Predictive Analytics for Big Data with Native Spark Modeling

Priti Mulchandani, Andreas Forster

September 2016

Trends in Data Science

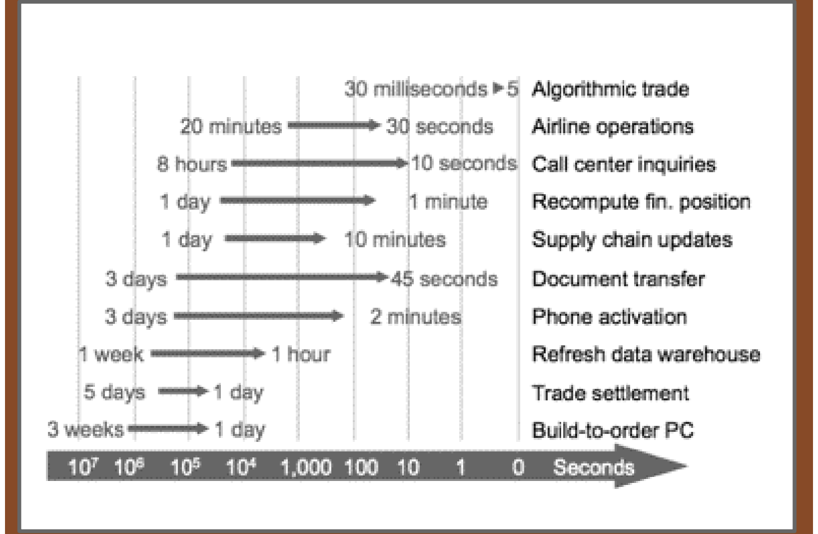
Massive Amount of Data

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



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 carreer 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

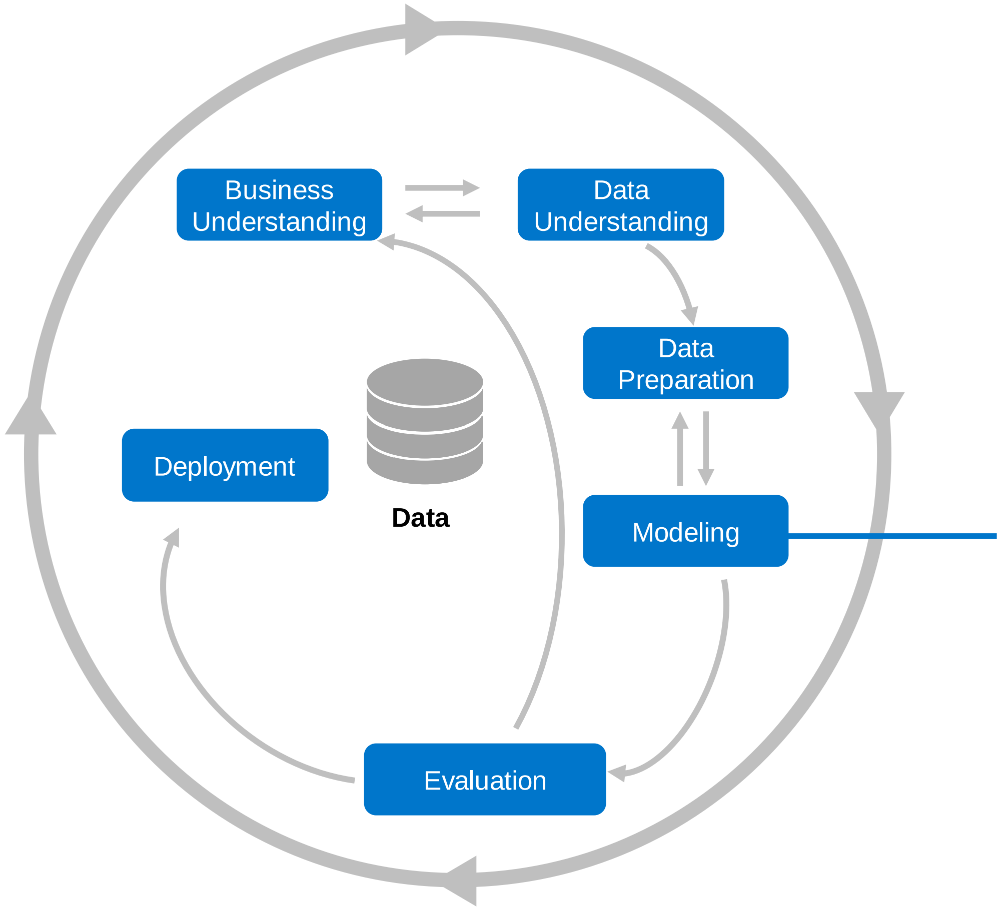
Apply the model on new data

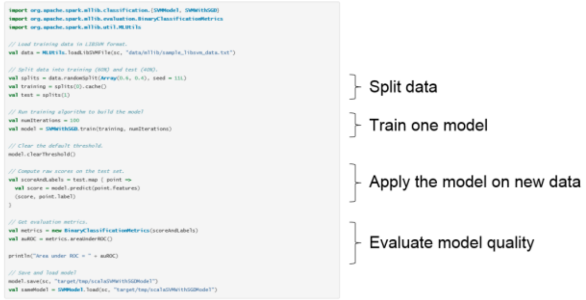
Evaluate model quality

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.

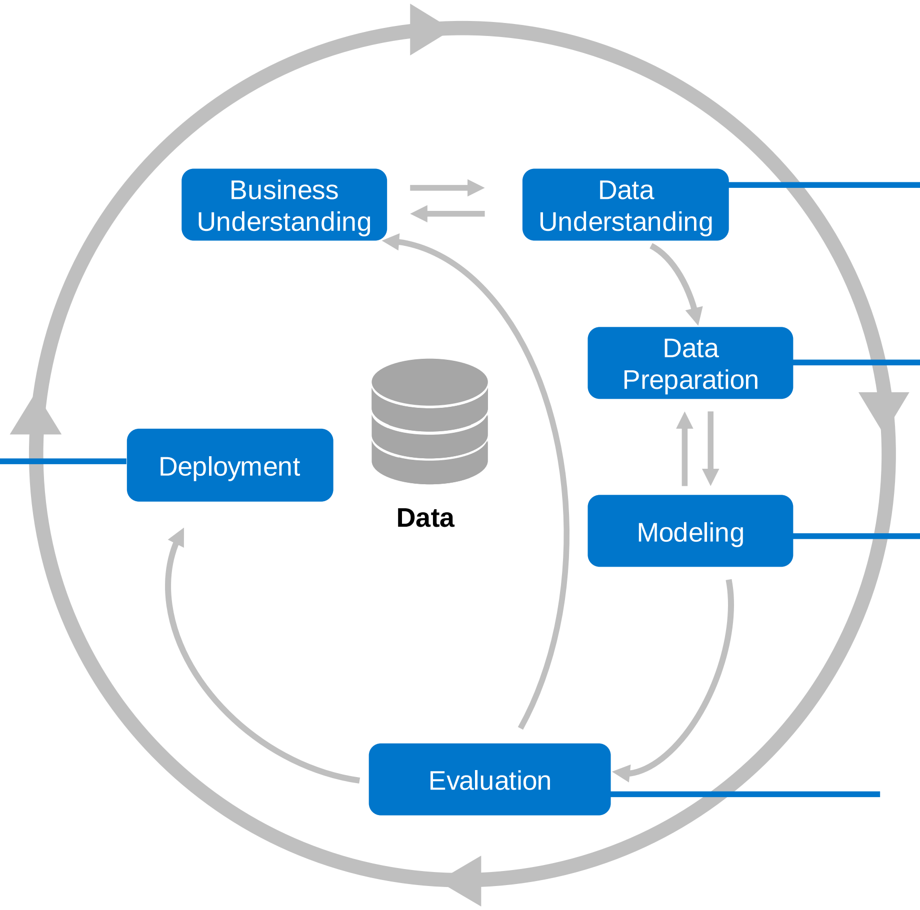




Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

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

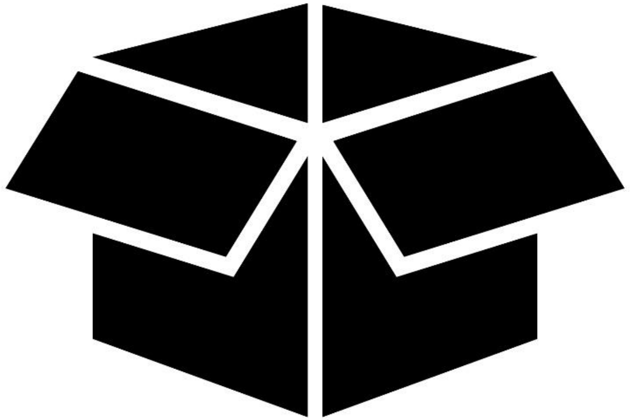
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups

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

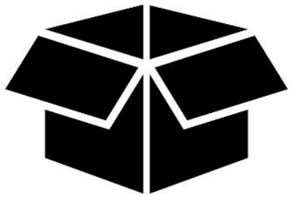
Automated Analytics



How?

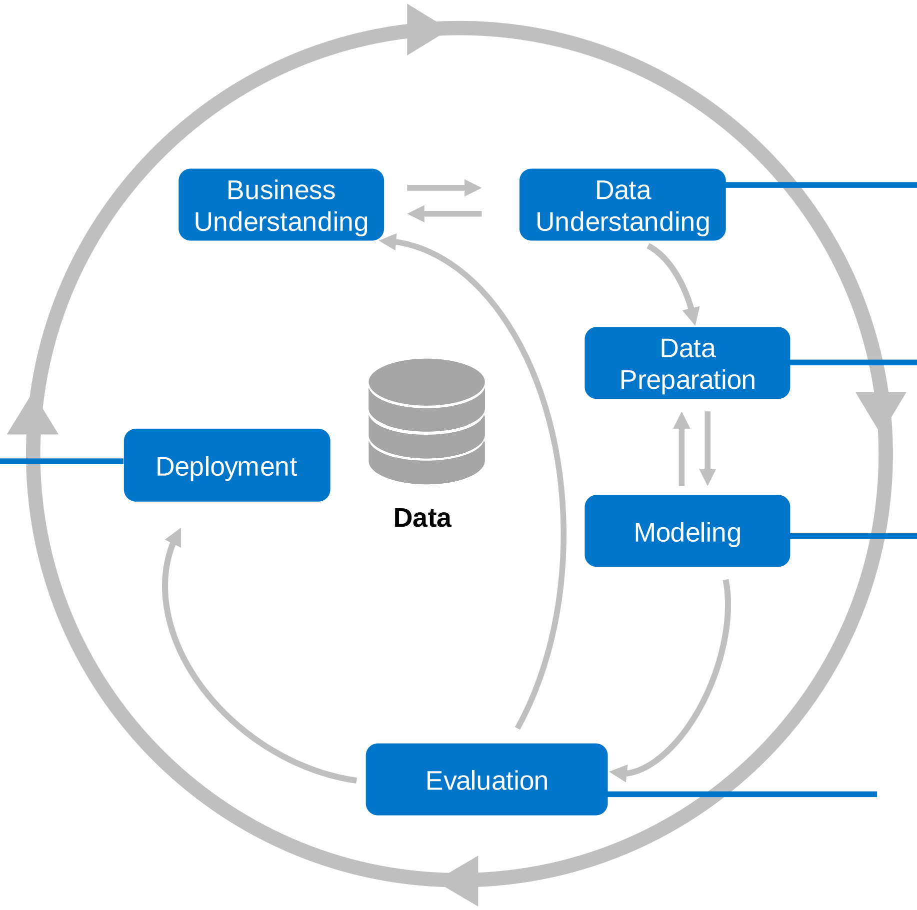
The Principles

- 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

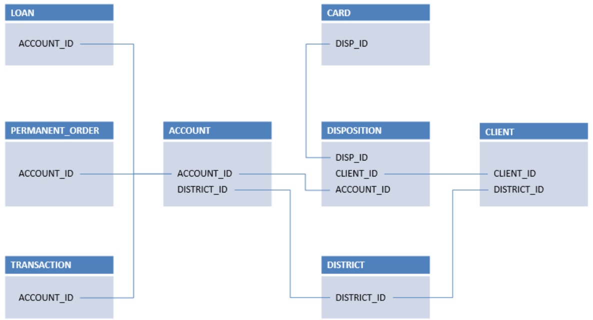
Calculate many different models

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

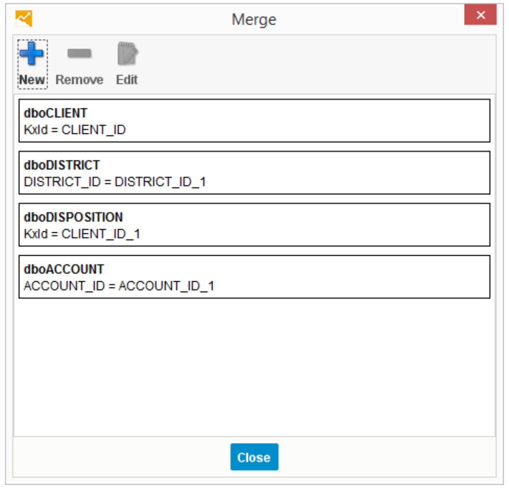
Source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Data Preparation
Turning raw data into wide descriptive datasets

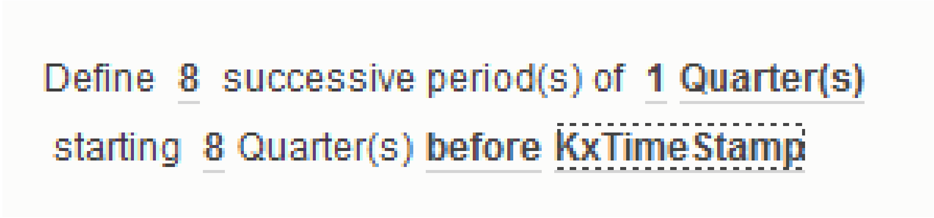
Creating a semantic layer. The structure does not have to be persistet.



Tables



Joins

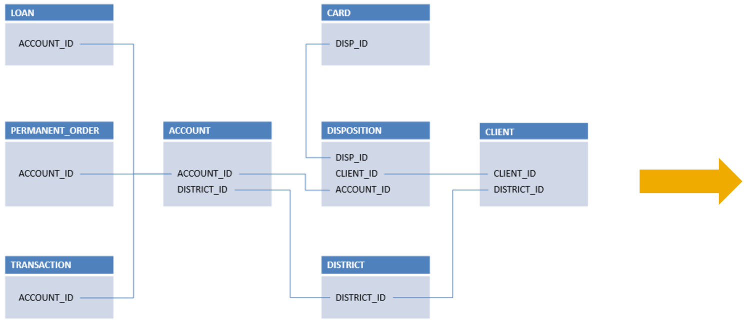


Aggregates

With understanding of time

Data Preparation
Turning raw data into wide descriptive datasets

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

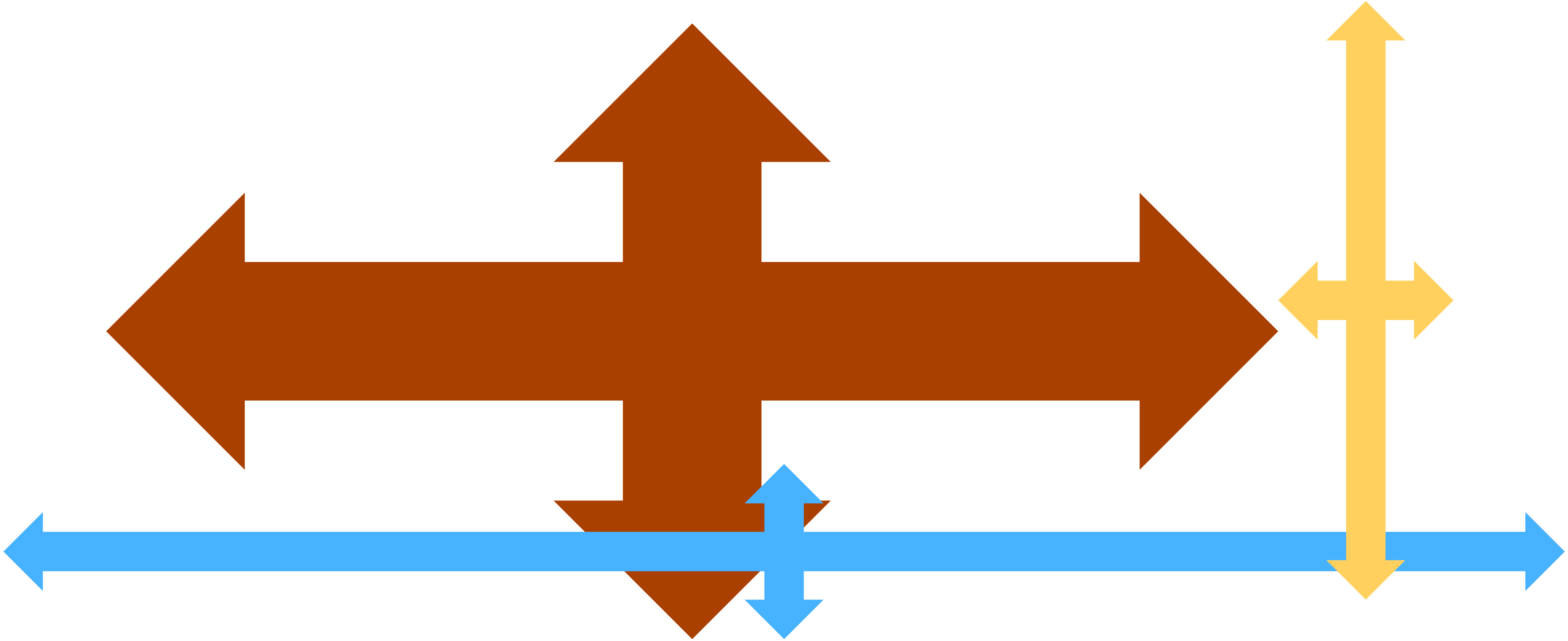


Tables

- 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

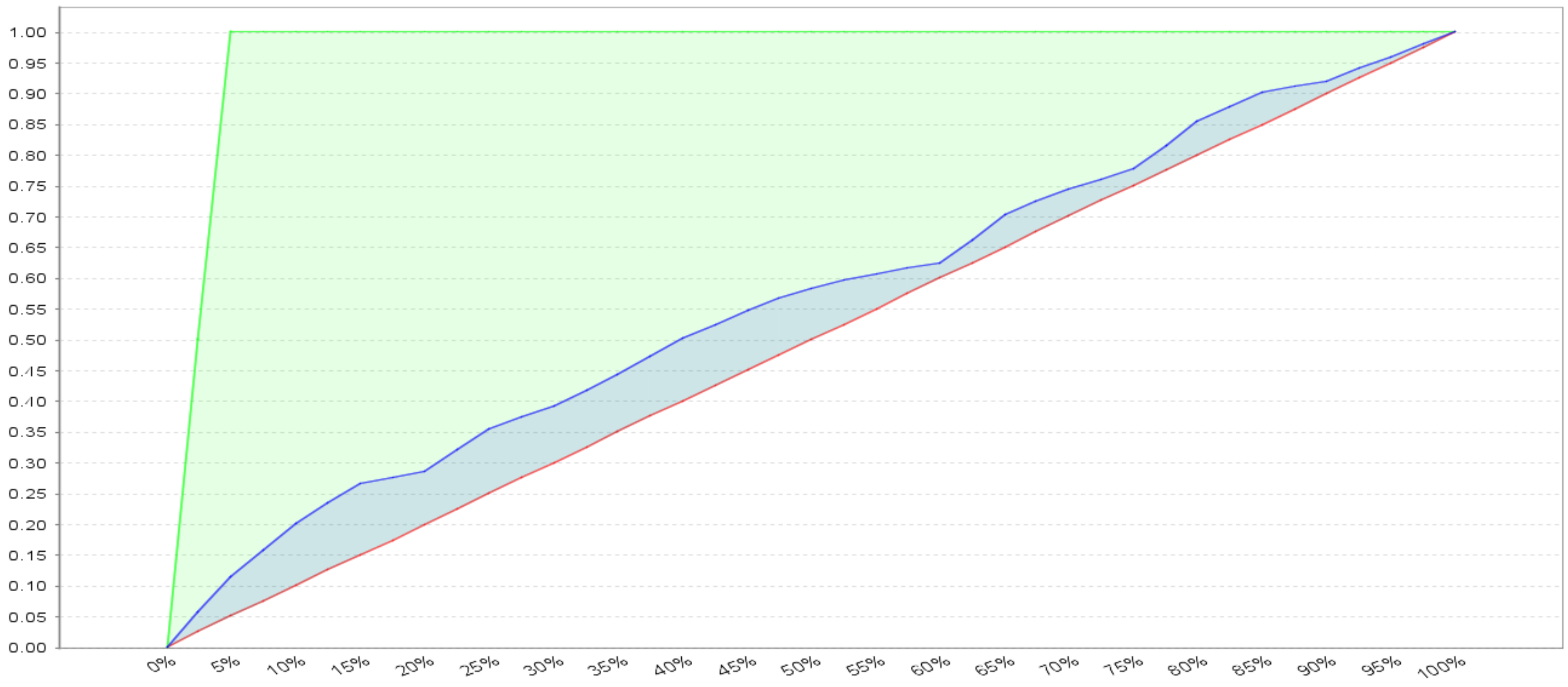
Big Data is not just big
Wide, or deep, or both



Why Big Data for Predictive?
Lift with Simple Aggregates

20 Variables

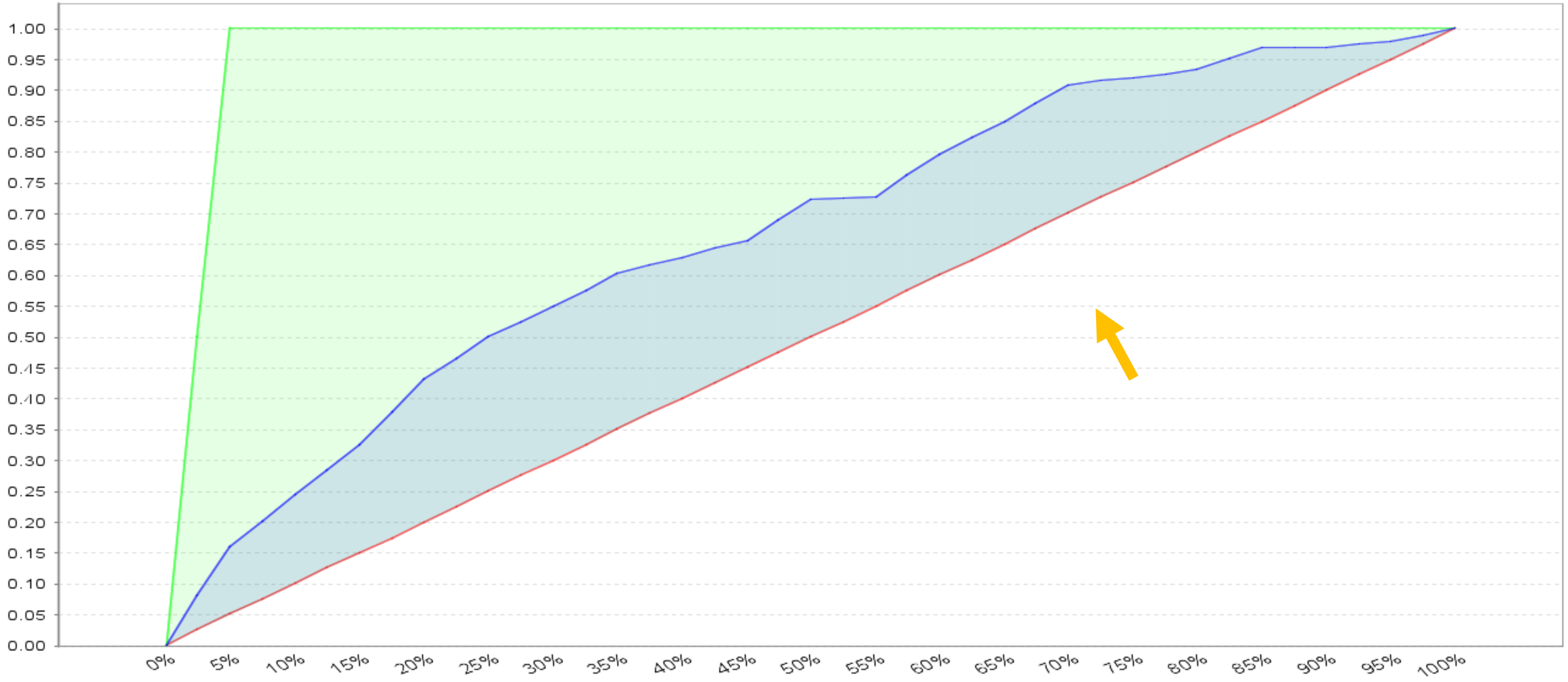
- Demographics / Account Information
- Simple Aggregates (e.g. Account Balance, Total Usage)



Why Big Data for Predictive?
Lift with Complex Aggregates

100 Variables

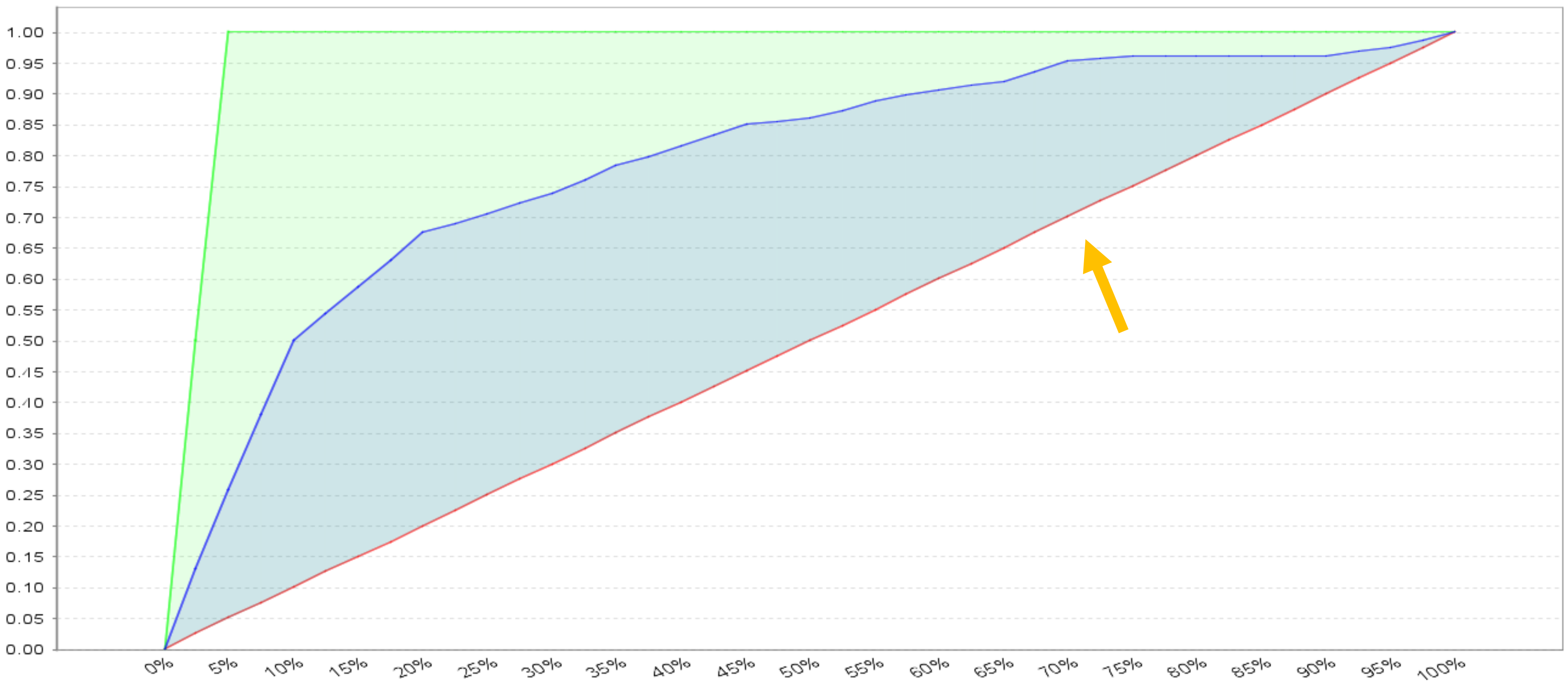
- Pivoting Transactions (e.g. Calls by Type)
- Time-Sensitive Aggregates (e.g. Calls by Week)



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.

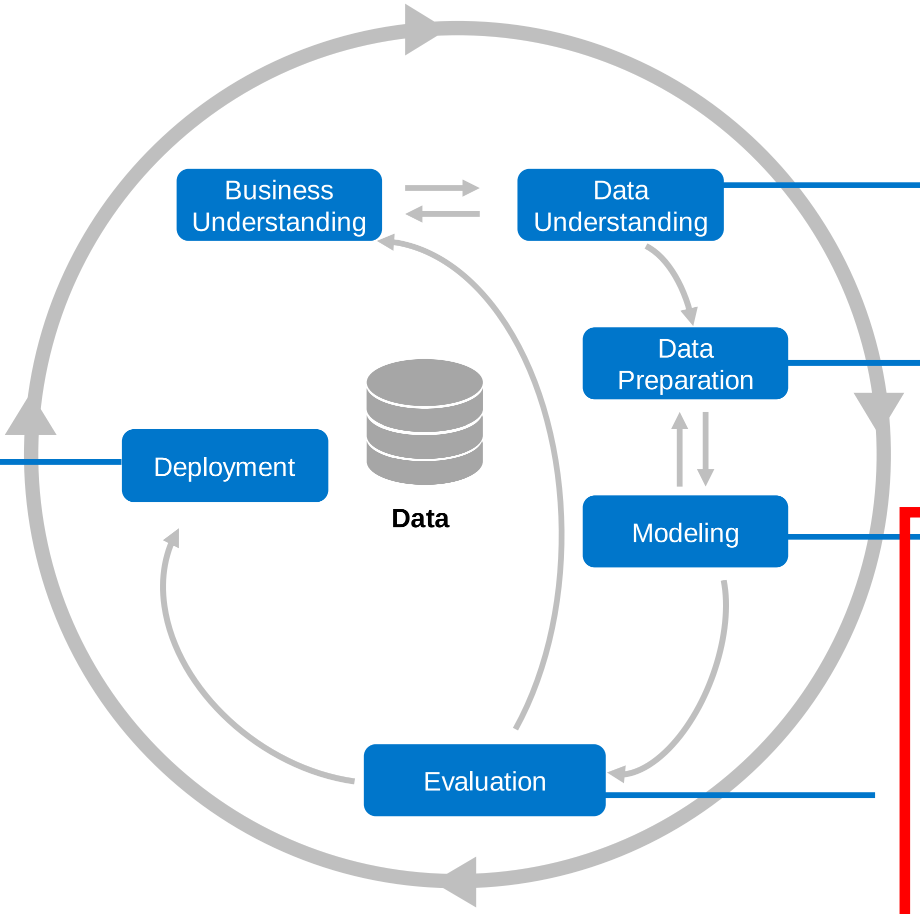
Category / Level	Percent of target variable in Estimation	Percent of target variable in Validation	Assigned value in encoded dataset	Grouping
A	0.1	0.1	A	A;D
B	0.2	0.2	B	B;KxMissing
C	0.15	0.3	KxOther	KxOther
D	0.1	0.1	D	A;D
E	0.35	0.15	KxOther	KxOther
NULL	0.2	0.2	KxMissing	B;KxMissing

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:

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

p: number of parameters

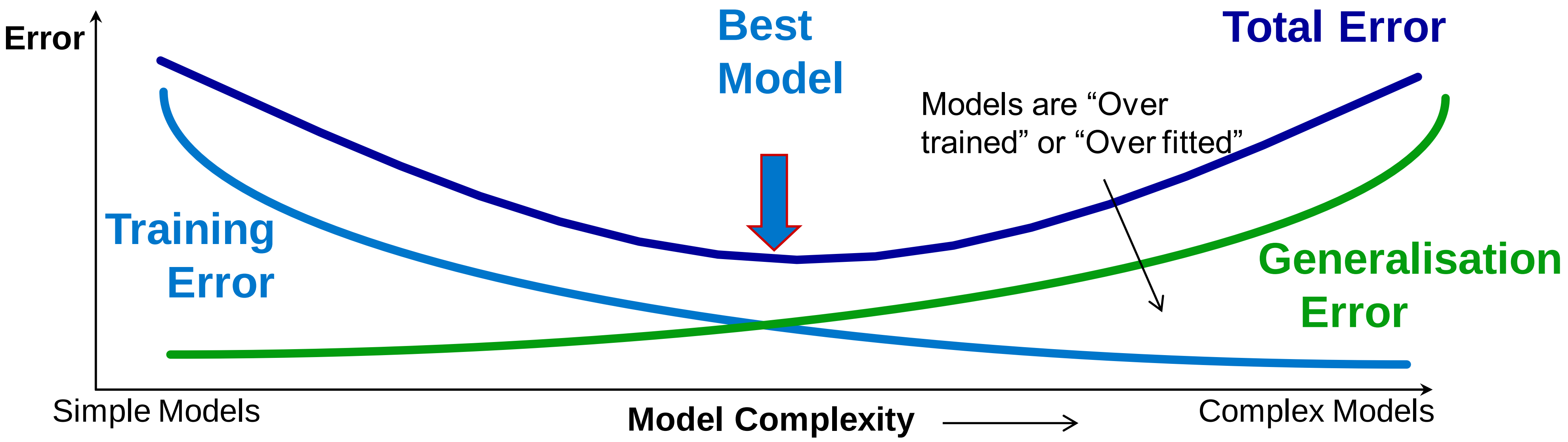
λ: Ridge Parameter

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

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

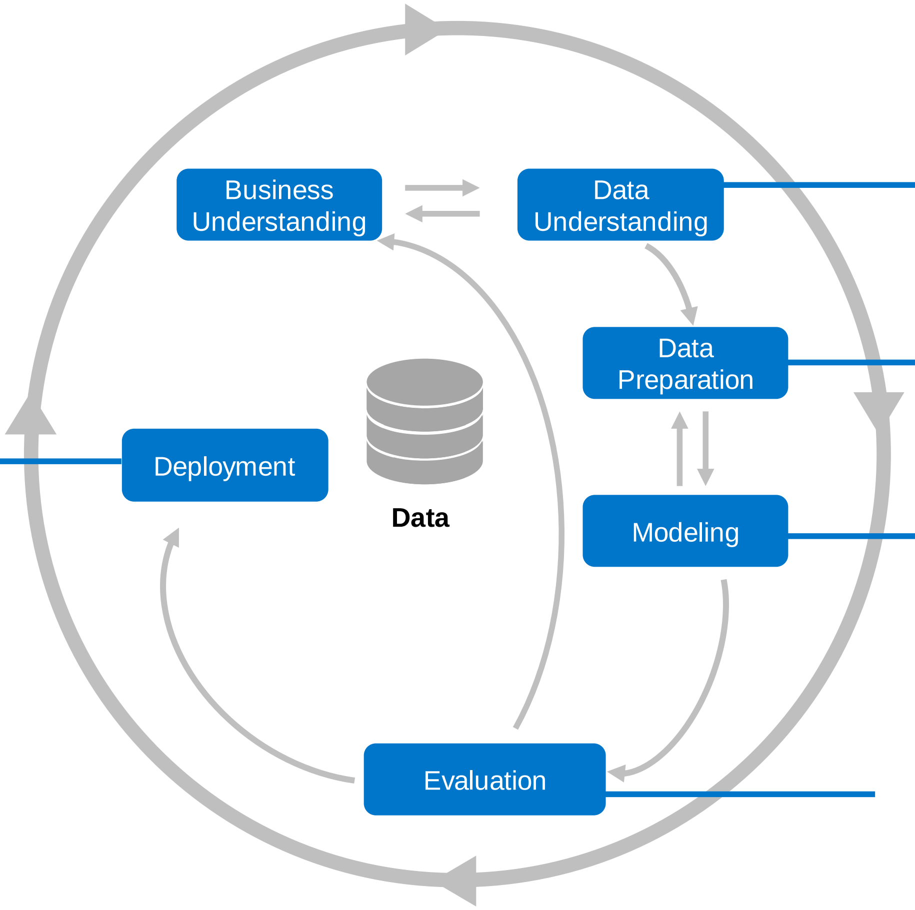
Modeling
Selecting the best model

- 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

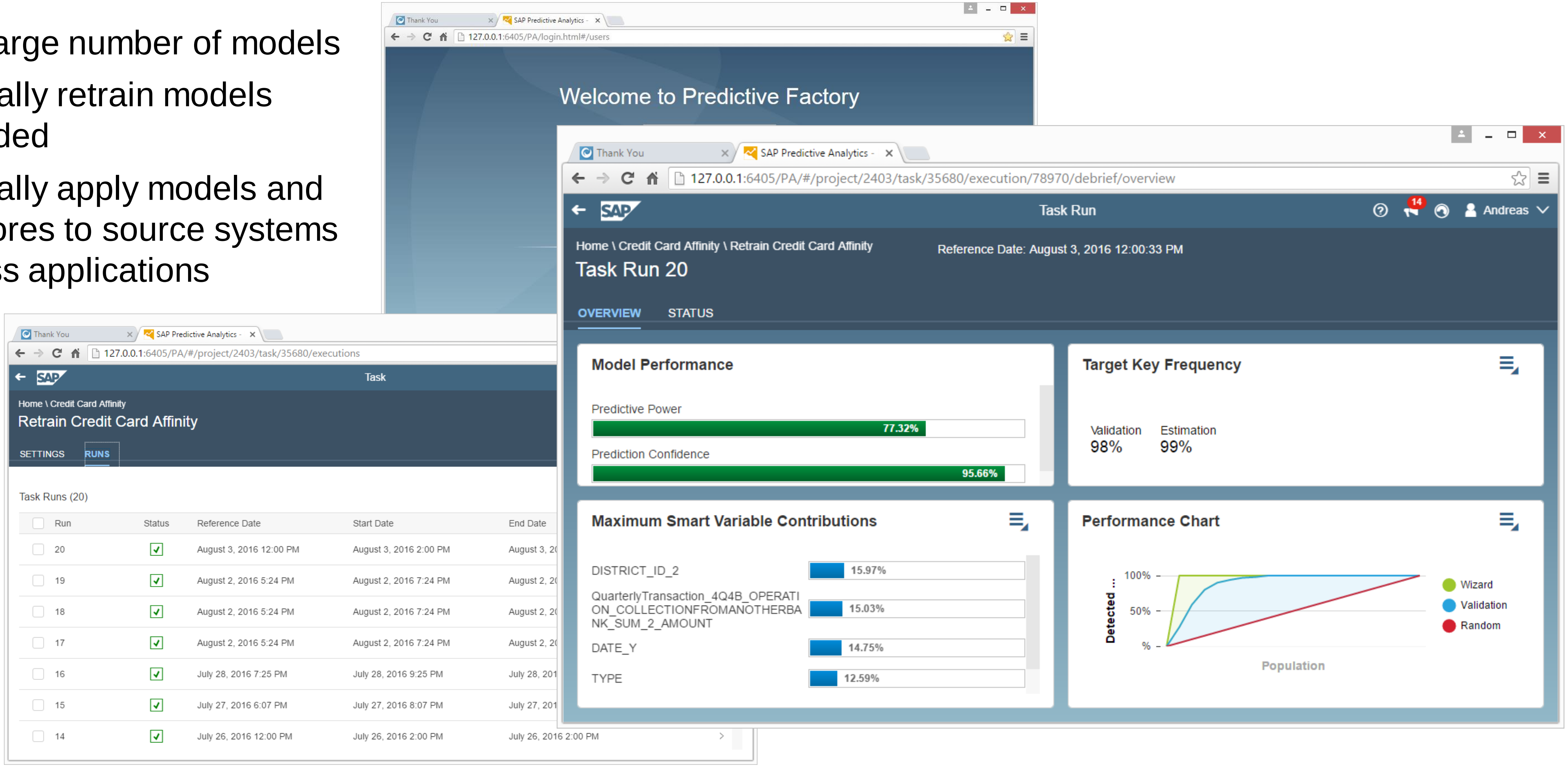
Calculate many different models

- 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

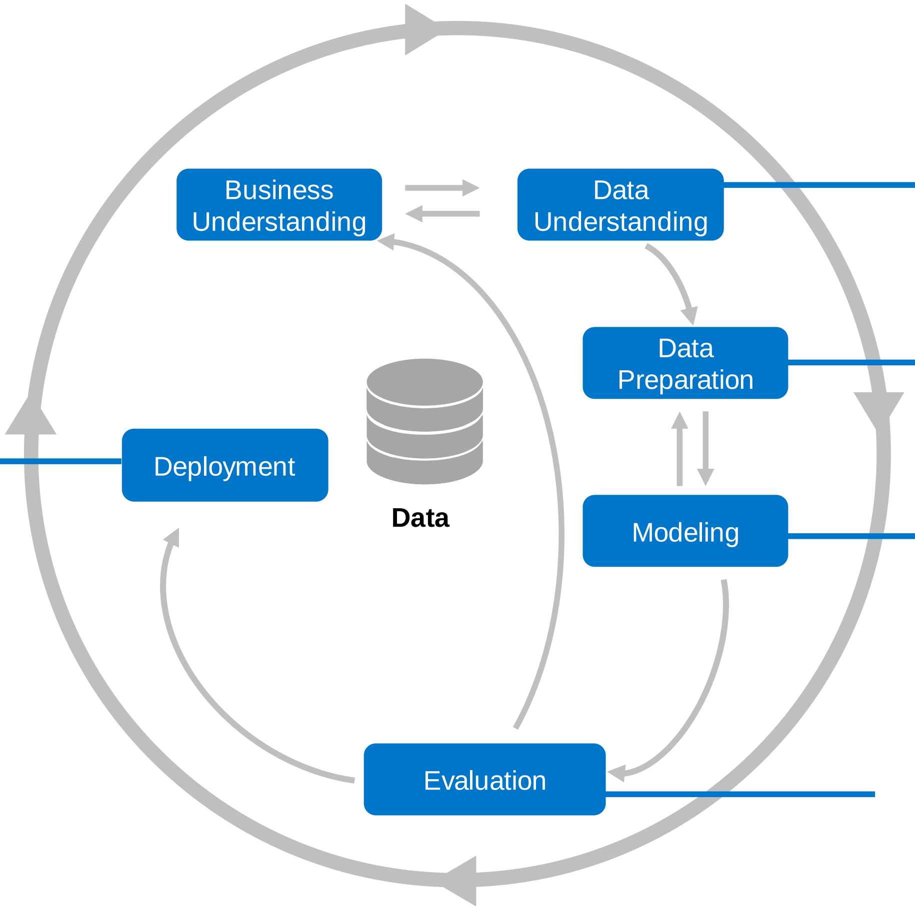
Closed Loop
Automtically Retrain and Apply Models

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



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

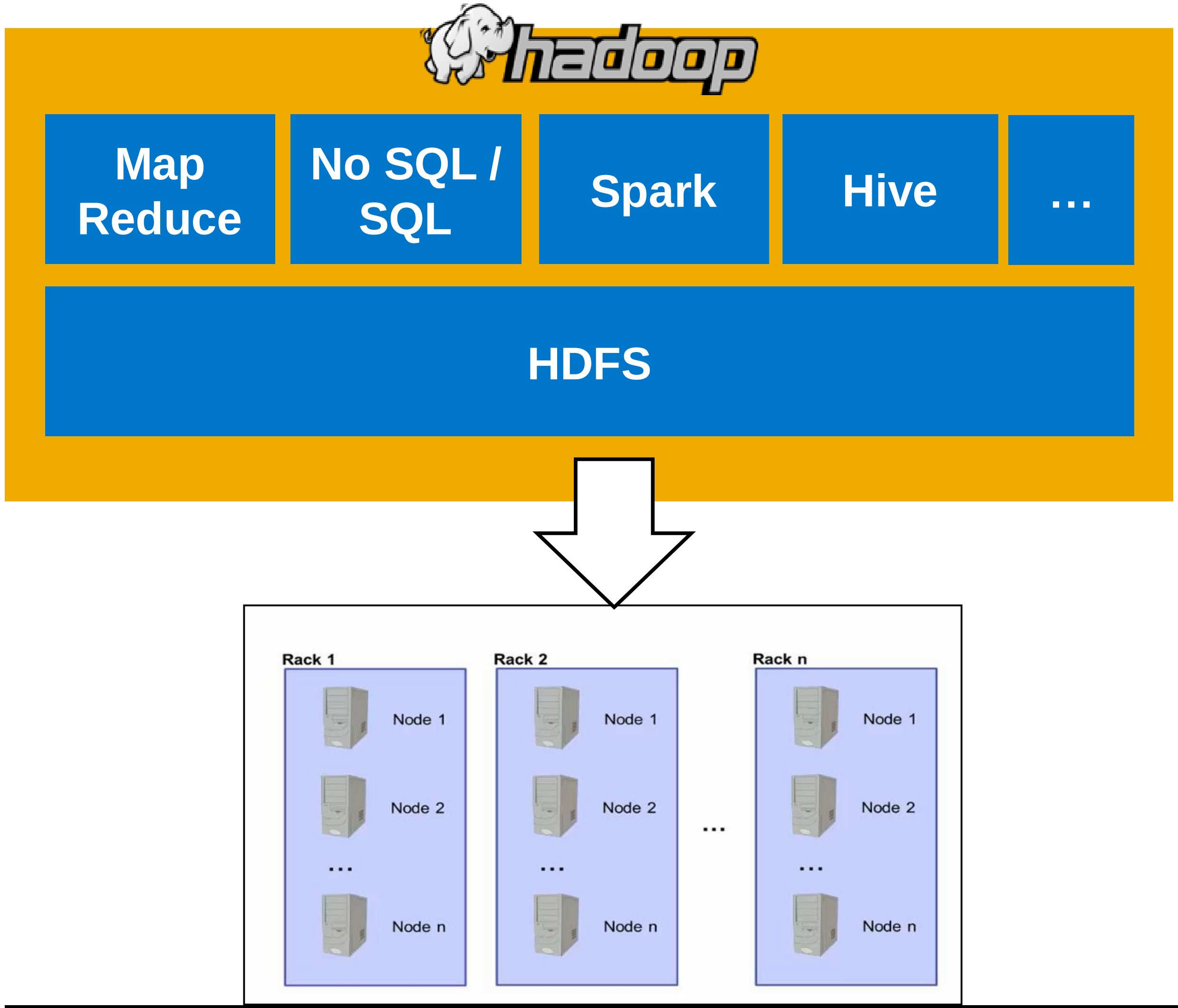
- Derive new variables in graphical interface that describe the subject
- Handle missing values and outliers
- Create robust groups

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

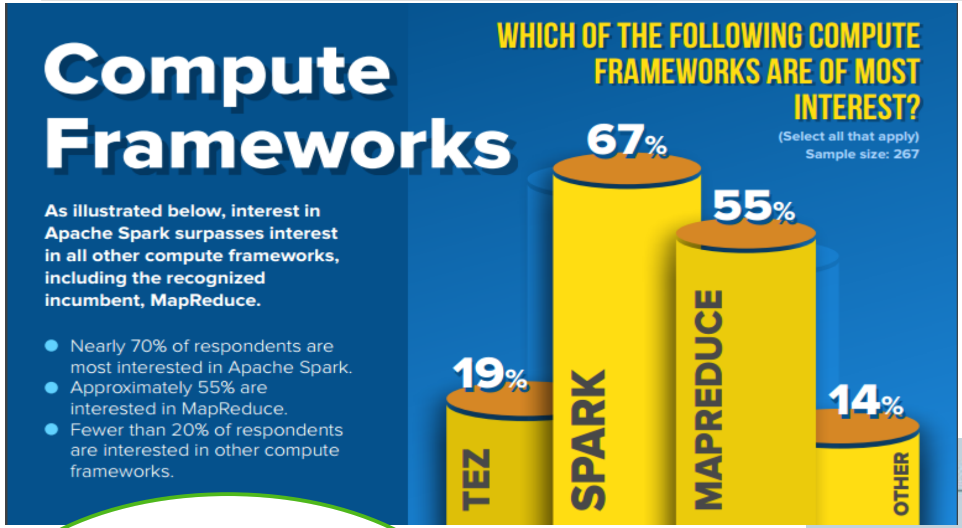
Big Data in Hadoop



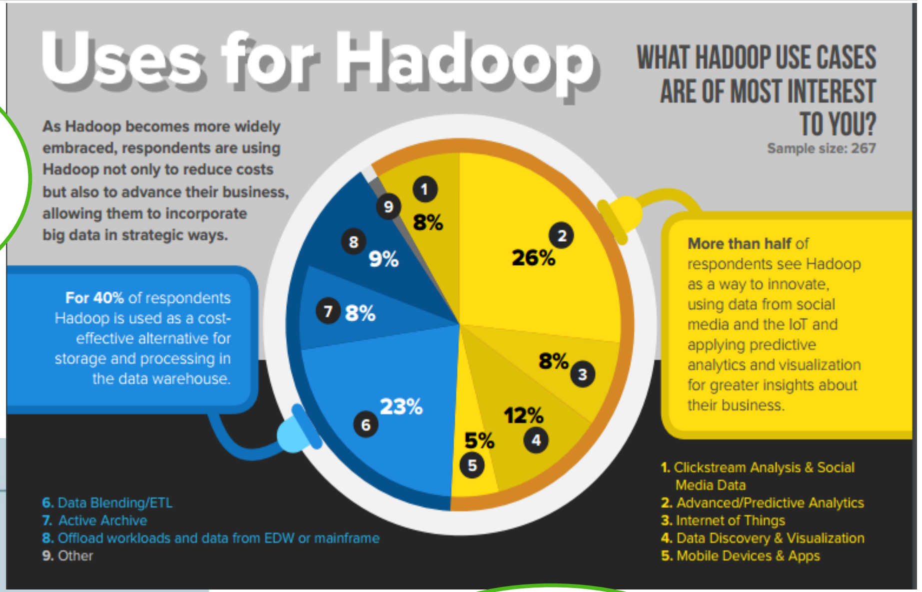
Features

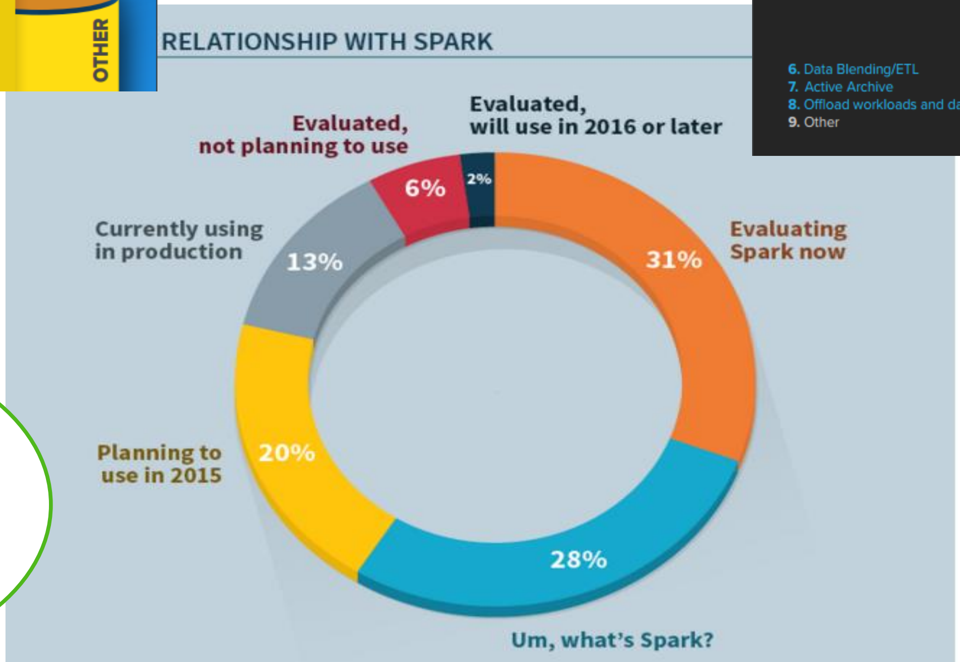
- Commodity Hardware ($1500/ TB)
- Open Source Stack (No Licensing fee)
- Elastic scaling
- scales linearly with # of nodes
- Easy to add 1000s of (cheap) nodes
- Code executes close to the data

Hadoop Perspective for 2016



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





Adoption interest for Spark has **topped** in Hadoop eco-system

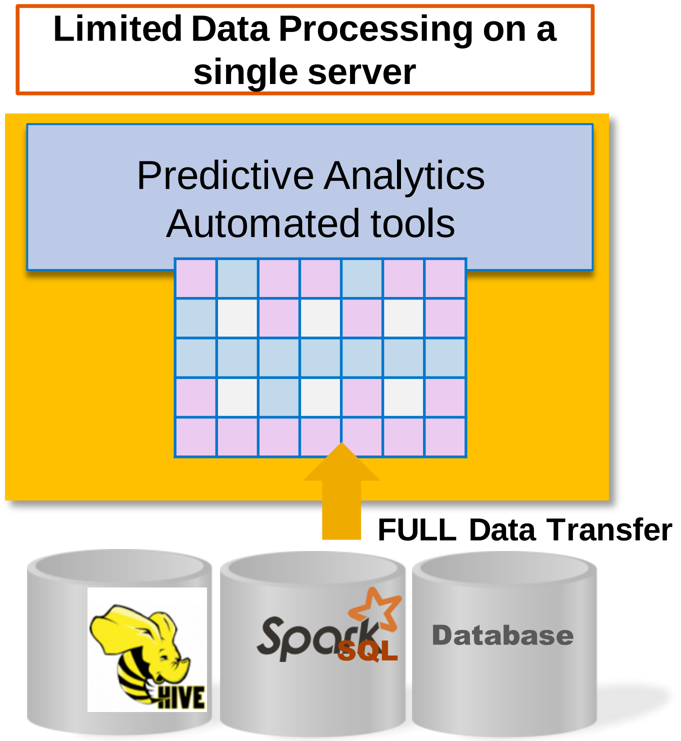
55% users want to leverage Hadoop for Business users and Advanced use cases

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

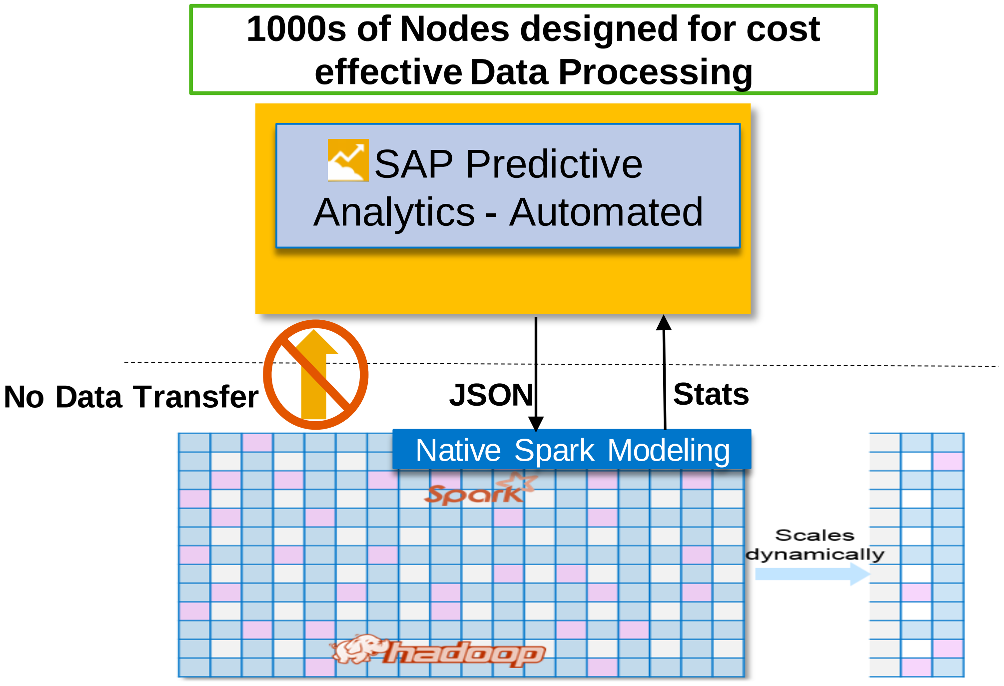
Source: http://www.syncsort.com/

Modeling for Big Data
Traditional Tiered Architecture vs. Native Spark Modeling

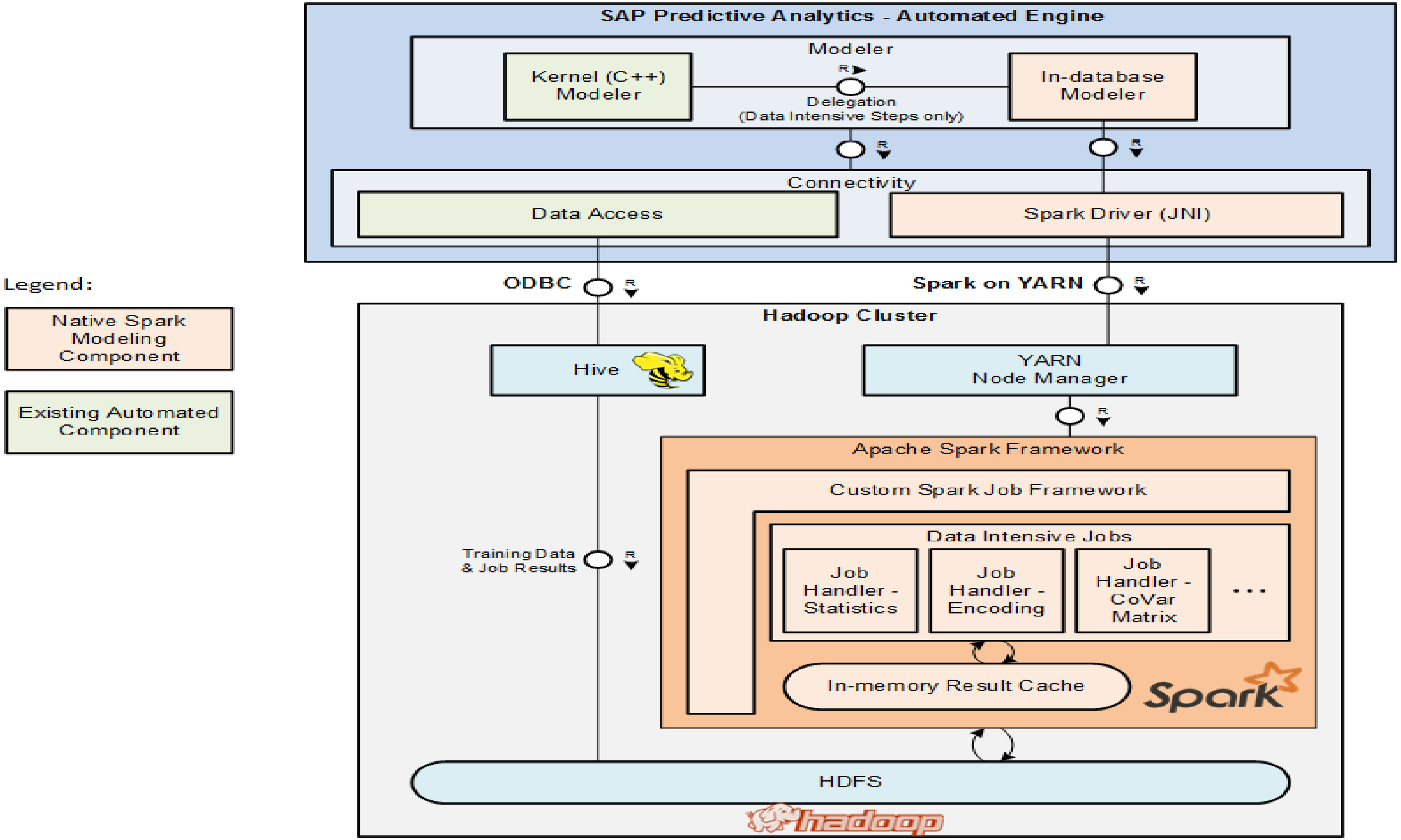
- Full dataset brought to application for processing
- Limited Performance, Scalability



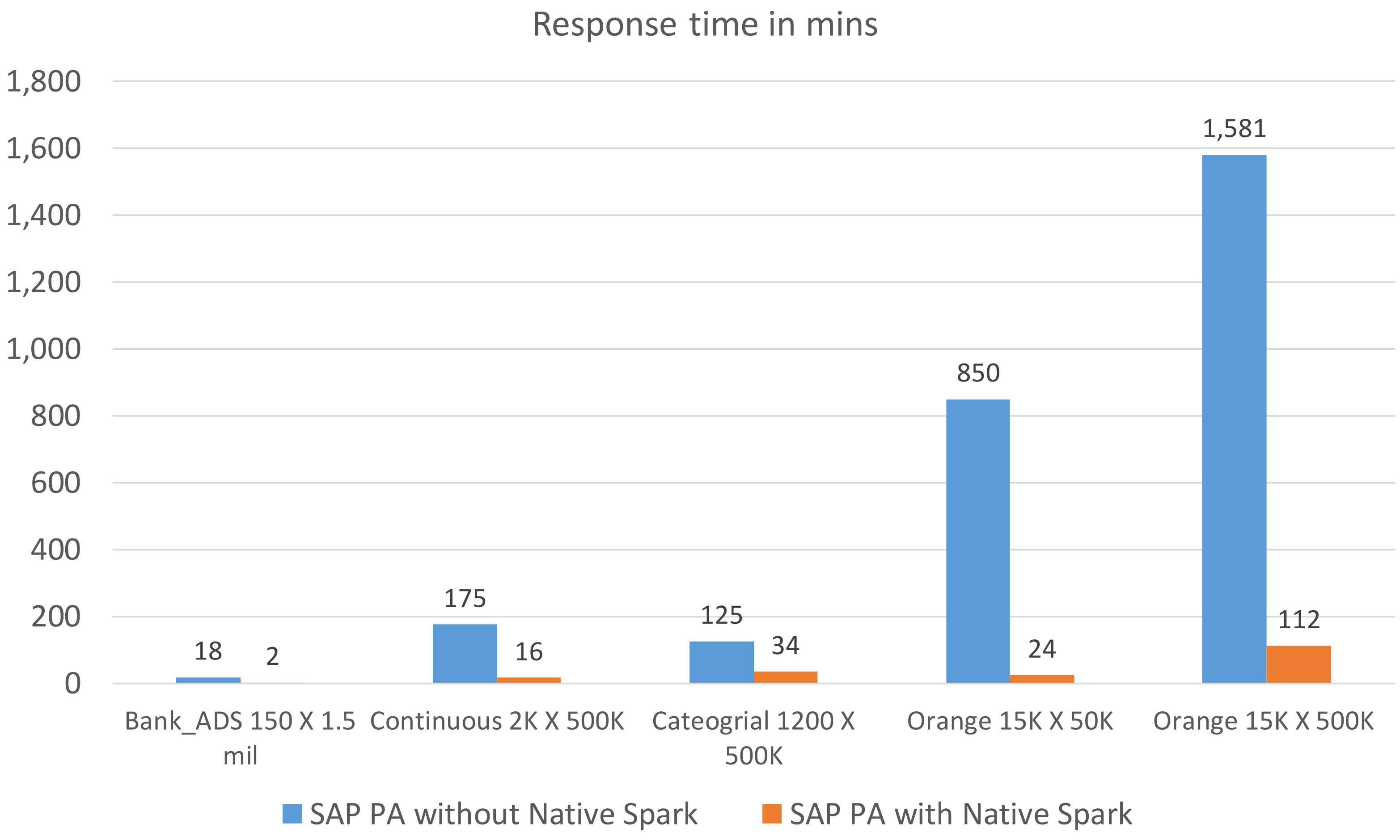
- Data processing beside data
- Performance and scalability built-in



Native Spark Modeling - Architecture



Performance and Scalability
With and Without Native Spark Modeling



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

- Increased Data Efficiency
- Predictive Analysis opened to Additional users
- Not just machine learning but a Framework
- Big Data automation shortens impl. time
- Higher Performance and Scalability
- No additional installation on Hadoop

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

