Quantmetry

Data Science Consulting

Project use case

Improve your marketing reach using large scale machine learning on Spark

June, 6th

Matthieu Vautrot
mvautrot@quantmetry.com
Nina Bertrand
nbertrand@quantmetry.com

Quantmetry – Data science consulting

Quantmetry
Data Science Consulting

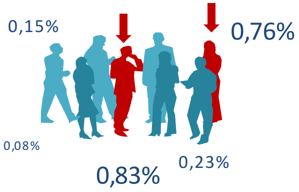
- Founded in 2011
- 25 consultants - Data Scientists & engineers

- @bertrand_nina
- @matthieuvautrot

Share of experience of an on-going project

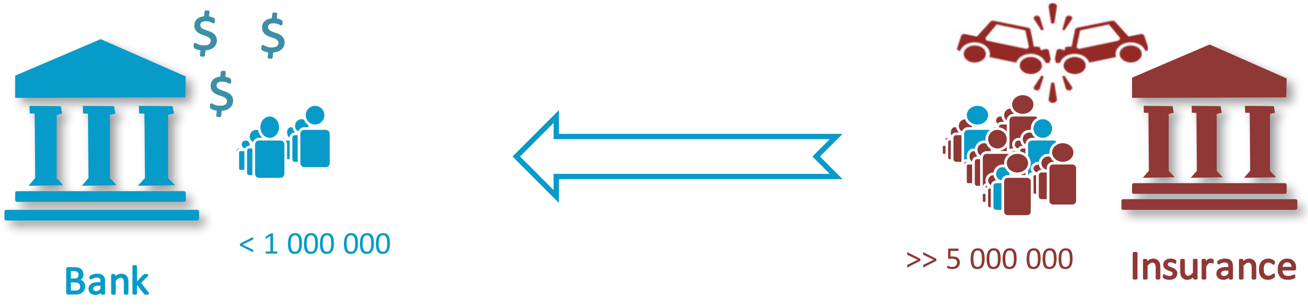
- Our client, a insurance and bank …

- … frequently build and run Marketing Campaigns to sell their products



- Can you do better than just “scoring” approaches with “Big Data” ?

The client’s new campaign



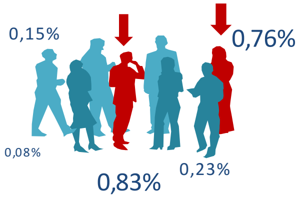
- Goal of the campaign : “Bancarize” as much Insurance clients as possible
- Can use multiple canals for it :

- Display (web adds through DMP and / or client web site tagging)

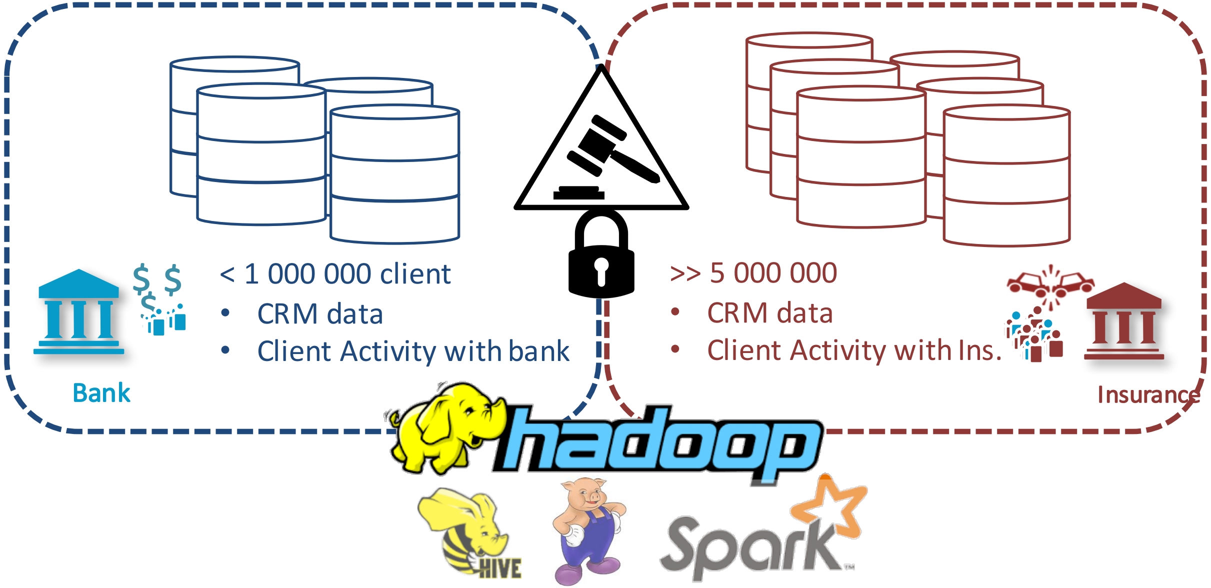
- Email

- Direct call

Existent data organization



- Can you do better than just “scoring” approaches with “Big Data” ?

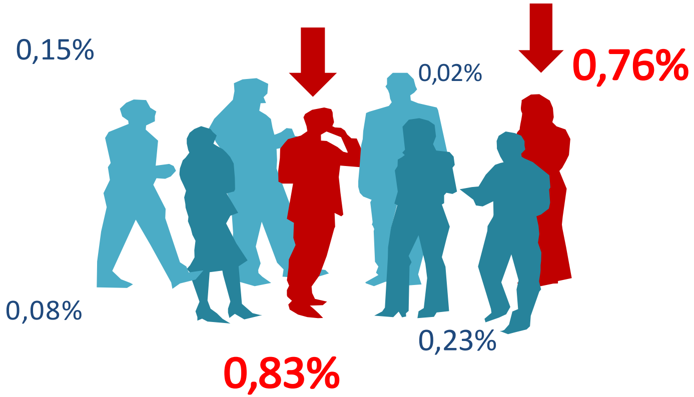


- A same Hadoop Cluster, 2 distinct tenants (usual legal stuff)

Of course we can !!

- Can you do better than just “scoring” approaches with “Big Data” ?

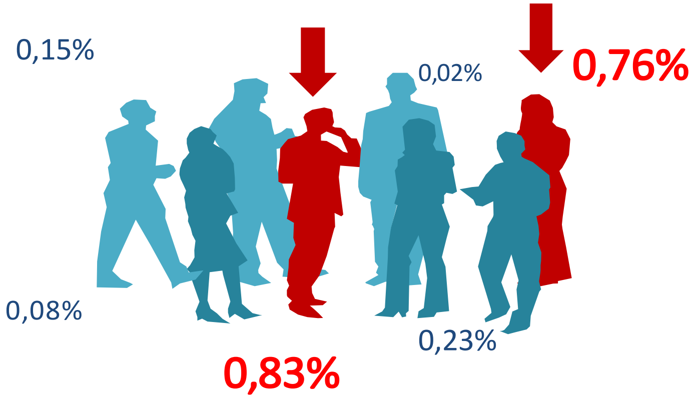
Can we really do better than regular scoring ?



A typical scoring campaign is generally built like that :

- Train predictive **model** on the client base : observed buyers of the product (or similar)
- Apply the **model** and score the whole eligible client data base
- Send to the **N top score** a marketing message

Observed scoring approach limitations

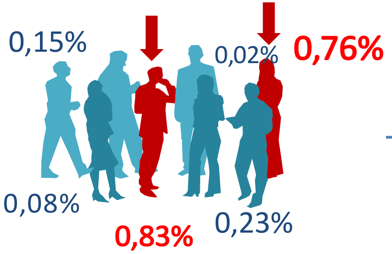


Two major limitations to this approach :

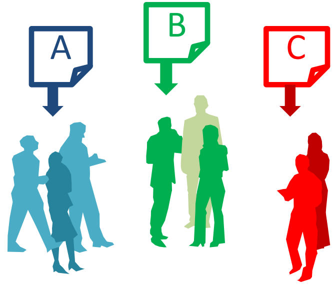
- **Lack of personalisation** : Same message is sent to the top scored group

- **Scoop natural noise :** Target who would buy the product anyway

When one message is not enough



Simple scoring campaign



Message personalisation campaign

Come because you can :

A : get money

B : we are better than the other Banks

C : win an ipad !!

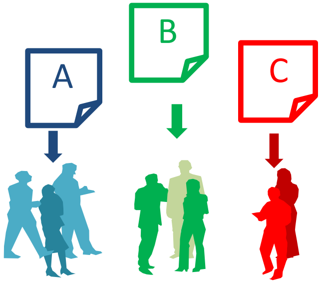
Two major limitations to this approach :

- **Lack of personalisation** : Same message is sent to the top scored group

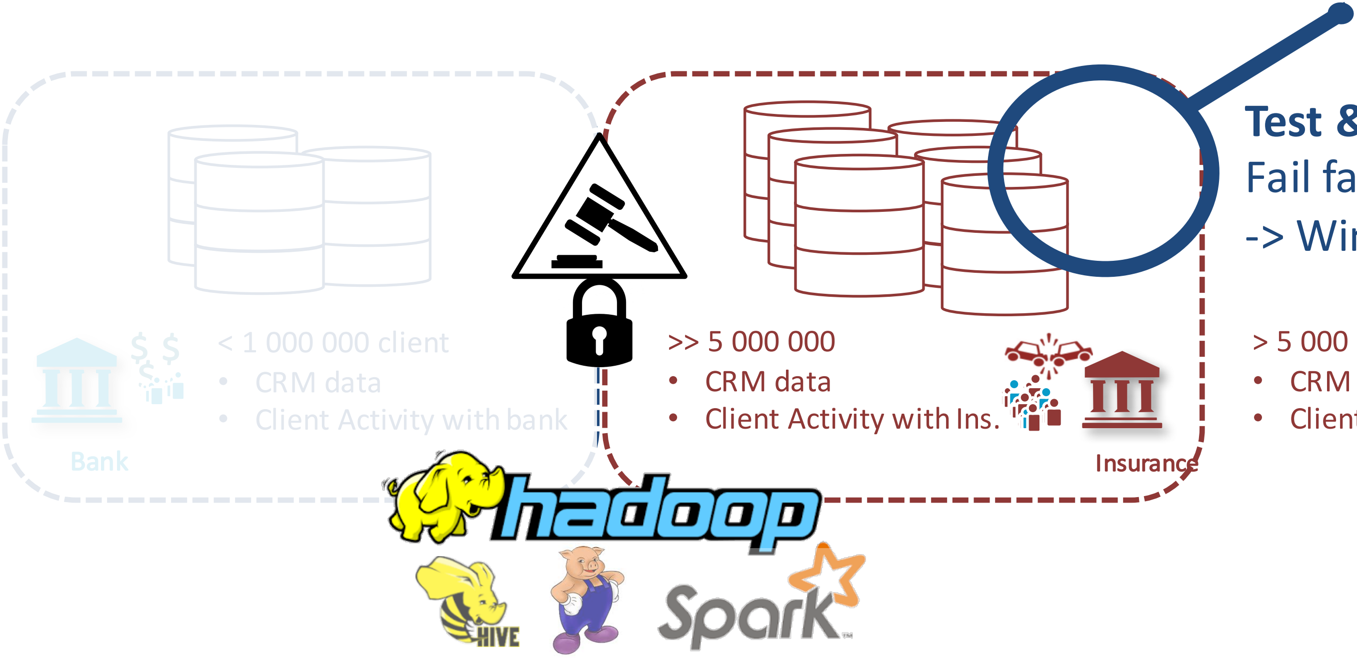
  -> Use multiple messages and find out who likes which one with ML

- **Scoop natural noise :** Target who would buy the product anyway

Test & learn on the client data base



-> Use multiple messages and find out who likes which one with ML

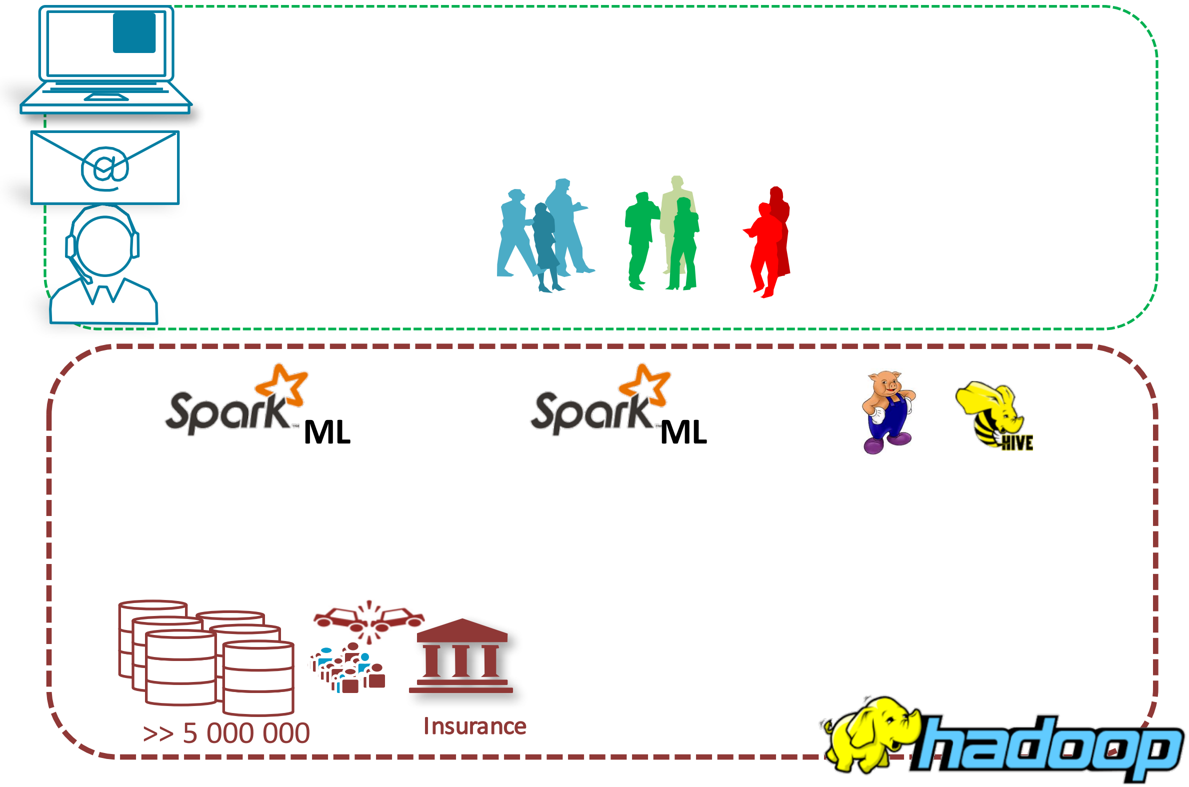


Test & learn :
Fail fast
-> Win quick

> 5 000 000
- CRM data
- Client Activity data

Test & learn the different messages directly from the insurance client data base

Test & learn multiple messages



Test & learn multiple messages

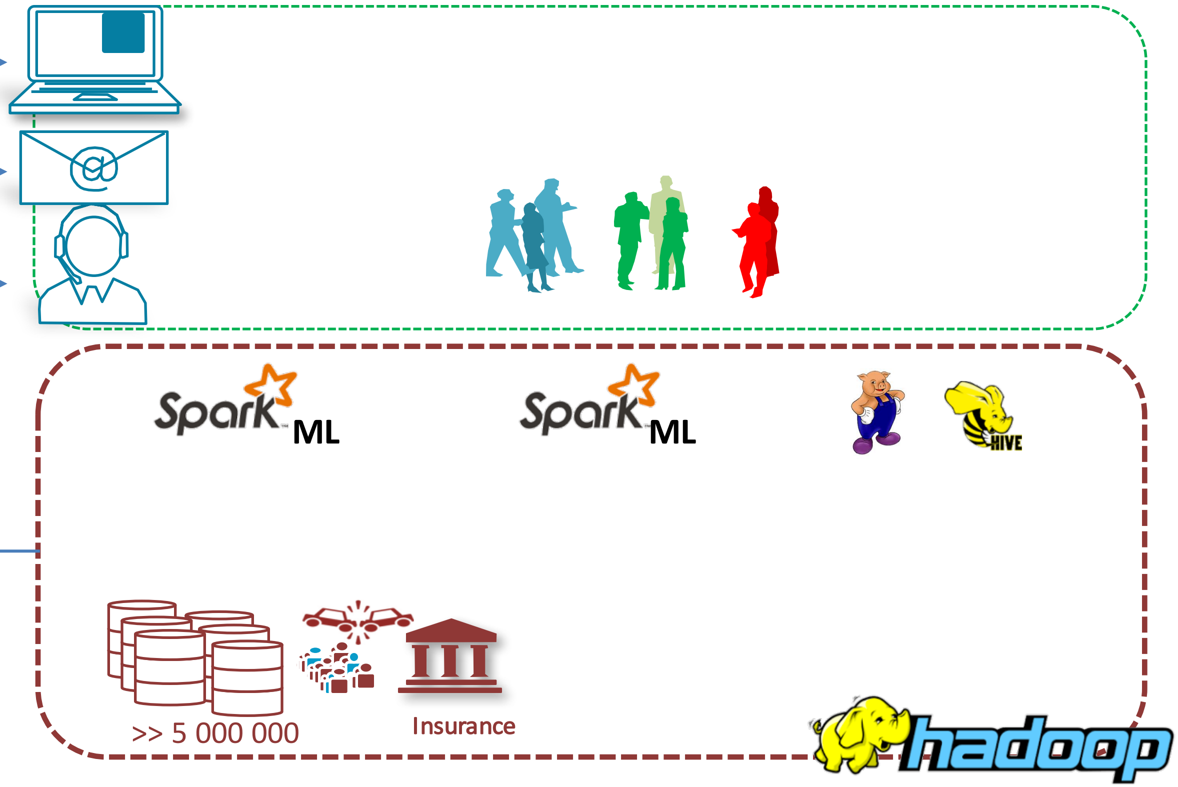
1

Every day :

- List of cookies to target (DMP)

Every couple weeks

- List of people to contact (email / tel)



- **At day 0** : random client sample / random message

Test & learn multiple messages

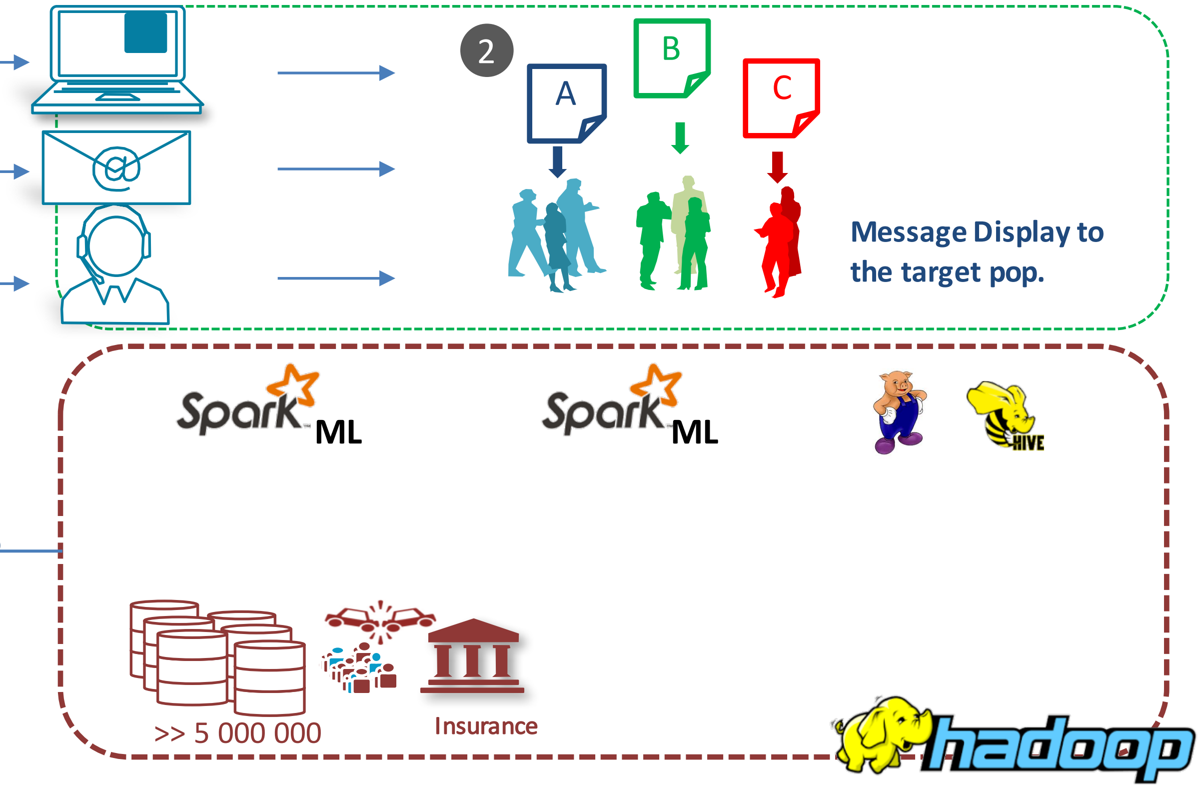
1

Every day :

- List of cookies to target (DMP)

Every couple weeks

- List of people to contact (email / tel)



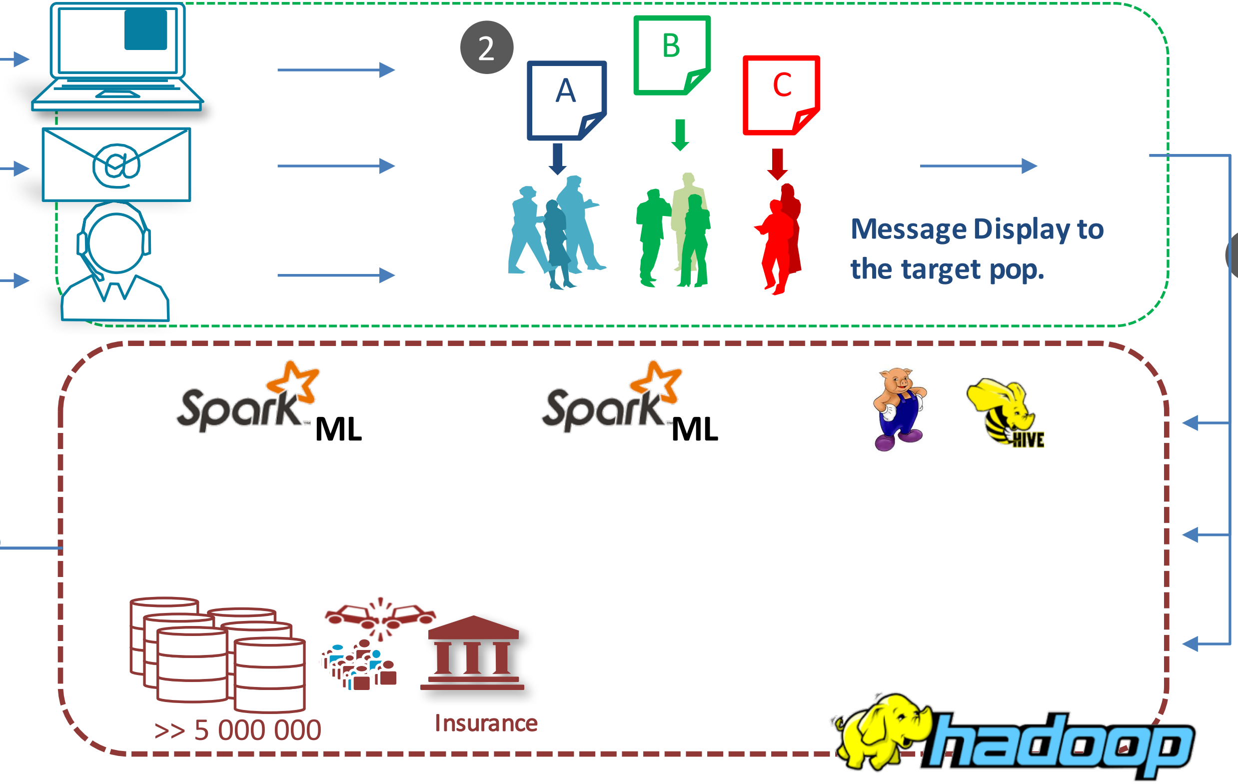
- **At day 0** : random client sample / random message

Test & learn multiple messages

1

Every day :
- List of cookies to target (DMP)

Every couple weeks
- List of people to contact (email / tel)

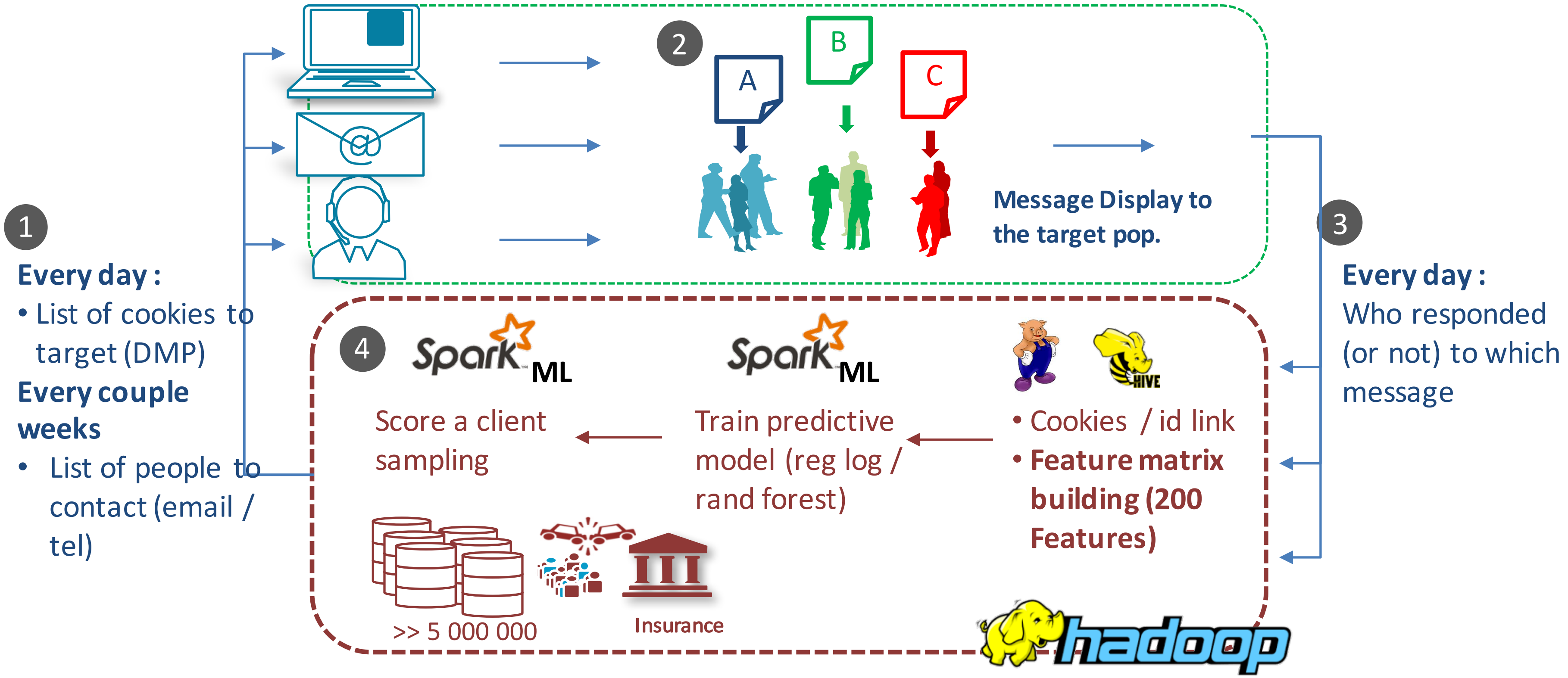


3

Every day :
Who responded (or not) to which message

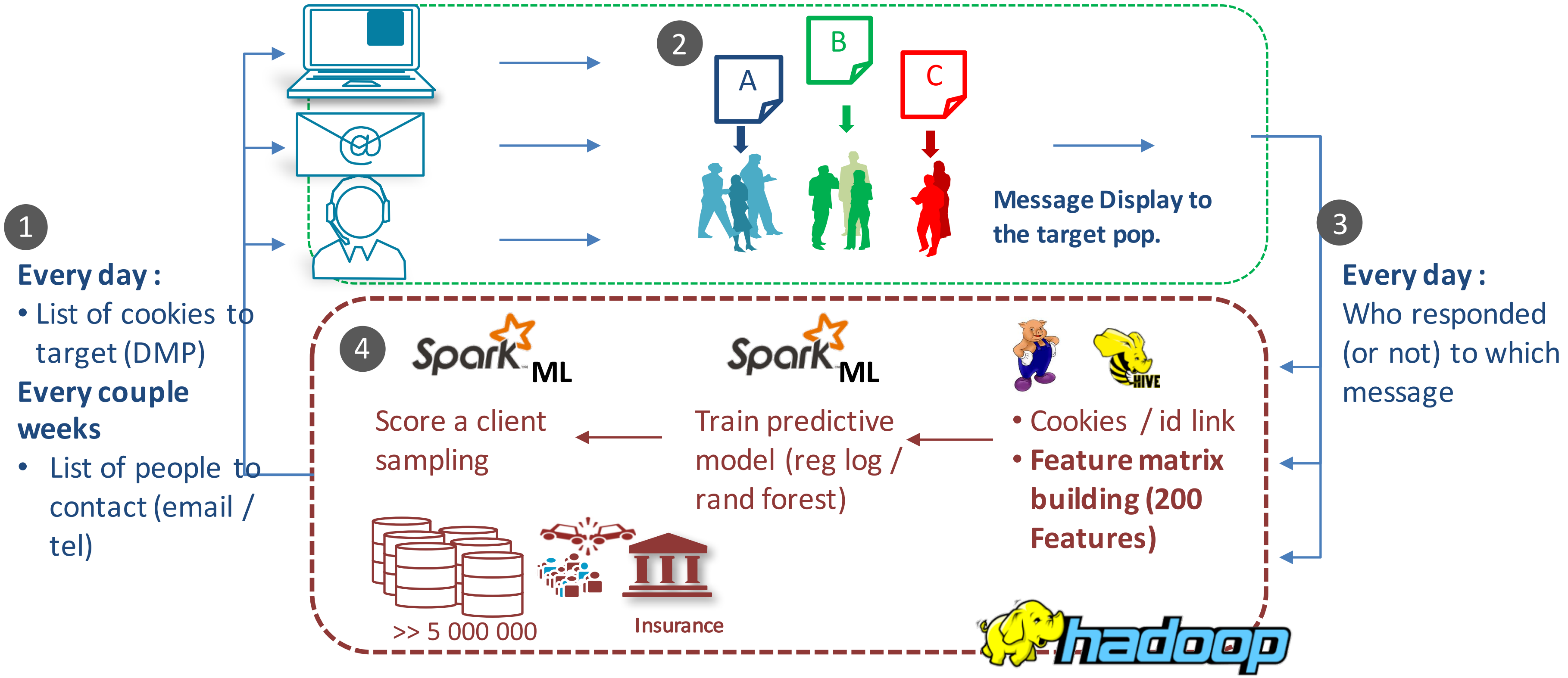
- **At day 0** : random client sample / random message

Test & learn multiple messages



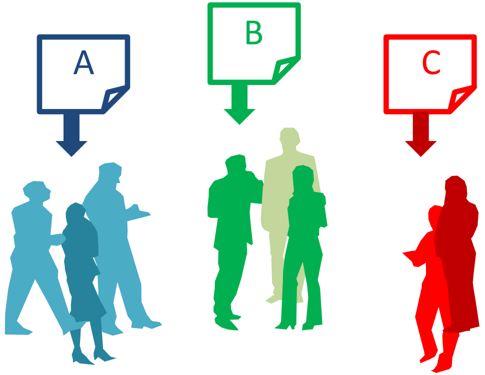
• **At day 0** : random client sample / random message

Test & learn multiple messages



- **At day 0** : random client sample / random message
- **At day N+1** : use **N** days results from learning

Can we really do better than scoring ?



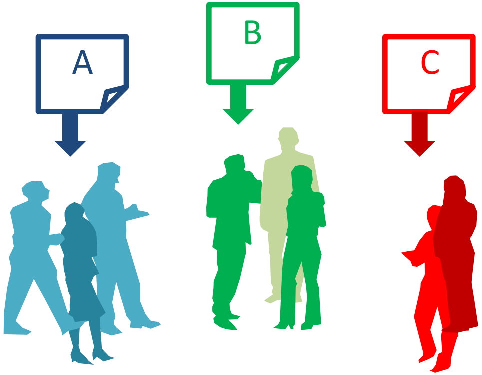
Two major limitations to this approach :

- **Lack of personalisation** : Same message is sent to the top scored group

  -> Use Multiple messages and find out who likes which one with ML

- **Scoop natural noise :** Target who would buy the product anyway

Let’s get rid off the natural noise



Two major limitations to this approach :

- **Lack of personalisation** : Same message is sent to the top scored group

  -> Use multiple messages and find out who likes which one with ML

- **Scoop natural noise :** Target who would buy the product anyway

  -> Use ML models that get rid off natural noise : uplift models

Uplift model : improve an effect treatment

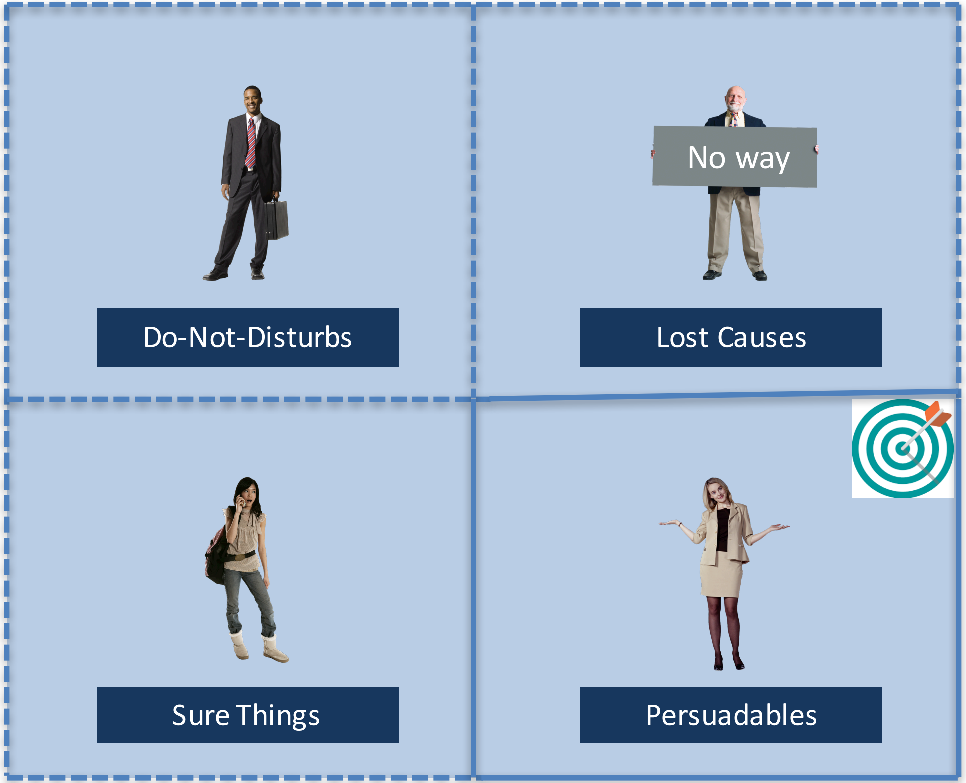
Uplift (or « True lift »)

Idea

✓ Describe **message effect** on target

Motivation

- Do not call self-converted-people
- Some customers are liable to buy but marketing phone call have a negative influence on them

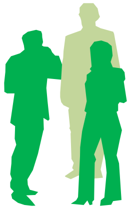


Methodology :

- Two samples should be regarded :

The treatment group (*T*)

The control one (*C*) who did not receive the treatment (placebo)



Different possible implementations :

- Independent models
- Regression with tuning parameters
- Sequential models

Uplift model : improve an effect treatment

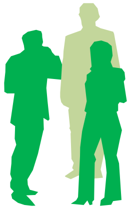
Implementation choice

Methodology :

- Two samples should be regarded :

The treatment group (*T*)

The control one (*C*) who did not receive the treatment (placebo)



Different possible implementations :

- **Independent models**
- Regression with tuning parameters
- Sequential models

Number of models : N messages X M canal

Spark

Train predictive model (reg log / rand forest)

- Requirement of Independent models

Random Forest / Log reg.
Controlled by **Uplift**

Spark ML

Every day :

- One predictive model is calculated for every Message X Canal
- Models as usual : random forest or logistic regression

Uplift model : improve an effect treatment

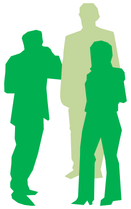
Implementation choice

Methodology :

- Two samples should be regarded :

The treatment group (*T*)

The control one (*C*) who did not receive the treatment (placebo)



Different possible implementations :

- **Independent models:** →

$$\mathbf{Uplift(x) = } \boldsymbol{P(S \mid x, T=1) - P(S \mid x, T=0)}$$

S the subscription event

- Regression with tuning parameters
- Sequential models

Uplift model : improve an effect treatment

Main difficulty

$$\text{Uplift}(x) = P(S \mid x, T=1) - P(S \mid x, T=0)$$

Difficulty :

- There is a predicted uplift by customer but no individual real uplift → no individual target..

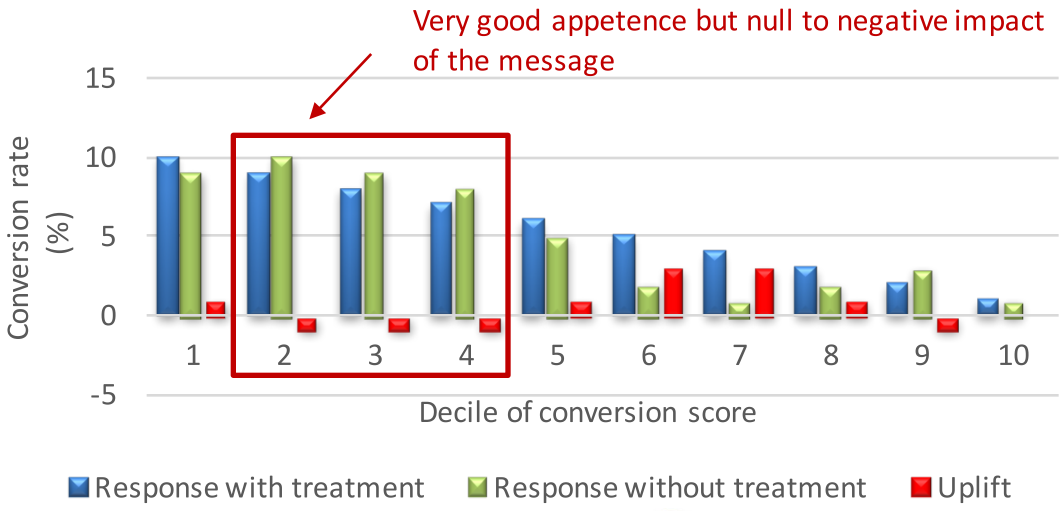
Solution :

- Sort customers by their uplift score in decreasing order
- Focus on quantile of customers
- Calculate difference between conversion rate of treated group and natural conversion rate

Uplift model : improve an effect treatment

Appetence VS Uplift

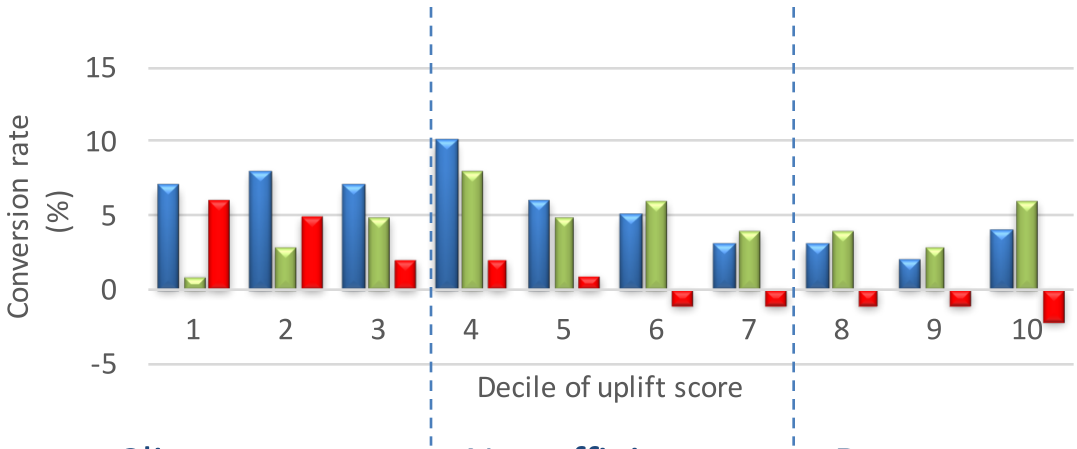
Appetence sorted by conversion probability



- Groups with highest conversion score has not necessarily been scored with the highest uplift.
- This people may have converted without any treatment.

Uplift model sorted by predicted uplift

« Perfect » uplift model



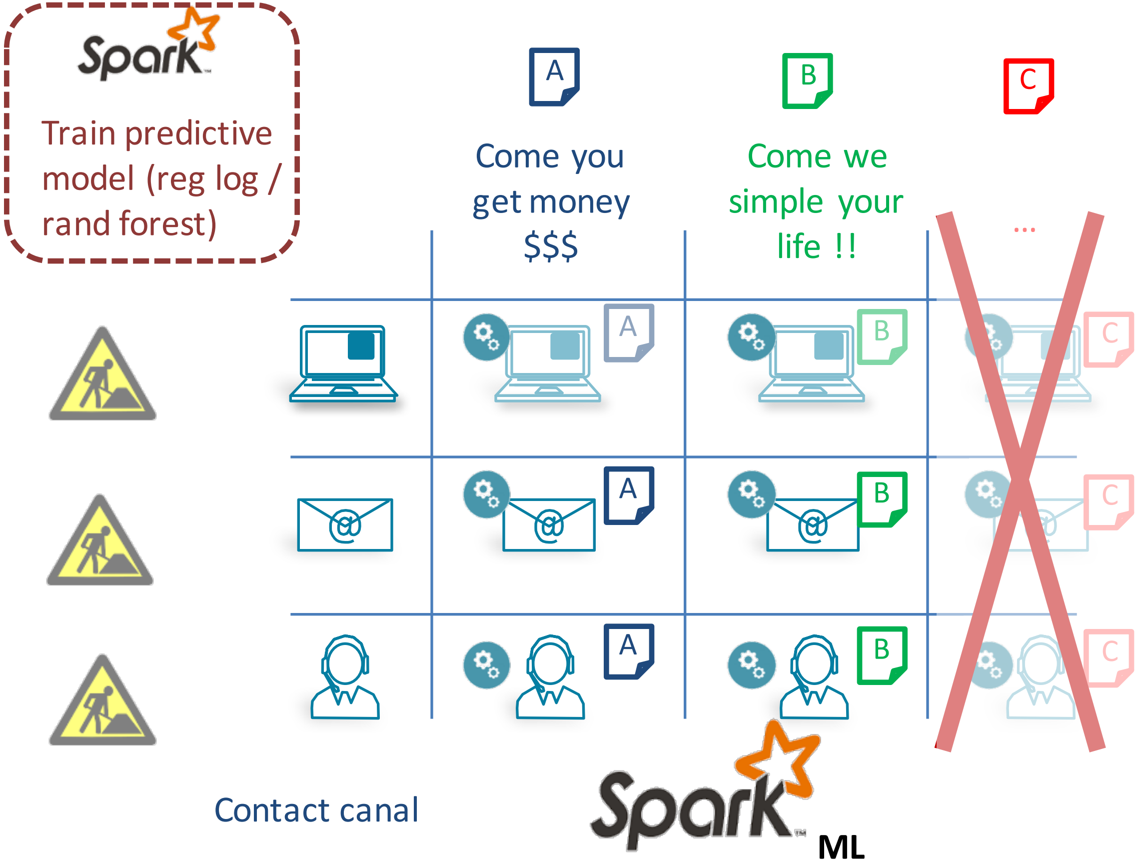
Client to contact

Not efficient to contact

Dangerous to contact

- … What about the real uplift ?
- How do you assess the performance ?

Actual state of the implementation



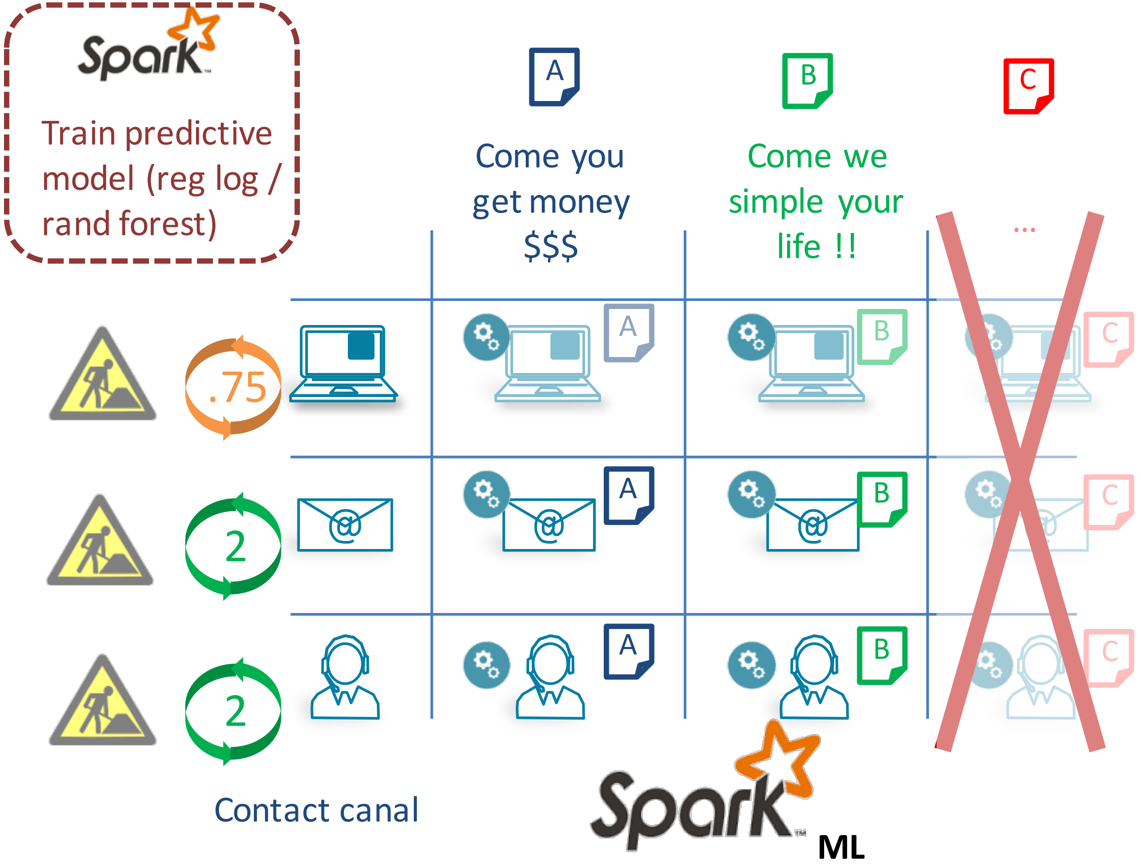
Need POC

- Quick agile POC iterations
- Limited to 2 messages to push

For all 3 canals

- Data preparation (Pig Hive) done
- Predictive Algorithms : done

Actual state of the implementation



Need POC

- Quick agile POC iterations
- Limited to 2 messages to push

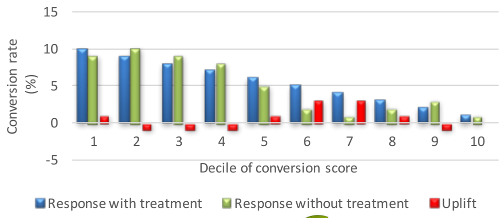
For all 3 canals

- Data preparation (Pig Hive) done
- Predictive Algorithms : done

- 2 waves already achieved in mail and tel
- DMP results analysis is on going

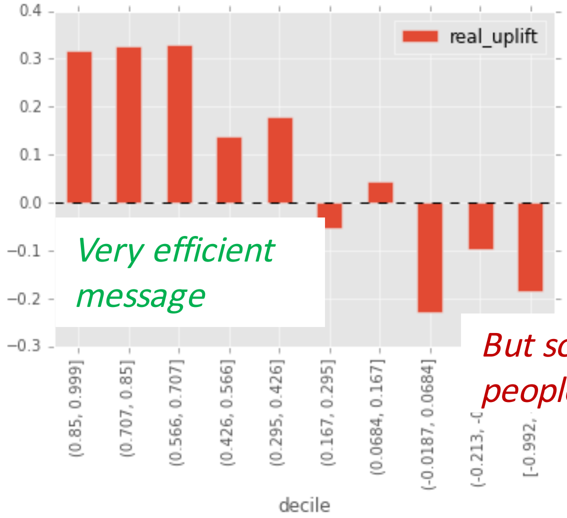
Uplift model : improve an effect treatment

Use case observed uplift and marketing insights



Observed uplift : for mail canal after 1rst wave

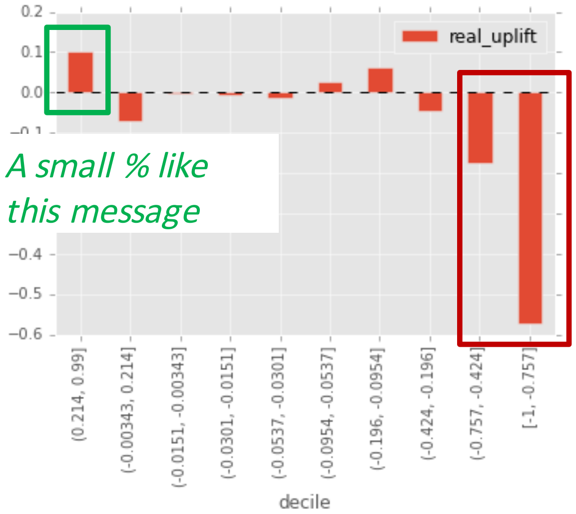
1st message : « Come you get money $$$ »



Very efficient message

But some people hate it !

2nd message : « Come we simple your life »



A small % like this message

Warning do not contact those guys

We just have to take best score between the 2 models

Data engineering the Marketing campaign

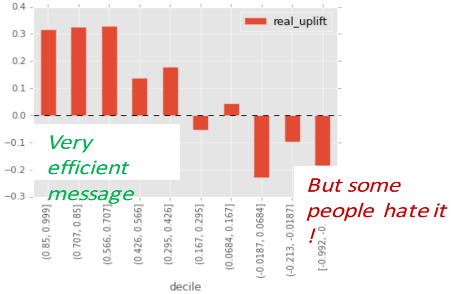
- Easy on paper but watch out to business and IT organization constraints (eg : DMP and Hadoop Cluster not easly linkable)
- Spark is good but sometimes Scikit learn can do the trick for first quicker ML iteration

POC → Indus. **ML**

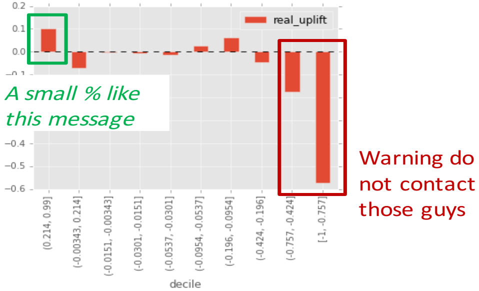
Uplift modeling

- Very efficient for marketing insight already on first waves -> Promising for the following up of the project !

1st message : « Come you get money $$$ »



2nd message : « Come we simple your life »



Quantmetry
Data Science Consulting

BERLIN BUZZWORDS 2016 JUNE 05-07

Q & A ?

Thank you

Nina Bertrand : nbertrand@quantmetry.com

Matthieu Vautrot : mvautrot@quantmetry.com