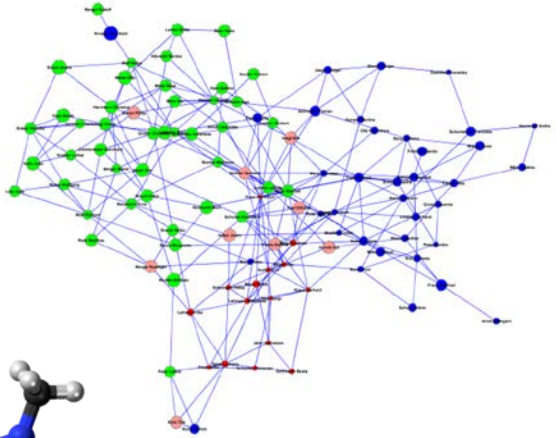
Open for Innovation ® KNIME

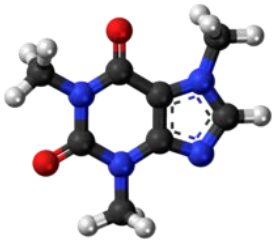
Data.

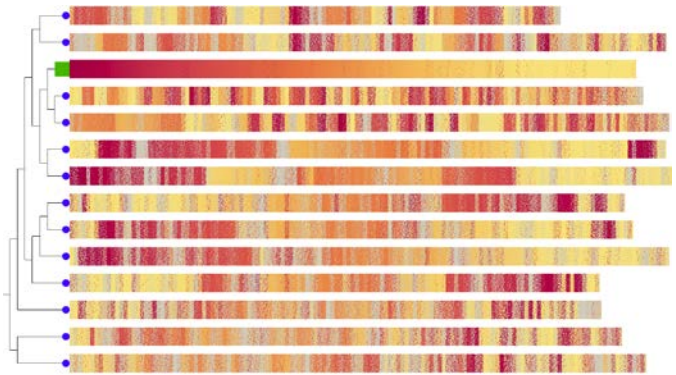
© 2016 KNIME.com AG. All Rights Reserved.

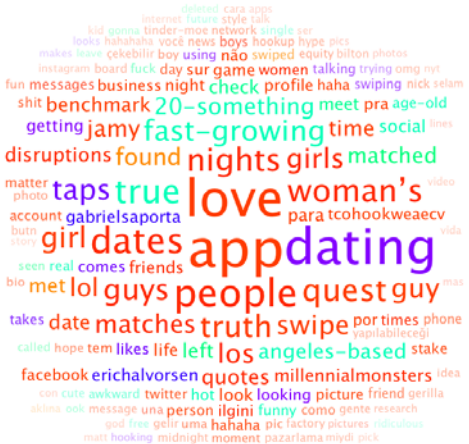
Data. The Past

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





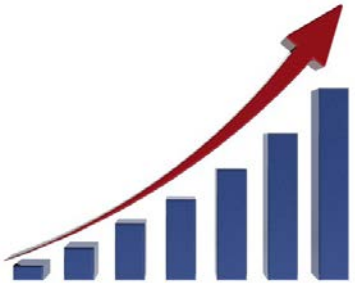


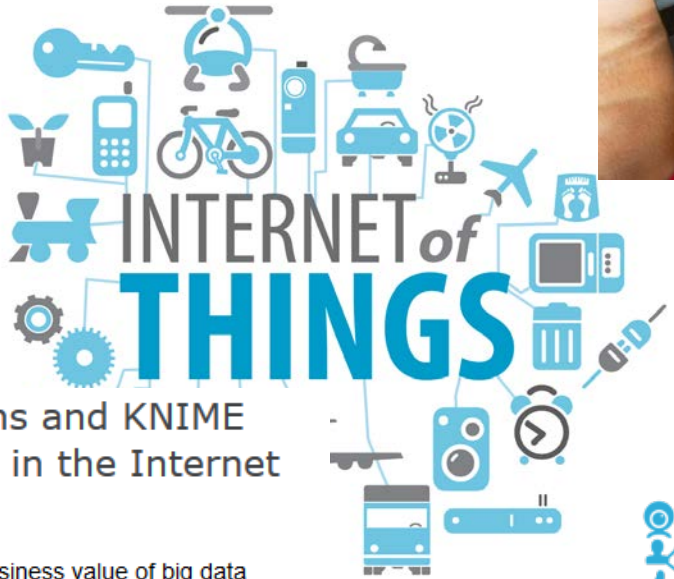


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Data. Current Trends





BOSCH
Invented for life

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.

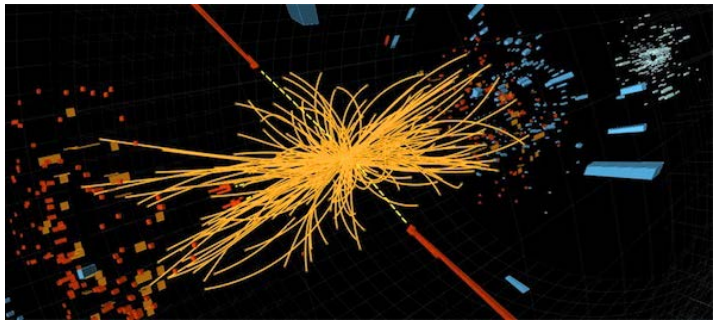
To leverage the business value of big data, Bosch Software Innovations will

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

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.
- ???

© 2016 KNIME.com AG. All Rights Reserved.

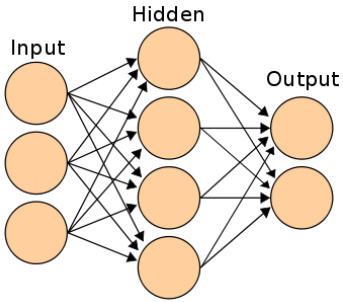
Open for Innovation ® KNIME

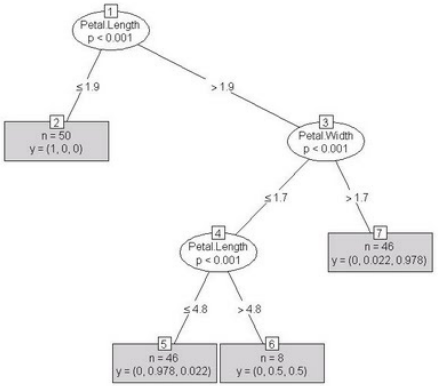
Methods.

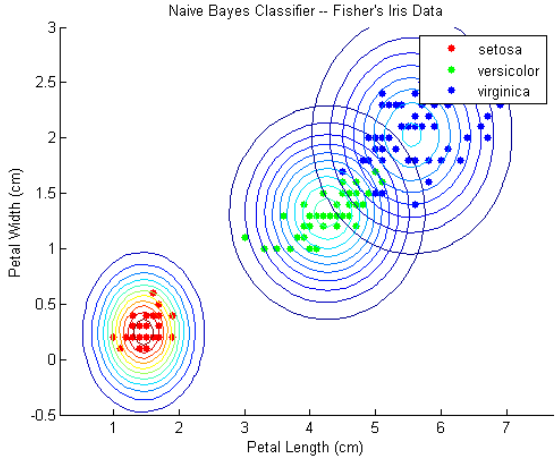
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

Methods. The Past



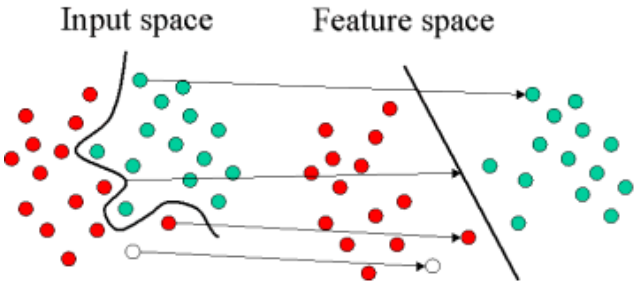


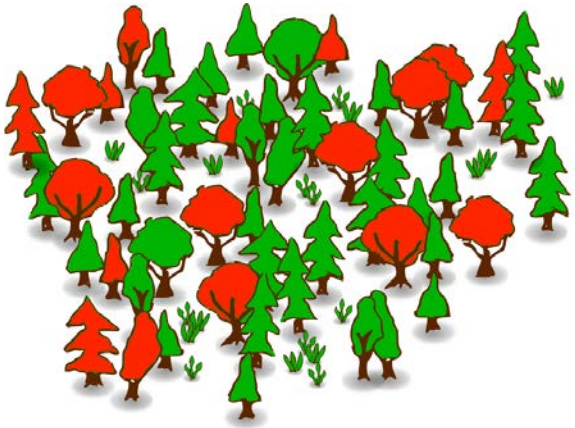


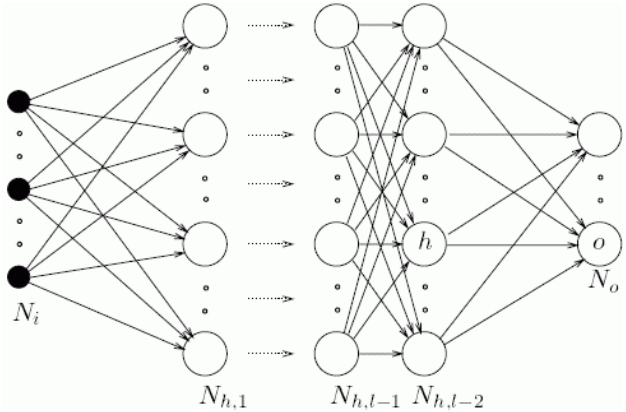
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

Methods. Current Trends





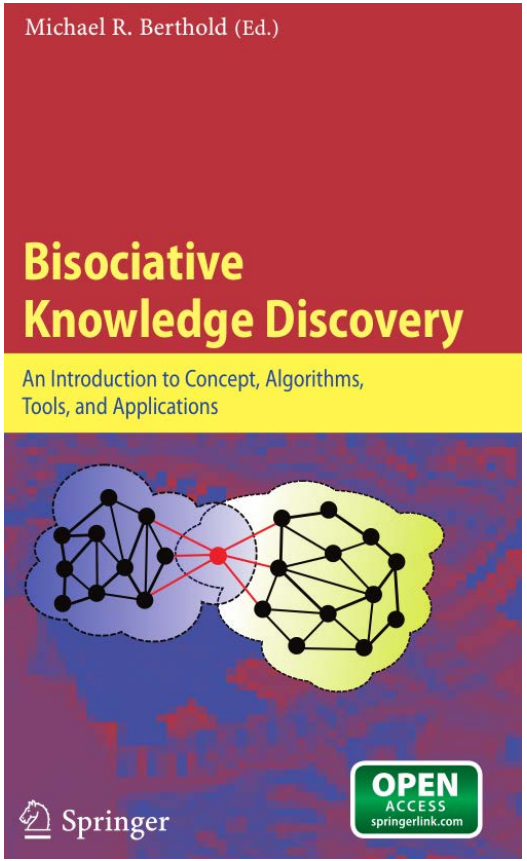


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Methods. What's up Next?

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



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

People.

© 2016 KNIME.com AG. All Rights Reserved.

People. The Past

Insights generated by Experts.

- Statistics
- Machine Learning

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

People. Current Trends

Insights generated
by more Experts.

- Data Mining
- Visual Exploration

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

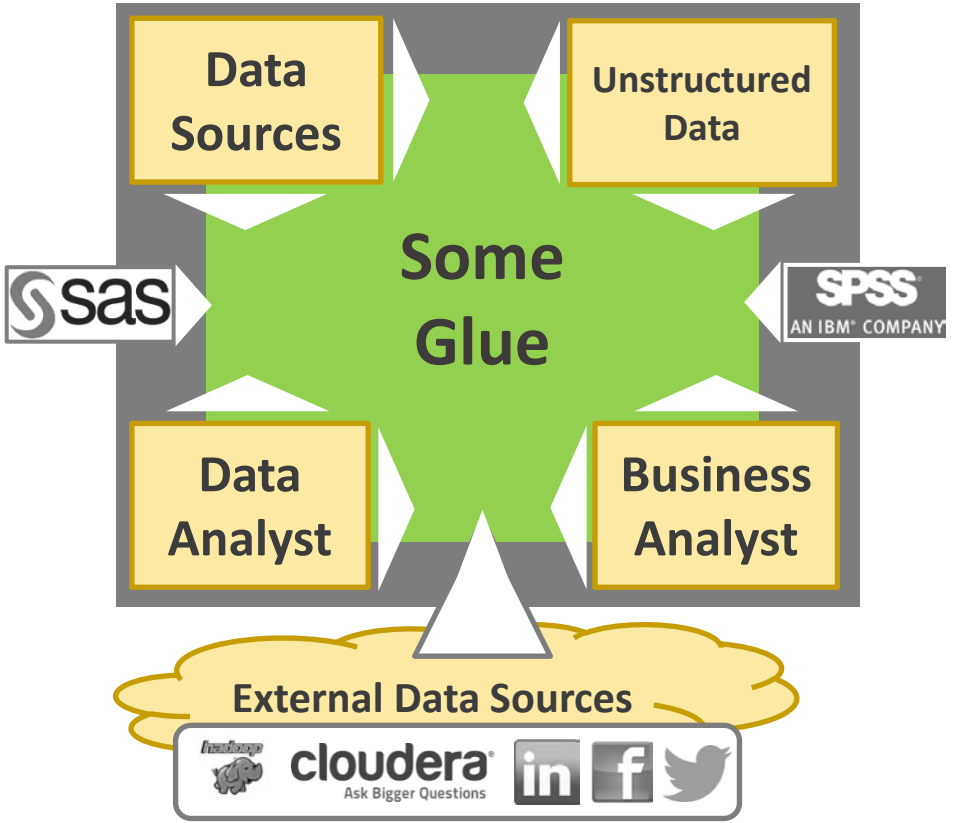
People. What's up Next?

Insights for Everybody,

Anytime.

- Casual Analysts
- Business Users

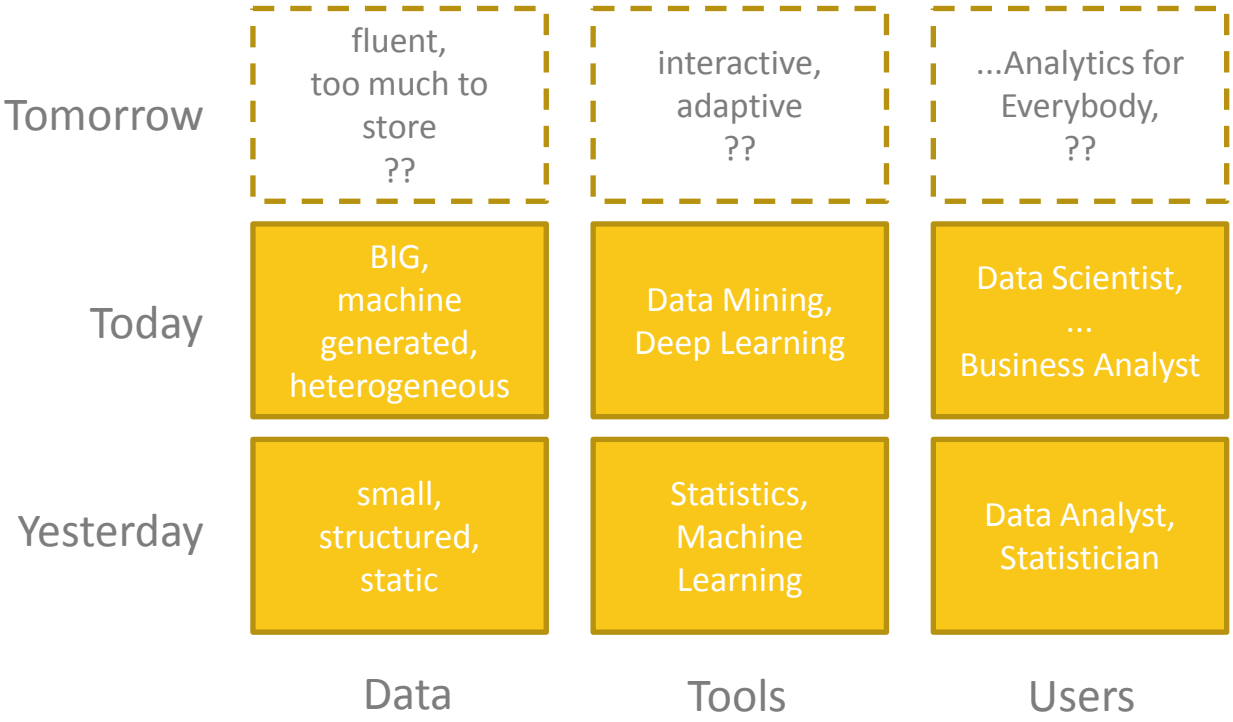
→ "Analytics for the Masses"...



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

Predicting the Future of Analytics is Hard



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.

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

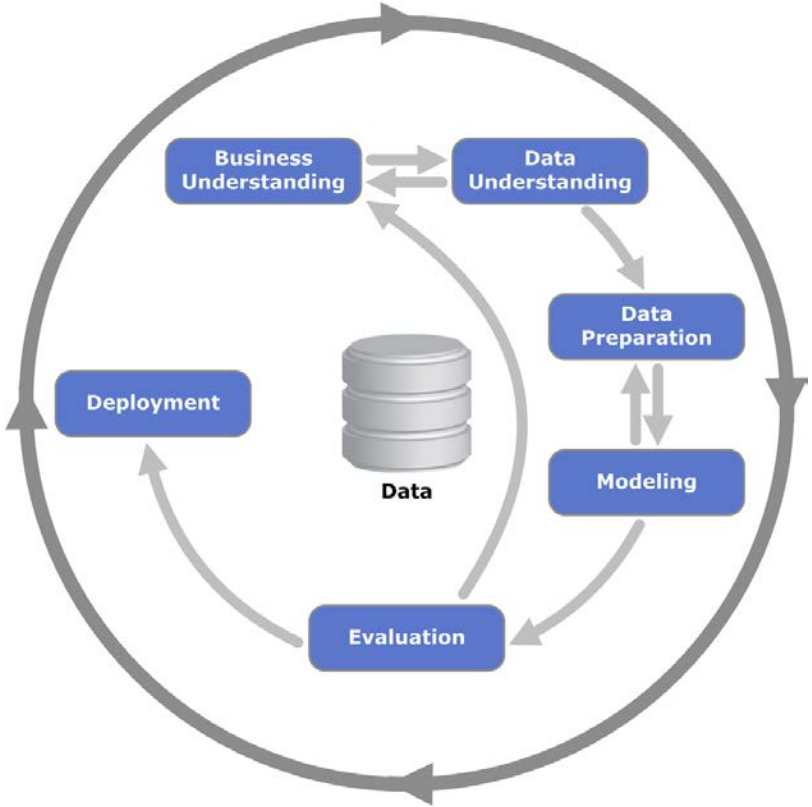
Data Science Processes.

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

CRISP-DM

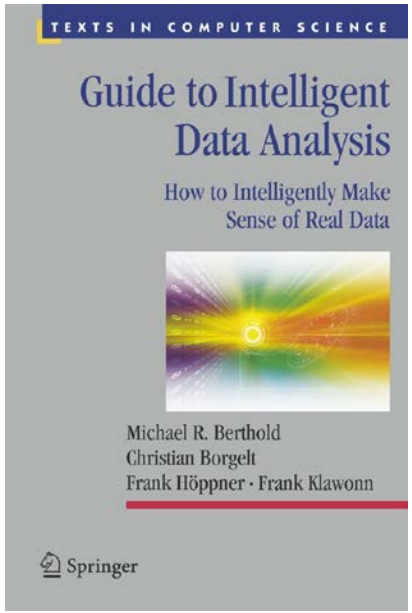
Cross Industry Standard Process for Data Mining

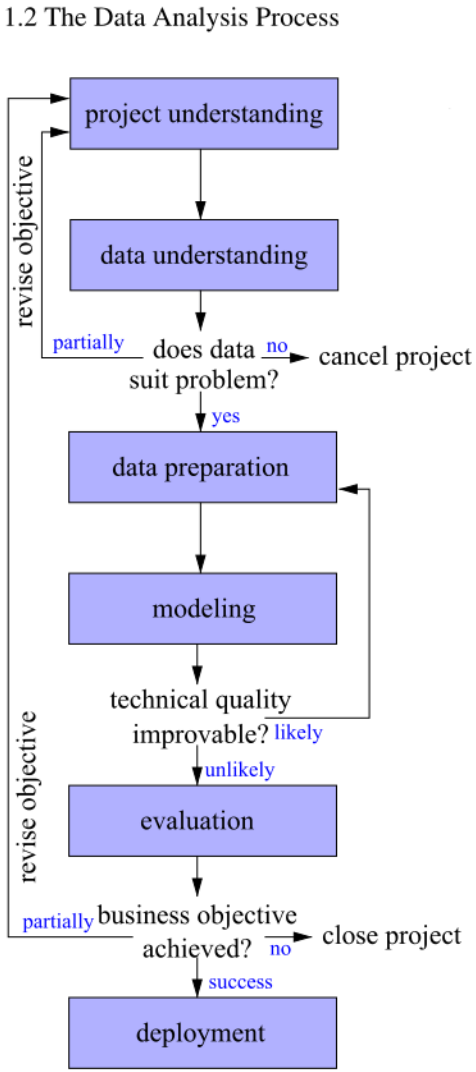


(source Wikipedia)

SEMMA

- Sample
- Explore
- Modify
- Model
- Assess

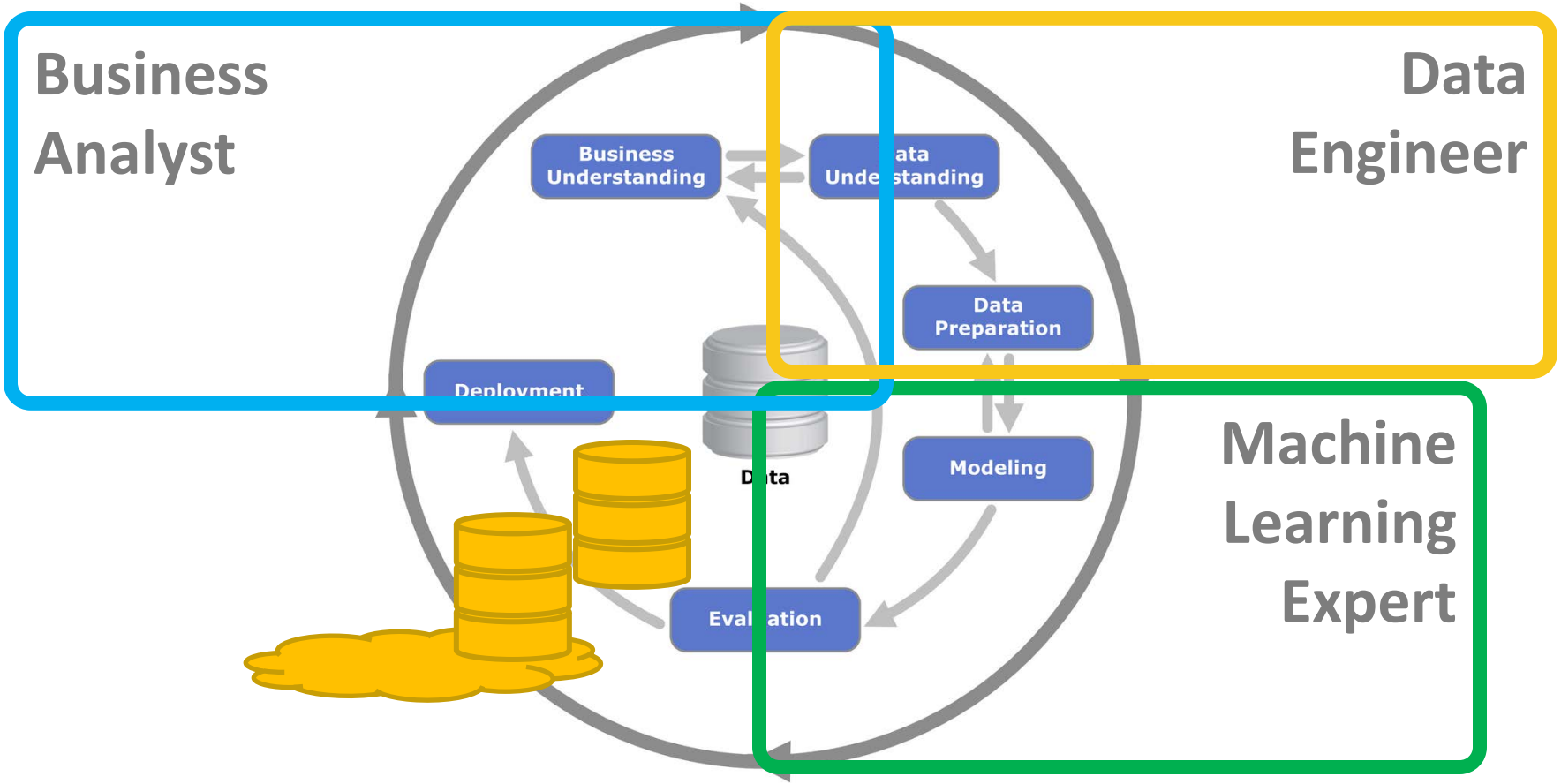




© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

CRISP-DM: one for all?



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

CRISP-DM...

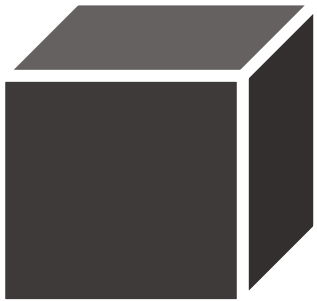
...and the famous magic black box

Step 1:

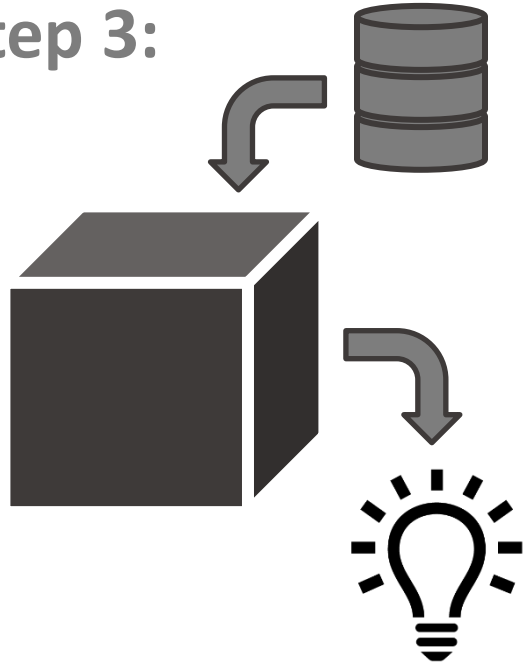
Data Scientist

+

Step 2:



Step 3:

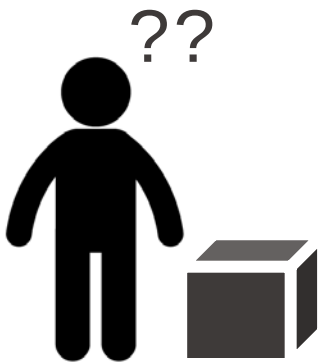


© 2016 KNIME.com AG. All Rights Reserved.

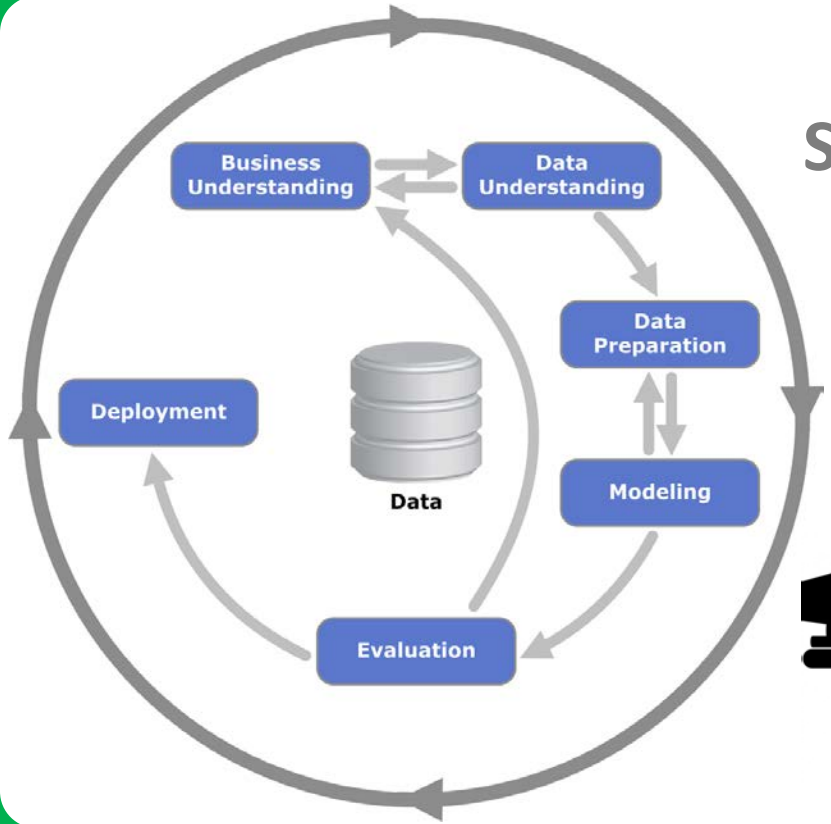
Open for Innovation ® KNIME

CRISP-DM and the Black Box

Business Analyst



Data Scientist



© 2016 KNIME.com AG. All Rights Reserved.

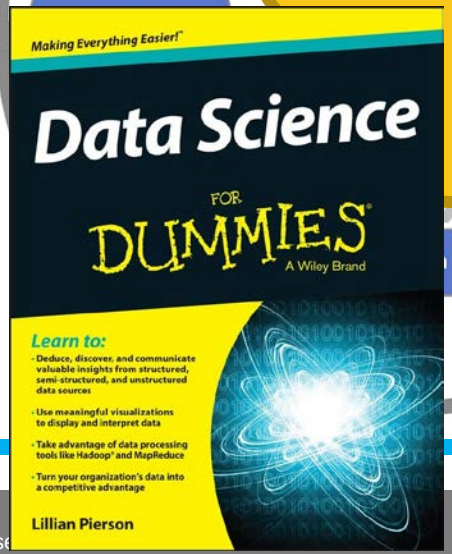
Open for Innovation ® KNIME

CRISP-DM and the Gartner View

~~Business~~
Analyst
Citizen
Data
Scientist

?

...and then some magic happens

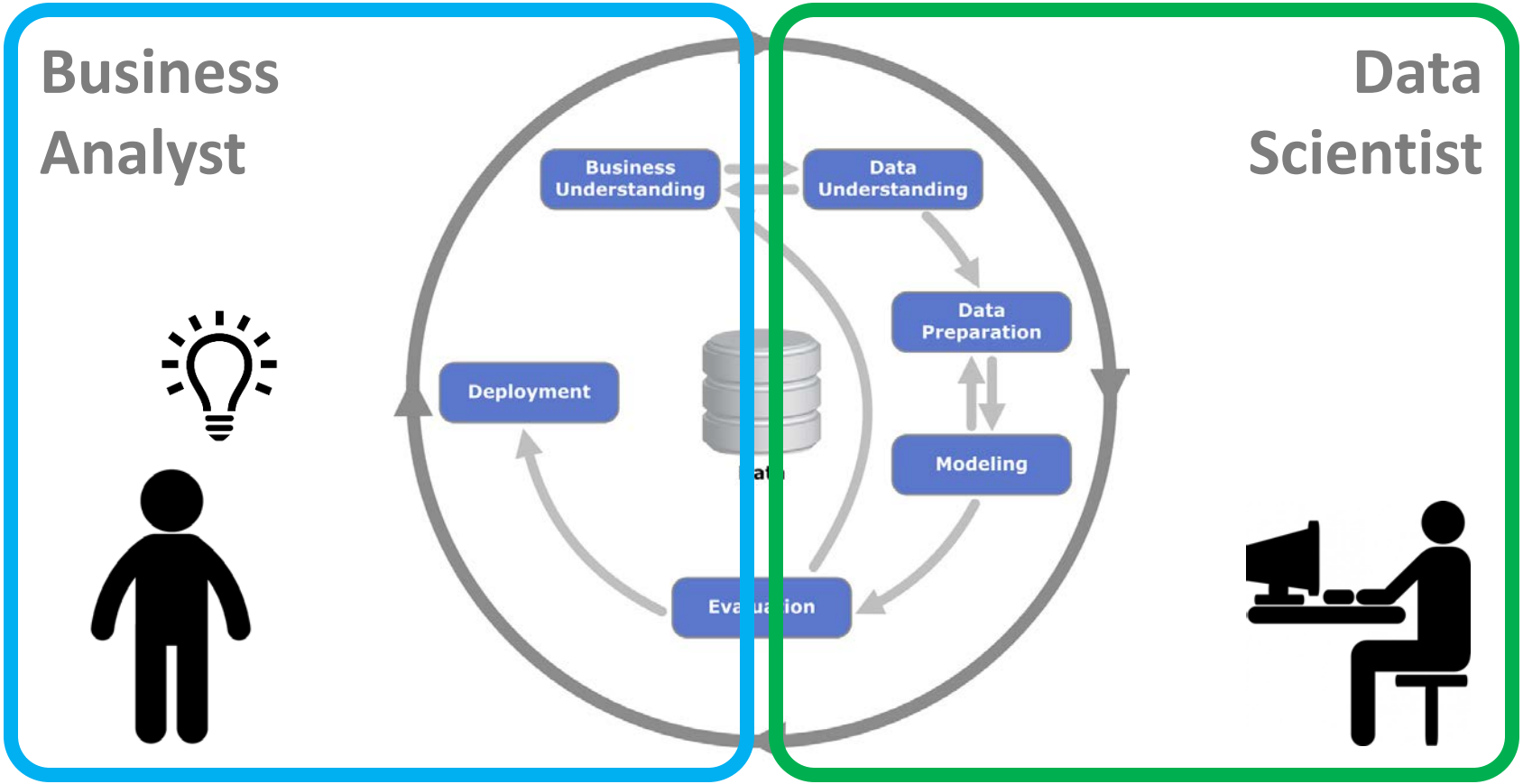


Data
Scientist

© 2016 KNIME.com AG. All Rights Rese

Open for Innovation ®
KNIME

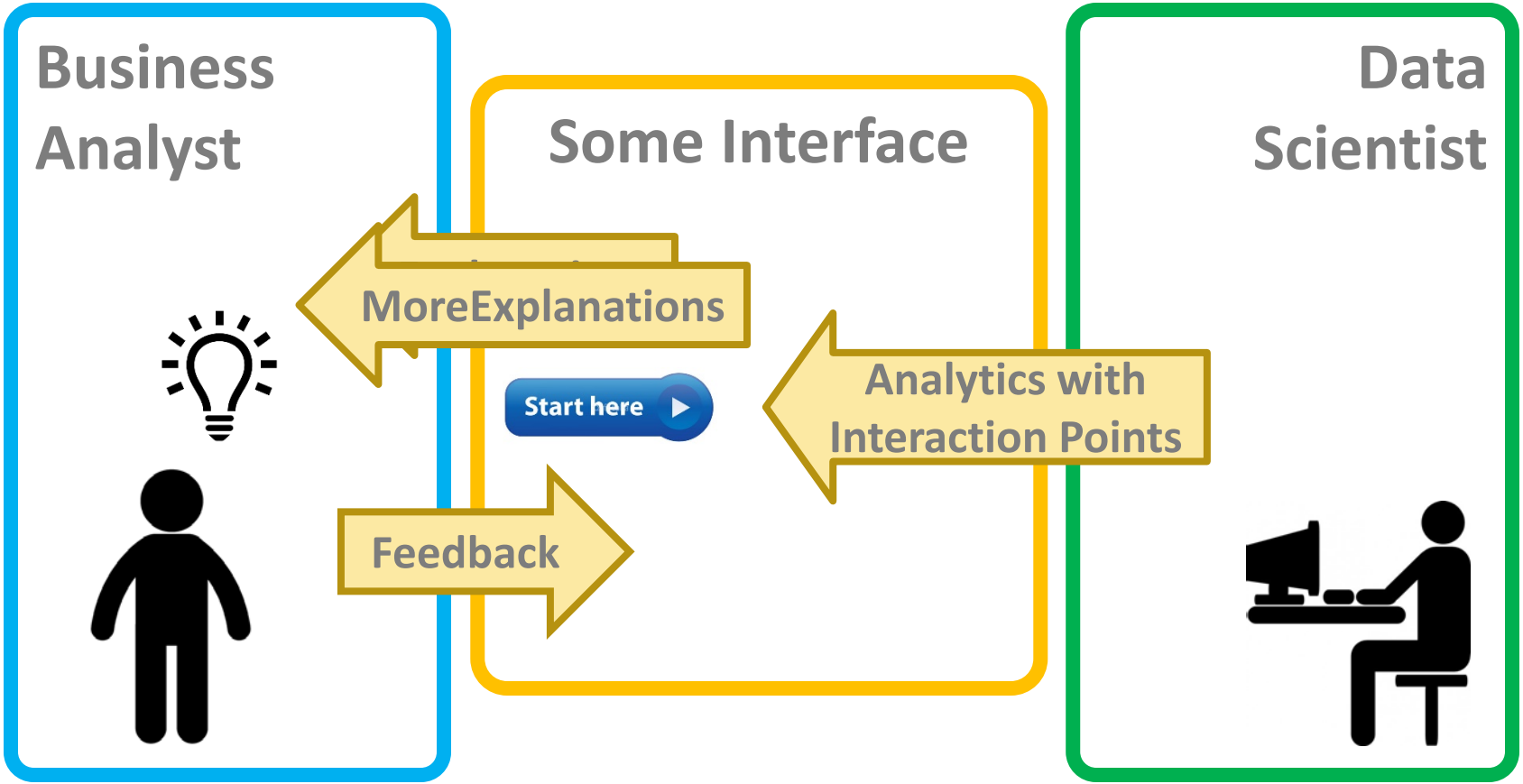
CRISP-DM in an Ideal World



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

CRISP-DM in the Real World?



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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

© 2016 KNIME.com AG. All Rights Reserved.

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 $\{(x_i, y_i)\}_{i=1}^{n}$, 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*:
    $$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$
    2. Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
    3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:
    $$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$
    4. Update the model:
    $$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$
3. Output $F_M(x)$.

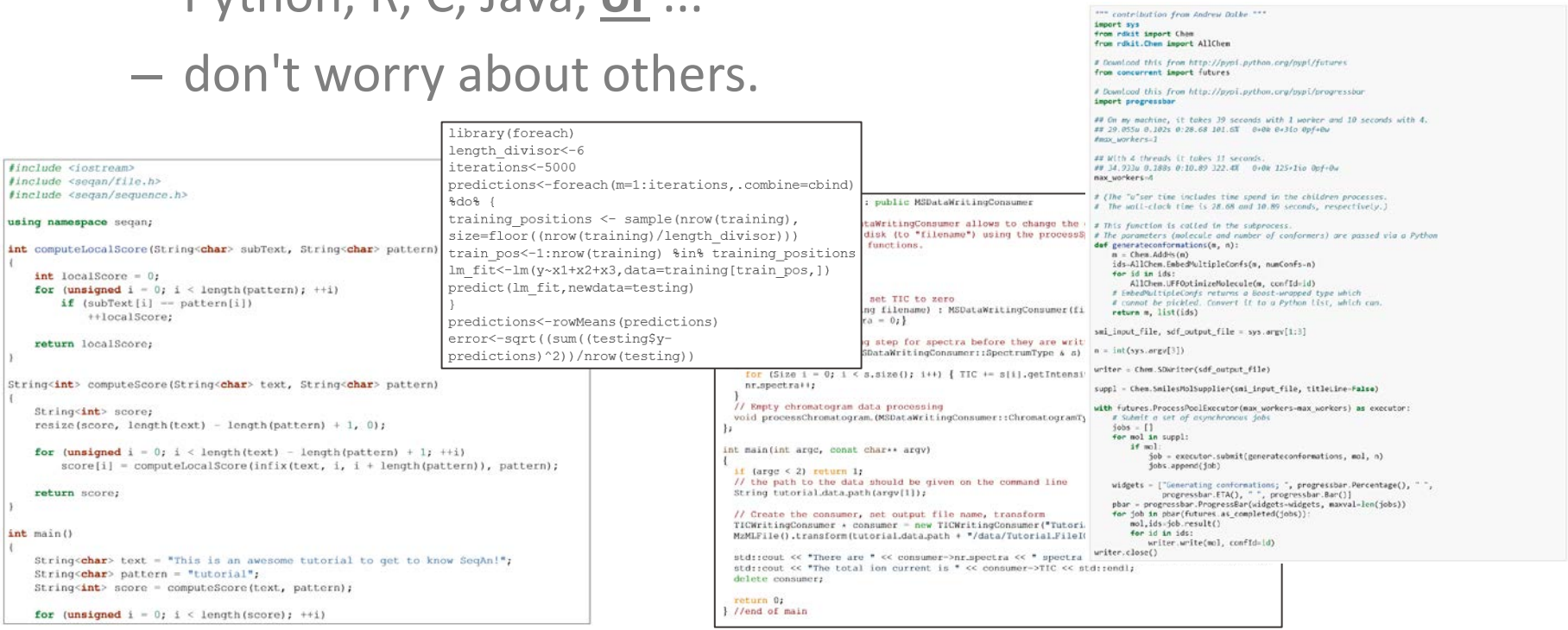
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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

taWritingConsumer allows to change the
disk (to "filename") using the process
functions.

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

g step for spectra before they are writ
SDataWritingConsumer::SpectrumType & s)

  for (Size i = 0; i < s.size(); i++) { TIC += s[i].getIntensi
  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
  MzMLFile().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.
## 29.055u 0.102s 0:28.68 101.6%   0+0k 0+31io 0pf+0w
#max_workers=1

## With 4 threads it takes 11 seconds.
## 34.933u 0.188s 0:10.89 322.4%   0+0k 125+1io 0pf+0w
max_workers=4

# (The "u"ser 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.UFFOptimizeMolecule(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)

smi_input_file, sdf_output_file = sys.argv[1:3]

n = int(sys.argv[3])

writer = Chem.SDWriter(sdf_output_file)

suppl = Chem.SmilesMolSupplier(smi_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()
```

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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*:

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

   2. Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. train it usin
   3. Compute multiplier γ_m by solving the following one-dimensior

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

   4. Update the model:

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

3. Output $F_M(x)$.

Methods & Algorithms (and even libraries)

≠

Tools

In [3]:
```
model1 <- gbm(medv ~ ., d
              distribution
              n.trees = 500
              interaction.d
              n.minobsinnod
              shrinkage = 0.001)
```

In [4]:
```
# summarize the model
options(repr.plot.width = 4, repr.plot.height = 4)
summary(model1)
```

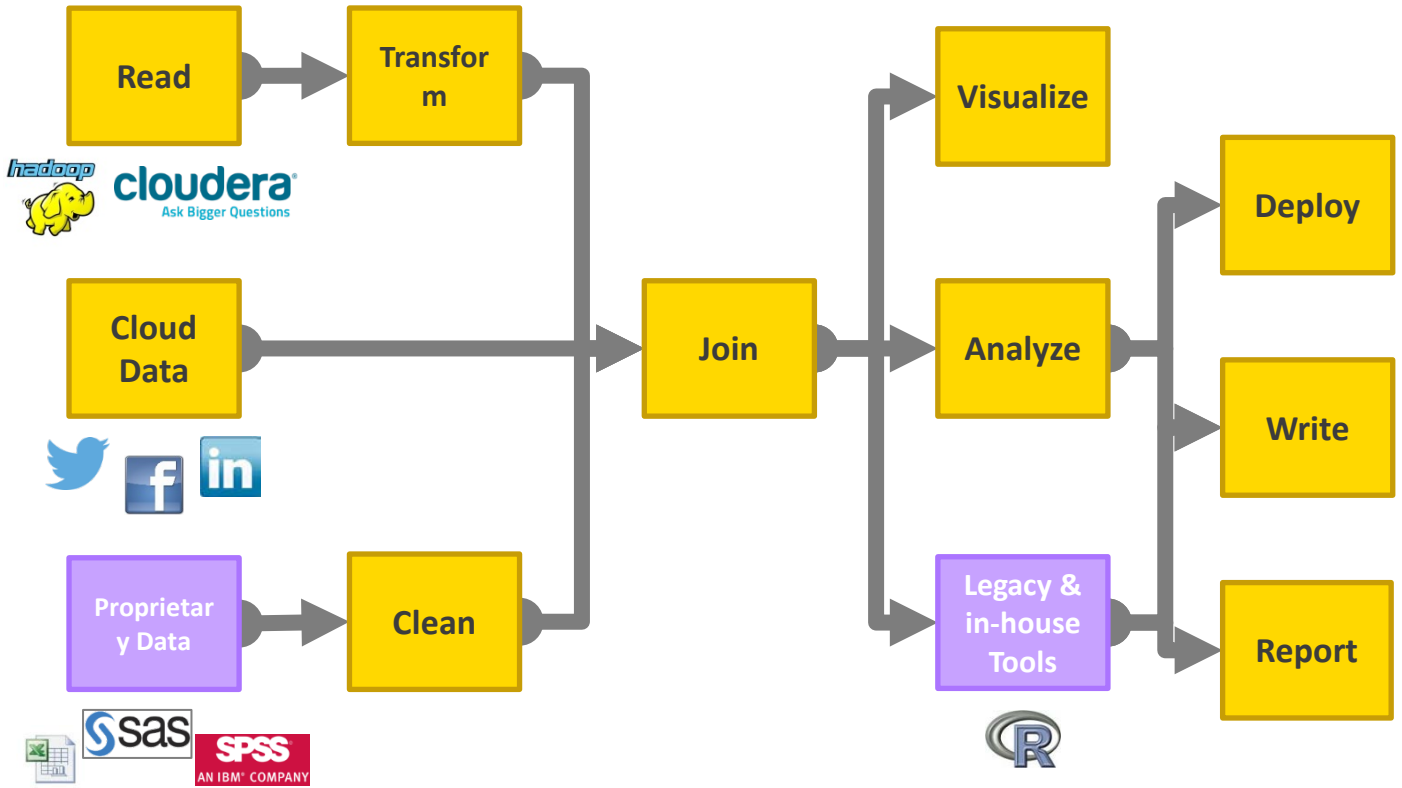
Out[4]:

	var	rel.inf
lstat	lstat	42.40815

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

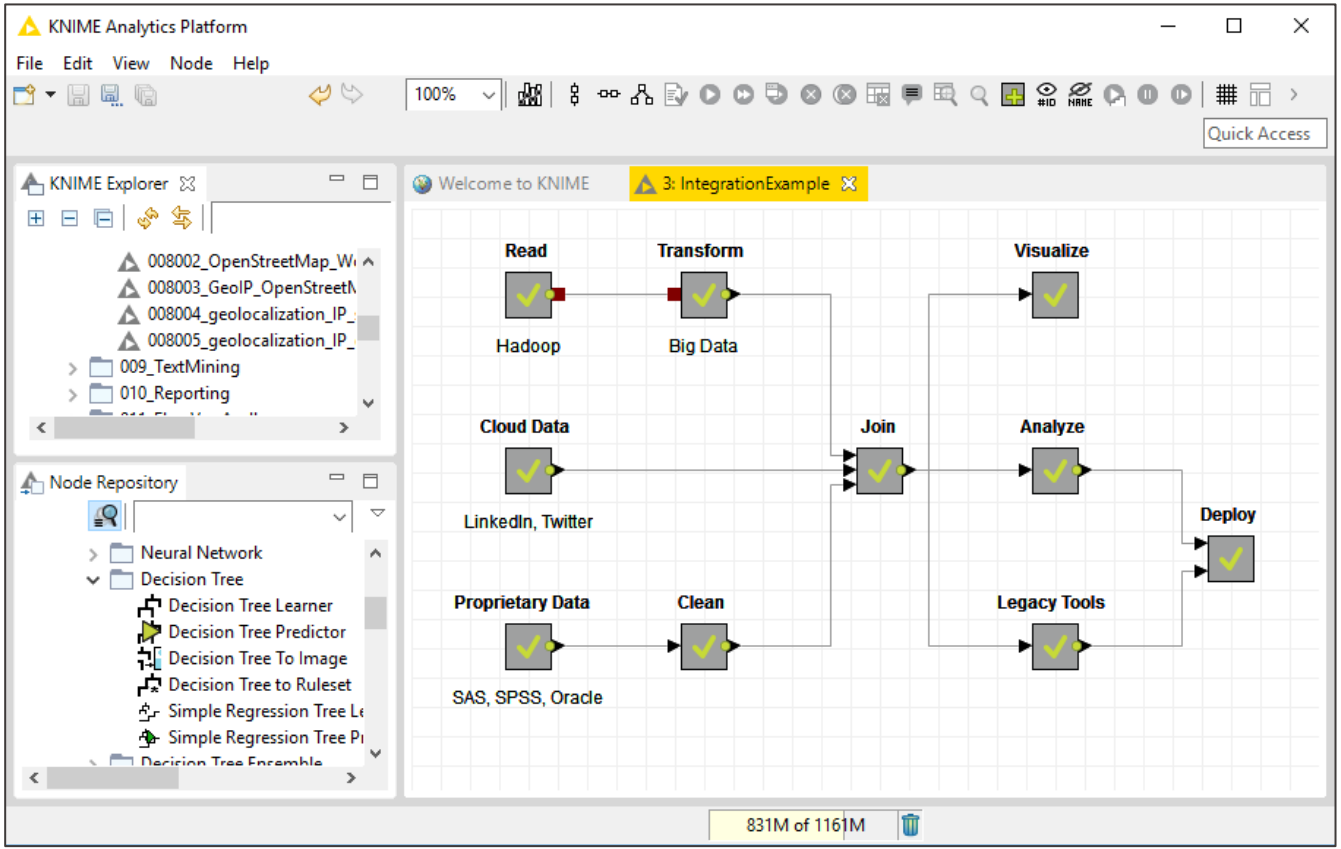
Visual Data Wrangling



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Visual Data Wrangling

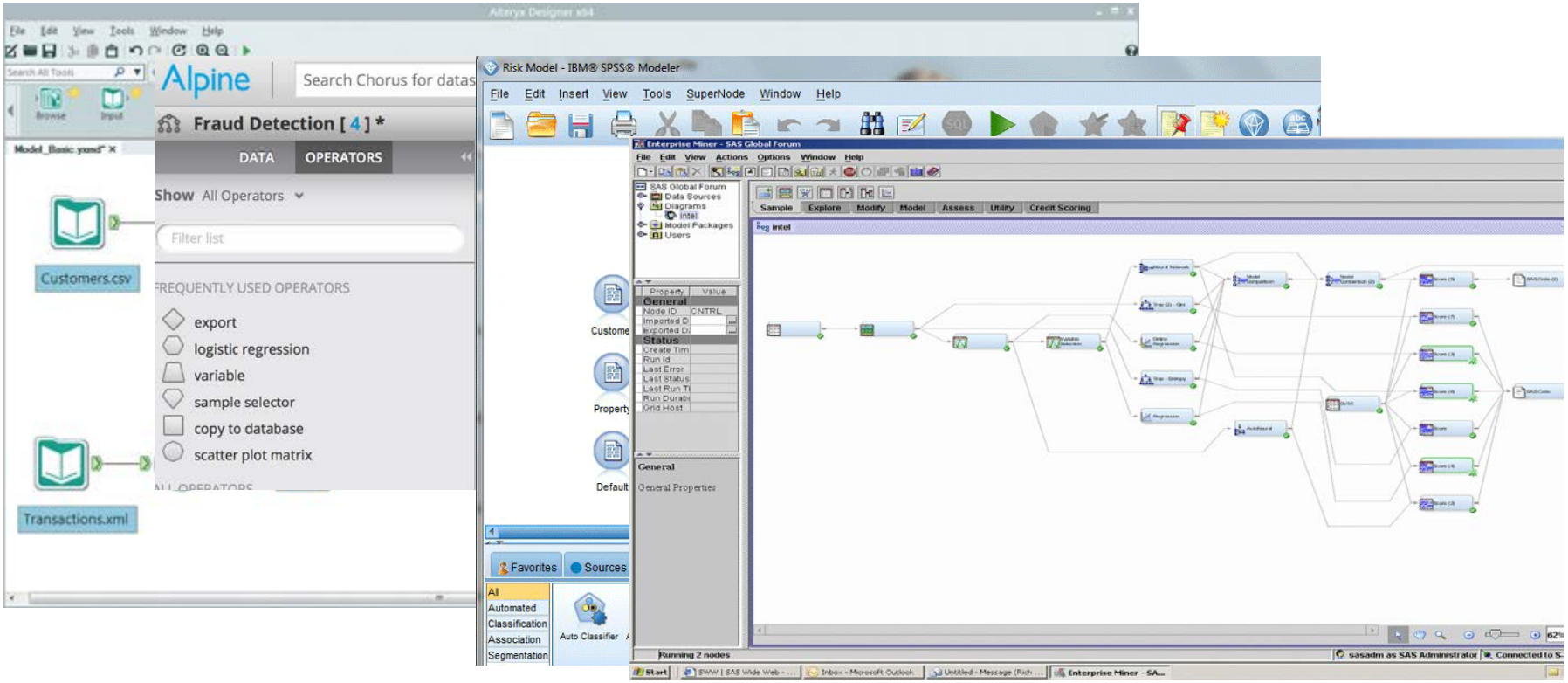


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Visual Workflows

KNIME is not the only one...



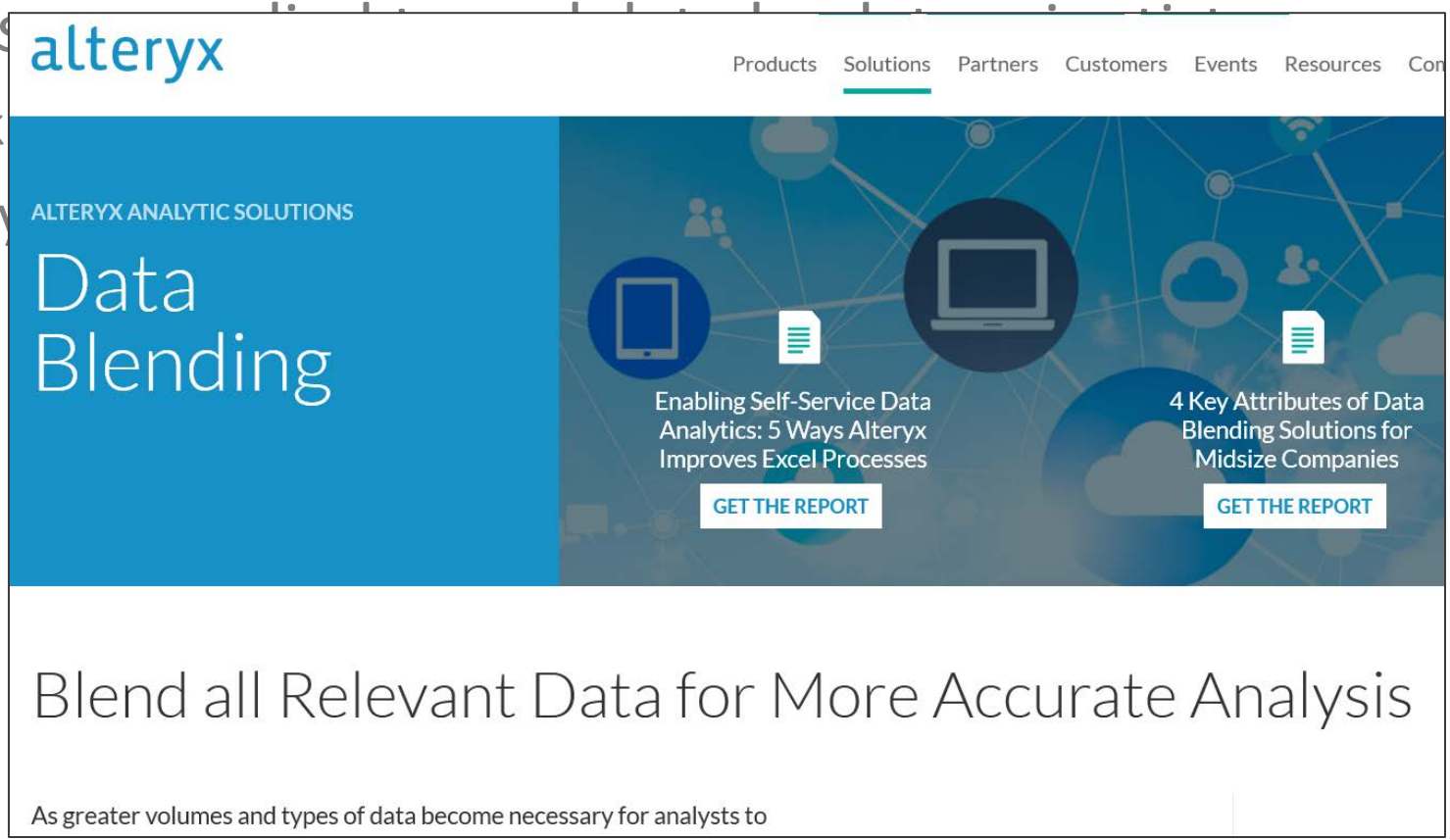
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Data Scientists

Algorithms

- may k
- mostly



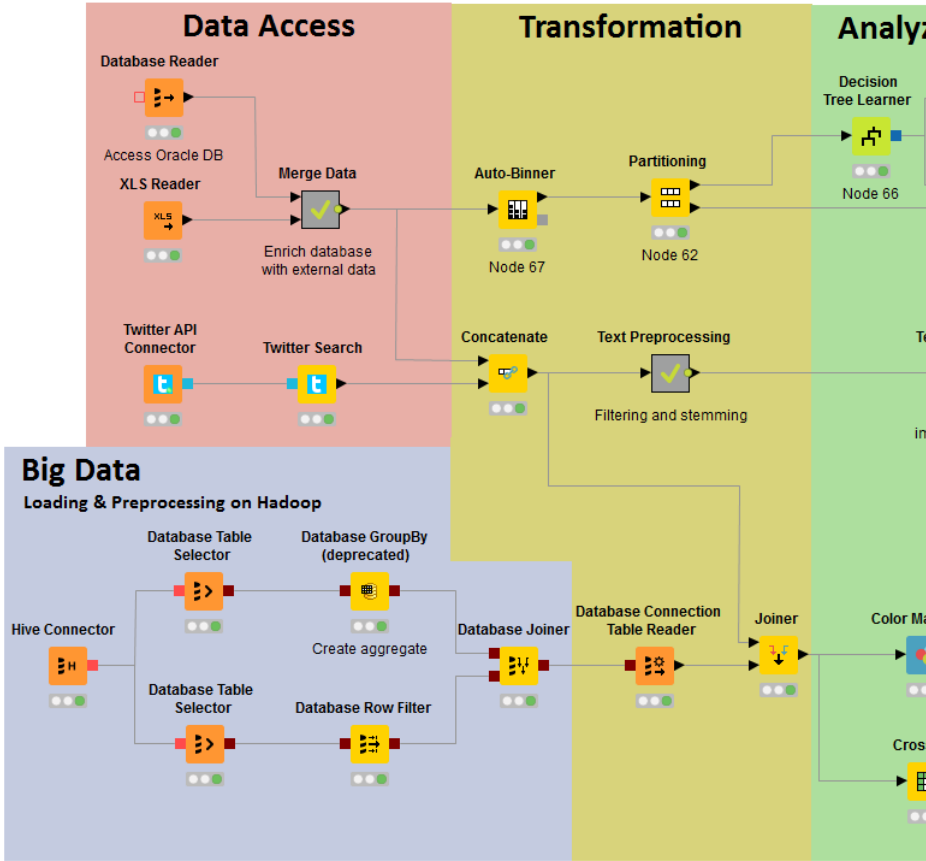
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Visual Data Blending

Data Blending:

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

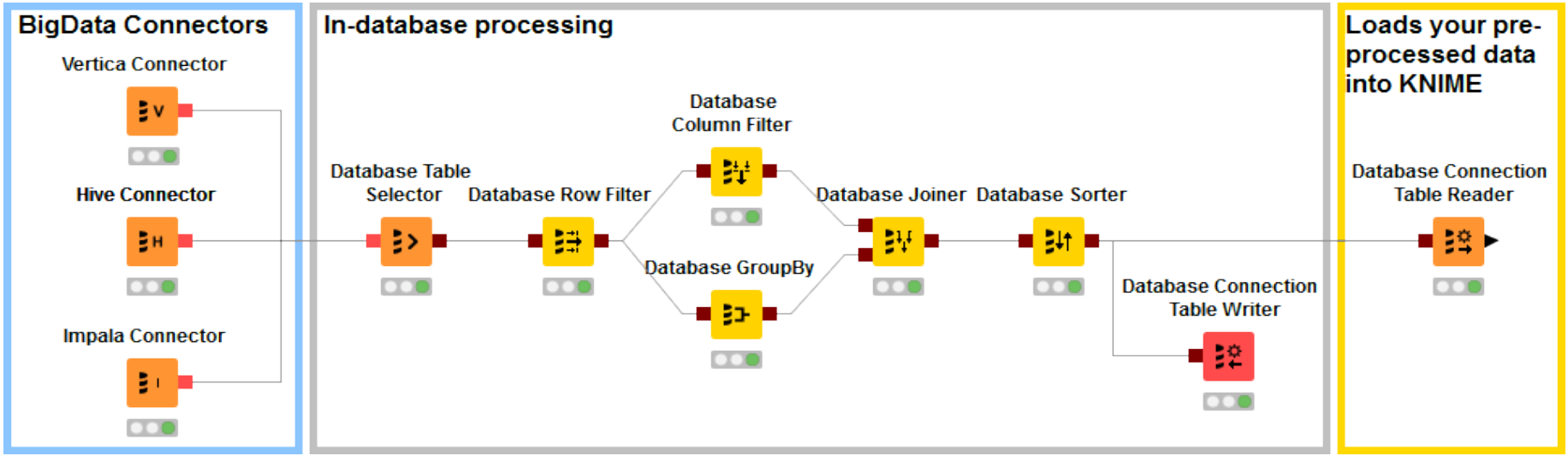


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Transforming Data

Visual SQL...

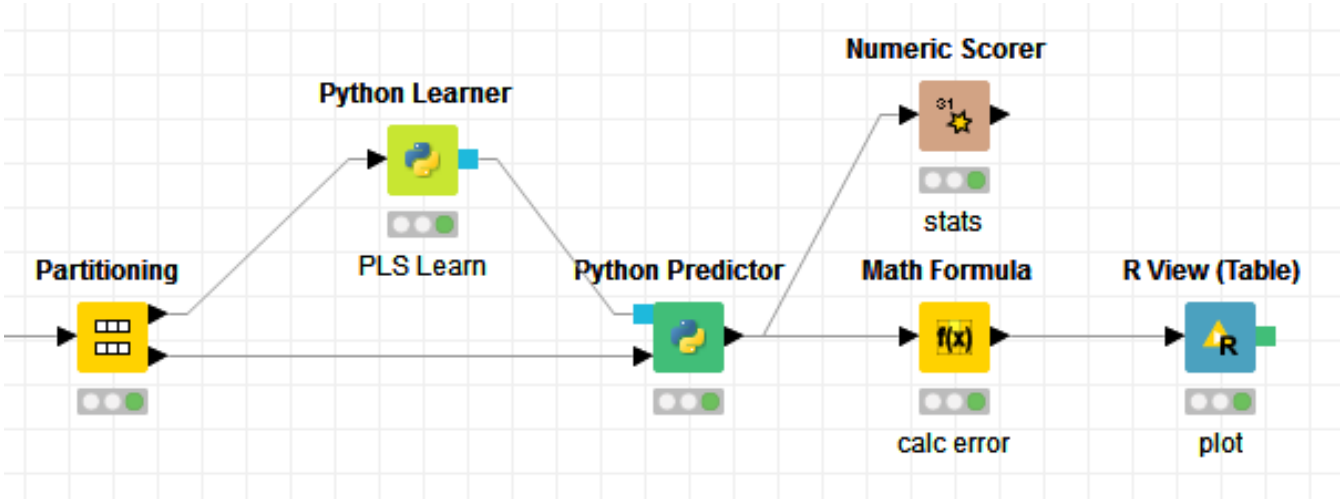


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Reusability = Collaboration

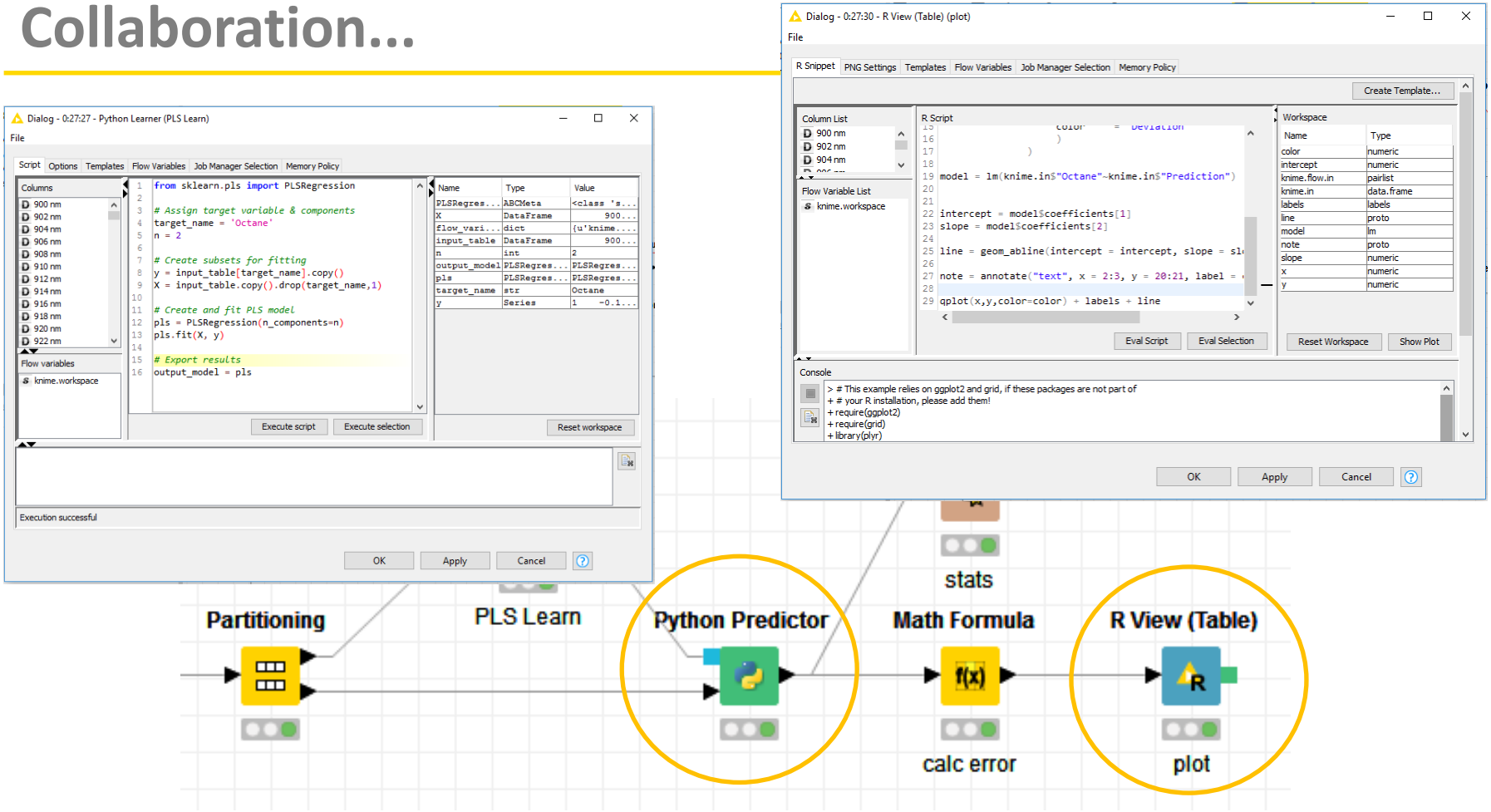
Combining and Reusing Expertise without having to care about the (to them: unimportant) details...



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Collaboration...

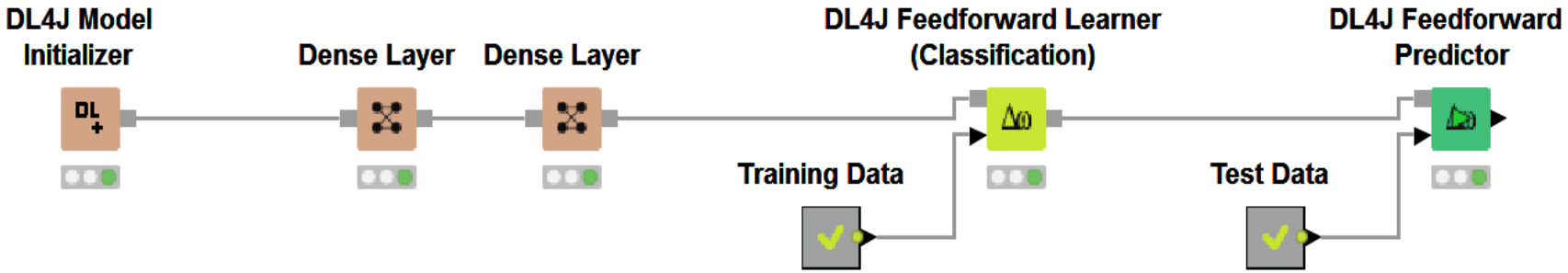


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

...Integration...

...getting quick access to new developments:

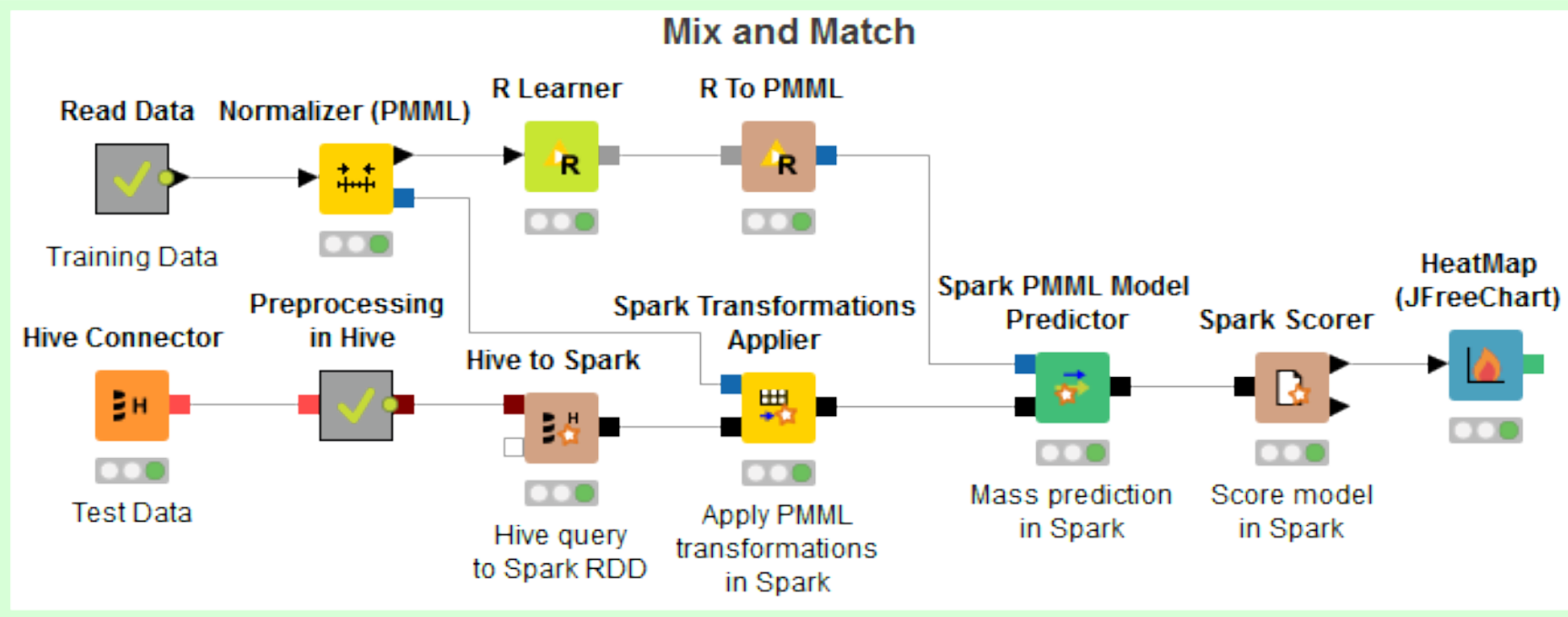


© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

...Mixing & Matching.

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



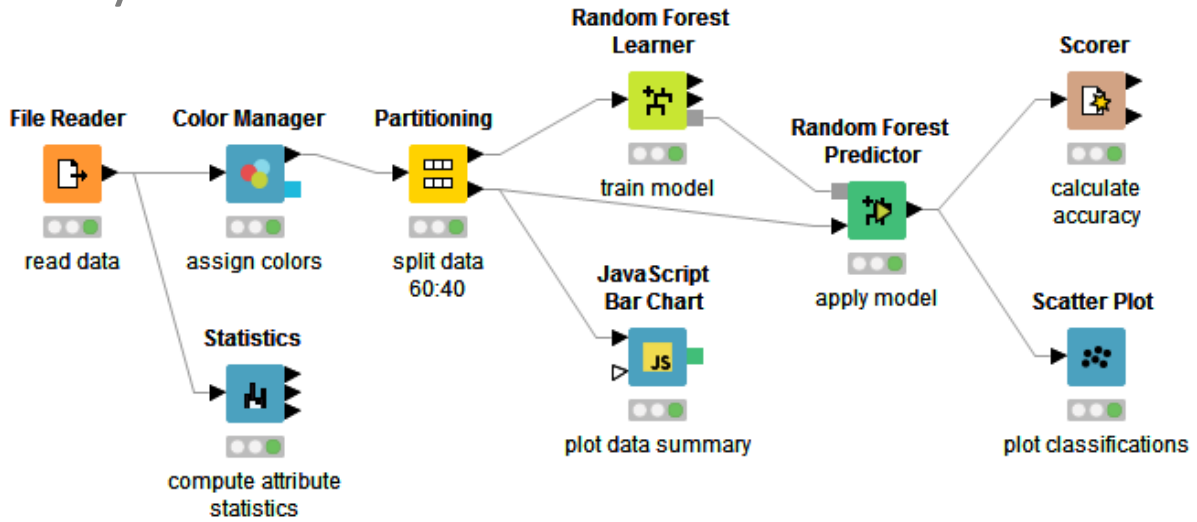
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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...



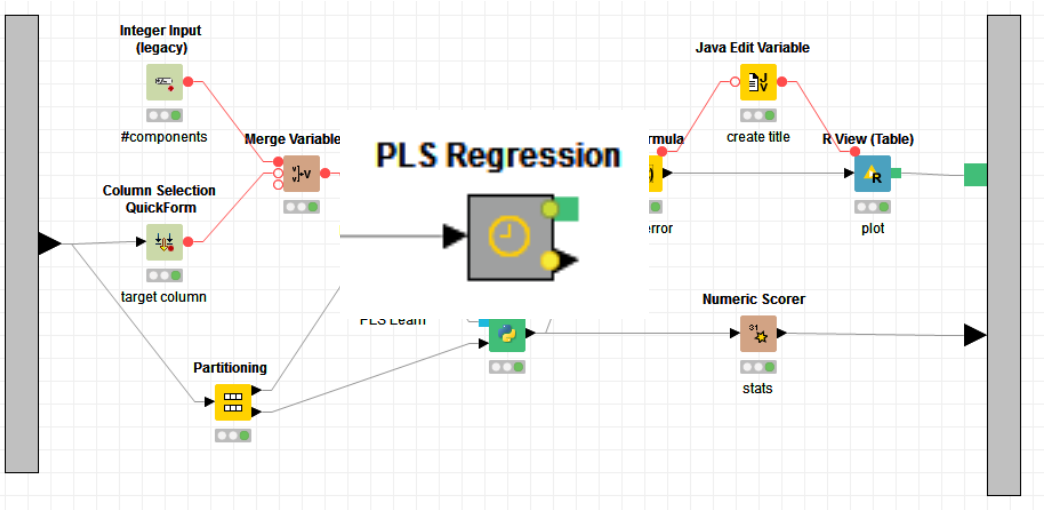
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Citizen Data Scientists

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

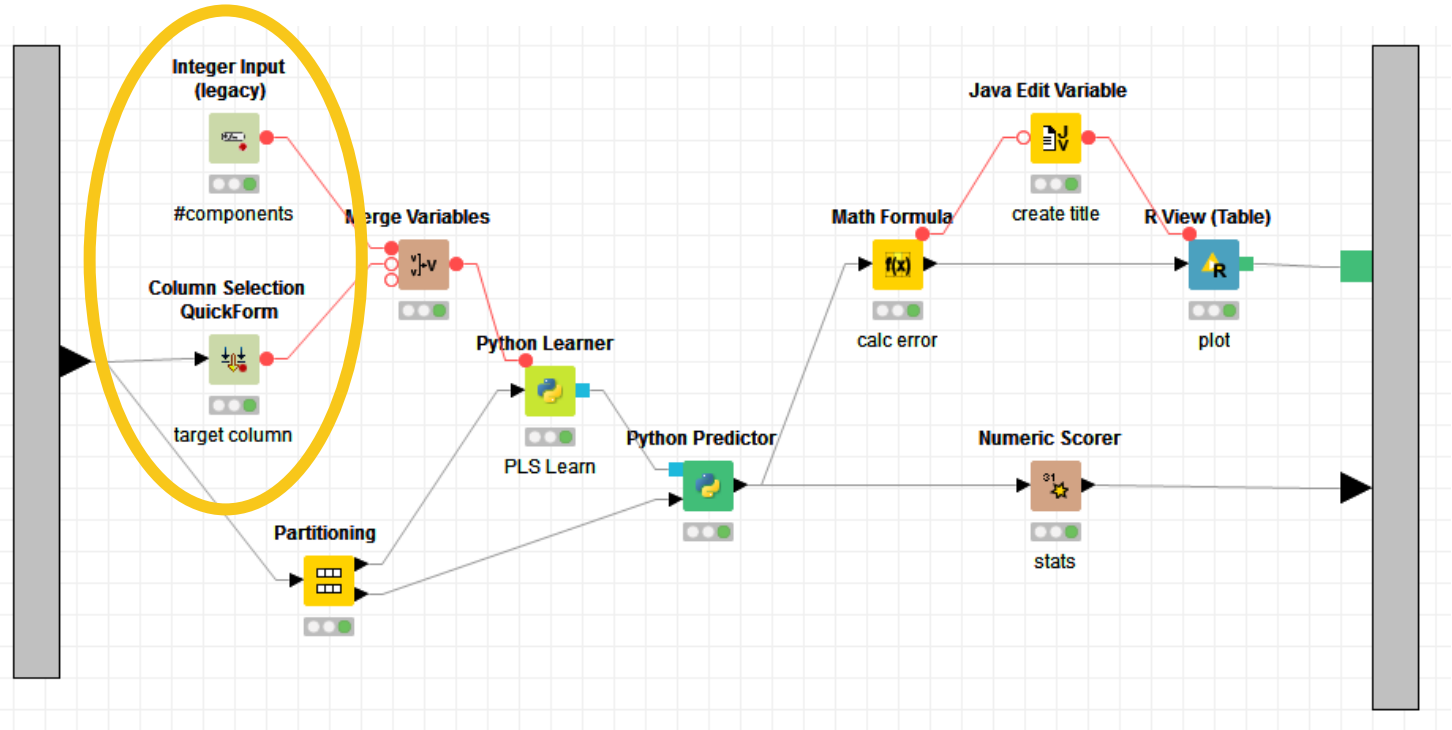
- maybe knows Java, Python, R, C, SQL or ...
- struggles with backwards compatibility, reproducibility, reusability...



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

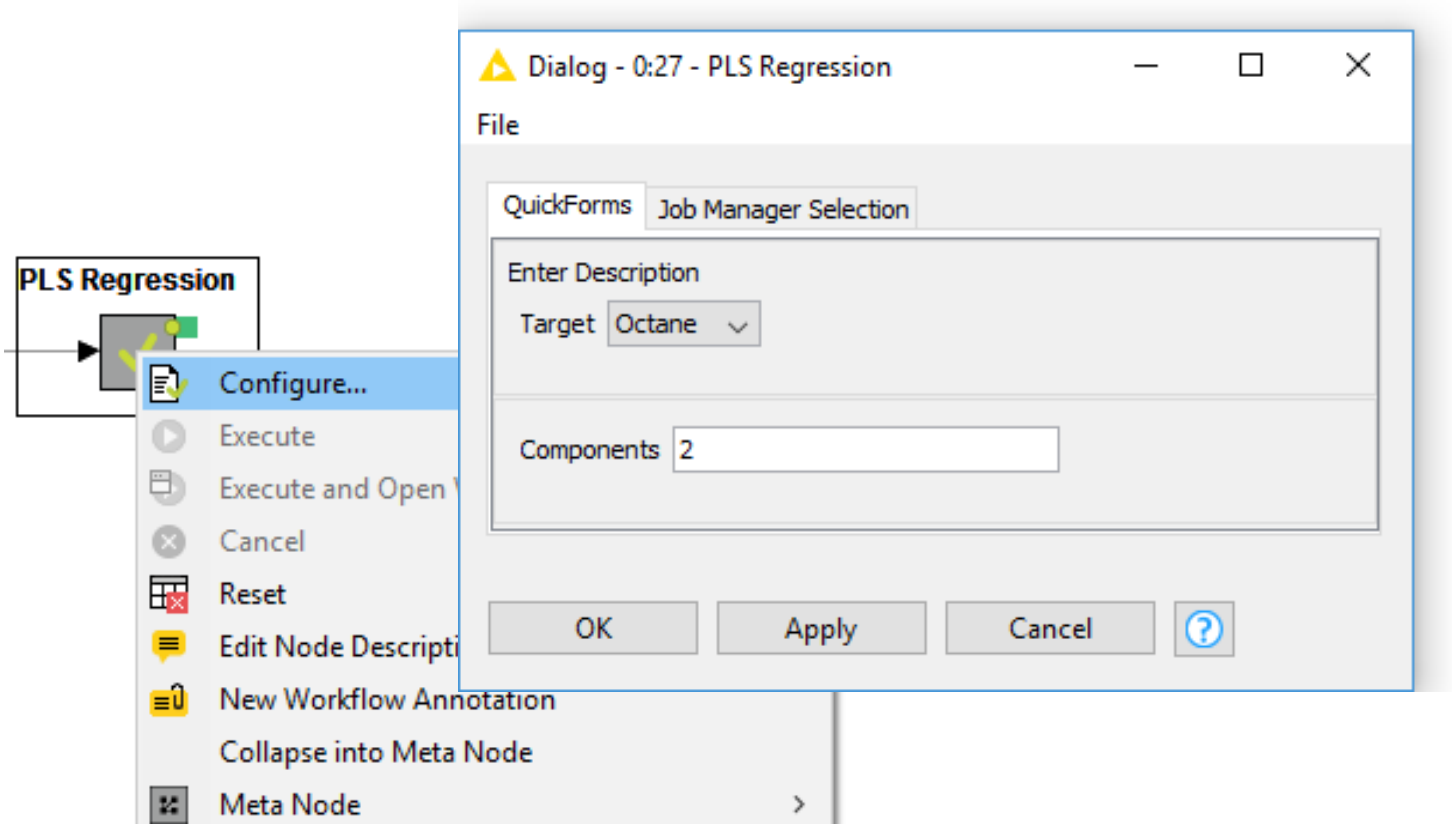
KNIME Metanodes: Controlling Parameters



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

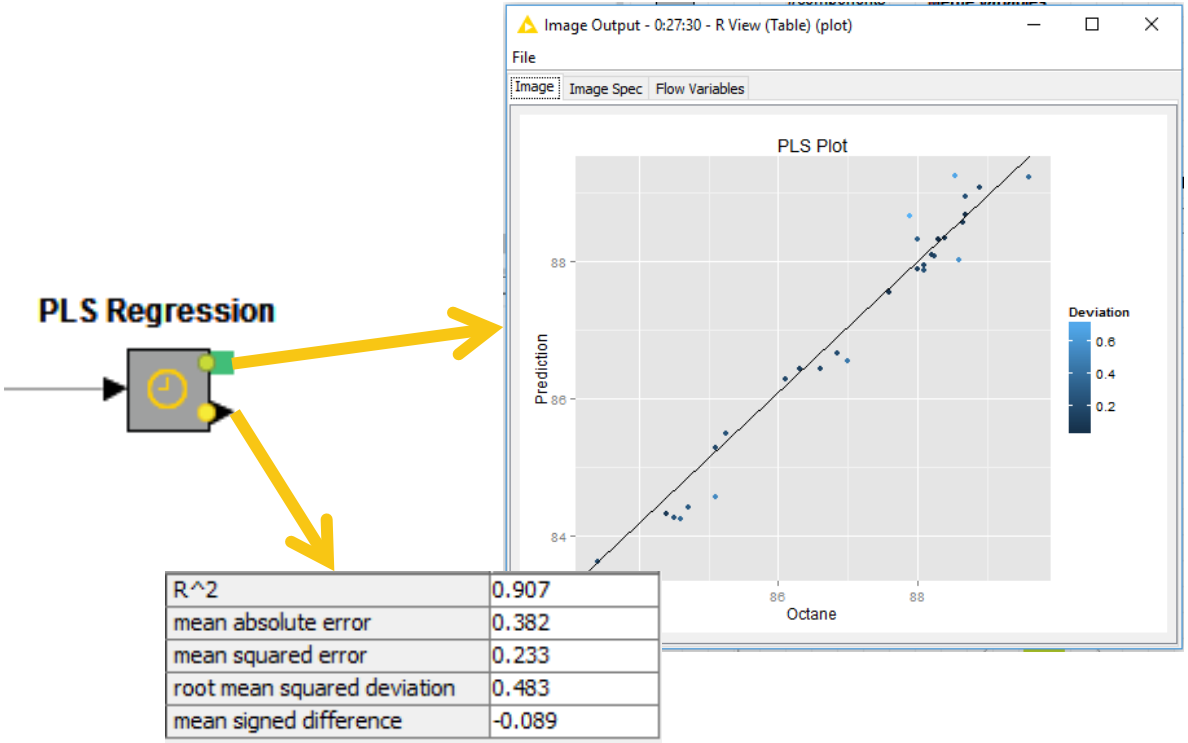
KNIME Metanodes: Exposing select parameters



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

KNIME Metanodes: Encapsulating Data Science

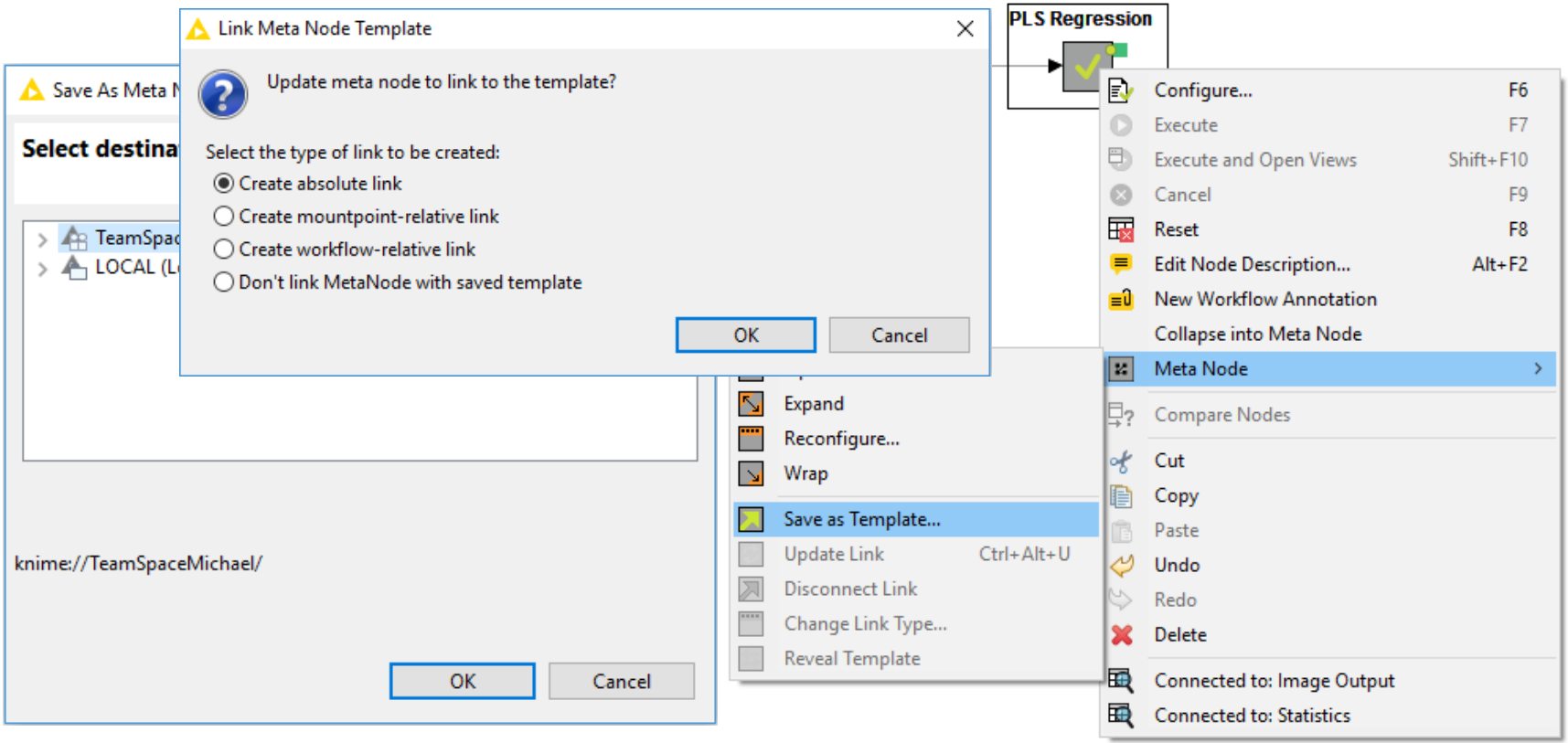


R^2	0.907
mean absolute error	0.382
mean squared error	0.233
root mean squared deviation	0.483
mean signed difference	-0.089

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Sharing Metanodes



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Workflow Templates

Visual Workflows...

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

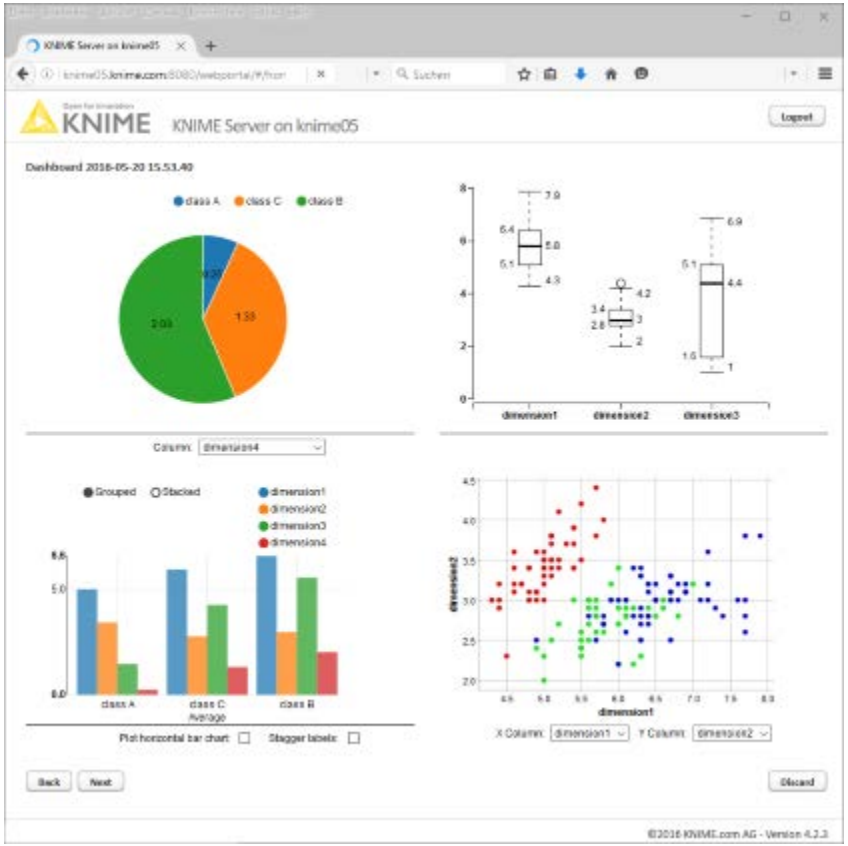
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Business Analyst

Ultimately the business analysts want to just use it:

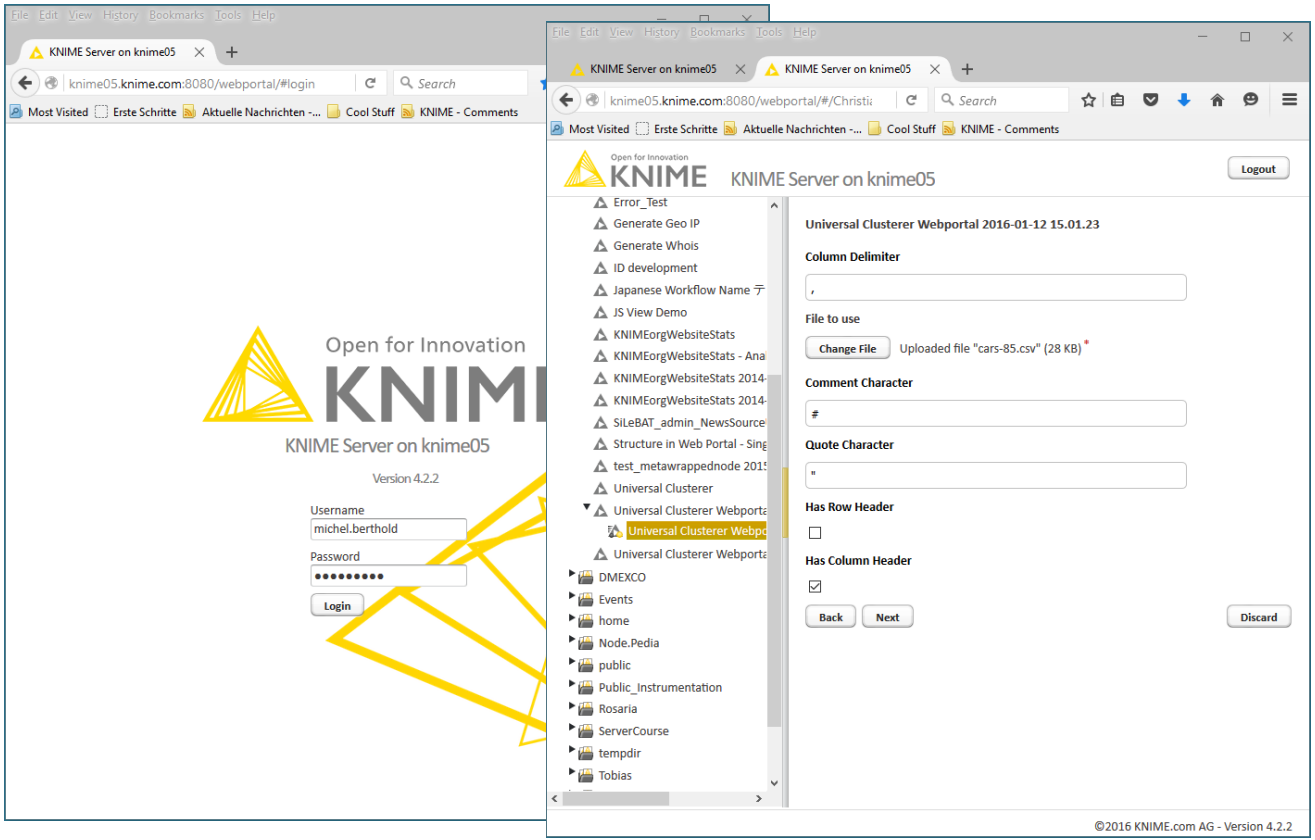
- they have domain expertise...
- ...but don't know (and should not know) C, R, Python, Java, or ...



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

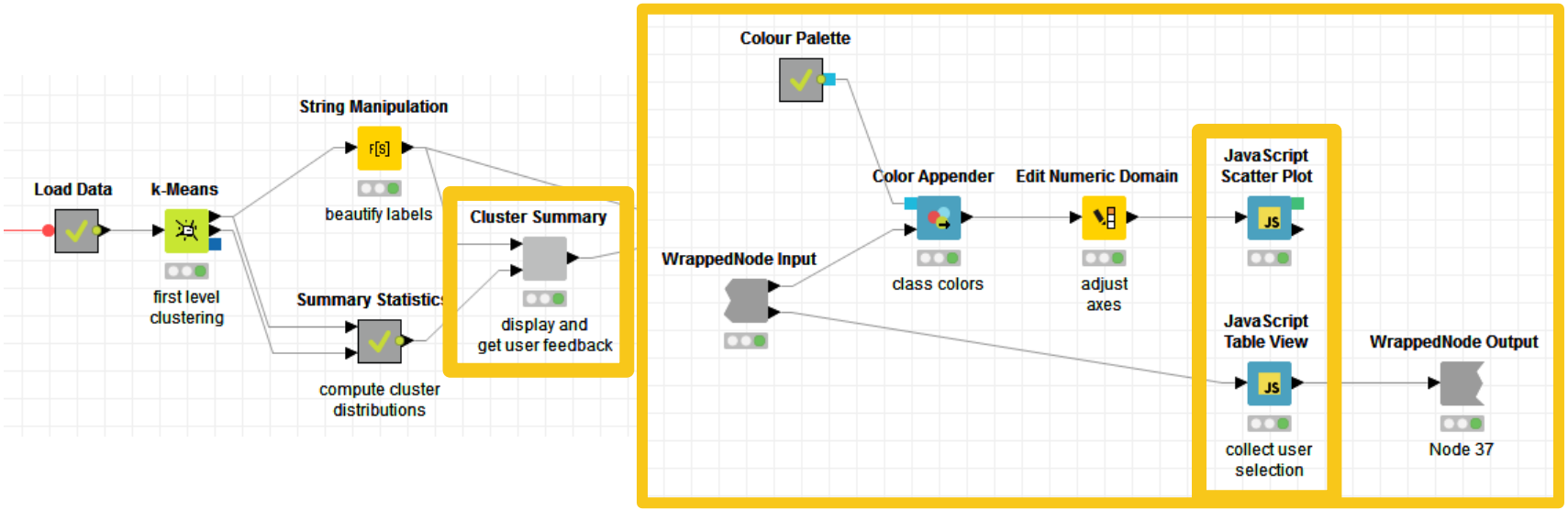
Guided Analytics



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

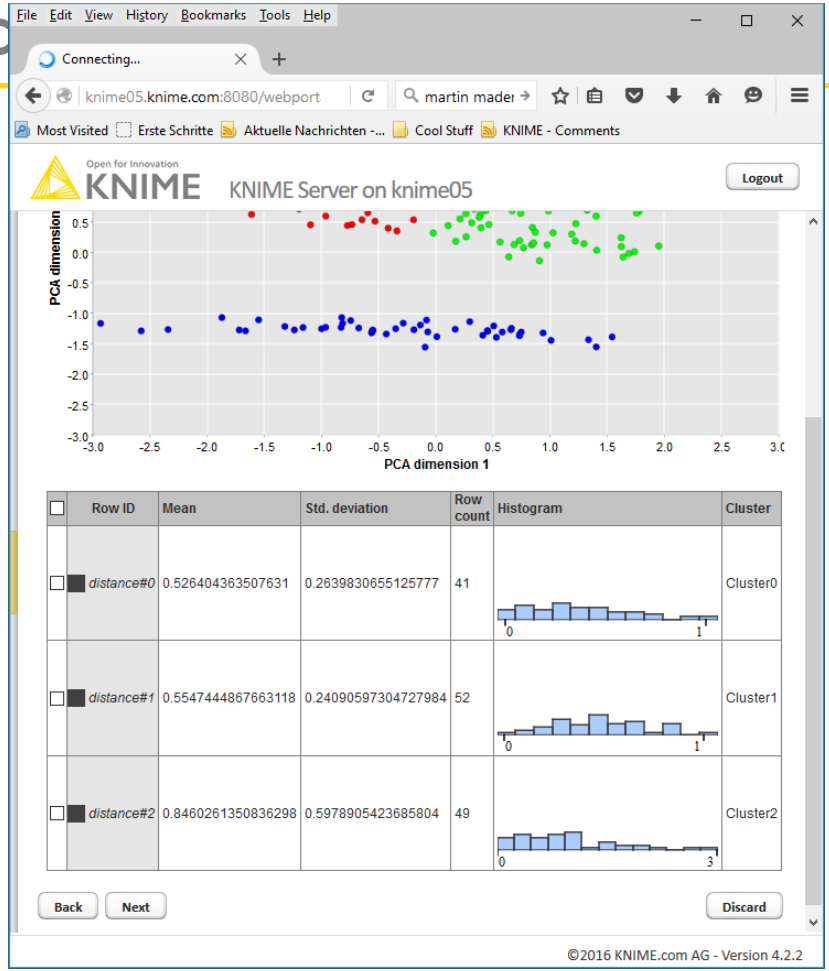
Guided Workflow Execution



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

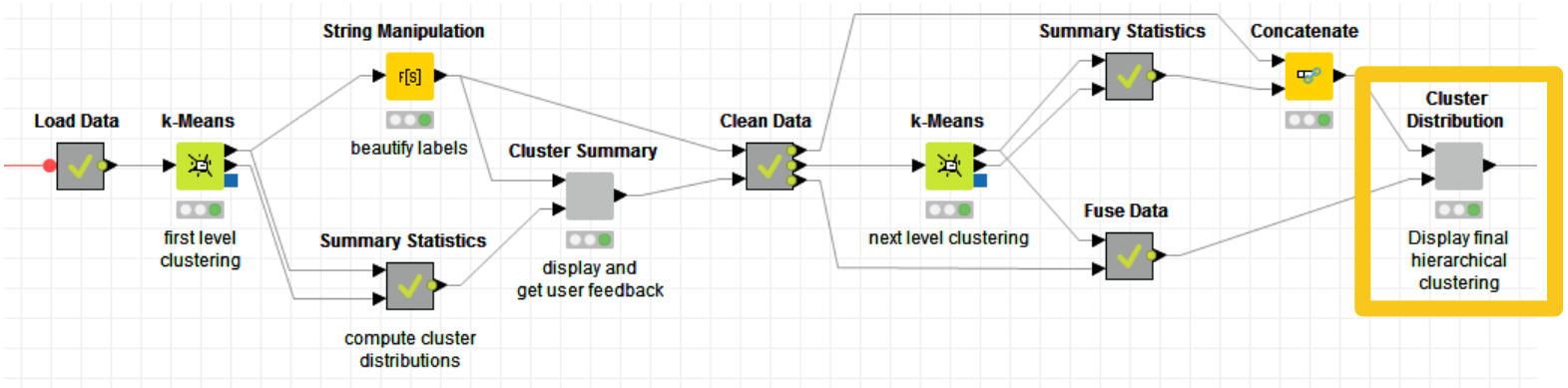
Guided Workflow Exec



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

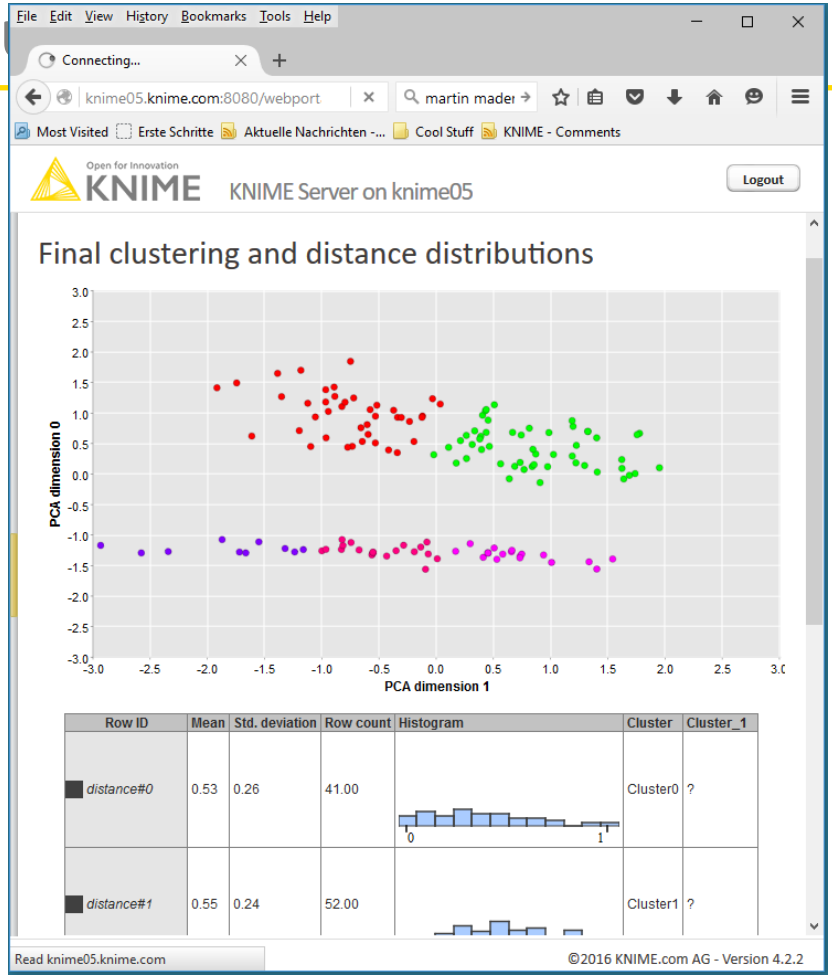
Guided Workflow Execution



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Guided Workflow Exec



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

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.

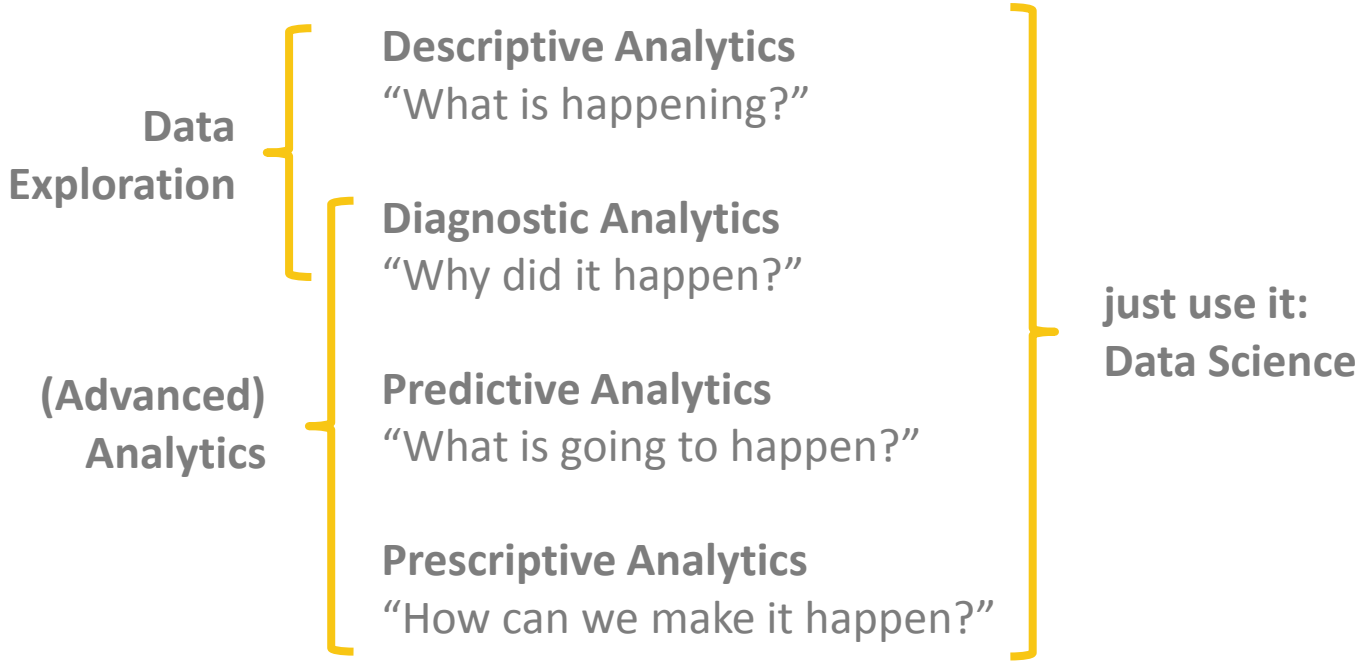
© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

Take Aways

© 2016 KNIME.com AG. All Rights Reserved.

1: Data Science and Flavors of Analytics



© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ®
KNIME

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.

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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!)

© 2016 KNIME.com AG. All Rights Reserved.

Open for Innovation ® KNIME

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.

© 2016 KNIME.com AG. All Rights Reserved.