Tagging Text in Money Transfers: A Use-Case of Spark in Banking

Luis Peinado Fuentes, BBVA Data & Analytics
Jose A. Rodriguez Serrano, BBVA Data & Analytics

BBVA
DATA & ANALYTICS

#EUds7 @BBVAData

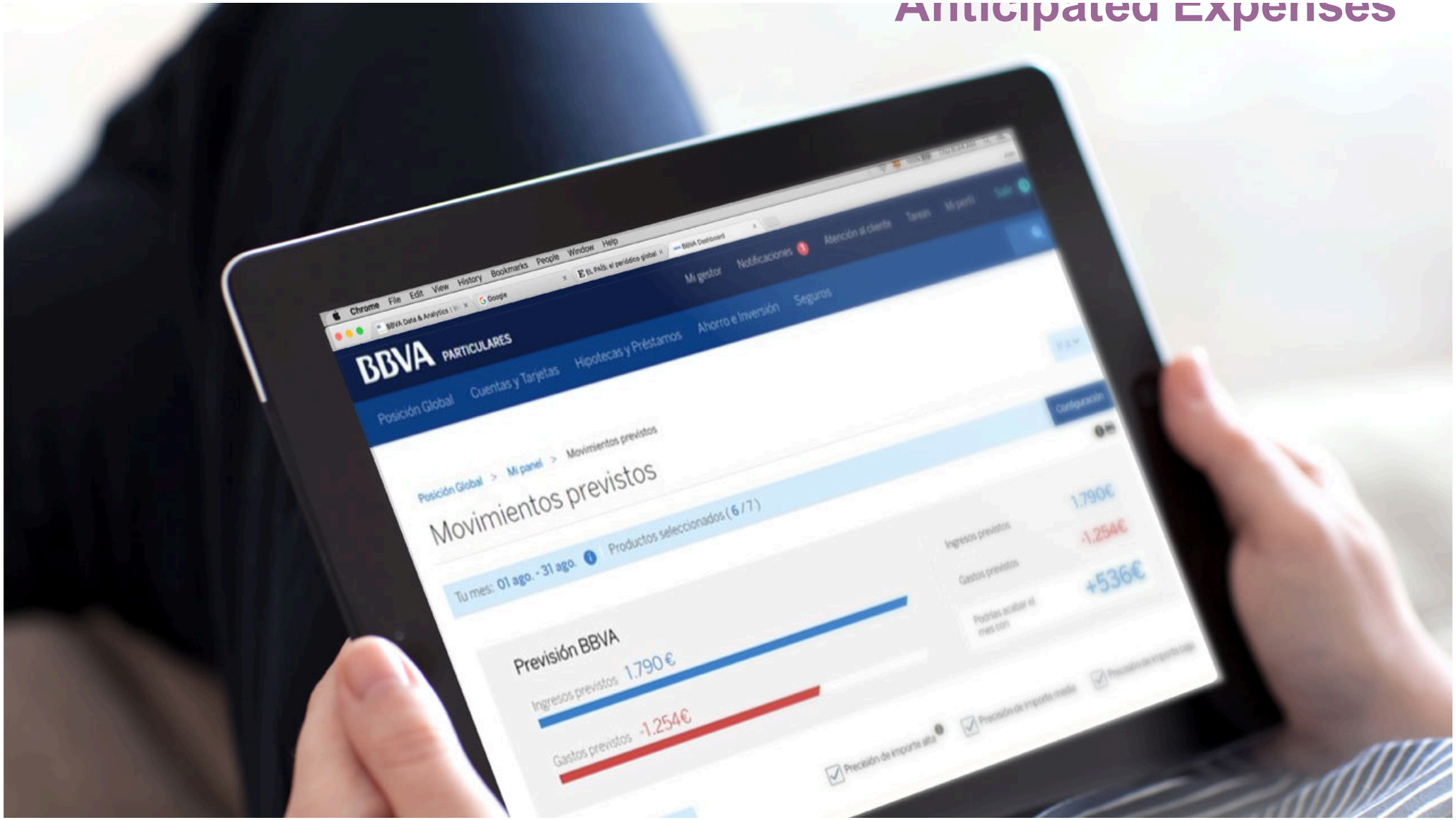
Our Goal:

Improving User Experience in Retail Banking Products with Machine Learning

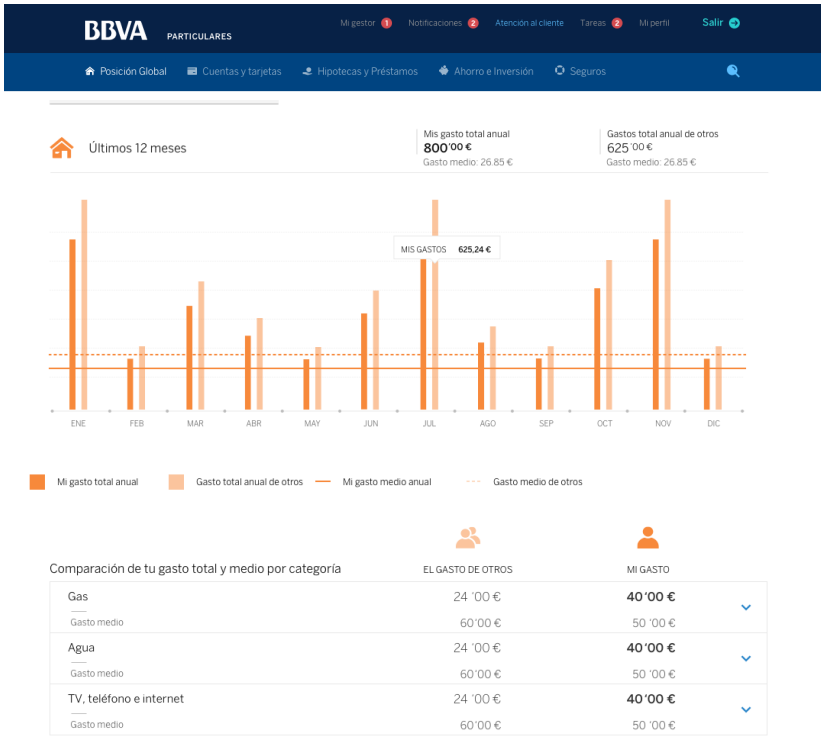
(Using Spark)

Dashboard View of Expenses and Income

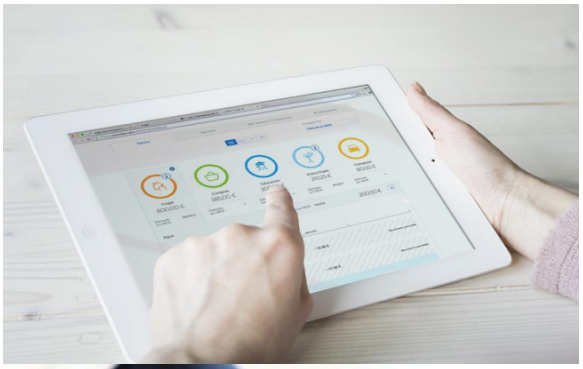
Anticipated Expenses

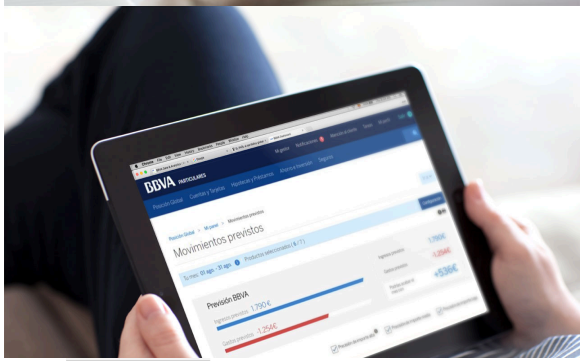


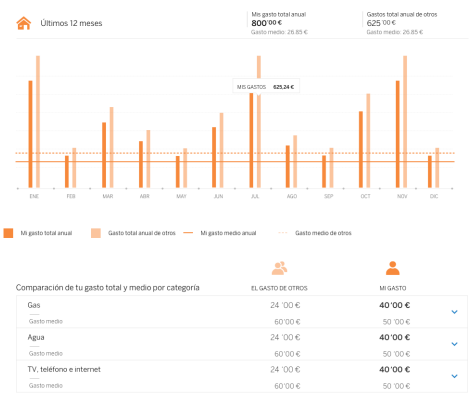
Customer Comparison Engine



All these examples require **categorized** bank transactions

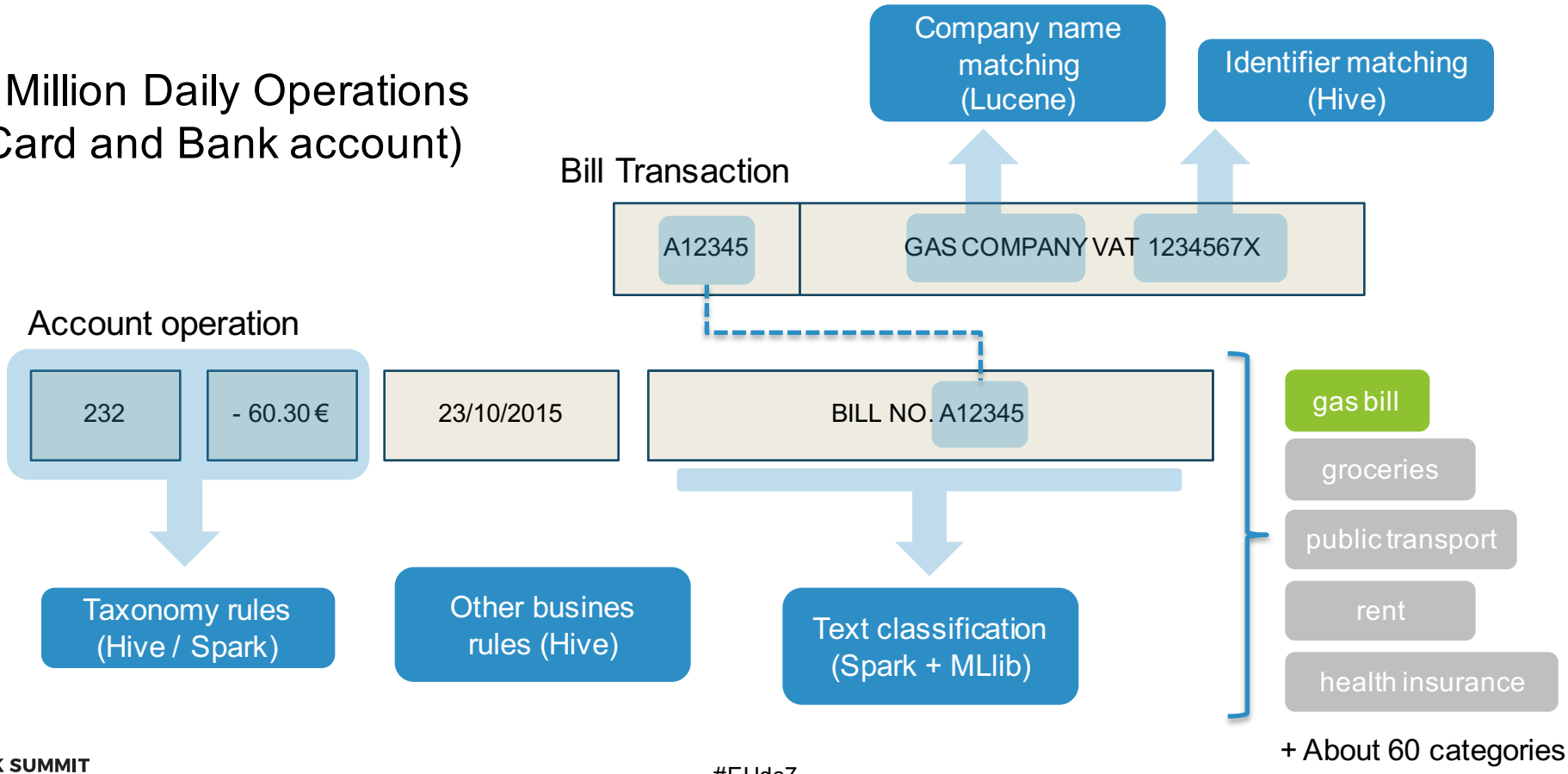




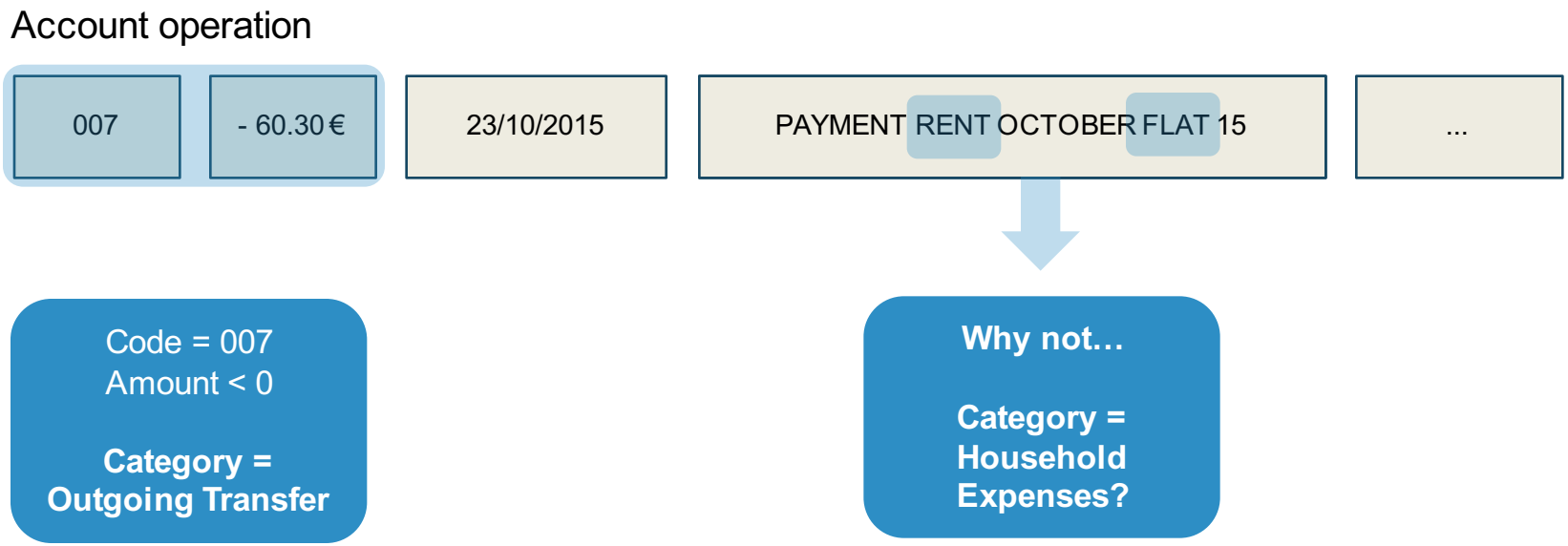


Categorizing Bank Transactions

7 Million Daily Operations
(Card and Bank account)



The Problem: Categorizing Transfers

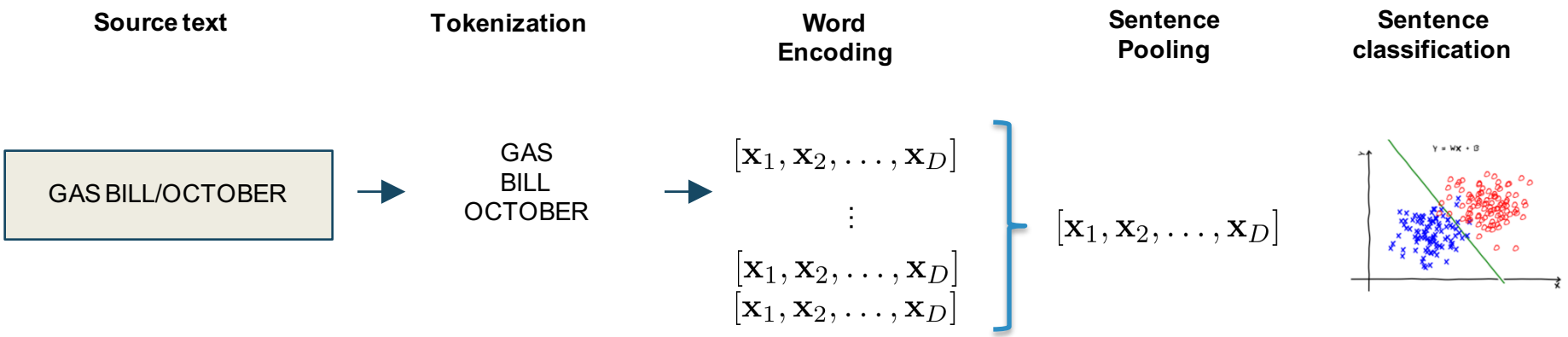


700,000 Daily Operations (Transfers)

Data Science Challenges

- We did not know the data source in advance
- We did not have a labeled set
- A fraction of texts is useless (“detection” rather than classification)
- Distribution of categories is imbalanced
- Prefer false negatives over false positives
- Very short text, language not even syntactically correct

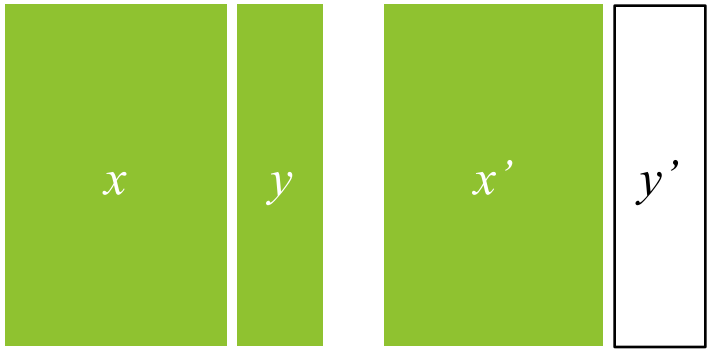
Basic Pipeline (in Pro since 2016)



- First implementation: TF-IDF features + linear classifier (98% precision, 21% recall
- Further tests with word2vec + Vector of Locally Aggregated Descriptors (VLAD)
- Implemented in Spark/Scala, using MLlib classes
- Own classes implemented for Multi-class Logistic Regression, VLAD
- Scala dependency injection useful to quickly setup variants of the above steps

Where Do We Get a Dataset From?

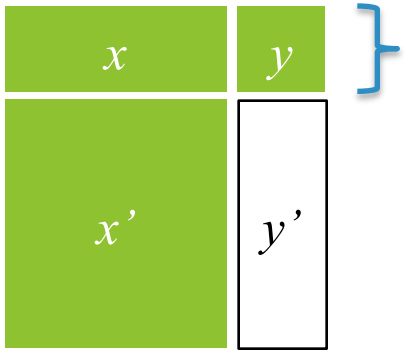
First Attempt:
Domain adaptation



Non-Transfers Transfers

But: $p(x, y) \neq p(x', y')$

Second Attempt:
"Assisted Dataset creation"



Transfers

$p(x, y) = p(x', y')$

Assisted annotation with tagging interface.

Tags suggested by rules, incomplete training, similar-string matches, and confirmed manually.

> 45K Transfers Tagged

Exploring better embeddings/poolings: w2v + VLAD

```
val word2vec = new Word2Vec()
val model = word2vec.fit(input)
val synonyms =
model.findSynonyms("alquiler", 10)
```

Word2vec "Synonyms"

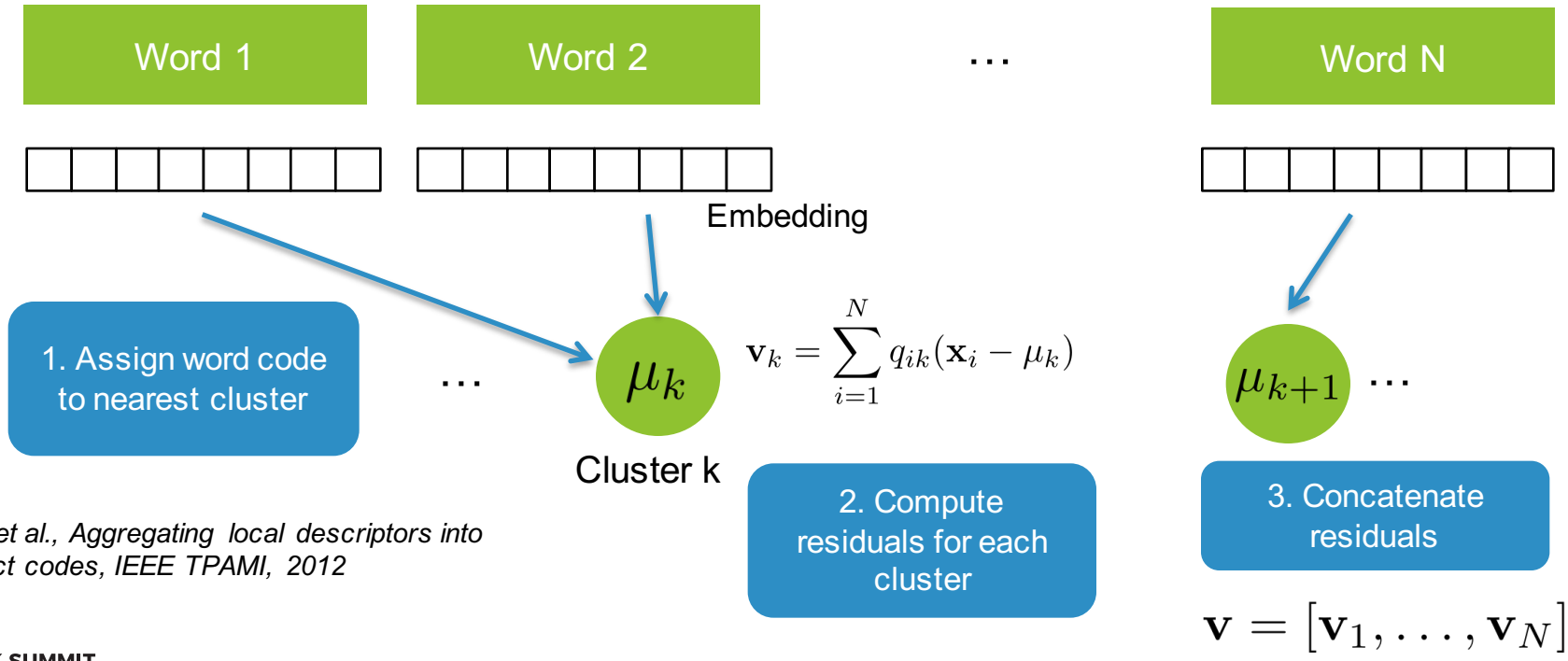
viaje → billete vuelo hotel reserva billetes gastos boda regalo casa

alquiler → piso cdad alq comunidad local garaje mes prop calle

pago → factura fra fras fact 14 resto fac 50 facturas

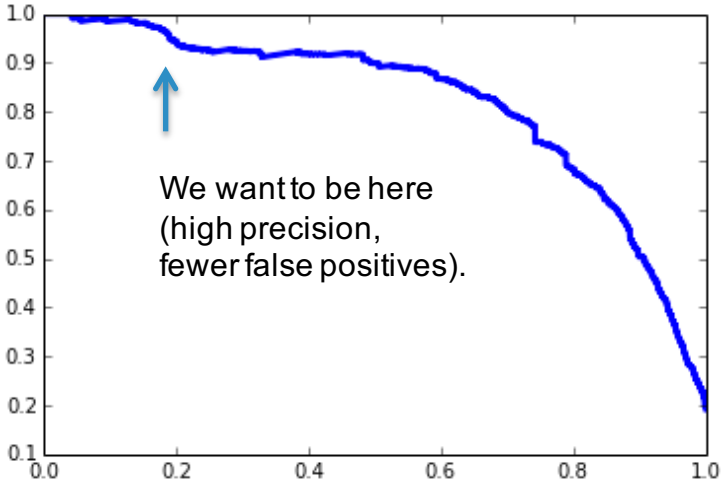
Vector of Locally Aggregated Descriptors (VLAD)

Own implementation in Spark/Scala



Jegou et al., Aggregating local descriptors into compact codes, IEEE TPAMI, 2012

Results of the Experiment



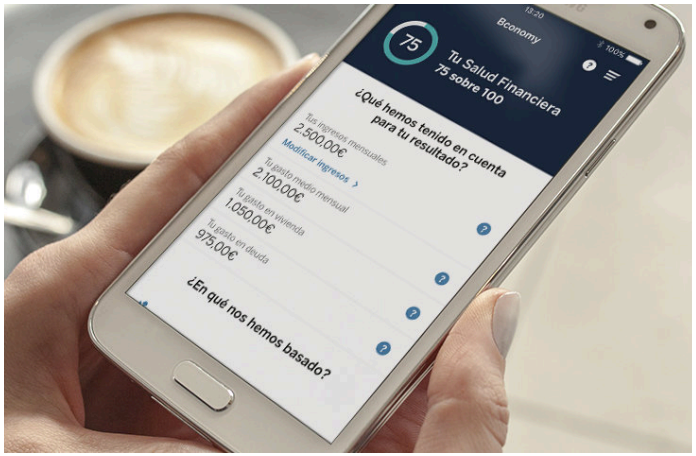
P/R curve for a tag

Method	Recall @ Prec=98%
TF-IDF	21.0 %
Word2vec (avg pool)	24.5 %
Word2vec (avg pool)+ Amount	26.9 %
Word2vec (VLAD pool)	37.3 %

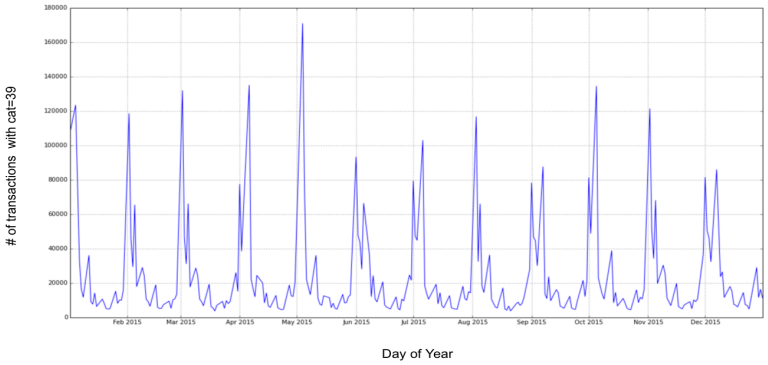
- The combination of w2v + VLAD is not in production
- Currently working on own multi-word embedding which yields similar results

Enablers

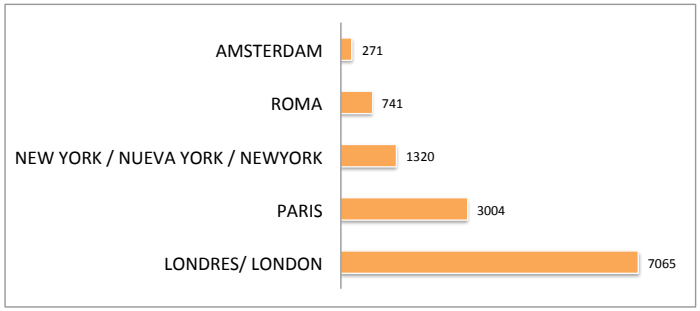
Bconomy –
Identifying Rental Expenses



When do household expenses happen?



Aggregated statistics about trip destinations



Other Challenges

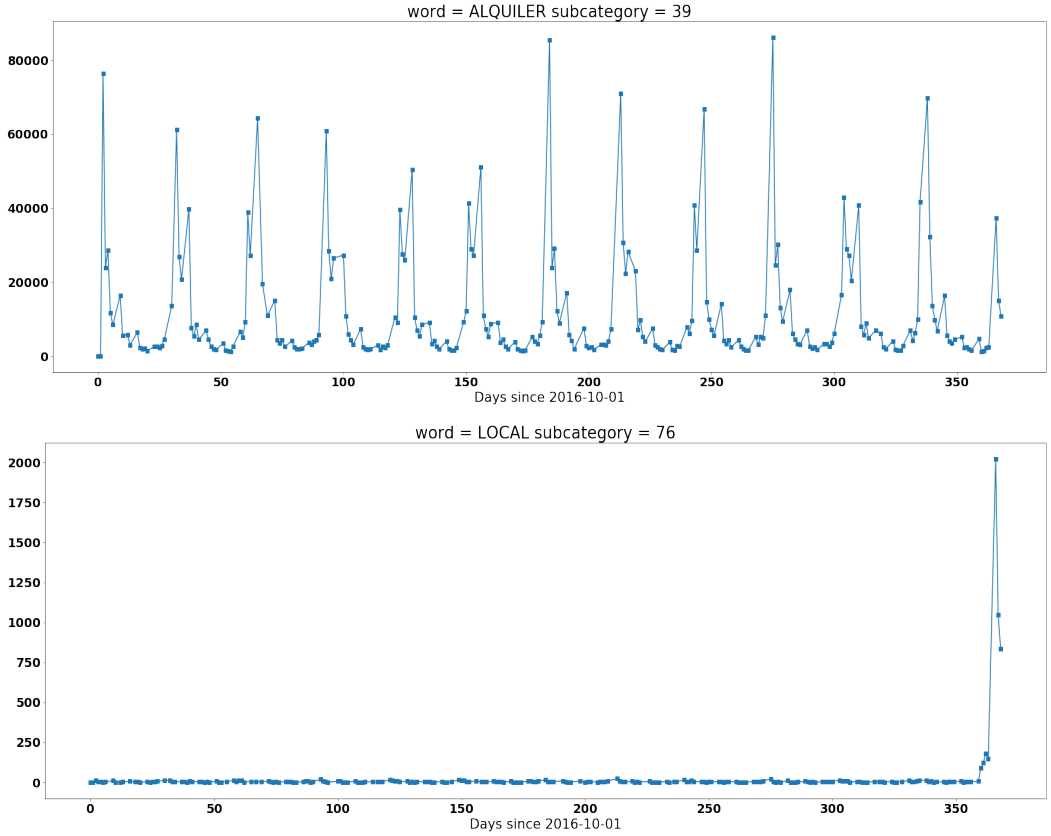
The system goes to the wild…

- Failure cases stay within the tolerated 2%

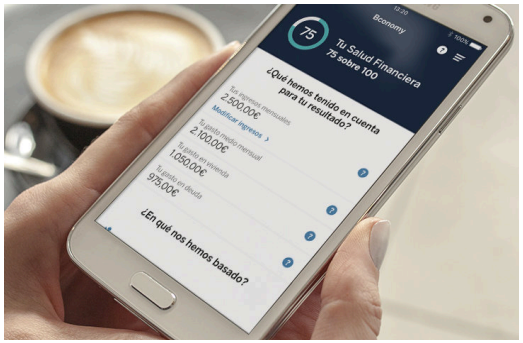
Open system

- Word black-list needed
- Proposal for daily quality checks
  (plots: daily evolution of number of word-tag pairs)

M. Zinkevich, Rules of ML:
Best Practices for ML Engineering



Conclusions



How Spark Helped Us

- Spark (+MLlib) was crucial in this use case.
- We complemented with own implementations of Multi-Class LR and Vector of Locally aggregated descriptors.
- We took advantage of Scala's code injection mechanisms to frame the classification process as a set of standardized steps.

Data Science Experience

- Running text classifier in a real production system
- Improving the standard components through experiments on word2vec and VLAD
- Ongoing work: own embedding method

Tagging Text in Money Transfers: A Use-Case of Spark in Banking

Luis Peinado Fuentes, BBVA Data & Analytics
Jose A. Rodriguez Serrano, BBVA Data & Analytics

Thanks to:

- Roberto Maestre and Advisory Team @ BBVA D&A for insightful comments
- Everyone at BBVA who made using Spark possible and disseminated Spark and Scala knowledge

BBVA
DATA & ANALYTICS

#EUds7 @BBVAData

(Just in case)

Appendix

Exploring better embeddings/poolings: w2v + VLAD

```
val word2vec = new Word2Vec()
val model = word2vec.fit(input)
val synonyms =
model.findSynonyms("alquiler", 10)
```

Word2vec "Synonyms"

```
viaje     →  billete vuelo hotel reserva billetes gastos boda regalo casa
alquiler  →  piso cdad alq comunidad local garaje mes prop calle

jose      →  antonio juan manuel francisco luis maria miguel carmen jesus
lloguer   →  pis pagament quota comunitat jordi josep despeses parking joan
2014      →  2013 06 07 08 04 05 02 09 03
madrid    →  viaje hotel barcelona sevilla malaga la san sl reserva
pago      →  factura fra fras fact 14 resto fac 50 facturas
enero     →  febrero marzo abril mayo septiembre noviembre junio agosto
```

References

Scientific literature

- *Jegou et al., Aggregating local descriptors into compact codes, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012.*
- *Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, ECML 1998.*
- *Mikolov et al., Distributed representation of words and phrases and their compositionality, NIPS 2013.*
- *Clinchant and Perronnin, Aggregating Continuous Word Embeddings for Information Retrieval, Workshop on Continuous Vector Space Models and Their Compositionality, 2013.*

Docs and posts

- *Feature Extraction and Transformation using the RDD API*
- *Word Embeddings in 2017: Trends and Future Directions*
- *M. Zinkevich, Rules of ML: Best Practices for ML Engineering*