Natural Language Processing Applied to Credit Risk Management

20/03/2018

Jean-Baptiste GHEERAERT – Jean-Baptiste JANVIER – Syrielle MONTARIOL

NLP Overview

Different Sources and Challenges

External Documents	• Pertinent synthesis for front office, credit analysts, economists, conformity warnings, …
Internal Documents	• Fraude detection • GDPR enforcement • Monitoring
Translation	• 76 countries, many subsidiaries • Understanding credit granting in subsidiaries • Subsidiaries monitoring • Security concerns: public API cannot be used

Many key questions shared by all entities

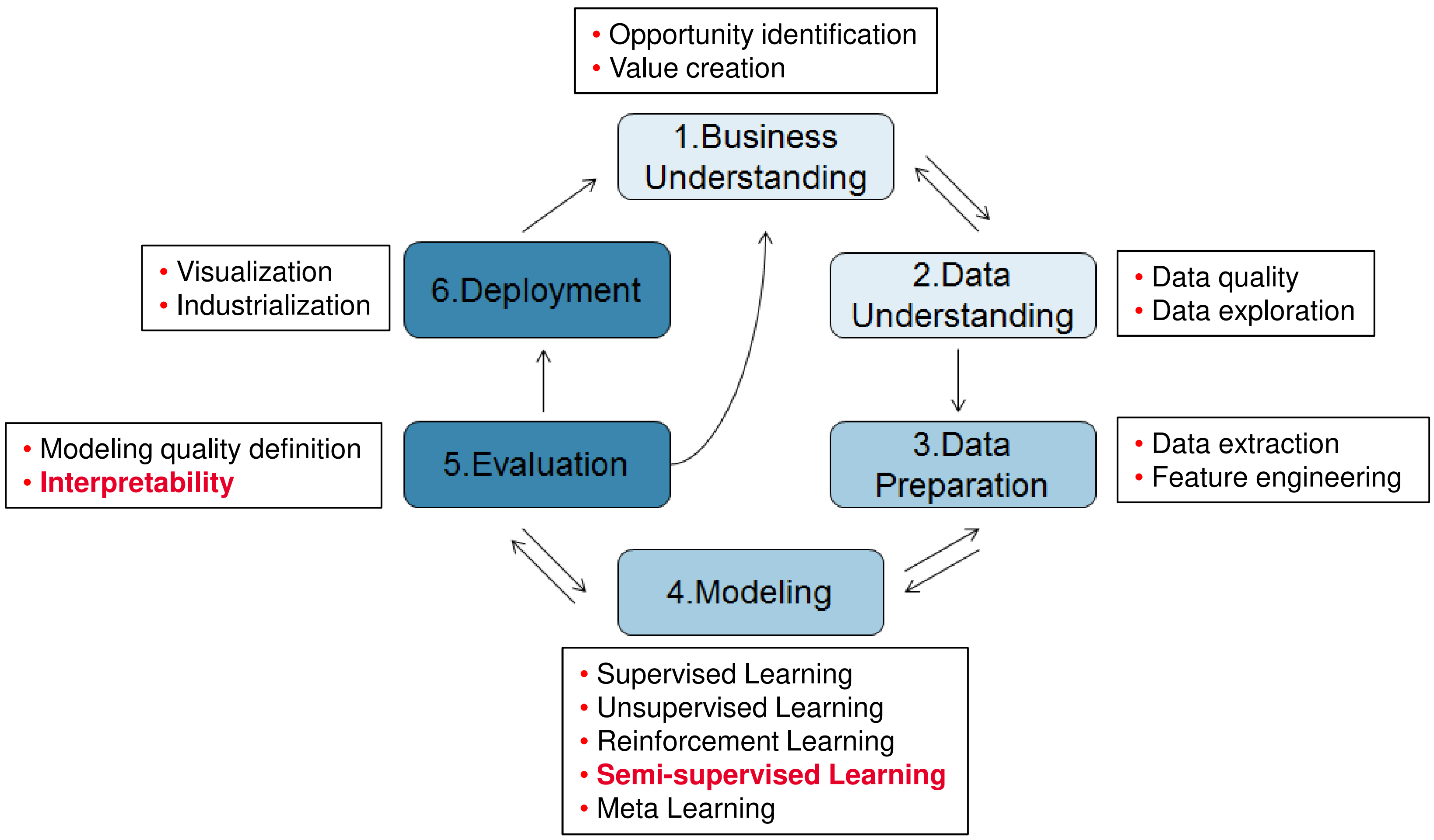
Raw data analysis
- Access
- Preprocessing
- Volume
- **Labelling**

Methodology
- Scope of the study
- Validation

Learning
- Similar algorithms but specifities to integrate for each application

Business feedbacks
- How to iterate smoothly ?
- How to assess performance ?

Data Modeling Workflow & Challenges



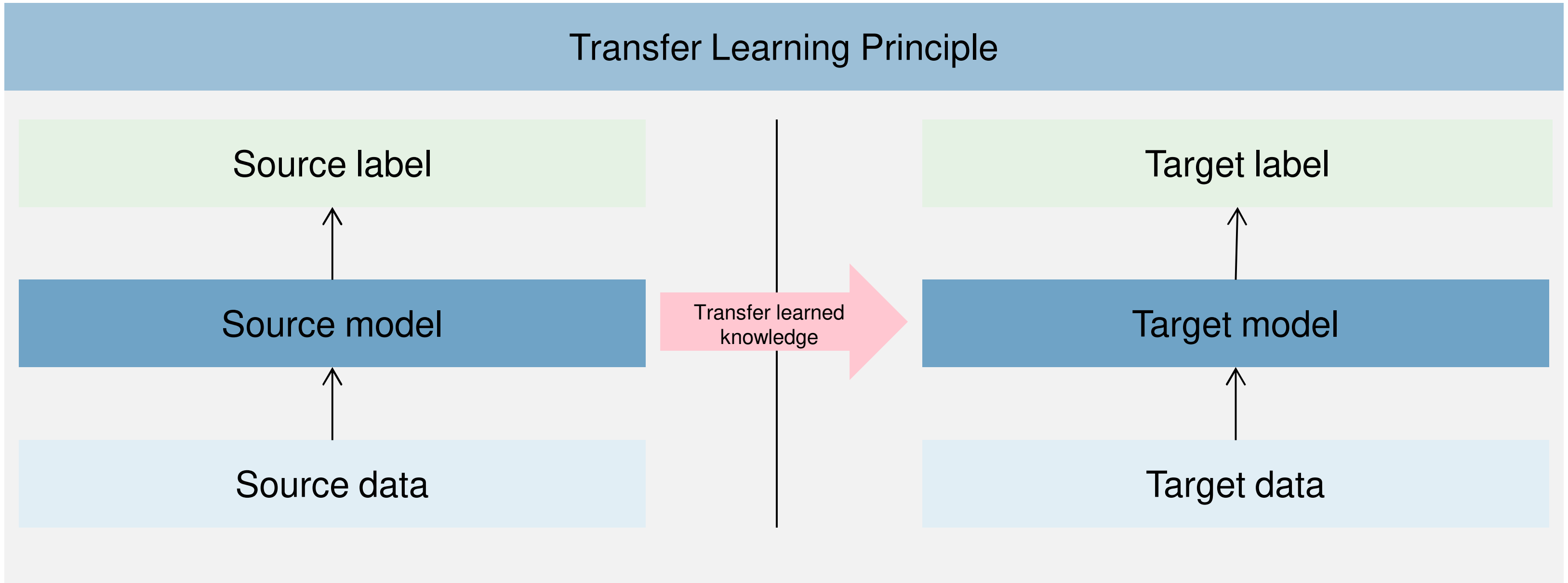
Transfer Learning

Definition

In a supervised model, Transfer Learning is the fact of using parameters trained on a dataset A for a task X (source model) on a dataset B for a task Y (target model).

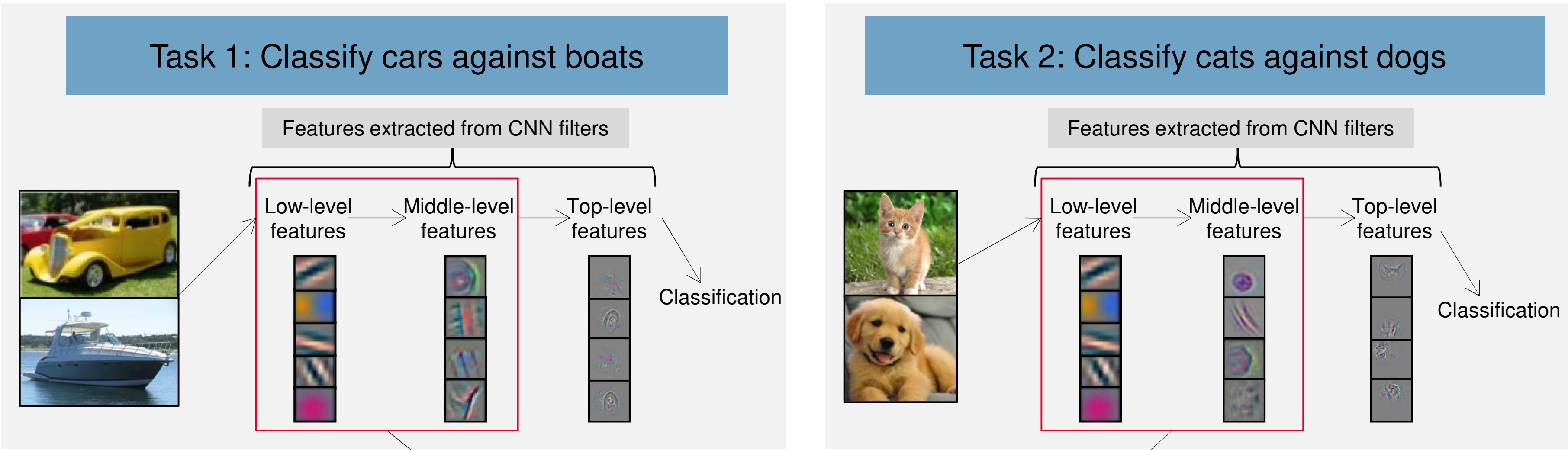
Main difficulties

- Distribution shift
- Different tasks



Transfer Learning

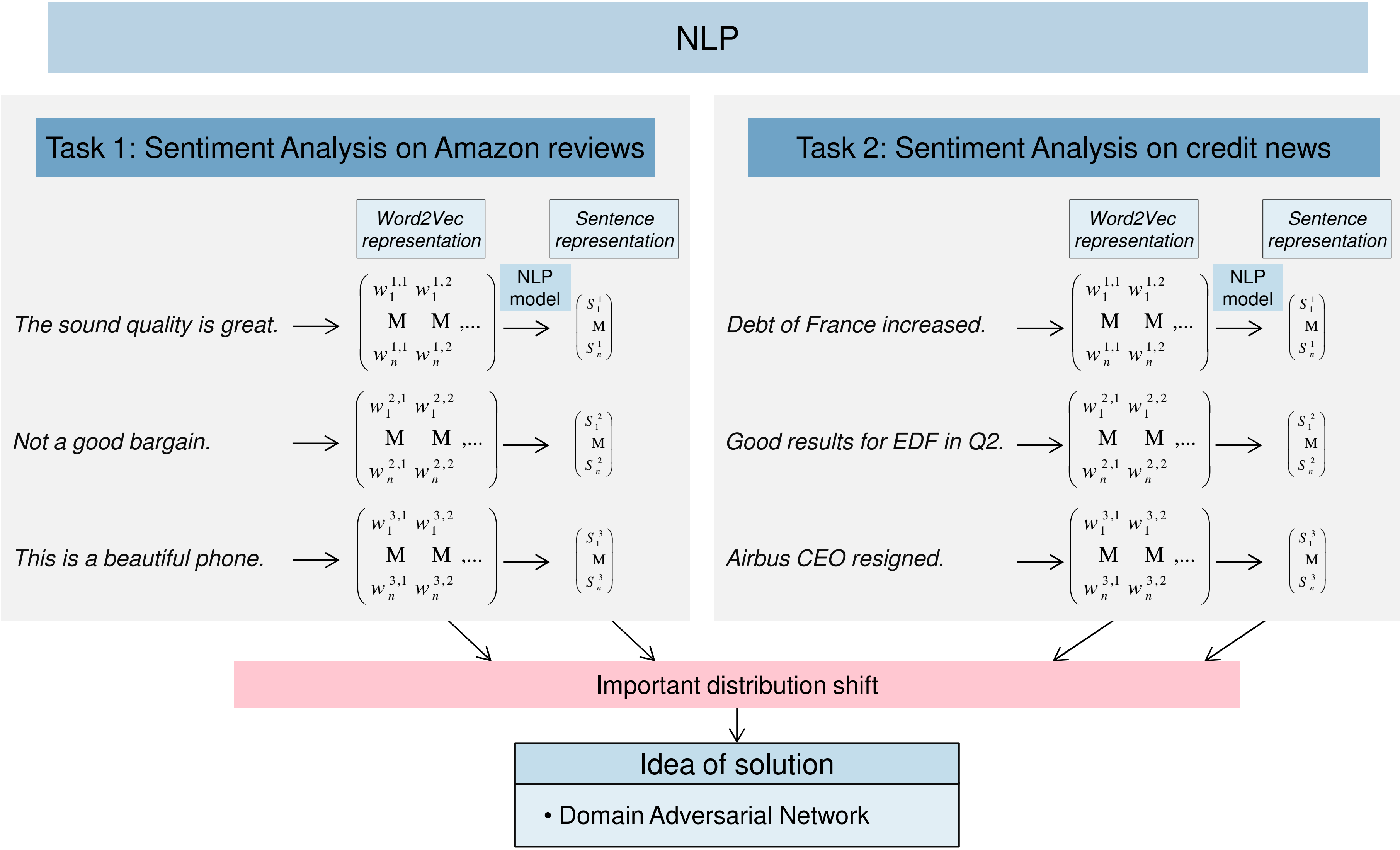
Computer Vision



Important distribution shift at an image level but similar low and middle levels features:
- Parameters of the first layers are easily transferable

***Visualizing and Understanding Convolutional Networks* by M.D. Zeiler and R. Fergus**

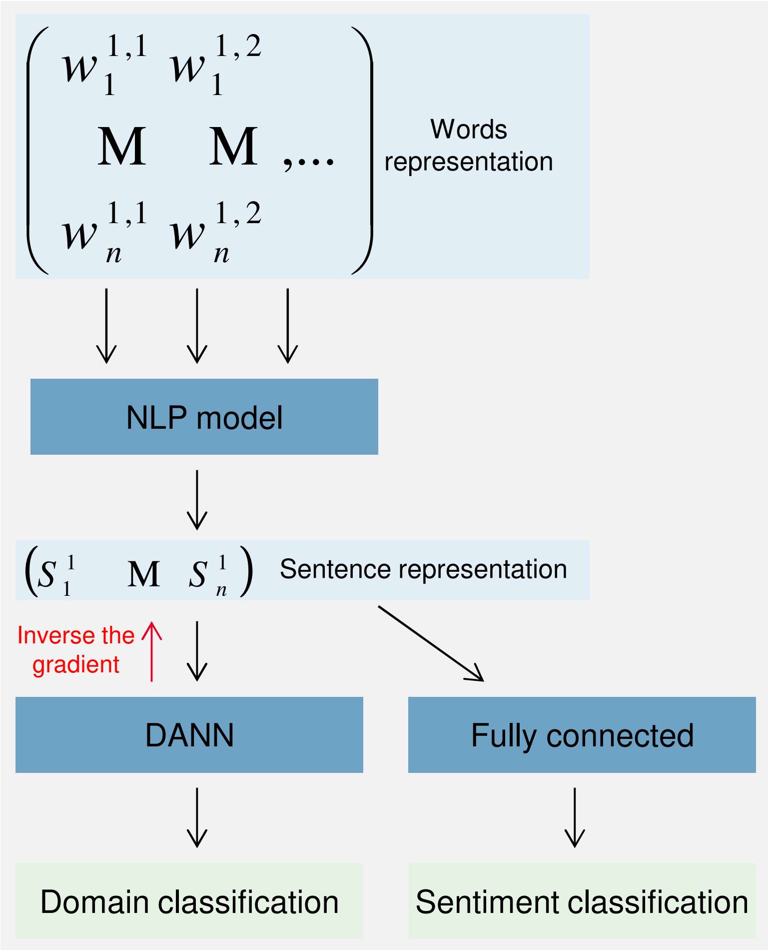
Transfer Learning



Transfer Learning

Idea

Try to find a joint represention of the source data and the target data



- Input data is:
  - labelled data from source dataset
  - unlabelled data from target dataset
- The DAN tries to classify the domain
- The inverse gradient updates the sentence representation to make it more difficult to determine the domain

Results in Computer Vision

Source	MNIST	SYN NUMBERS	SVHN	SYN SIGNS
Target	MNIST-M	SVHN	MNIST	GTSRB
Source only	0.5225	0.8674	0.5490	0.7900
DANN	**0.7666**	**0.9109**	**0.7385**	**0.8865**

Results in NLP

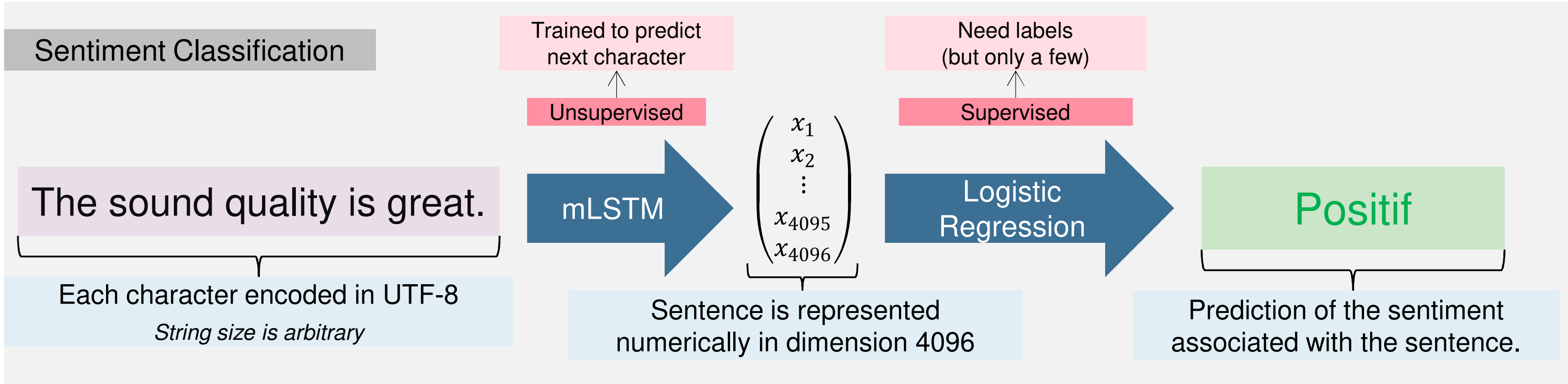
Source	Target	DANN	NN
DVD	Electronics	**0,754**	0,732
Electronics	DVD	**0,738**	0,733
Electronics	Kitchen	**0,854**	0,854
Kitchen	Electronics	**0,843**	0,841

***Domain-Adversarial Training of Neural* Networks by Y. Ganin & al.**

A semi-supervised method for Sentiment Analysis

Principle

1. Apply a first model (trained without supervision) to represent the sentence as a vector.
2. Use those vectors as input data of a logistic regression.

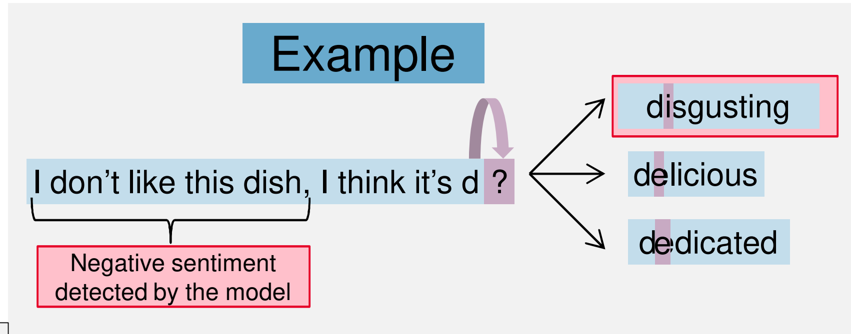


Interpretation

• Knowledge of the sentence sentiment helps for next caracter prediction.

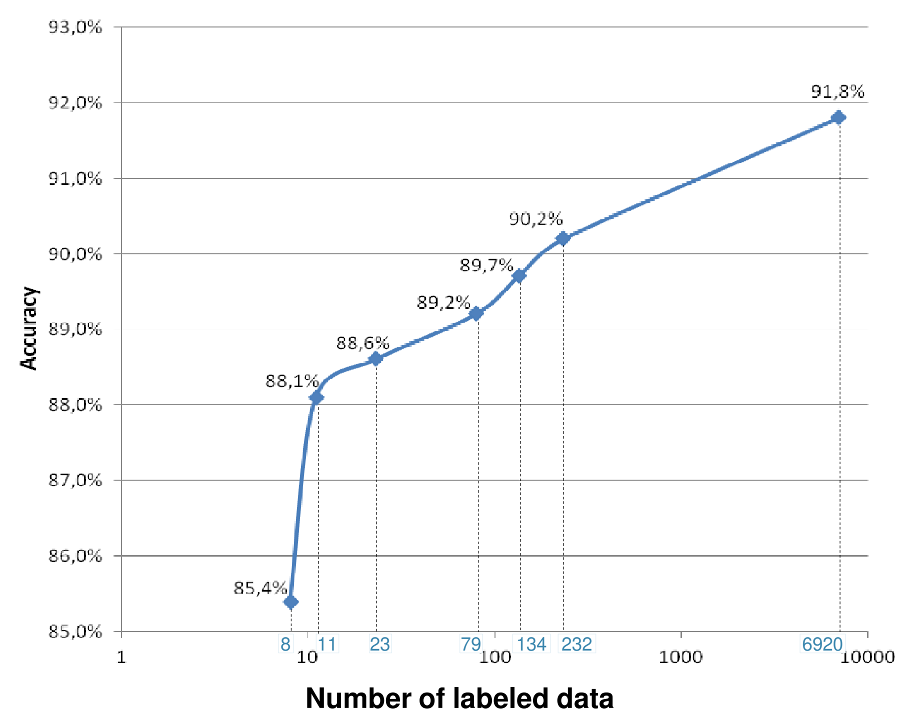
➔ Model tries to catch sentiment representation

Learning to Generate Reviews and Discovering Sentiment, A. Radford & al.



Results from the article

With few labeled data, performance are already good



State-of-the-Art sur le Stanford Sentiment Treebank dataset

Comparision with other models

Modèle	Accuracy
RNTN	85,4%
CNN	88,1%
DMN	88,6%
Paragram SL-999 LSTM	89,2%
NSE	89,7%
CT-LSTM together	90,2%

All trained on 6920 labeled data

Explanability

- Complex neural models come at the cost of interpretability
- Applications often need interpretable justifications : rationales.
- Approach evaluated on Multi-aspect sentiment analysis

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter.

Ratings

Look: 5 stars

Aroma: 2 stars

Beer review with rationales

Rationalizing Neural Prediction **by T. Lei & al.**

Rationals

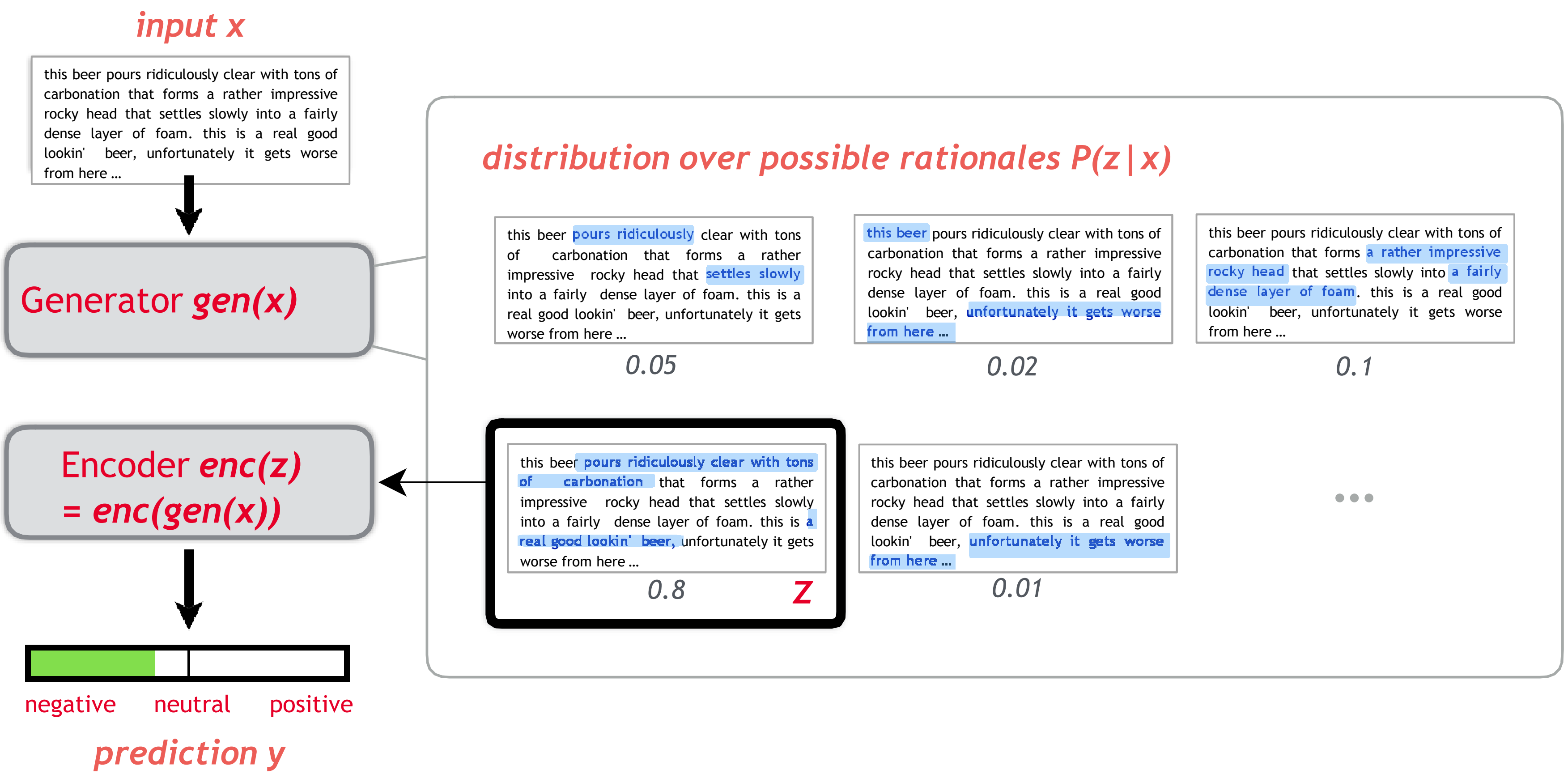
Goal: make powerful models more interpretable by learning **rationales** behind the prediction

Rationals are simply subsets of the words from the input text that satisfy two key properties:

1. Selected words represent **short and coherent** pieces of text (phrases);

2. Selected words must alone suffice for prediction as a **substitute** of the original text.

Model Architecture



two modular components optimized jointly

Training Objective

- **From an input X of length l, it generates l binary variables z. The generator estimate P(Z|X).**
- **The authors first define a cost function as follows:**

$$cost(\mathbf{x}, \mathbf{z}, \mathbf{y}) = ||\mathbf{y} - f_{\theta_e}(\mathbf{x}, \mathbf{z})||^2 + \lambda_1 ||\mathbf{z}|| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

$$P(\mathbf{Z} = \mathbf{z} | \mathbf{X} = \mathbf{x}) = g_{\theta_g}(\mathbf{x})$$

The cost function depends therefore on the value of z in three ways:

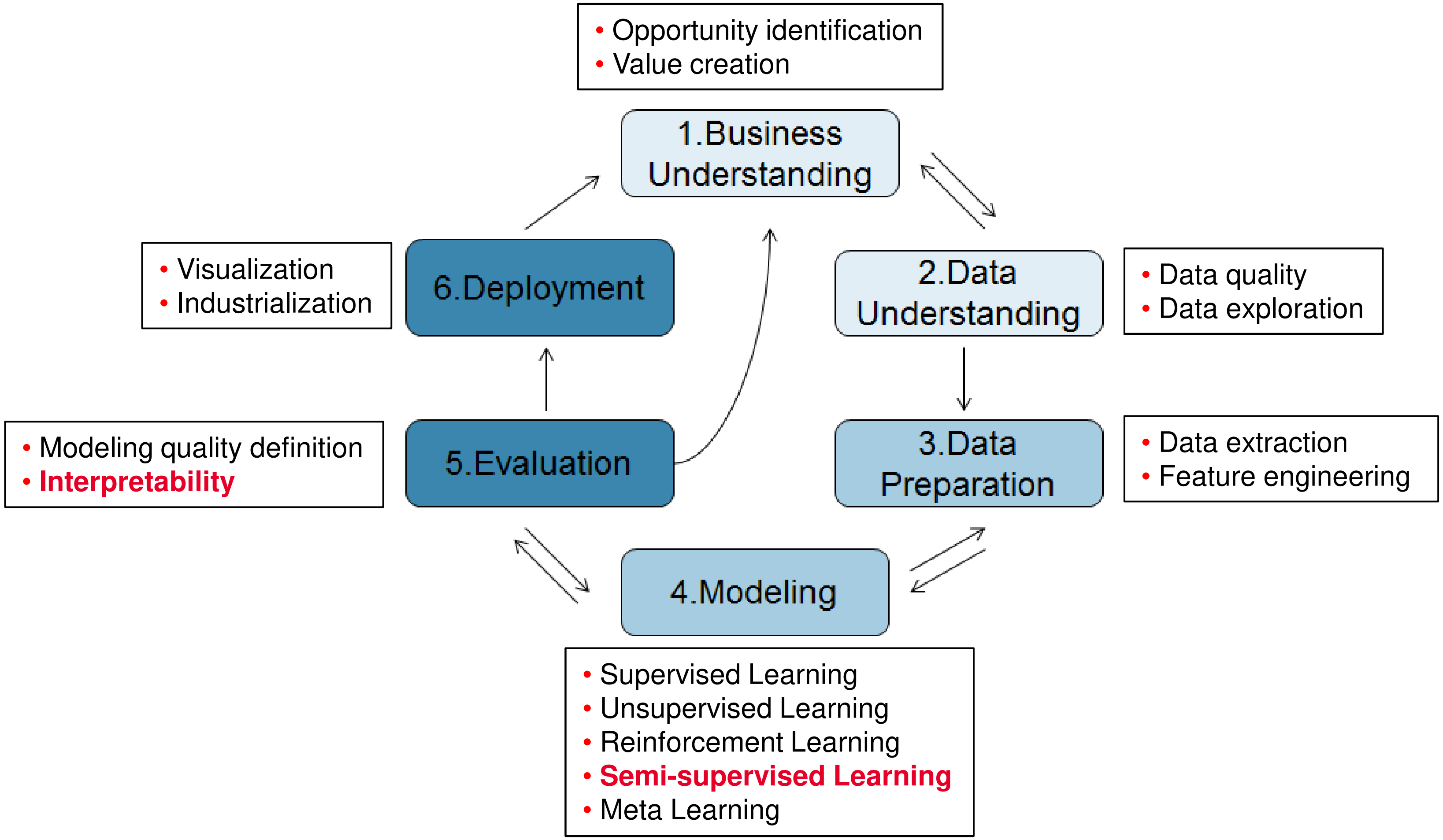
- The first term $||y-f(x, z)||^2$ is the reconstruction error.
  The target is y while the encoder predicts f(x, z) → **Sufficiency**
- The term ||z|| ensures that the selection (the number of z set to one) is as small as possible → **Sparsity**
- The last term ∑| $\mathbf{z}_t - \mathbf{z}_{t-1}$ |favors contiguous selection (phrases). → **Coherency**

The loss function to be optimized for each training example is :

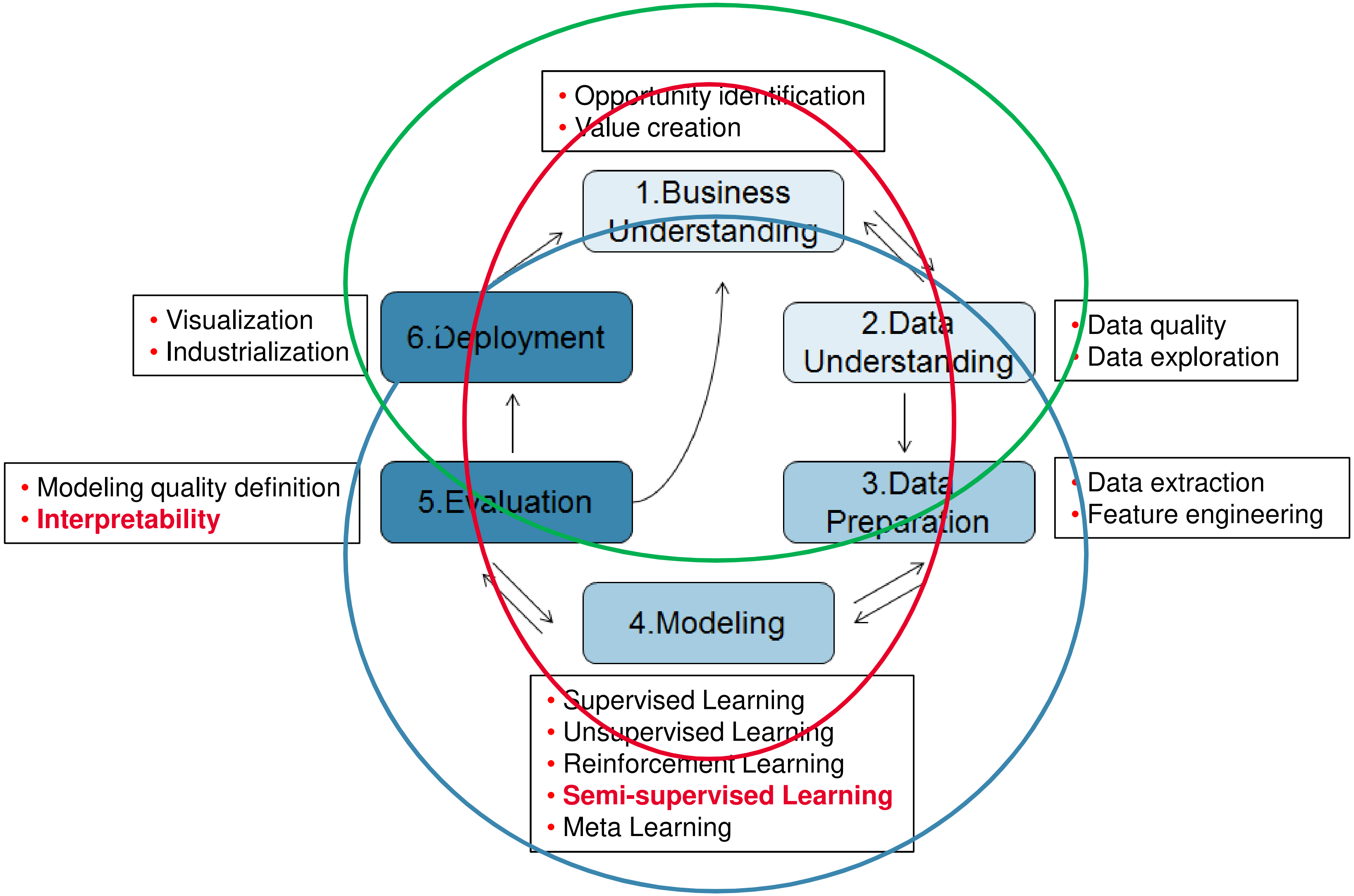
$$\begin{aligned} \mathcal{L}(\theta_g, \theta_e, \mathbf{x}, \mathbf{y}) &= E_{\mathbf{z} \sim P(\mathbf{Z}|\mathbf{X})} cost(\mathbf{x}, \mathbf{y}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z} | \mathbf{X} = \mathbf{x}) cost(\mathbf{x}, \mathbf{y}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} g_{\theta_g}(\mathbf{x}) cost(\mathbf{x}, \mathbf{y}, \mathbf{z}) \end{aligned}$$

→ Computing optimization performed because summation over z is exponential

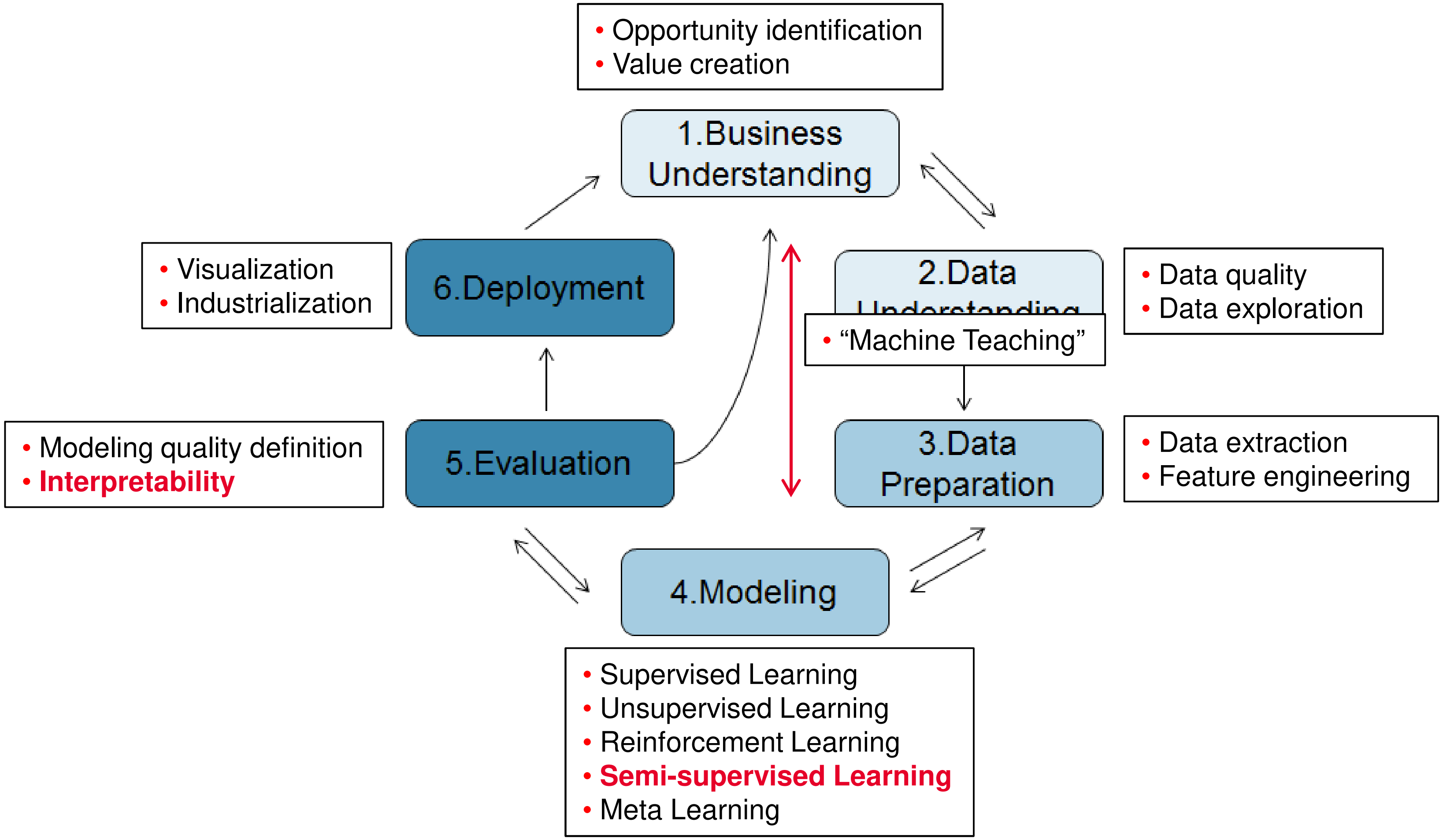
Data Modeling Workflow & Challenges



Data Modeling Workflow & Challenges



Data Modeling Workflow & Challenges



Q&A