Analítica Avanzada: Un Enfoque Metodológico usando KNIME y CRISP-DM

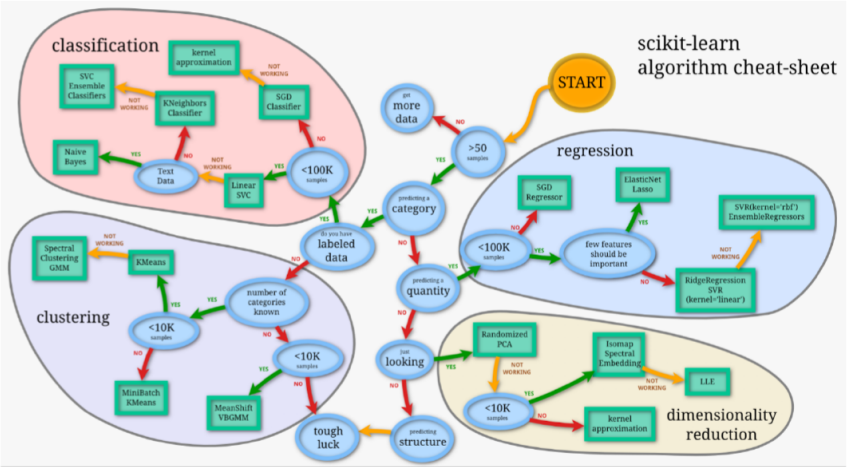
victor.toledo@coredevx.com
R+D Leader

Outline

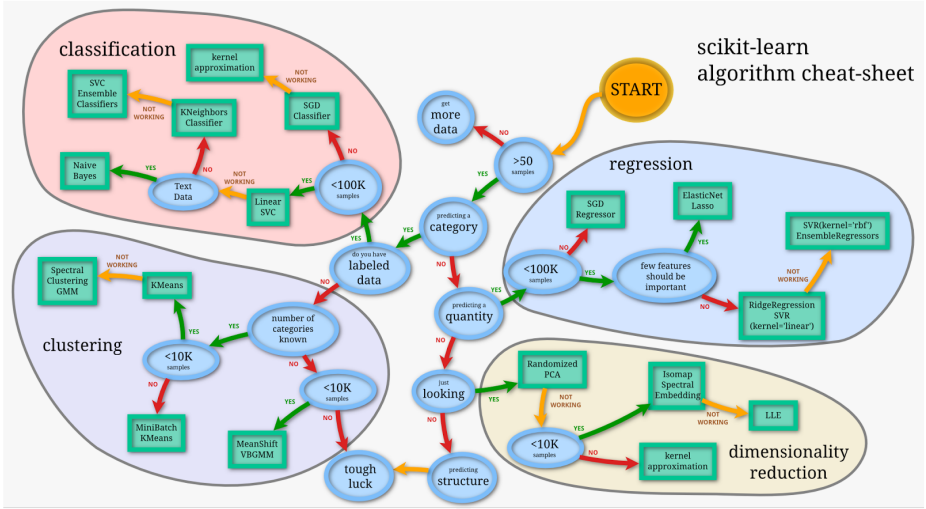
- Qué (creemos) es Analítica Avanzada (AA)
- Que Desafíos / Dimensiones existen en AA (Talento, Implementación, Infraestructura)
- Cuál es el Retorno con AA: Ecuación de ROI
- Aproximaciones a Desafíos en cada Dimensión (KNIME, CRISP-DM, DS-CI/CD)

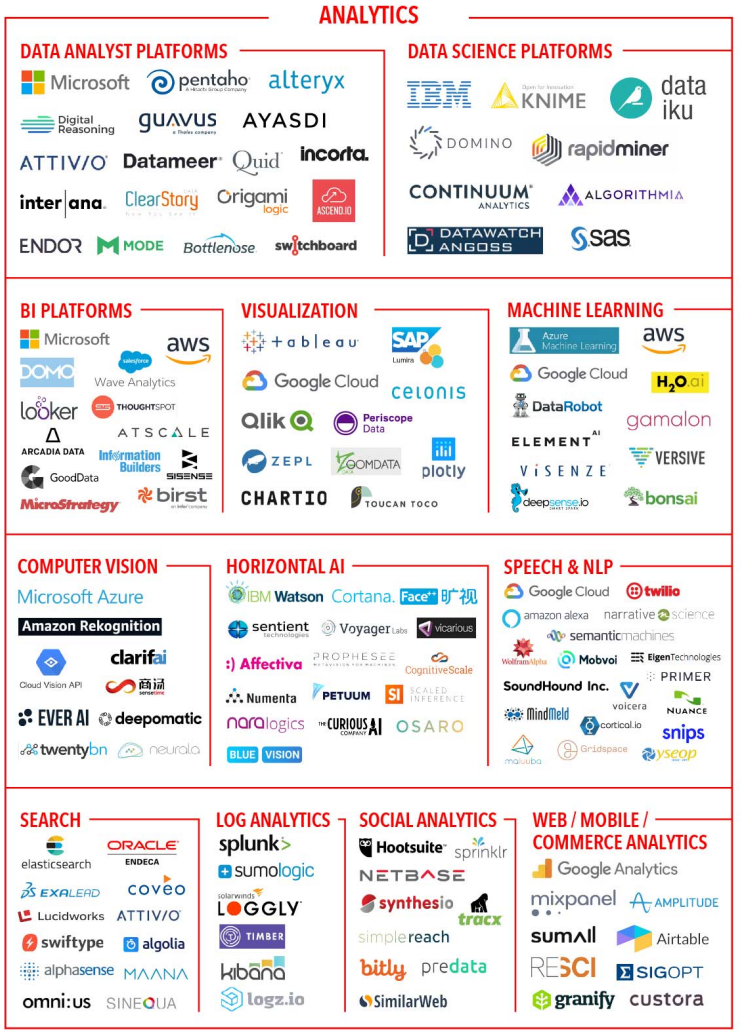
Analítica Avanzada (AA)

	Business Intelligence	Advanced Analytics
Orientation	Rearview	Future
Types of questions	What happened When, who, how many	What will happen? What will happen if we change this one thing? What's next?
Methods	Reporting (KPIs, metrics) Automated Monitoring/Alerting (thresholds) Dashboards Scorecards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Predictive Modeling Data Mining Text Mining Multimedia Mining Descriptive Modeling Statistical / Quantitative Analysis Simulation & Optimization
Big Data	Yes	Yes
Data types	Structured, some unstructured	Structured and Unstructured
Knowledge Generation	Manual	Automatic
Users	Business Users	Data scientists, Business analysts, IT, Business Users
Business Initiatives	Reactive	Proactive

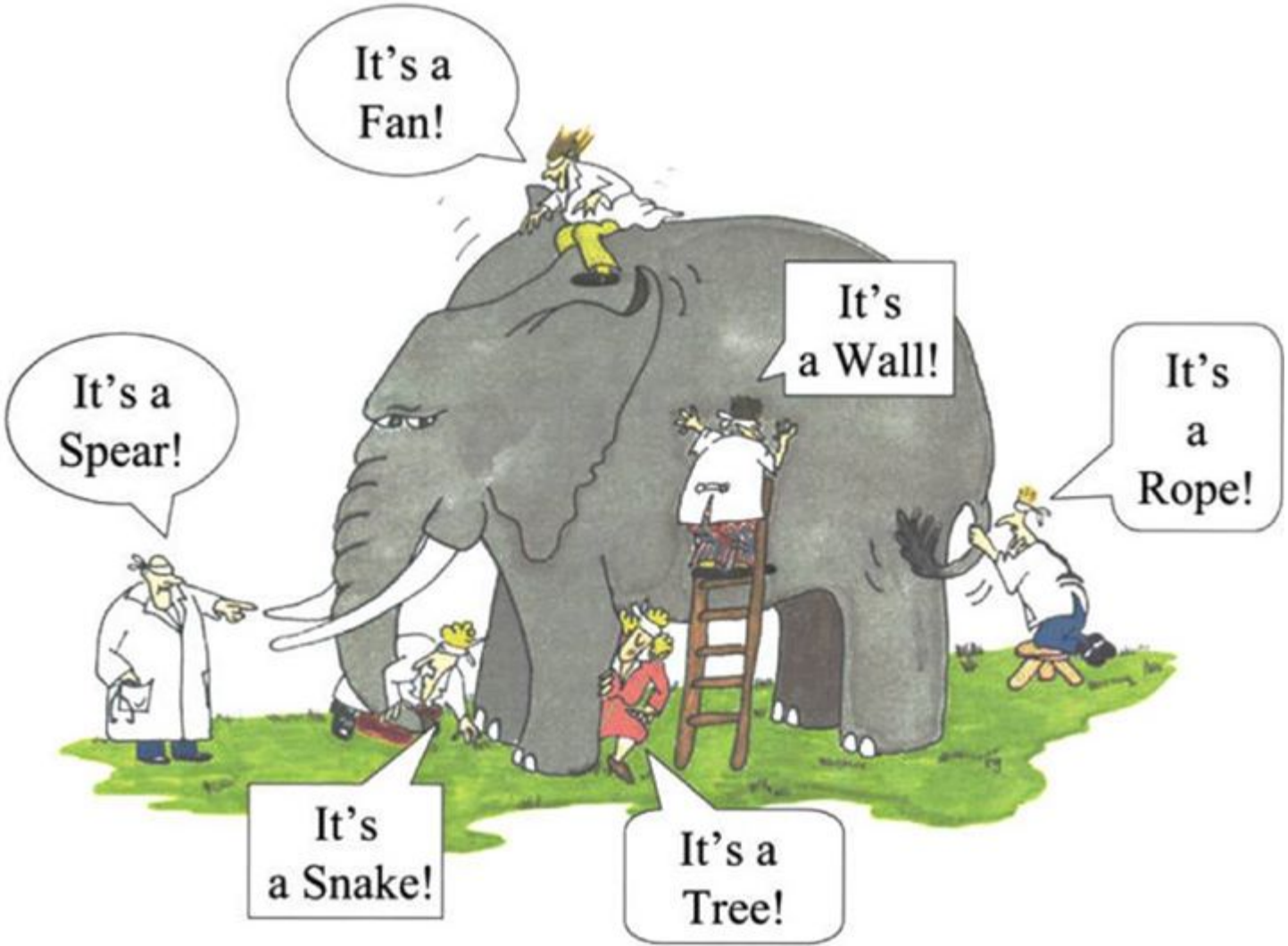


AA: Landscape





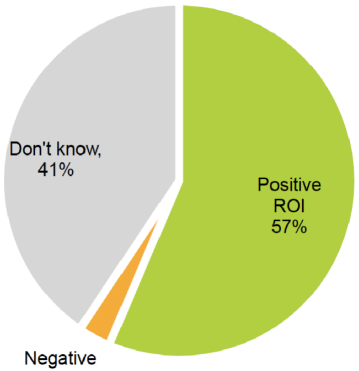
Hay un Elefante en la Sala!


It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!

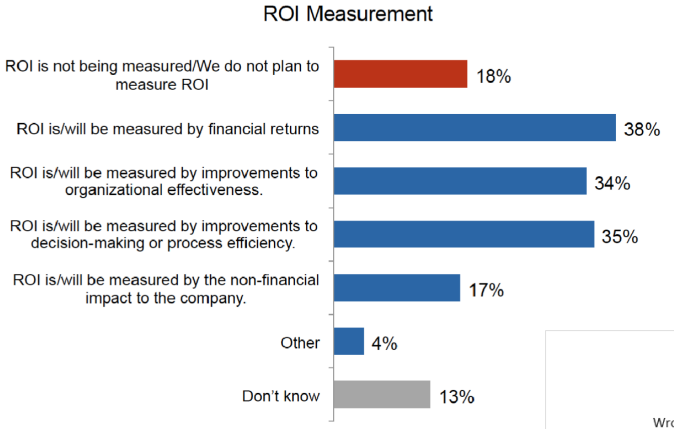
El Desafío:
Entregar valor a la Industria usando AA en tiempo y recursos definidos

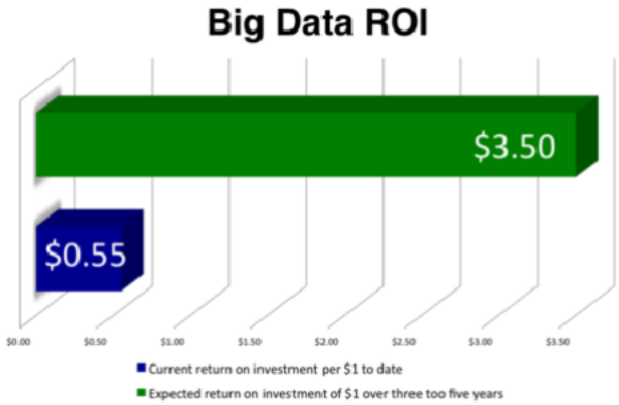
41% of Organizations Don't Know if Big Data ROI Will Be Positive or Negative

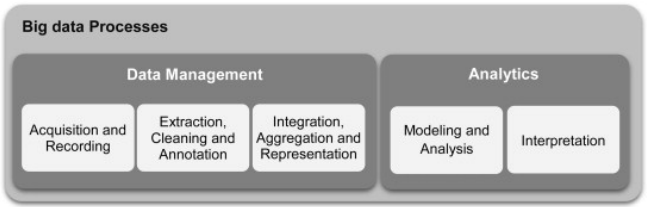
Please indicate whether your organization's big data investment has /is expected to have - a positive or negative ROI?

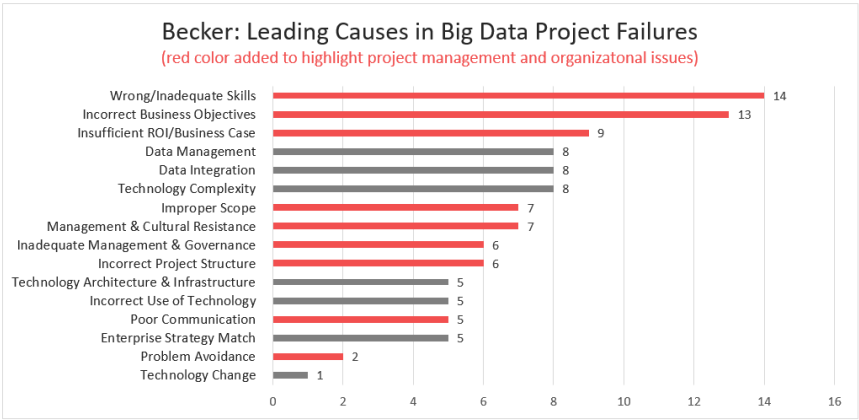


How is ROI being measured/How will ROI be measured for your organization's big data investment?









AA: ROI Equation

[Cost of Talent] x [Time Spent] + [Overall Infrastructure]

- Requires Specialists
- Requires Coding
- Requires an Ecosystem of Tools

First, identify the talents you need to have access to:

- Project management
- Data wrangling
- Data analysis
- Subject expertise
- Design
- Storytelling

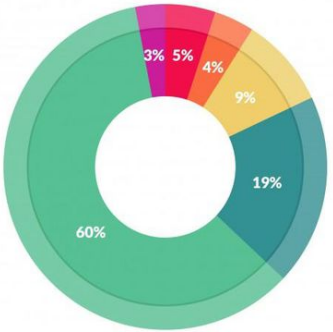
Next, map talents to team members:

Person	Talent	Person	Talent
Anand		Roberto	
Cameron		Stephani	
Emily		Susan	
Kevin		Xia-Li	

Finally, assess how much depth you have for each type of talent:

Talent	Depth
Project management	
Data wrangling	
Data analysis	
Subject expertise	
Design	
Storytelling	

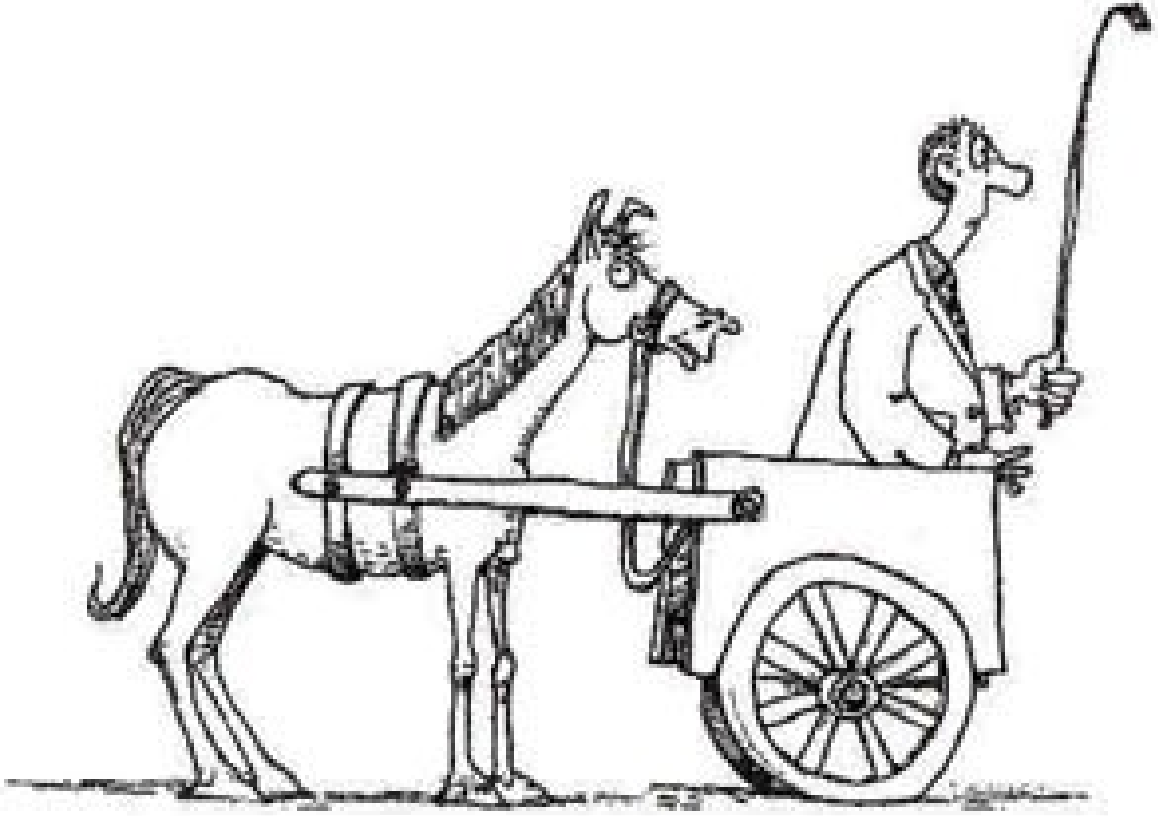
From: "Data Science and the Art of Persuasion," by Scott Berinato, January–February 2019 HBR



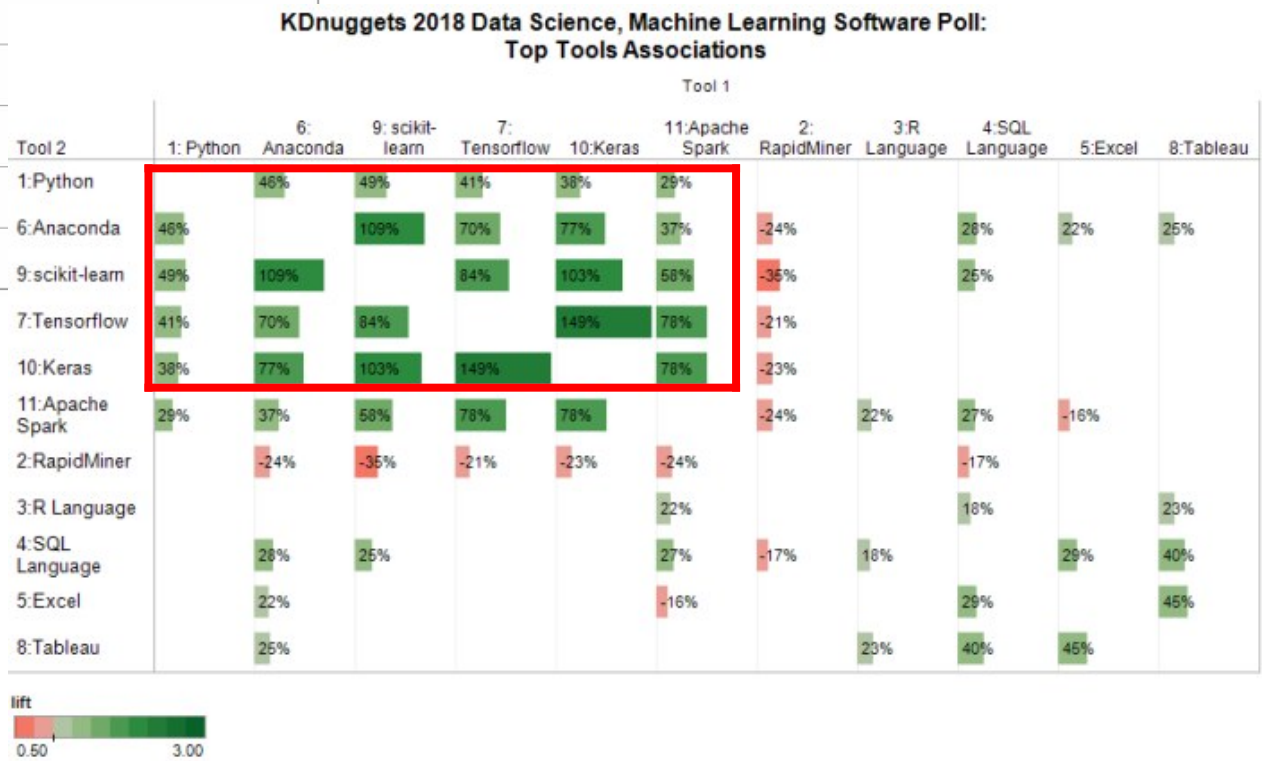
What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Una Respuesta: Definir un Framework Robusto, y actuar de acuerdo a éste



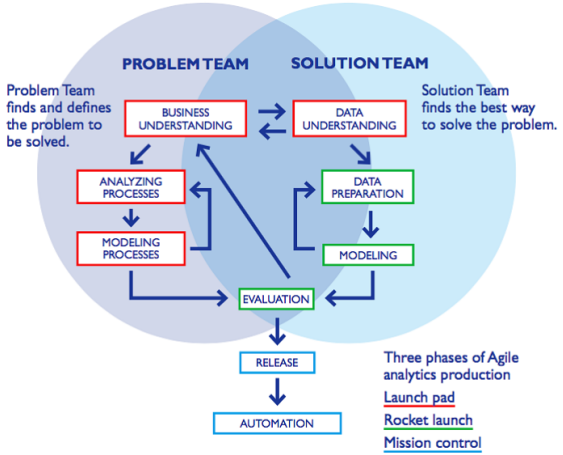
What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total] 2014 poll 2007 poll	
CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

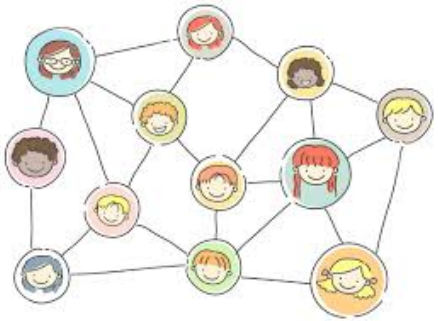


Framework

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

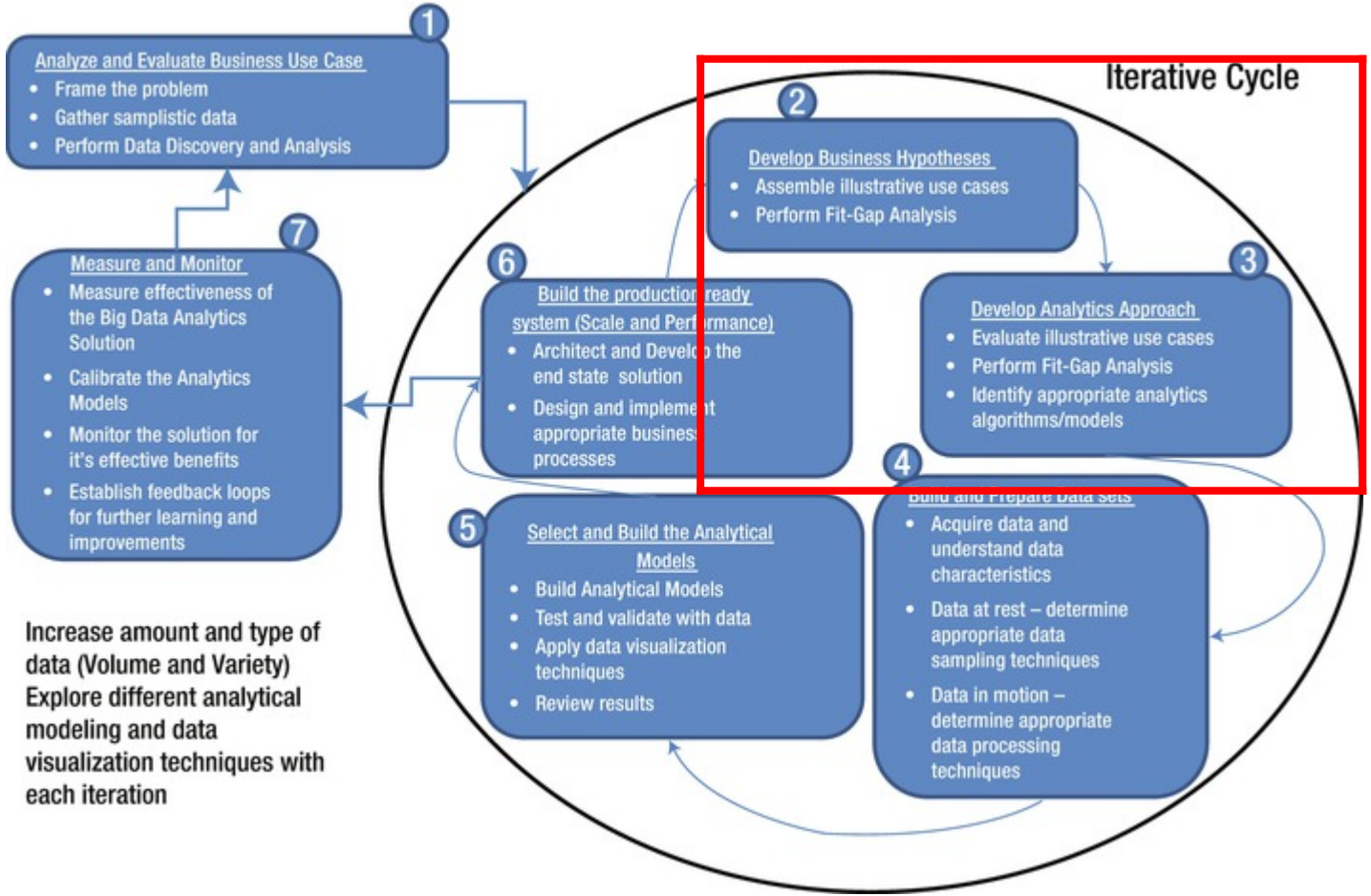




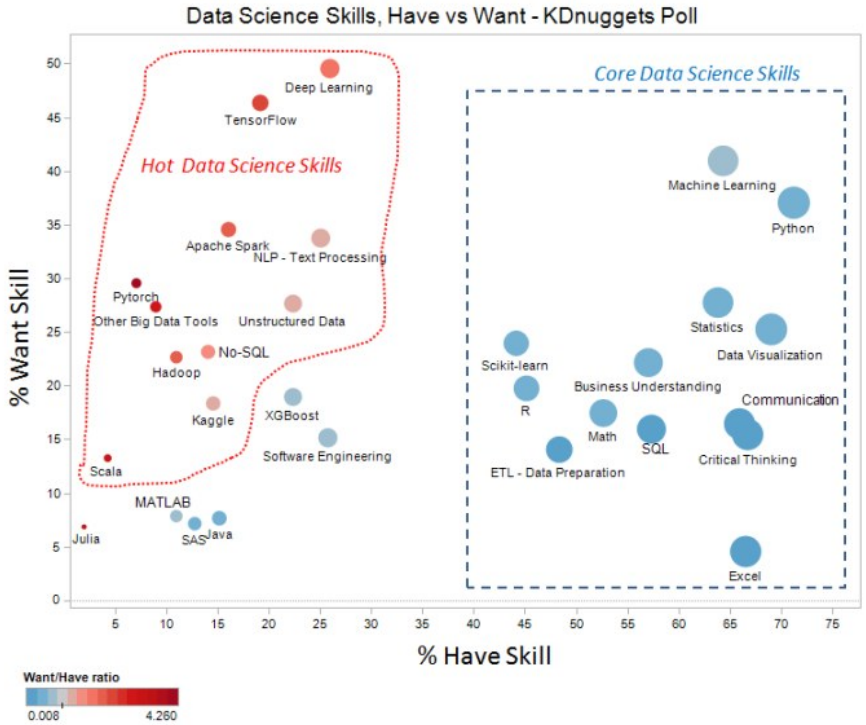
Metodología: CRISP-DM

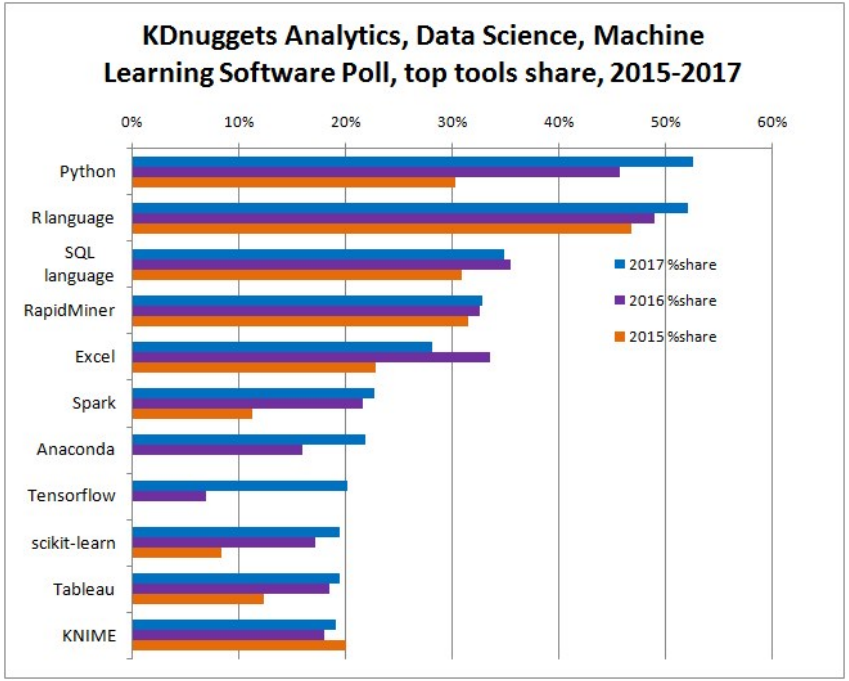
- Define pasos generales (As-Appropiate)
- No específica en Entregables.
- No específica en Artefactos
- No específica en Roles y Resposabilidades

Requerimos especificarlo para asegurar tiempos y recursos involucrados


1
Analyze and Evaluate Business Use Case
• Frame the problem
• Gather samplistic data
• Perform Data Discovery and Analysis
Iterative Cycle
2
Develop Business Hypotheses
• Assemble illustrative use cases
• Perform Fit-Gap Analysis
3
Develop Analytics Approach
• Evaluate illustrative use cases
• Perform Fit-Gap Analysis
• Identify appropriate analytics algorithms/models
4
• Acquire data and understand data characteristics
• Data at rest – determine appropriate data sampling techniques
• Data in motion – determine appropriate data processing techniques
5
Select and Build the Analytical Models
• Build Analytical Models
• Test and validate with data
• Apply data visualization techniques
• Review results
6
• Architect and Develo the end state solution
• Design and implemen appropriate busines processes
7
Measure and Monitor
• Measure effectiveness of the Big Data Analytics Solution
• Calibrate the Analytics Models
• Monitor the solution for it's effective benefits
• Establish feedback loops for further learning and improvements
Increase amount and type of data (Volume and Variety)
Explore different analytical modeling and data visualization techniques with each iteration

AA: Skills+Tools

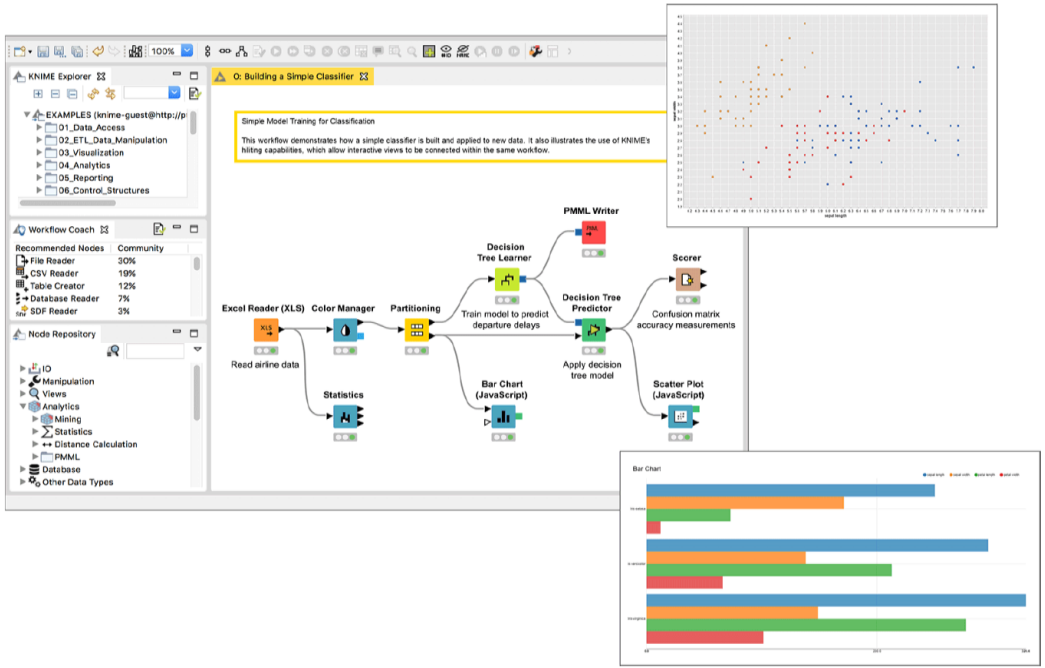




What is KNIME Analytics Platform?

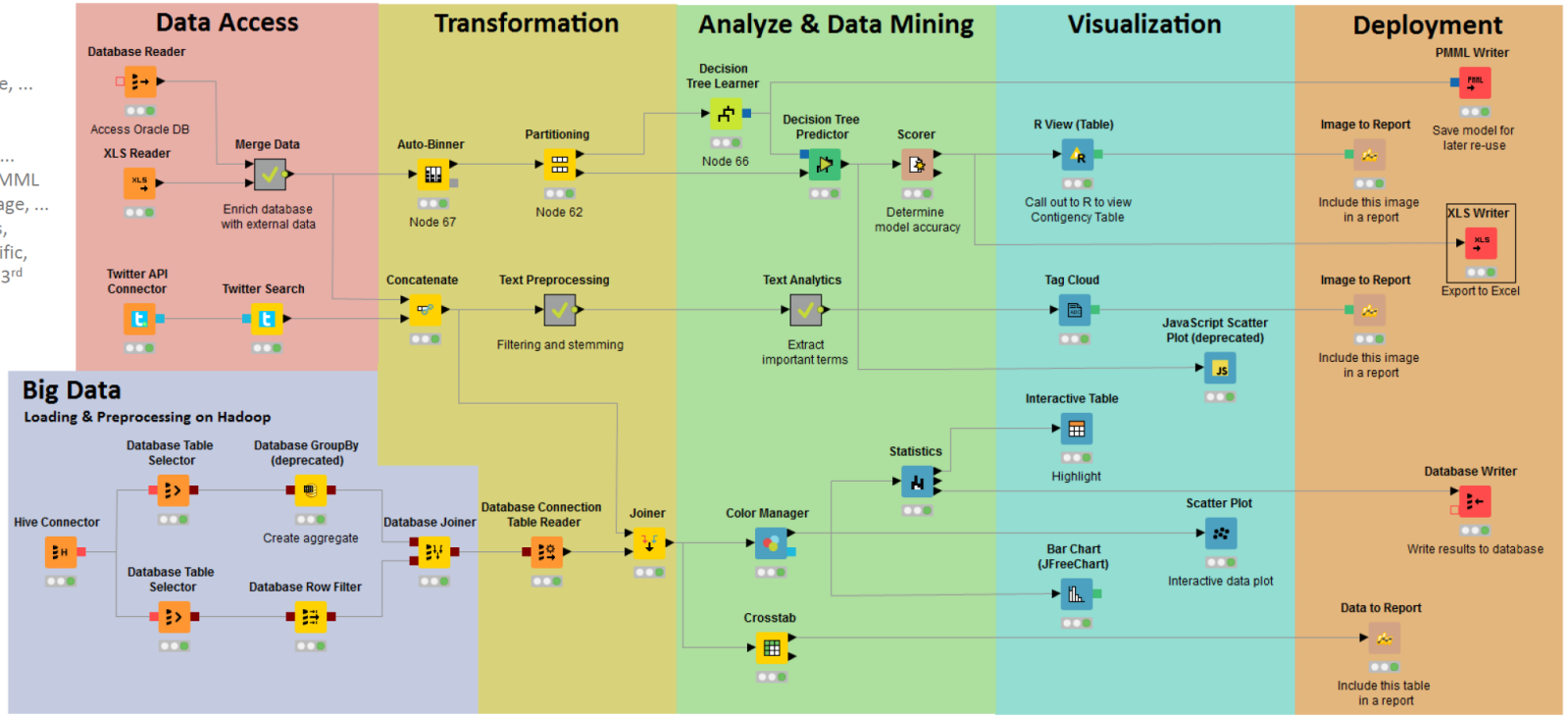
- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Big Data
  - Many integrations,

such as Java, R, Python, Weka, H2O, etc.



Over 2000 native and embedded nodes included:

Data Access
MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers,
Industry Specific,
Community / 3rd party ...



Big Data
Hive, Impala, HDFS Vertica,
Teradata/Aster, Spark, MLlib,
Community / 3rd party, ...

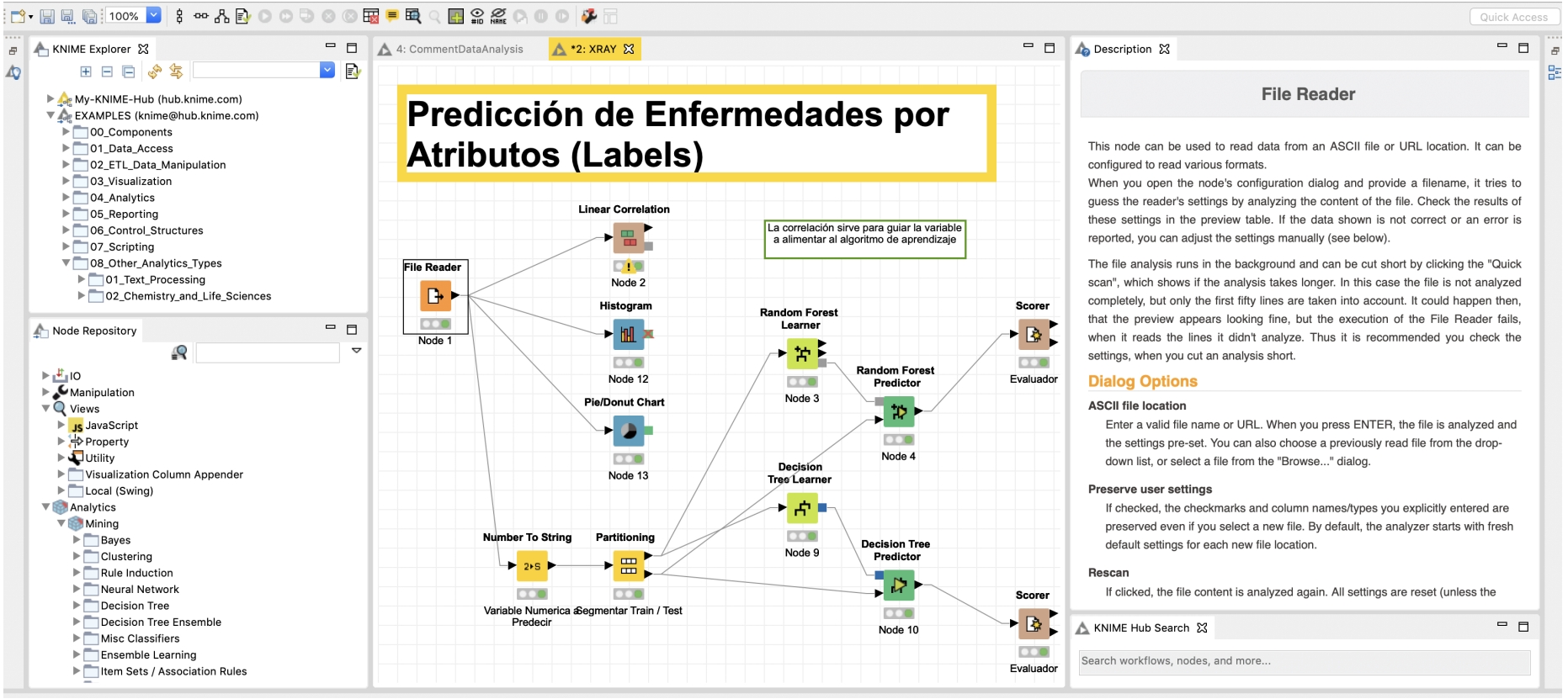
Transformation
Row, Column, Matrix
Text, Image, Networks, Time
Series, Java, Python,
Community / 3rd party, ...

Analysis & Mining
Statistics, Machine Learning, Data
Mining, Web Analytics, Text
Mining, Network Analysis, Social
Media Analysis, R, Weka, Python,
Community / 3rd party, ...

Visualization
R, Python,
JFreeChart,
JavaScript,
Community / 3rd party, ...

Deployment
via BIRT
PMML, XML, JSON
Databases, Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd party, ...

Implementación: Reducir Codificación



Implementación: Code Oriented

Deep Learning

Assignment 1

The objective of this assignment is to learn about simple data curation practices, and familiarize you with some of the data we'll be reusing later.

This notebook uses the notMNIST dataset to be used with python experiments. This dataset is designed to look like the classic MNIST dataset, while looking a little more like real data: it's a harder task, and the data is a lot less 'clean' than MNIST.

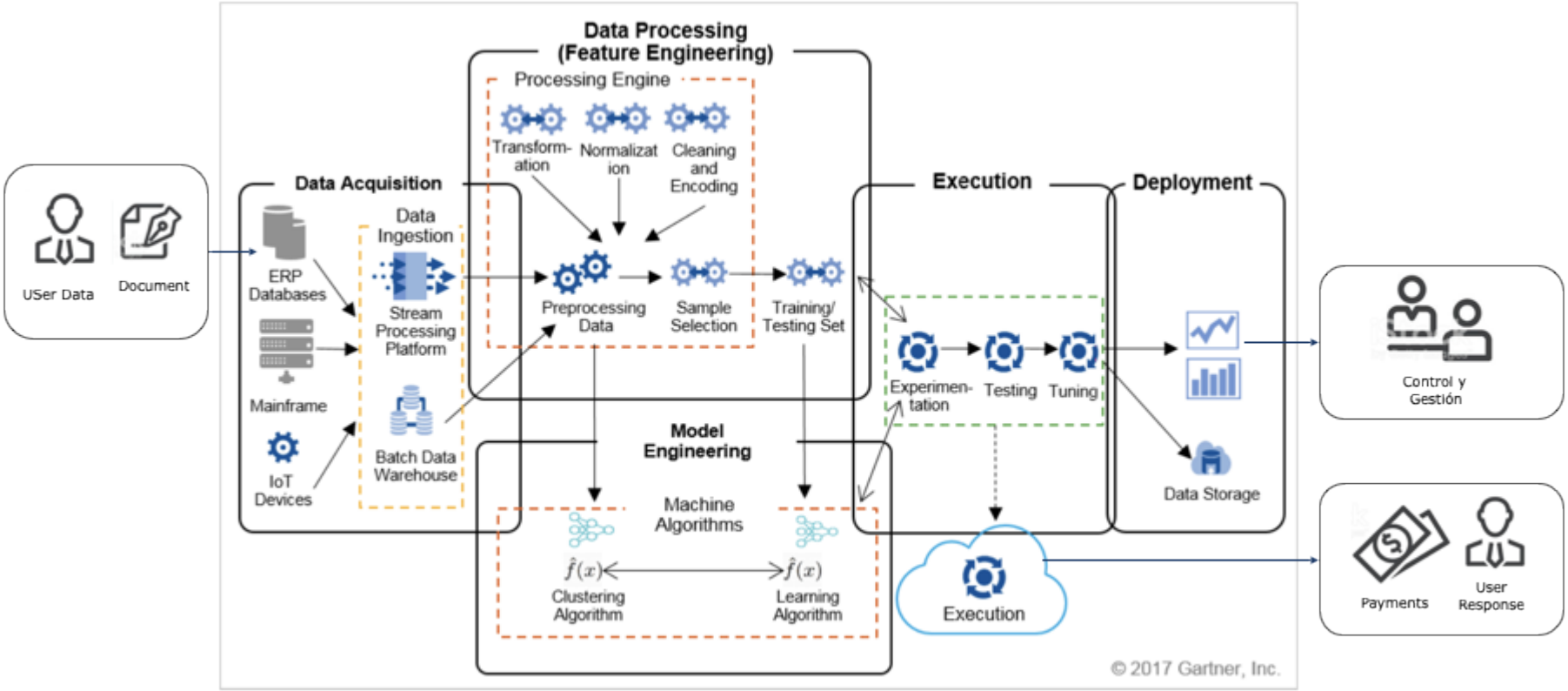
In [0]:
```
# These are all the modules we'll be using later. Make sure you can import them
# before proceeding further.
import matplotlib.pyplot as plt
import numpy as np
import os
import sys
import tarfile
from IPython.display import display, Image
from scipy import ndimage
from sklearn.linear_model import LogisticRegression
from six.moves.urllib.request import urlretrieve
from six.moves import cPickle as pickle
```

First, we'll download the dataset to our local machine. The data consists of characters rendered in a variety of fonts on a 28x28 image. The labels are limited to 'A' through 'J' (10 classes). The training set has about 500k and the testset 19000 labelled examples. Given these sizes, it should be possible to train models quickly on any machine.

In [0]:
```
url = 'http://yaroslavvb.com/upload/notMNIST/'

def maybe_download(filename, expected_bytes):
  """Download a file if not present, and make sure it's the right size."""
  if not os.path.exists(filename):
    filename, _ = urlretrieve(url + filename, filename)
  statinfo = os.stat(filename)
  if statinfo.st_size == expected_bytes:
```

AA: Infrastructure



KNIME PLATFORM

Conclusiones

- En que patrón de AA estamos? Orientado a la Programación? Orientado al Modelamiento?
- Necesitamos una **metodología** clara desde la iniciación de una idea de AA hasta la puesta en producción.
- Necesitamos medir cuantitativamente los éxitos del proyecto : ROI + Metodología.
- Necesitamos definir una **comunidad** y **prácticas locales** (No Deus Ex-Machina).

¡Gracias!
Preguntas?

victor.toledo@coredevx.com