

Predictive Analytics for Big Data with Native Spark Modeling

Priti Mulchandani, Andreas Forster
September 2016



Trends in Data Science

0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1

Massive Amount of Data

Conversations

Transactions Machines

Analytical Skill Gap

“Demand for deep analytical talent in the US could be 50 to 60% greater than its projected supply by 2018”

McKinsey Global Institute

Ever Faster Decision Cycle

Current Duration	Future Duration	Process
30 milliseconds	5 seconds	Algorithmic trade
20 minutes	30 seconds	Airline operations
8 hours	10 seconds	Call center inquiries
1 day	1 minute	Recompute fin. position
1 day	10 minutes	Supply chain updates
3 days	45 seconds	Document transfer
3 days	2 minutes	Phone activation
1 week	1 hour	Refresh data warehouse
5 days	1 day	Trade settlement
3 weeks	1 day	Build-to-order PC

10⁷ 10⁶ 10⁵ 10⁴ 1,000 100 10 1 0 Seconds

Gartner

So how does Automated Analytics help?

You are a Data Scientist

- Automate the recurring tasks and **save time**
- **Get inspiration** on which direction to investigate manually
- Help **structure your data** sets for manual approach
- **Deploy** models into production with ease
- Have **additional functionality** in your portfolio to tackle day to day challenges

Support Productivity

You are an Analyst

- **Get access** to the world of Predictive Analytics / Machine Learning
- **Deliver new benefits** by providing Predictive Models in addition to Business Intelligence
- Build on **existing analytical skillset**
- Find a **new career path**

Enable Users

You are a Company

- **Benefit** from Predictive Insight where needed in **business processes**
- **Scale the use of predictive models** without manual bottlenecks
- **Accelerate** your path to a **digital business**

Scale

But I am a Data Scientist, and I am efficient «by hand»

A logistic regression only takes a few lines of code in MLlib.

```
import org.apache.spark.mllib.classification.{SVMModel, SVMWithSGD}
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics
import org.apache.spark.mllib.util.MLUtils

// Load training data in LIBSVM format.
val data = MLUtils.loadLibSVMFile(sc, "data/mllib/sample_libsvm_data.txt")

// Split data into training (60%) and test (40%).
val splits = data.randomSplit(Array(0.6, 0.4), seed = 11L)
val training = splits(0).cache()
val test = splits(1)

// Run training algorithm to build the model
val numIterations = 100
val model = SVMWithSGD.train(training, numIterations)

// Clear the default threshold.
model.clearThreshold()

// Compute raw scores on the test set.
val scoreAndLabels = test.map { point =>
  val score = model.predict(point.features)
  (score, point.label)
}

// Get evaluation metrics.
val metrics = new BinaryClassificationMetrics(scoreAndLabels)
val auROC = metrics.areaUnderROC()

println("Area under ROC = " + auROC)

// Save and load model
model.save(sc, "target/tmp/scalaSVMWithSGDModel")
val sameModel = SVMModel.load(sc, "target/tmp/scalaSVMWithSGDModel")
```

} Split data

} Train one model

} Apply the model on new data

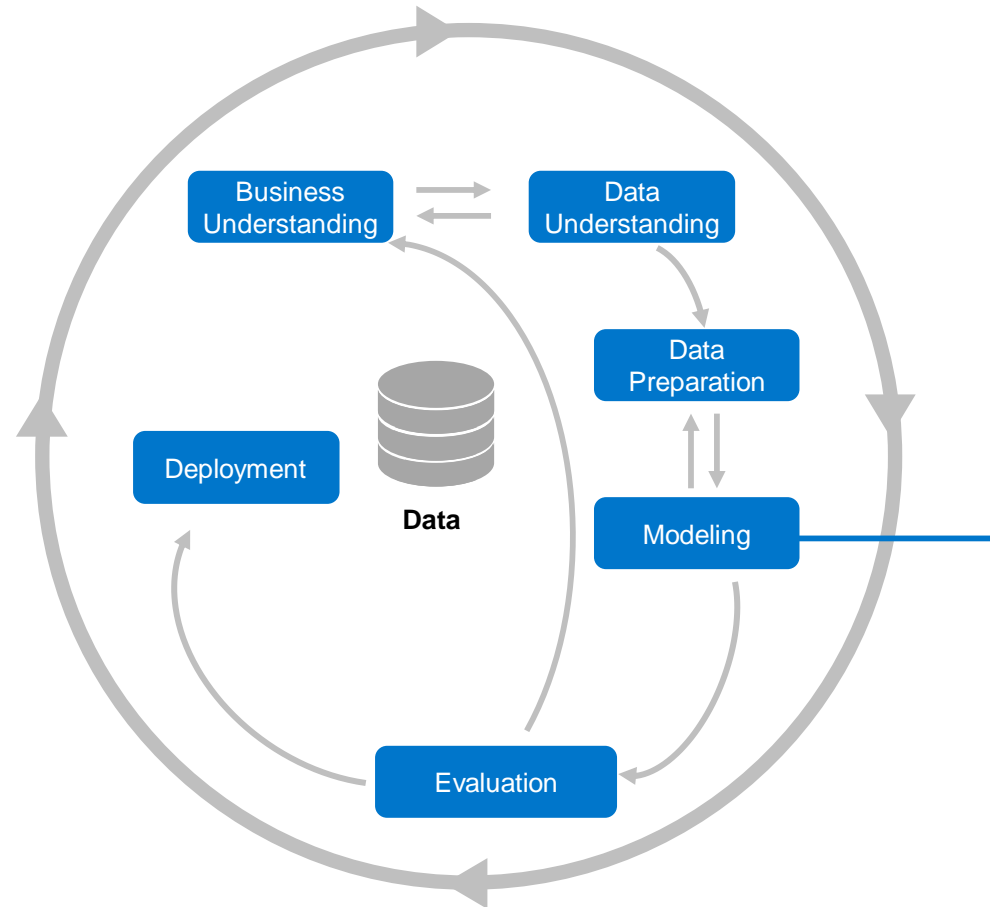
} Evaluate model quality

Source: <http://spark.apache.org/docs/latest/mllib-linear-methods.html>

However, most projects are more complex

The Cross Industry Standard Process for Data Mining (CRISP-DM)

The previous code only creates 1 model. The remaining aspects are not addressed yet.



```
import sys, os, random, numpy as np, pandas as pd, sklearn as skl, sklearn.metrics as skl_metrics, sklearn.cross_validation as skl_cv
import sys, os, random, numpy as np, pandas as pd, sklearn as skl, sklearn.metrics as skl_metrics, sklearn.cross_validation as skl_cv

# Load training data in a pandas format
url = "http://www.kddcup.org/data/2008/2008.t1000000.txt"
data = pd.read_csv(url)

# Split data into training (80%) and test (20%)
skl_cv = skl.cross_validation.KFold(n=len(data), k=10, seed=10)
skl_train, skl_test = skl_cv.split(data)

# Use training data to train the model
skl_model = skl.LinearSVC()
skl_model.fit(skl_train)

# Apply the model on the test set
skl_score = skl_model.score(skl_test)
print("Score: %f" % skl_score)

# Evaluate model quality
skl_metrics = skl.metrics.mean_squared_error(skl_model.predict(skl_test), skl_test['target'])
print("MSE: %f" % skl_metrics)
```

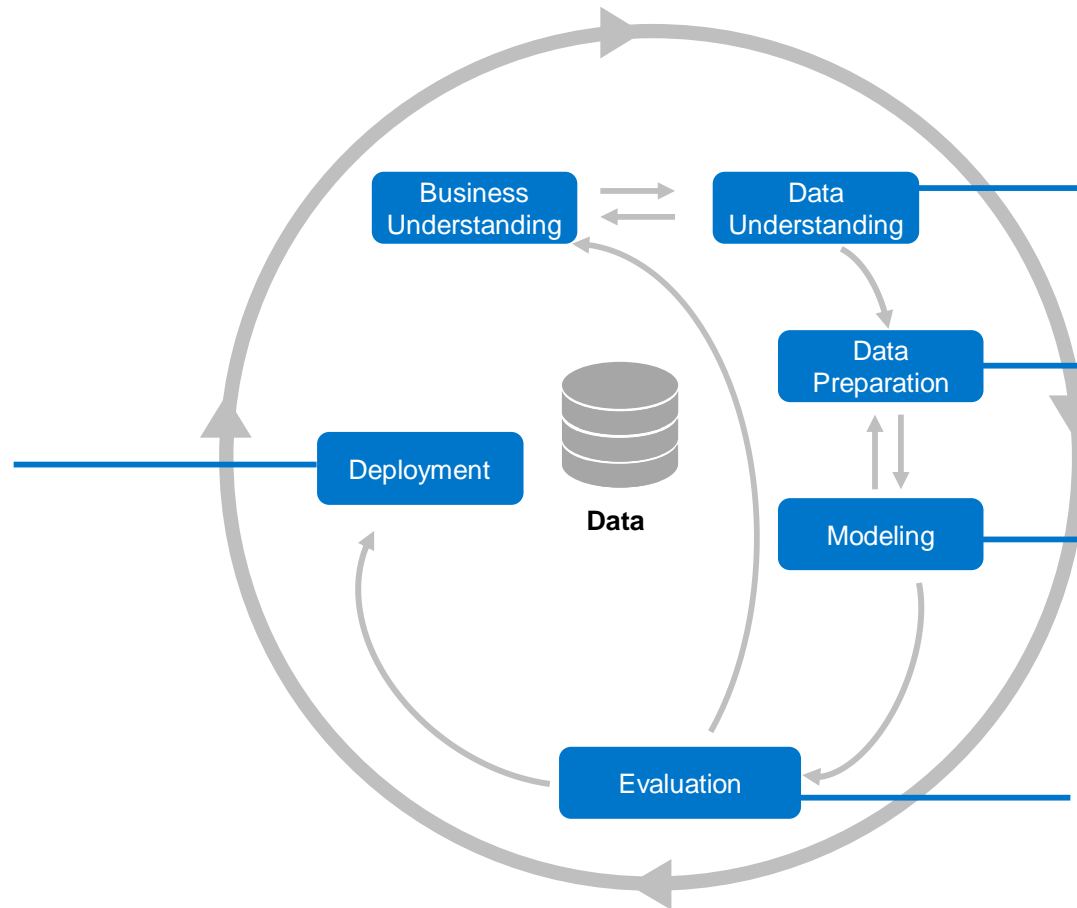
- } Split data
- } Train one model
- } Apply the model on new data
- } Evaluate model quality

Source: [https://en.wikipedia.org/wiki/Cross Industry Standard Process for Data Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

Automated Predictive Analytics

The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications



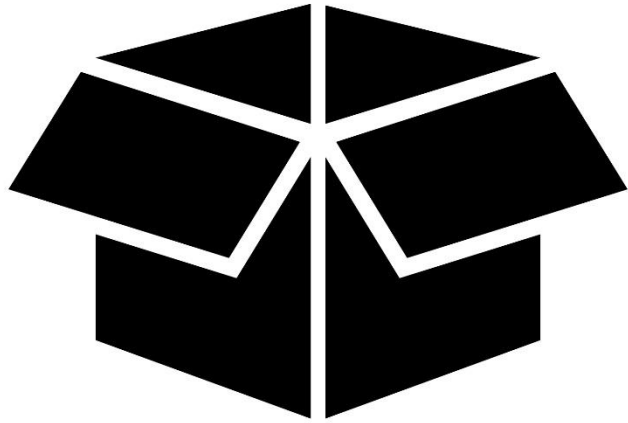
Explorative / Agile BI frontend

- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

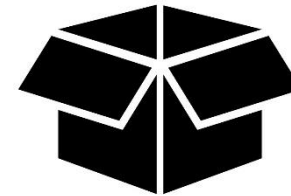
Automated Analytics



How?

The Principles

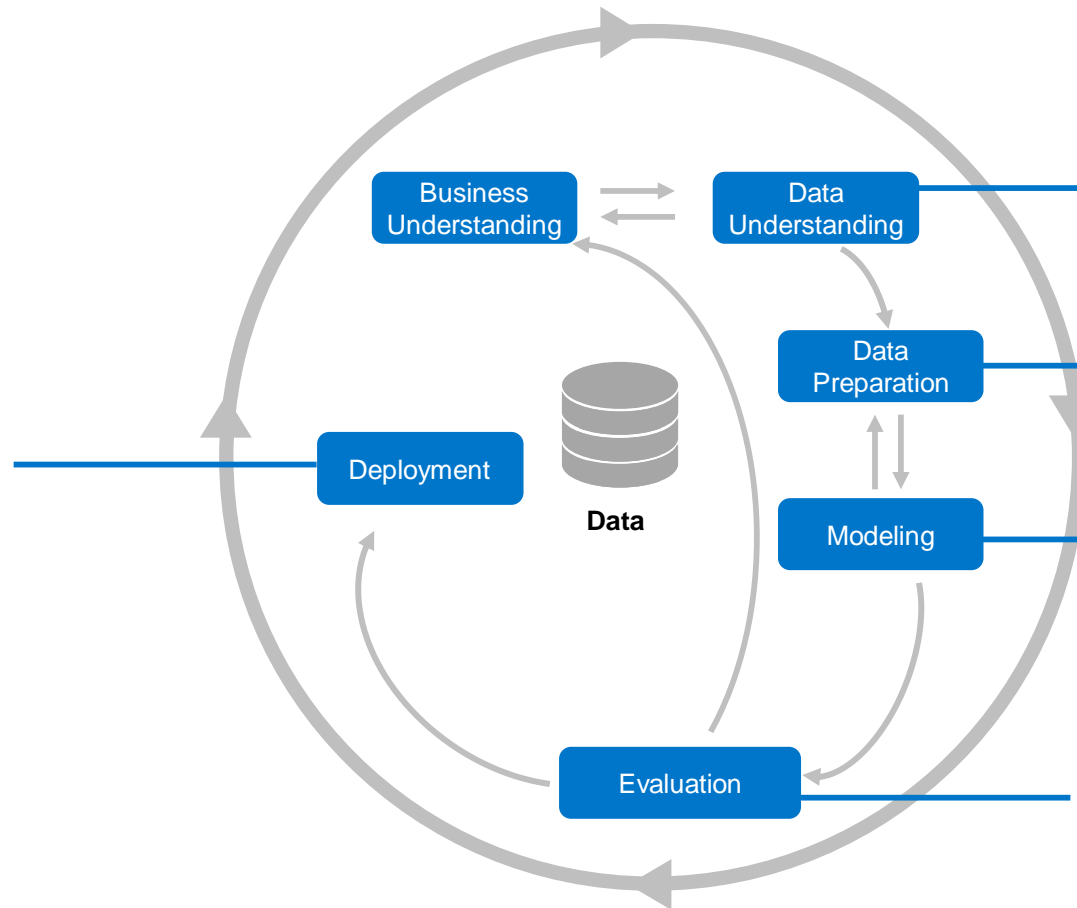
- The technology used in the Automated Mode of SAP Predictive Analytics is an implementation of the theory of statistical learning from [Vladimir Vapnik](#). SAP obtained this technology with the acquisition of a company called [KXEN](#) in 2013.
- Some principles are key:
 - [No hypothesis whatsoever](#), no testing of them
 - No required distribution of the predictors
 - Ability to handle large number of predictors
 - No assumption on relationships between predictors
 - The user has control of the process
- The process is 2 steps:
 - Preparation of the data for further processing / encoding
 - Algorithmics
- It relies on [Structured Risk Minimization \(SRM\)](#) which is implemented in the encoding but also in all steps of model building. The algorithmics is Ridge Regression.



Automated Predictive Analytics

The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications



Explorative / Agile BI frontend

- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups

Calculate many different models

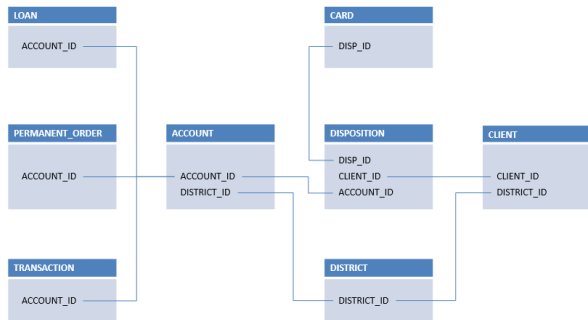
- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

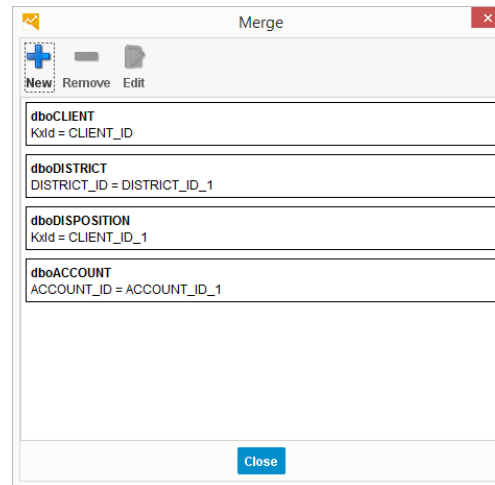
Data Preparation

Turning raw data into wide descriptive datasets

Creating a semantic layer. The structure does not have to be persistent.



Tables



Joins



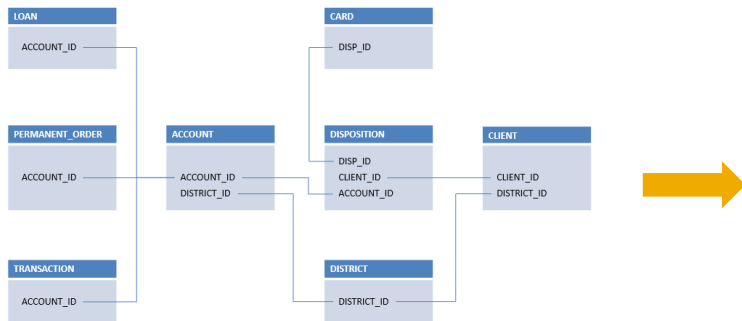
Define 8 successive period(s) of 1 Quarter(s)
starting 8 Quarter(s) before KxTimeStamp

Aggregates
With understanding of time

Data Preparation

Turning raw data into wide descriptive datasets

Creating a semantic layer. The structure does not have to be persistent.

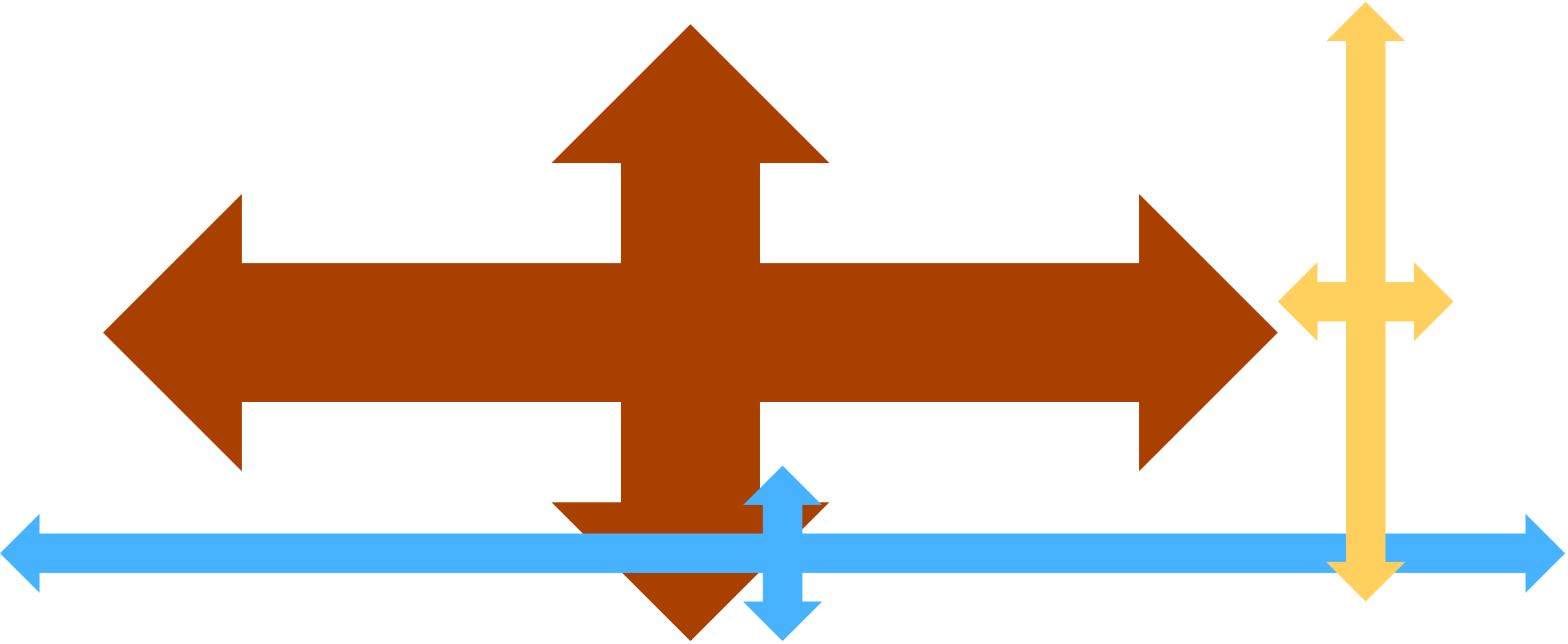


Tables

- Name
- Age
- Martial status
- Account Balance today
- Average Account Balance -1 Quarter
- Average Account Balance -2 Quarters
- Average Account Balance -3 Quarters
- Differences in Avg Account Balance in Euro
- Differences in Avg Account Balance in %
- Average Account Balance -1 Year
- Average Account Balance -2 Years
- Average Account Balance -3 Years
- Differences in Avg Account Balance in Euro
- Differences in Avg Account Balance in %
- Maximum Account Balance -1 Quarter
- Maximum Account Balance -2 Quarters
- Maximum Account Balance -3 Quarters
- Differences in Max Account Balance in Euro
- Differences in Max Account Balance in %
- Maximum Account Balance -1 Year
- Maximum Account Balance -2 Years
- Maximum Account Balance -3 Years
- Differences in Max Account Balance in Euro
- Differences in Max Account Balance in %
- ...
- ...
- ... and thousands of further columns...

Wide descriptive datasets

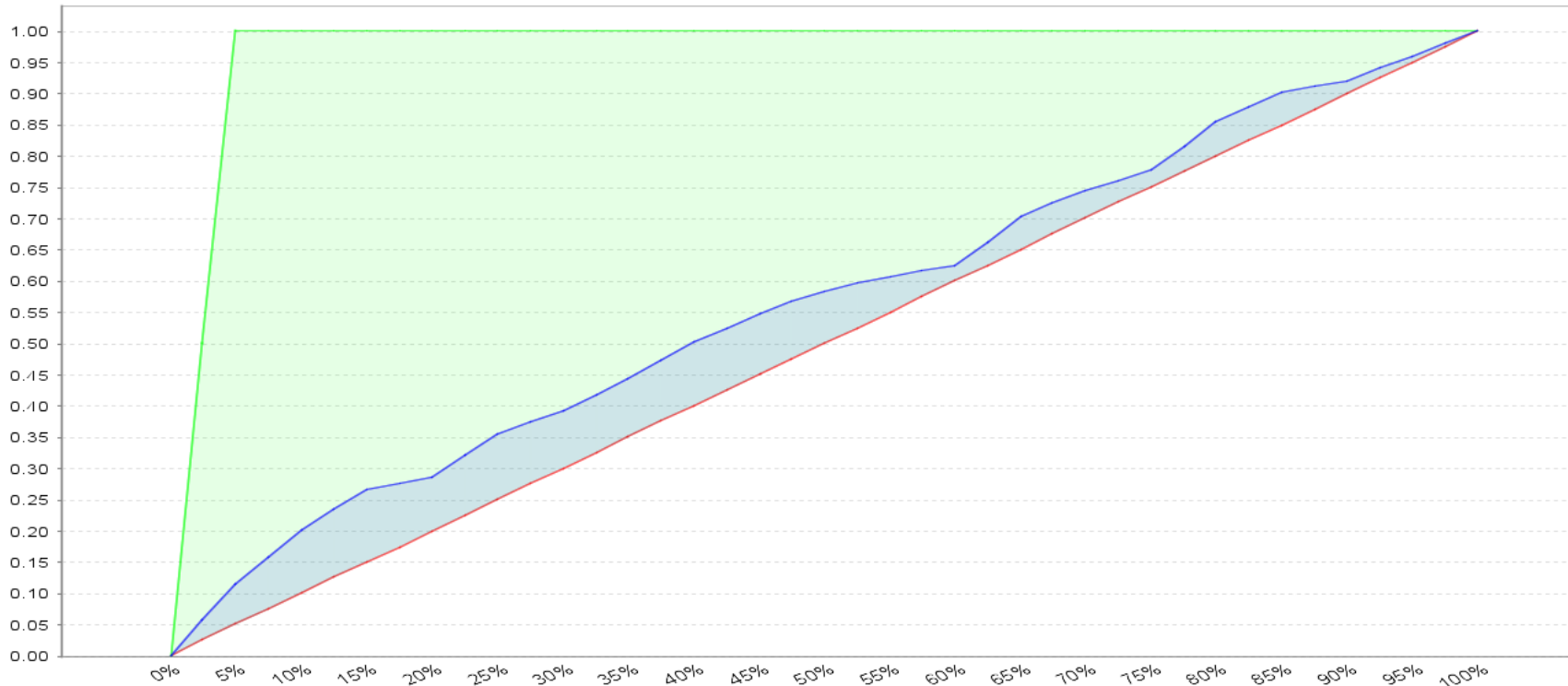
Big Data is not just big Wide, or deep, or both



Why Big Data for Predictive? Lift with Simple Aggregates

20 Variables

- Demographics / Account Information
- Simple Aggregates (e.g. Account Balance, Total Usage)

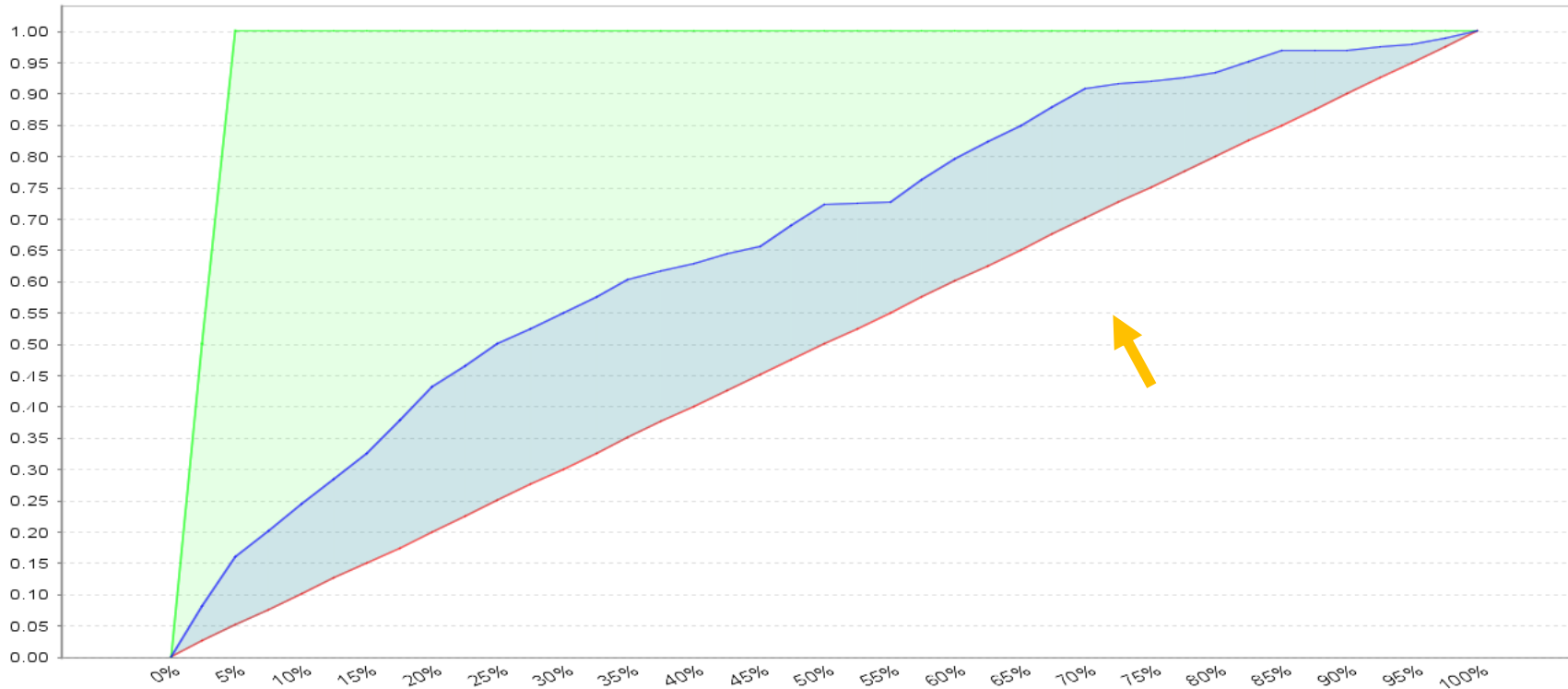


Why Big Data for Predictive?

Lift with Complex Aggregates

100 Variables

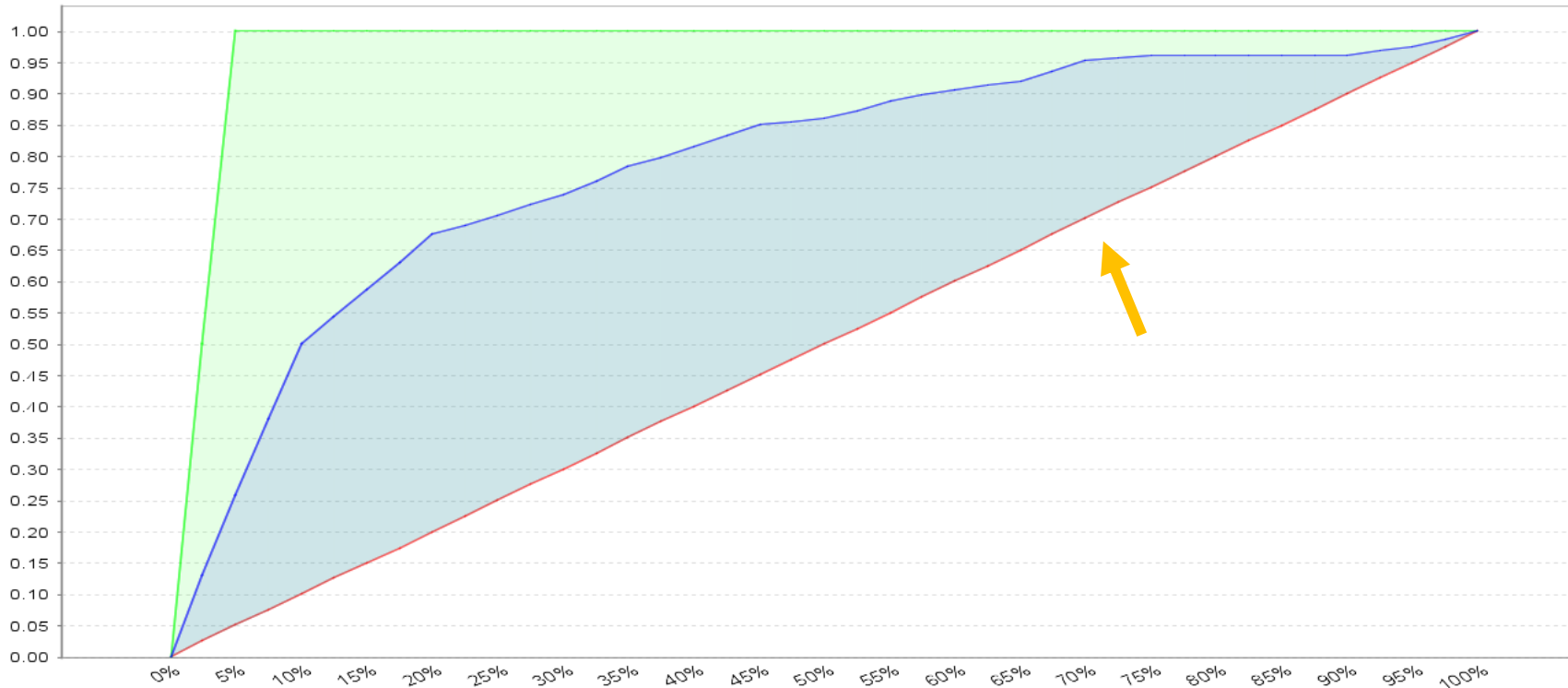
- Pivoting Transactions (e.g. Calls by Type)
- Time-Sensitive Aggregates (e.g. Calls by Week)



Why Big Data for Predictive? Lift with Social Network Analysis

200 Variables

- Social Network Analysis (e.g. Calls in First Circle)
- Community Detection (e.g. Community Churn Rate)



Data Preparation

Encoding the columns, Nominal and Ordinal columns

Example: Let's consider a Variable V1 with 4 categories A, B, C and D and some missing values.

Category / Level	Percent of target variable in Estimation	Percent of target variable in Validation	Assigned value in encoded dataset
A	0.1	0.1	A
B	0.2	0.2	B
C	0.15	0.3	KxOther
D	0.1	0.1	D
E	0.35	0.15	KxOther
NULL	0.2	0.2	KxMissing

Categories with low frequency (outliers) are put together in a noise category called KxOther. It contains as well categories that are not robust i.e. that don't have the same target rate between Estimation and Validation (tested with a Chi Square Test of Independence).

Data Preparation

Binning to obtain robust groups

- Grouping can help to increase robustness. Categories are grouped depending on the target encoding.

Category / Level	Percent of target variable in Estimation	Percent of target variable in Validation	Assigned value in encoded dataset	Grouping
A	0.1	0.1	A	A;D
B	0.2	0.2	B	B;KxMissing
C	0.15	0.3	KxOther	KxOther
D	0.1	0.1	D	A;D
E	0.35	0.15	KxOther	KxOther
NULL	0.2	0.2	KxMissing	B;KxMissing

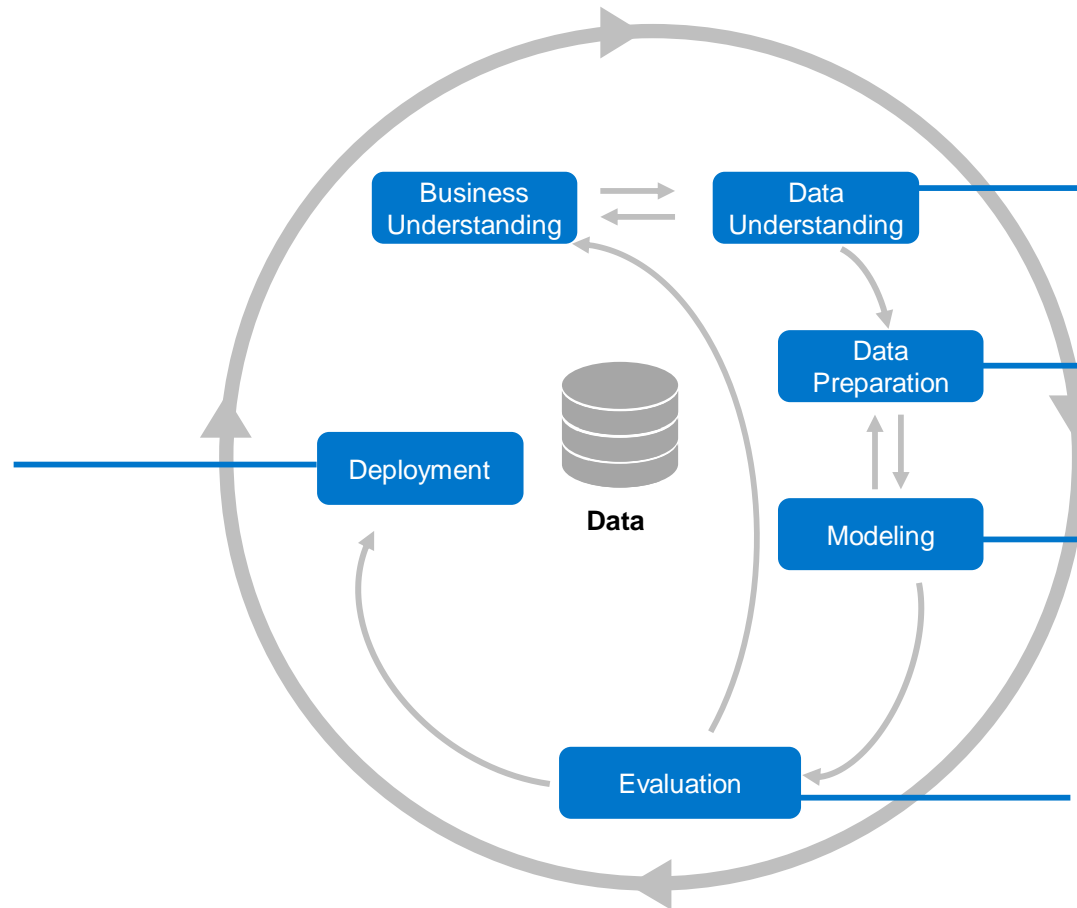
From the encoding we can expect that A and D could be regrouped as well as B and NULL (as they have similar . This is done iteratively:

- by calculating $KI+KR$ for the non-regrouped categories and the regrouped ones
- If $KI+KR$ doesn't decrease (with a tolerance), the group is kept
- Further grouping is tried to the point where $KI + KR$ decreases

Automated Predictive Analytics

The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications



Explorative / Agile BI frontend

- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups

Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Modeling

Ridge Regression

The Ridge Regression penalizes the size of the coefficients by minimizing this extended term:

$$\left(\sum_{i=1}^n (y_i - x_i^T \boldsymbol{\beta})^2 \right) + \lambda \sum_{j=1}^p \beta_j^2$$

p : number of parameters

λ : Ridge Parameter

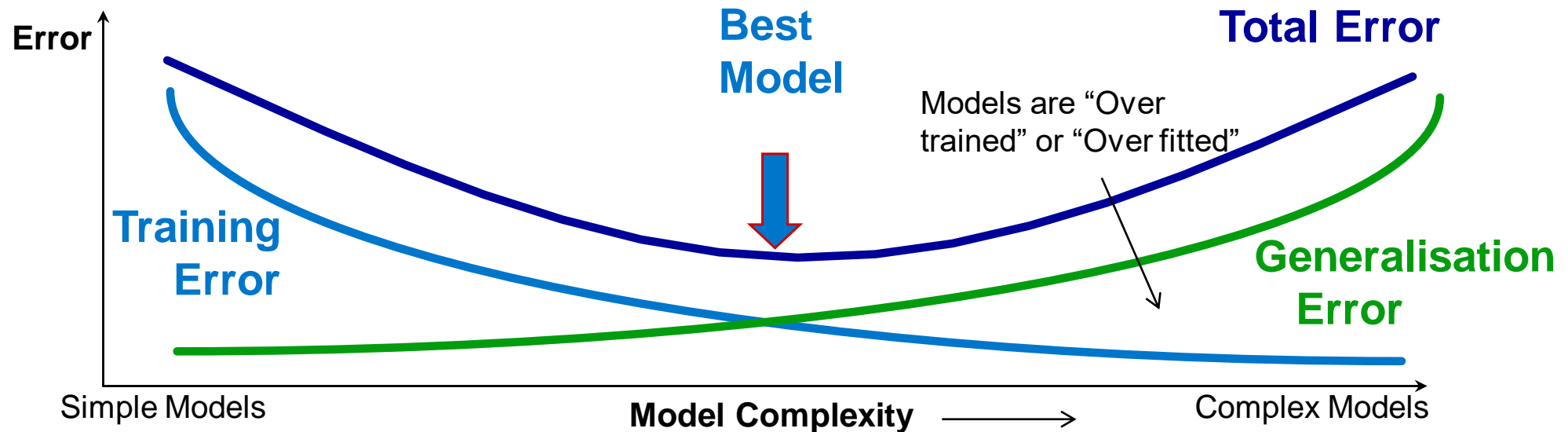
The coefficients that minimize that error are estimated with: $\hat{\boldsymbol{\beta}} = (X^T X + \lambda I)^{-1} X^T y$

Source: <http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/9-1.pdf>

Modeling

Selecting the best model

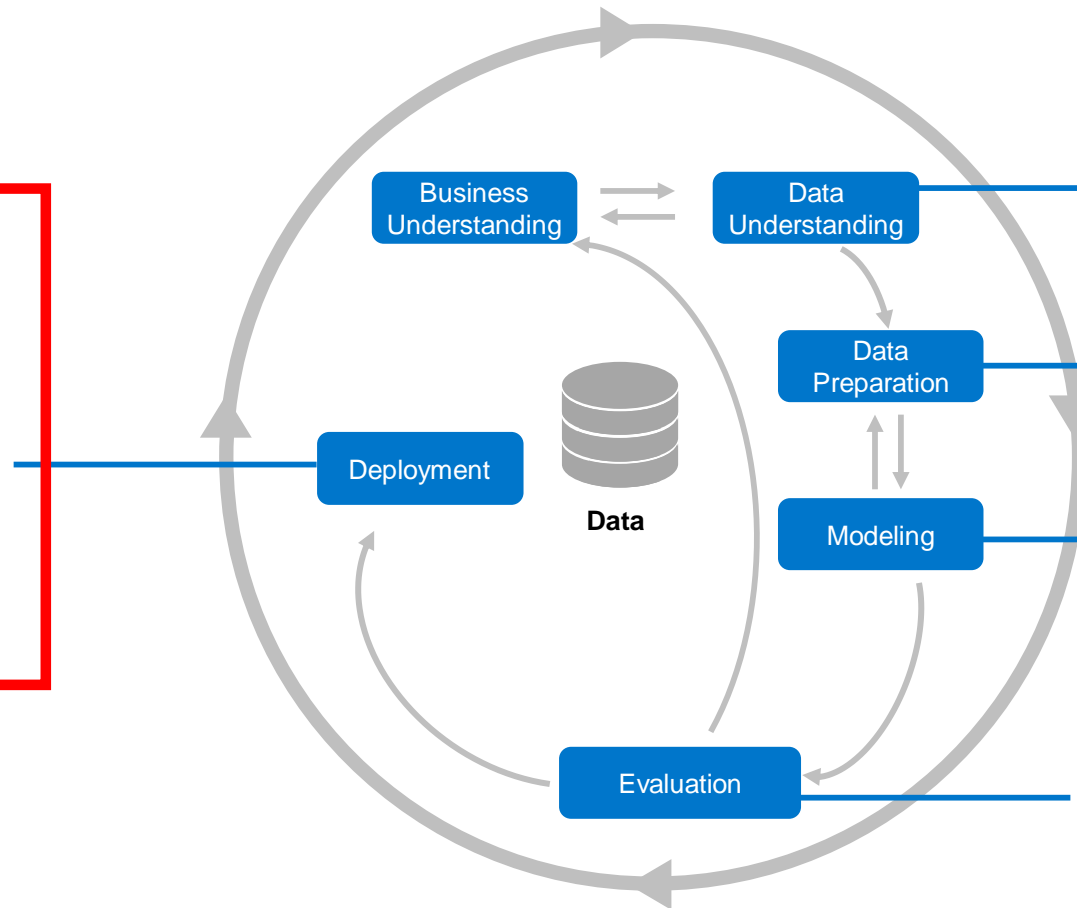
- By playing with λ , more or less constraint is applied on the coefficients of the regression.
 - If a lot of constraint is applied, the Training error (ε_t) is high but the Generalization error (ε_g) is low
 - Inversely, if little constraint is applied, the Training error (ε_t) is low but the Generalization (ε_g) is is high



Automated Predictive Analytics

The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications



Explorative / Agile BI frontend

- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Closed Loop Automatically Retrain and Apply Models

- Maintain large number of models
- Automatically retrain models when needed
- Automatically apply models and persist scores to source systems or business applications

The screenshot displays the SAP Predictive Analytics interface. The top window shows a 'Welcome to Predictive Factory' message. The main window, titled 'Task Run', provides a detailed overview of a specific task run. It includes a breadcrumb trail: 'Home \ Credit Card Affinity \ Retrain Credit Card Affinity'. The task run is identified as 'Task Run 20' with a reference date of 'August 3, 2016 12:00:33 PM'. The interface is divided into several sections:

- Model Performance:** Shows 'Predictive Power' at 77.32% and 'Prediction Confidence' at 95.66%, both represented by green progress bars.
- Target Key Frequency:** Displays 'Validation' at 98% and 'Estimation' at 99%.
- Maximum Smart Variable Contributions:** Lists variables and their contributions: 'DISTRICT_ID_2' (15.97%), 'QuarterlyTransaction_4Q4B_OPERATI ON_COLLECTIONFROMANOTHERBA NK_SUM_2_AMOUNT' (15.03%), 'DATE_Y' (14.75%), and 'TYPE' (12.59%).
- Performance Chart:** A line chart showing 'Detected ...' on the y-axis (0% to 100%) and 'Population' on the x-axis. It compares three models: 'Wizard' (green), 'Validation' (blue), and 'Random' (red). The 'Wizard' model shows the highest detection rate, followed by 'Validation', and 'Random' shows the lowest.

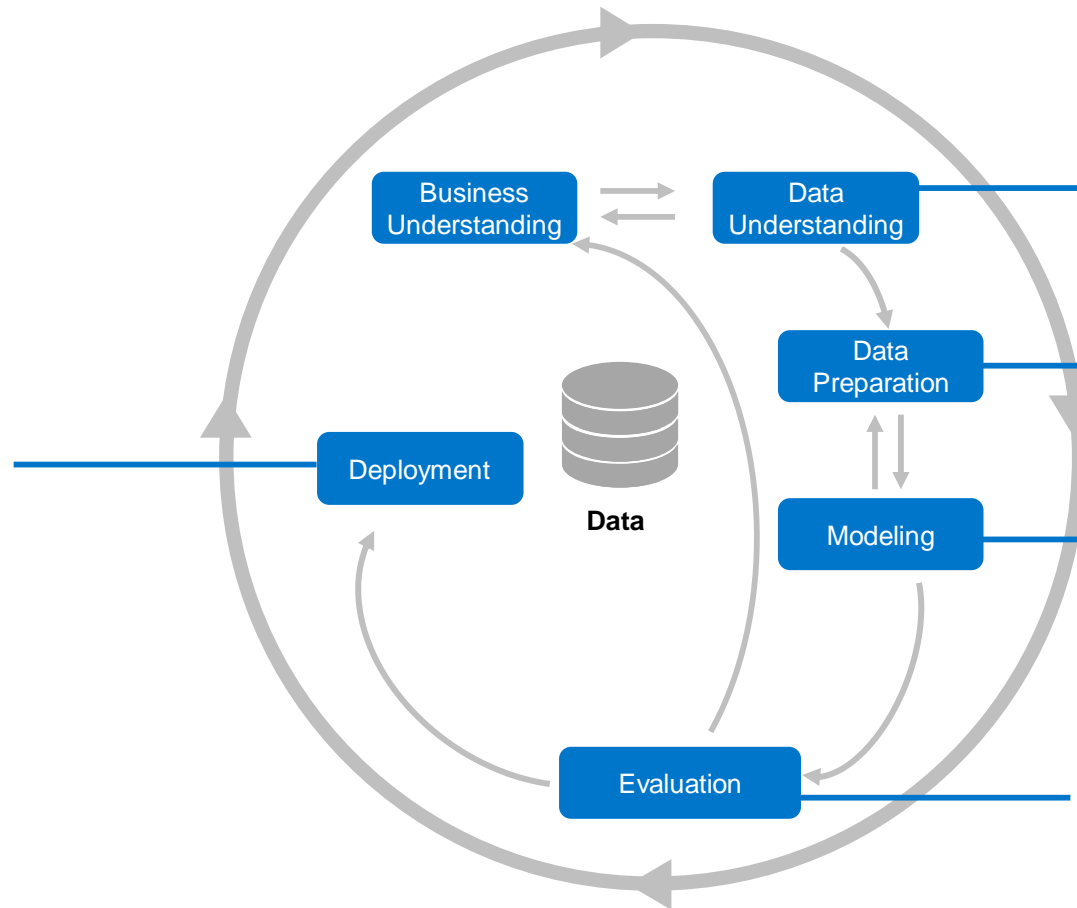
Below the main overview, a table lists 'Task Runs (20)'. The table has columns for 'Run', 'Status', 'Reference Date', 'Start Date', and 'End Date'. All listed runs have a status of 'Completed' (indicated by a green checkmark).

Run	Status	Reference Date	Start Date	End Date
20	Completed	August 3, 2016 12:00 PM	August 3, 2016 2:00 PM	August 3, 2016 2:00 PM
19	Completed	August 2, 2016 5:24 PM	August 2, 2016 7:24 PM	August 2, 2016 7:24 PM
18	Completed	August 2, 2016 5:24 PM	August 2, 2016 7:24 PM	August 2, 2016 7:24 PM
17	Completed	August 2, 2016 5:24 PM	August 2, 2016 7:24 PM	August 2, 2016 7:24 PM
16	Completed	July 28, 2016 7:25 PM	July 28, 2016 9:25 PM	July 28, 2016 9:25 PM
15	Completed	July 27, 2016 6:07 PM	July 27, 2016 8:07 PM	July 27, 2016 8:07 PM
14	Completed	July 26, 2016 12:00 PM	July 26, 2016 2:00 PM	July 26, 2016 2:00 PM

Automated Predictive Analytics

The Cross Industry Standard Process for Data Mining (CRISP-DM)

- Mass-produce such best-performing models
- Monitor these models on their predictive quality
- Retrain if needed
- Calculate new scores and write back or into business applications



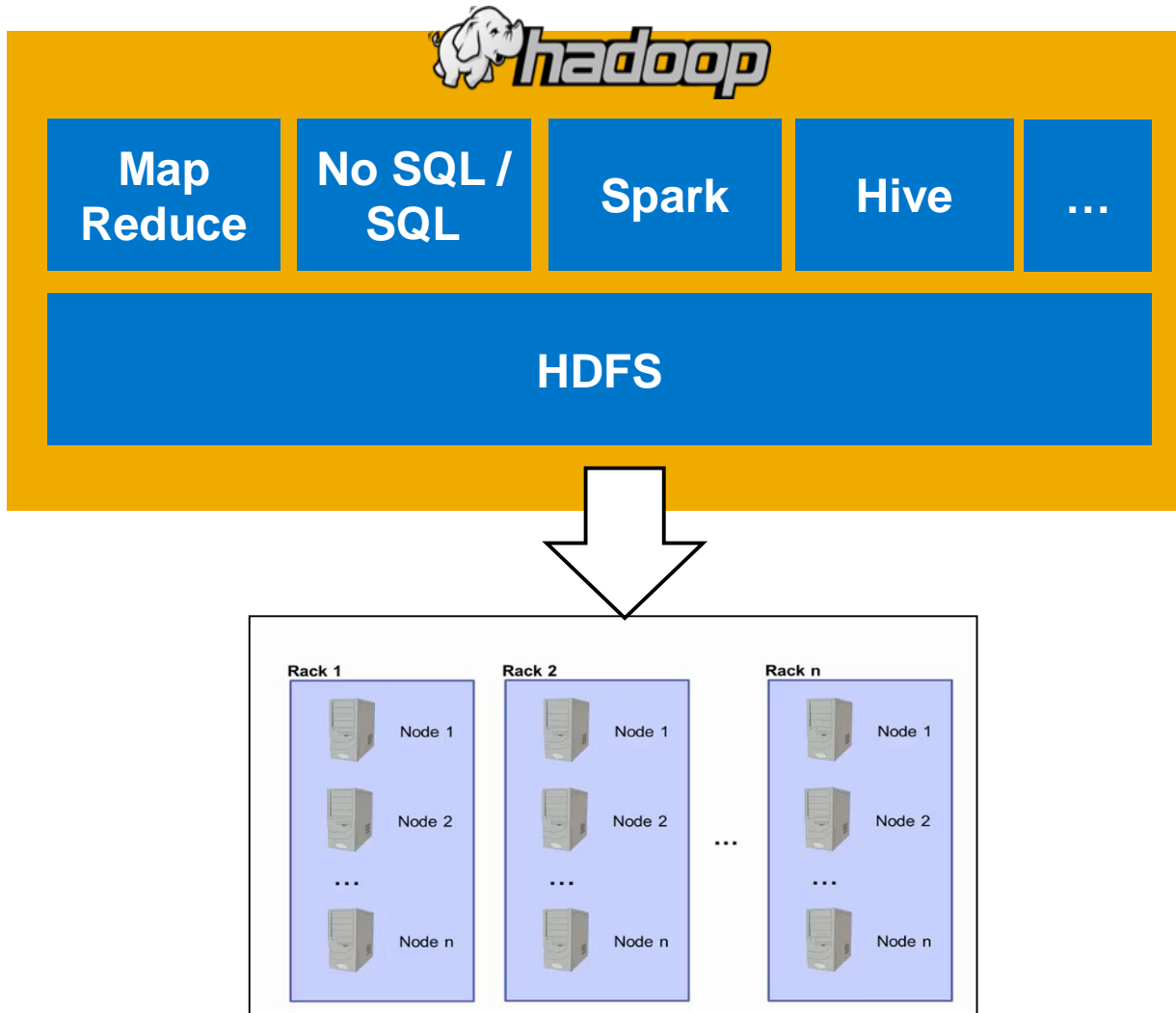
Explorative / Agile BI frontend

- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups
- Calculate many different models

- Evaluate models on unseen data and select the best-performing
- Interpret model and discuss insight with the business department

Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Big Data in Hadoop



Features

- Commodity Hardware (\$1500/ TB)
- Open Source Stack (No Licensing fee)
- Elastic scaling
- scales linearly with # of nodes
- Easy to add 1000s of (cheap) nodes
- Code executes close to the data

Hadoop Perspective for 2016

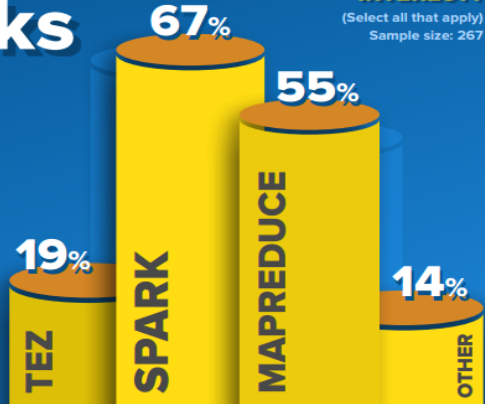
Compute Frameworks

As illustrated below, interest in Apache Spark surpasses interest in all other compute frameworks, including the recognized incumbent, MapReduce.

- Nearly 70% of respondents are most interested in Apache Spark.
- Approximately 55% are interested in MapReduce.
- Fewer than 20% of respondents are interested in other compute frameworks.

WHICH OF THE FOLLOWING COMPUTE FRAMEWORKS ARE OF MOST INTEREST?

(Select all that apply)
Sample size: 267

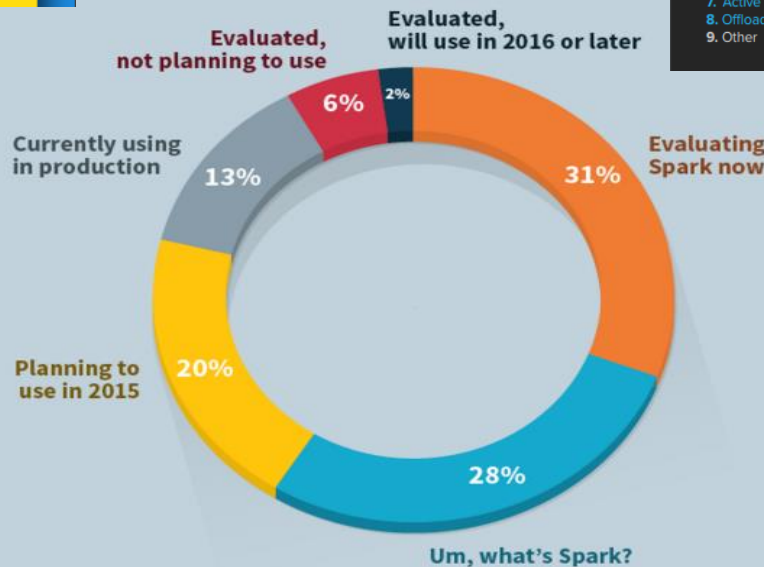


Adoption interest for Spark has topped in Hadoop eco-system

2016's #1 trend: Apache Spark will move from talking point into deployment

Big data workloads in production jumped by nearly 30% from 2014 to 2015

RELATIONSHIP WITH SPARK



Um, what's Spark?

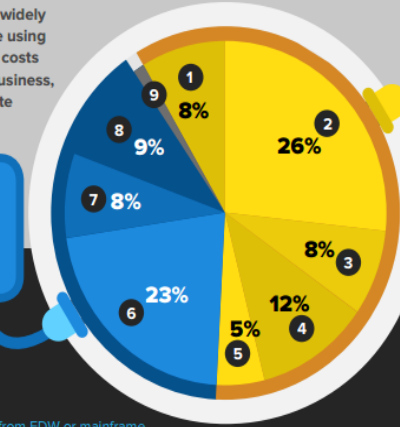
Uses for Hadoop

As Hadoop becomes more widely embraced, respondents are using Hadoop not only to reduce costs but also to advance their business, allowing them to incorporate big data in strategic ways.

For 40% of respondents Hadoop is used as a cost-effective alternative for storage and processing in the data warehouse.

WHAT HADOOP USE CASES ARE OF MOST INTEREST TO YOU?

Sample size: 267



More than half of respondents see Hadoop as a way to innovate, using data from social media and the IoT and applying predictive analytics and visualization for greater insights about their business.

1. Clickstream Analysis & Social Media Data
2. Advanced/Predictive Analytics
3. Internet of Things
4. Data Discovery & Visualization
5. Mobile Devices & Apps

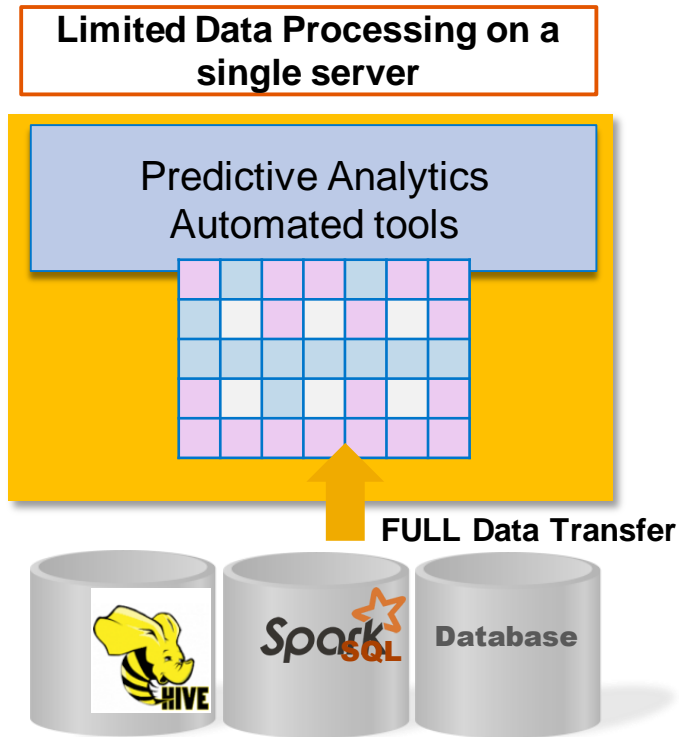
55% users want to leverage Hadoop for Business users and Advanced use cases

Source: <http://www.syncsort.com/>

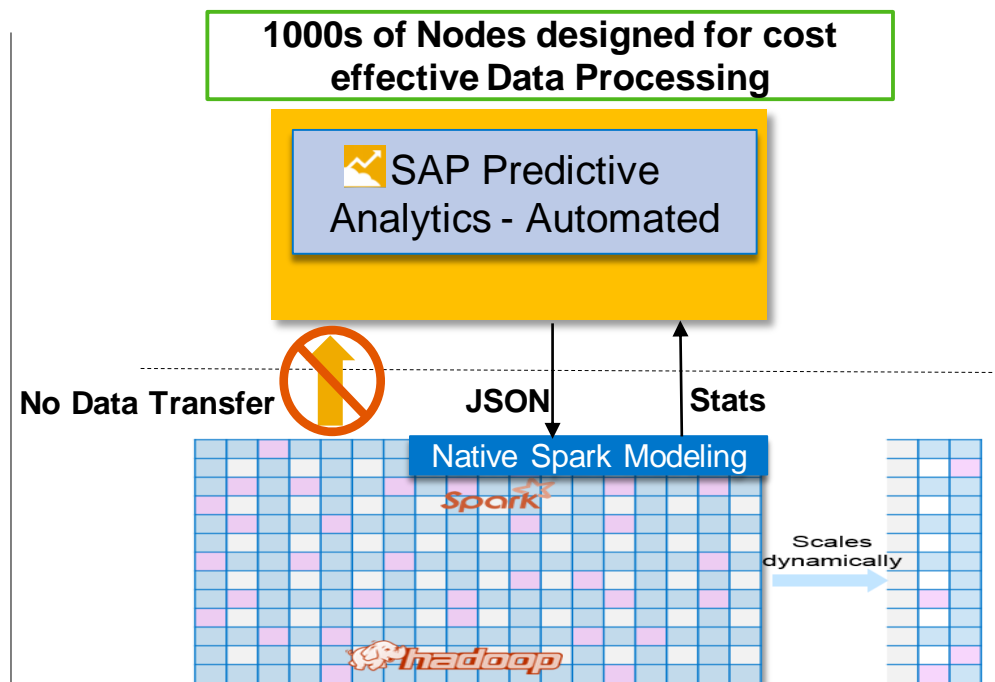
Modeling for Big Data

Traditional Tiered Architecture vs. Native Spark Modeling

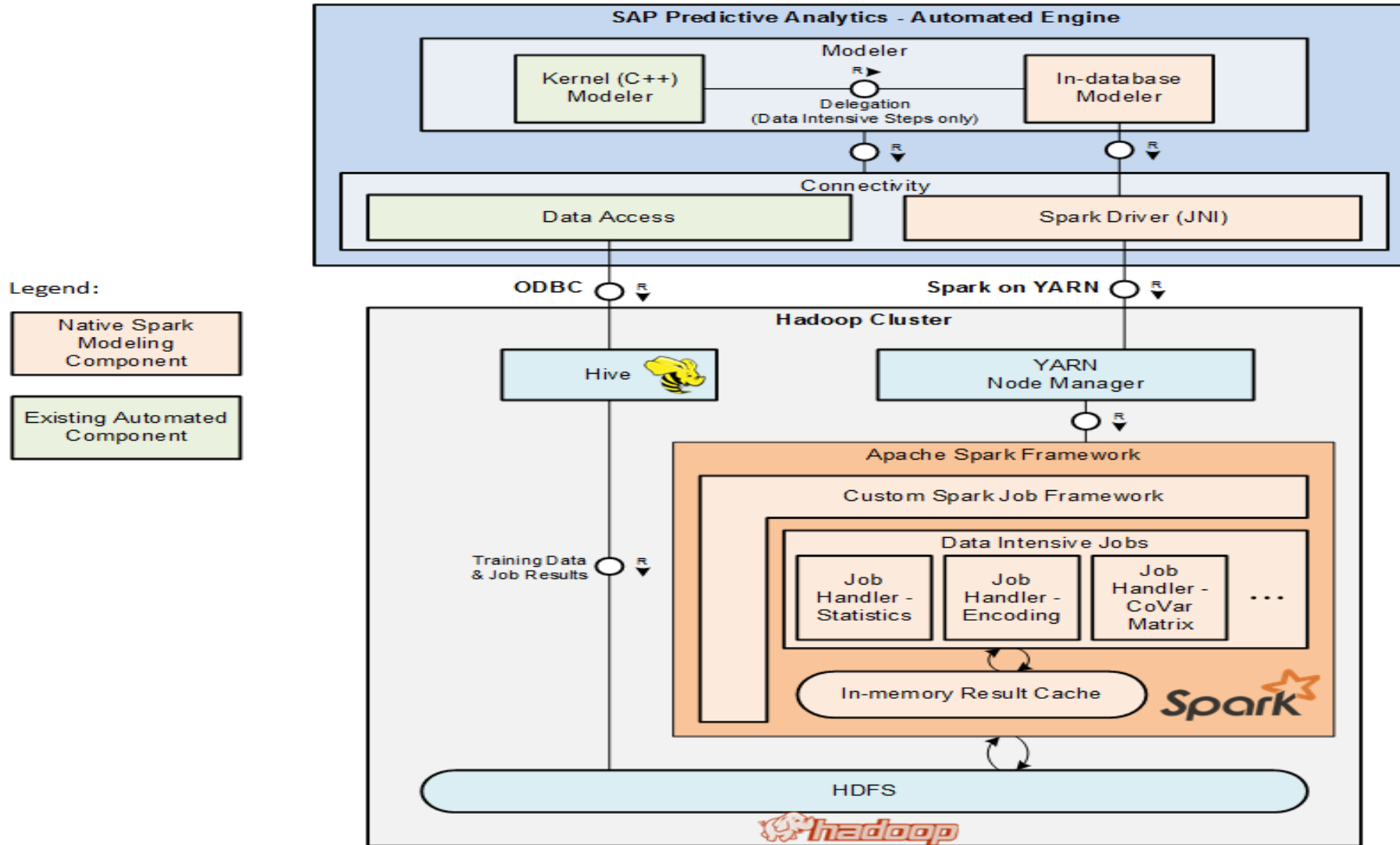
- Full dataset brought to application for processing
- Limited Performance, Scalability



- Data processing beside data
- Performance and scalability built-in

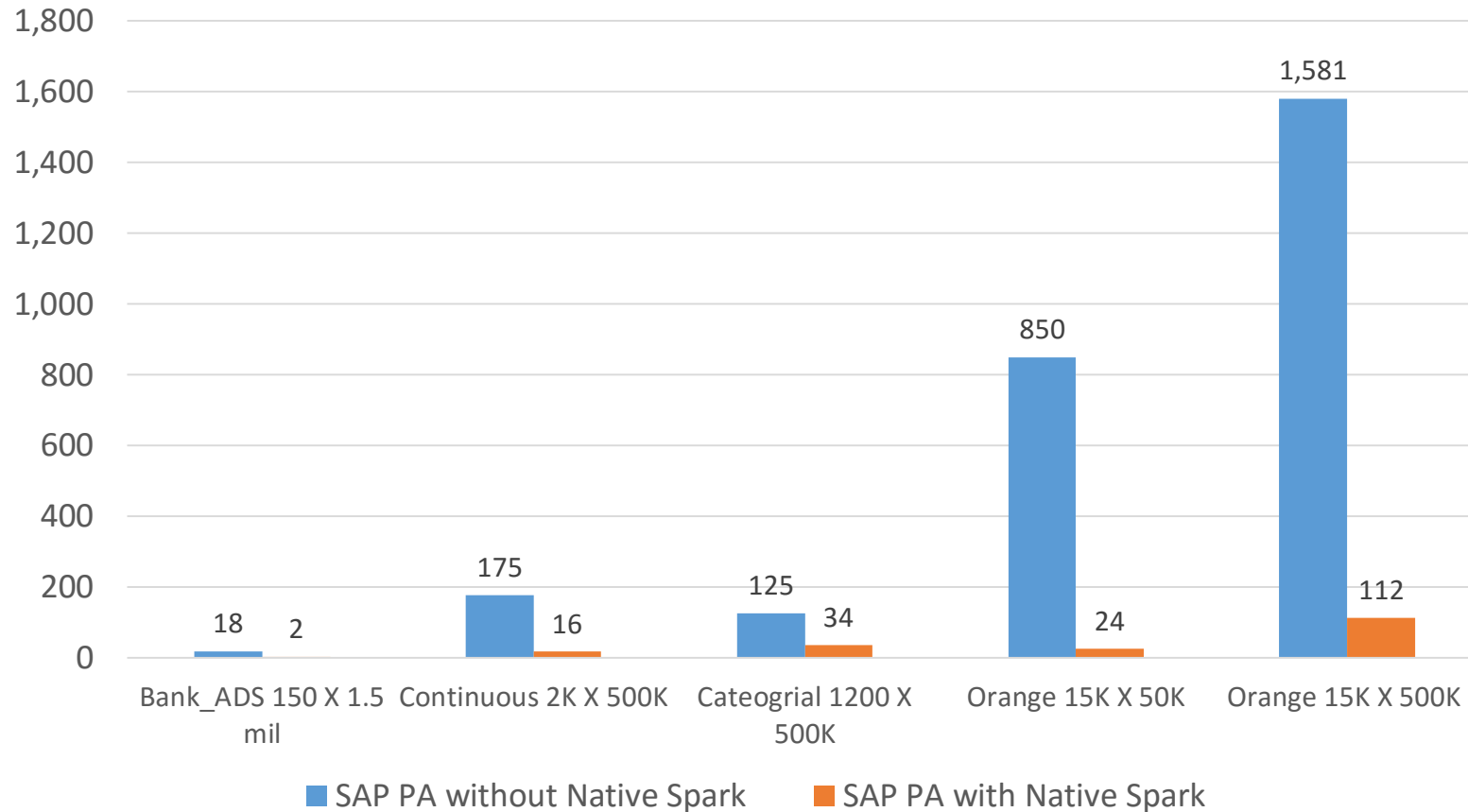


Native Spark Modeling - Architecture



Performance and Scalability With and Without Native Spark Modeling

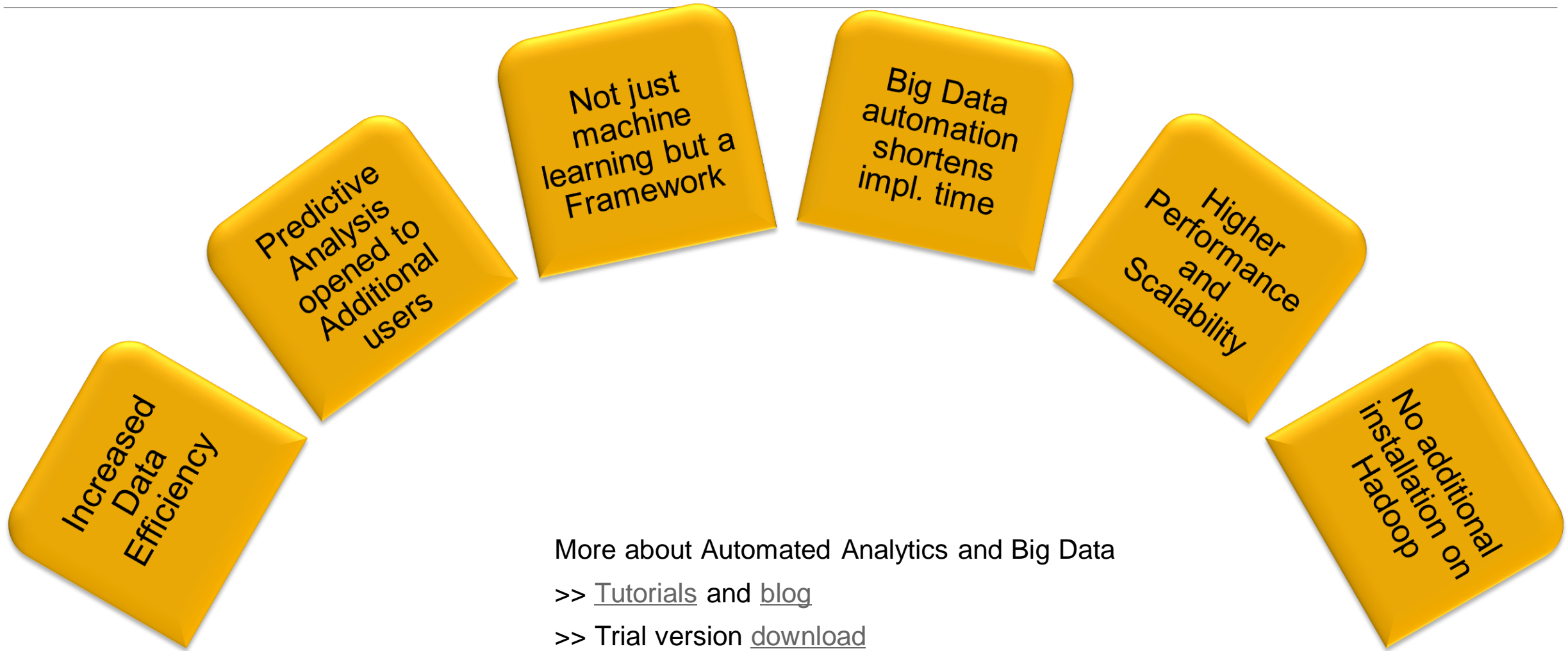
Response time in mins



Summary

- **14 times faster for 15K var dataset**
- **10 times faster for 2K var dataset**
- Native Spark Modelling performance is better with bigger and wider datasets
- Scalability = quadratic $O(n^2)$ of matrix operations

Summary



More about Automated Analytics and Big Data

>> [Tutorials](#) and [blog](#)

>> Trial version [download](#)



Thank you

Contact information:

Priti Mulchandani
Product Manager for Big Data Analytics
p.mulchandani@sap.com

Andreas Forster
Global Center of Excellence
andreas.forster@sap.com