Towards Automating Data Science Workflows at the Scale of Banking

Jose A. Rodríguez-Serrano
BBVA Data & Analytics

DataEngConf
Barcelona, September 2018

Data Science +
Data Engineering

=

Reducing Inefficiencies in
Products or Services…

Example 1: Connect retailers to customers

BBVA
DATA & ANALYTICS

BBVA

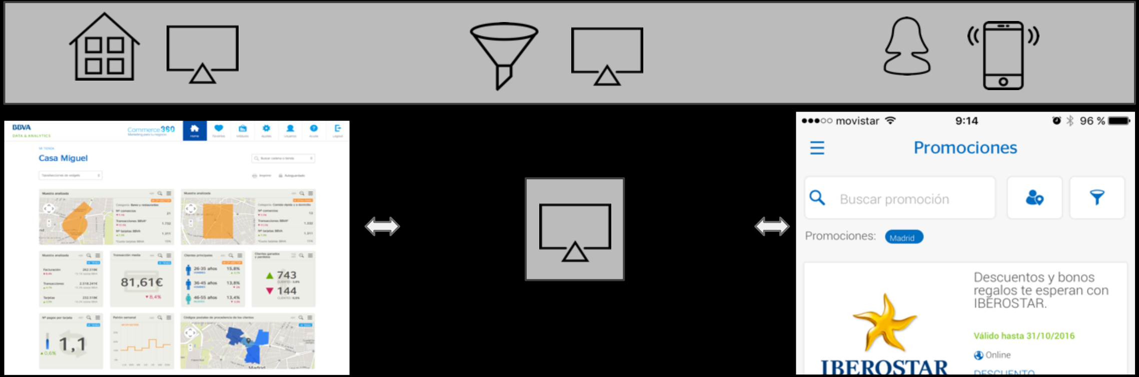
Ahora que sé que nuevos clientes están interesados en mi negocio, les atraigo sin esfuerzo con **promociones personalizadas**

Con ***One Click Campaign*** BBVA identifica cuales son los clientes más afines a tu negocio y te facilita enviarles promociones con un click pagando sólo por los que consigas atraer.

www.bbva.es

Designed for expansion™
Diseñado para la expansión™

Advertising Platform



●●●○○ movistar 9:14 96 %

Promociones

Buscar promoción

Promociones: Madrid

Descuentos y bonos regalos te esperan con IBEROSTAR.

Válido hasta 31/10/2016

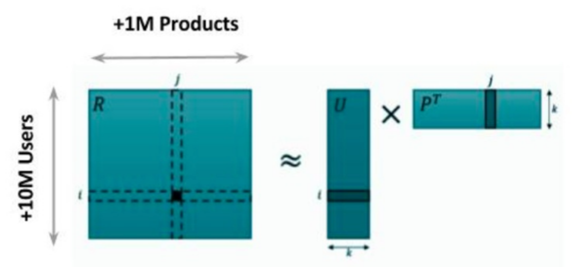
Online

IBEROSTAR

Retailer Dashboard

Campaign manager

Client App

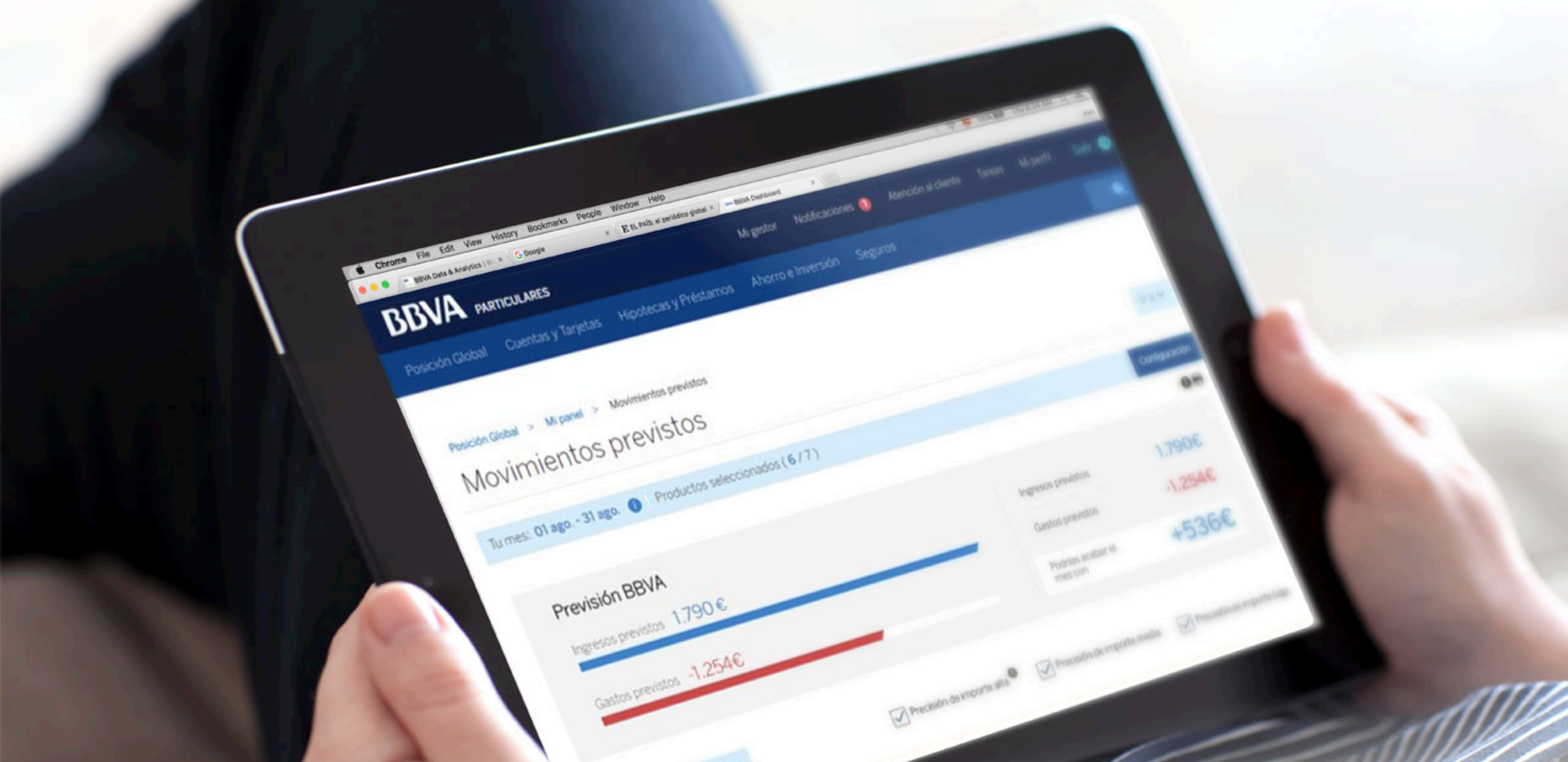


https://www.bbvadata.com/cost-effective-scalable-collaborative-filtering-based-recommender-system/

Example 2: Browse expenses more meaningfully

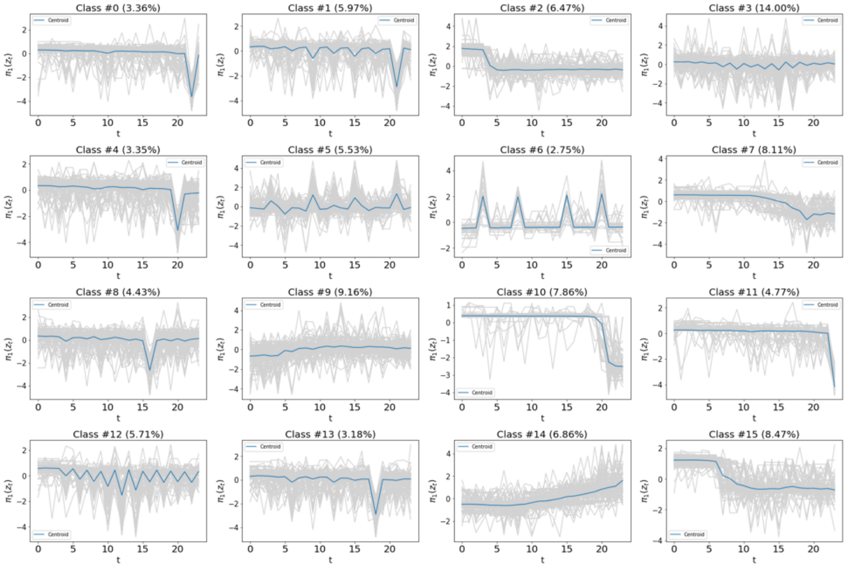
https://databricks.com/session/classifying-text-in-money-transfers-a-use-case-of-apache-spark-in-production-for-banking

Example 3: Forecast Expenses

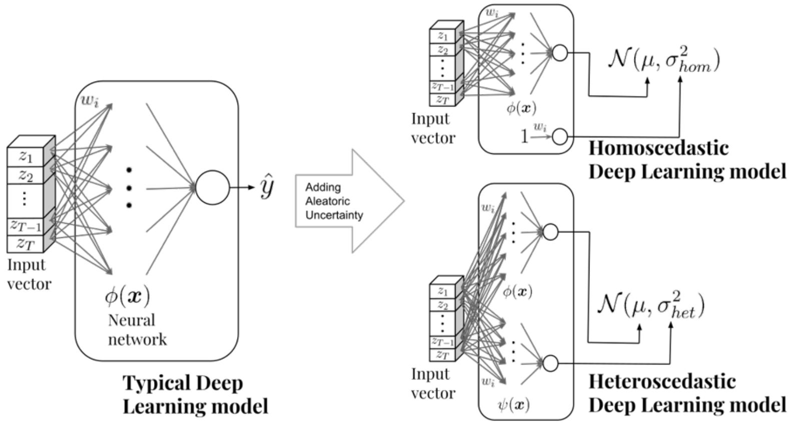

Chrome File Edit View History Bookmarks People Window Help
BBVA Data & Analytics
Google
EL PAÍS: el periódico global
BBVA Dashboard
BBVA PARTICULARES
Mi gestor
Notificaciones
Atención al cliente
Tareas
Mi perfil
Salir
Posición Global
Cuentas y Tarjetas
Hipotecas y Préstamos
Ahorro e Inversión
Seguros
Configuración
Posición Global > Mi panel > Movimientos previstos
Movimientos previstos
Tu mes: 01 ago. - 31 ago.
Productos seleccionados (6 / 7)
Previsión BBVA
Ingresos previstos 1.790 €
Gastos previstos -1.254€
Ingresos previstos 1.790€
Gastos previstos -1.254€
Podrías acabar el mes con +536€
Precisión de importe alta
Precisión de importe media
Precisión de importe baja

Behind the Scenes:
Expense Forecasting Research

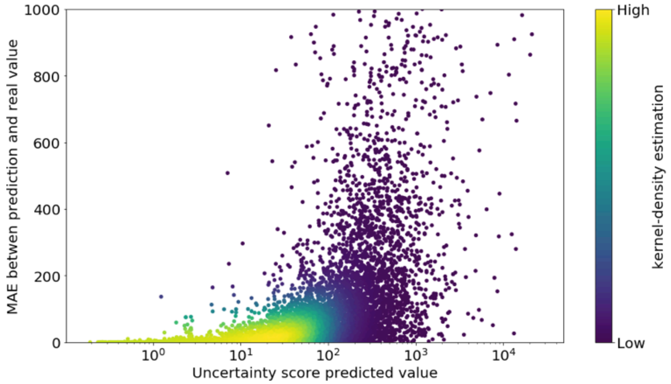
Real-world dataset of human expenses
> 100M time series / month



Heteroscedastic Neural Network

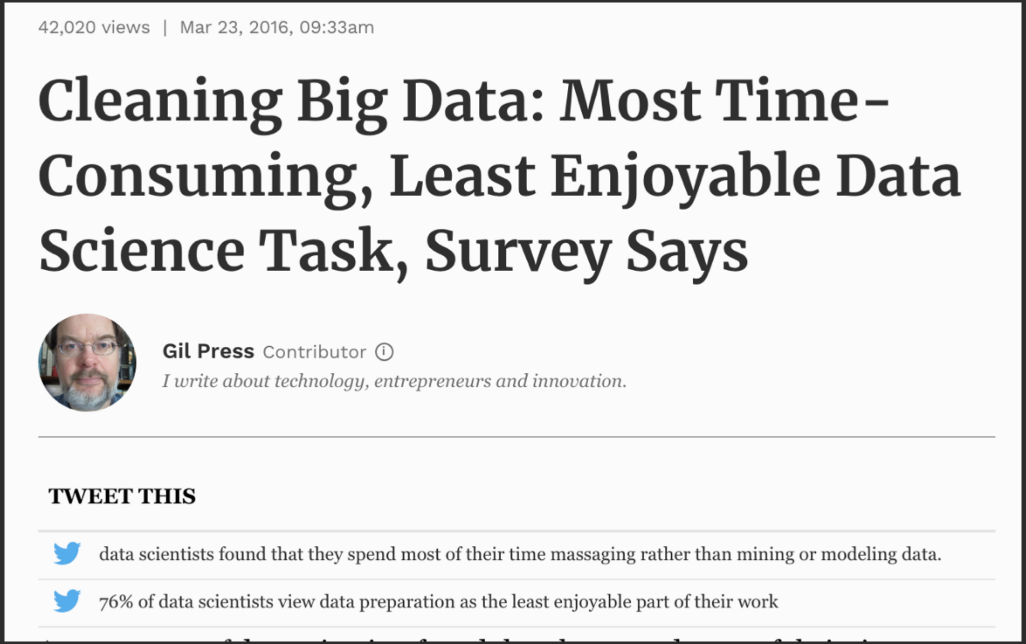


z_1
z_2
z_{T-1}
z_T
Input vector
w_i
$\phi(\boldsymbol{x})$
Neural network
$\hat{y}$
Typical Deep Learning model
Adding Aleatoric Uncertainty
$\mathcal{N}(\mu, \sigma^2_{hom})$
1
Homoscedastic Deep Learning model
$\psi(\boldsymbol{x})$
$\mathcal{N}(\mu, \sigma^2_{het})$
Heteroscedastic Deep Learning model



Brando et al., *Uncertainty Modelling in Deep Networks: Forecasting Short and Noisy Series*, ECML 2018 but…

Data Science Workflows have inefficiencies themselves



42,020 views | Mar 23, 2016, 09:33am

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

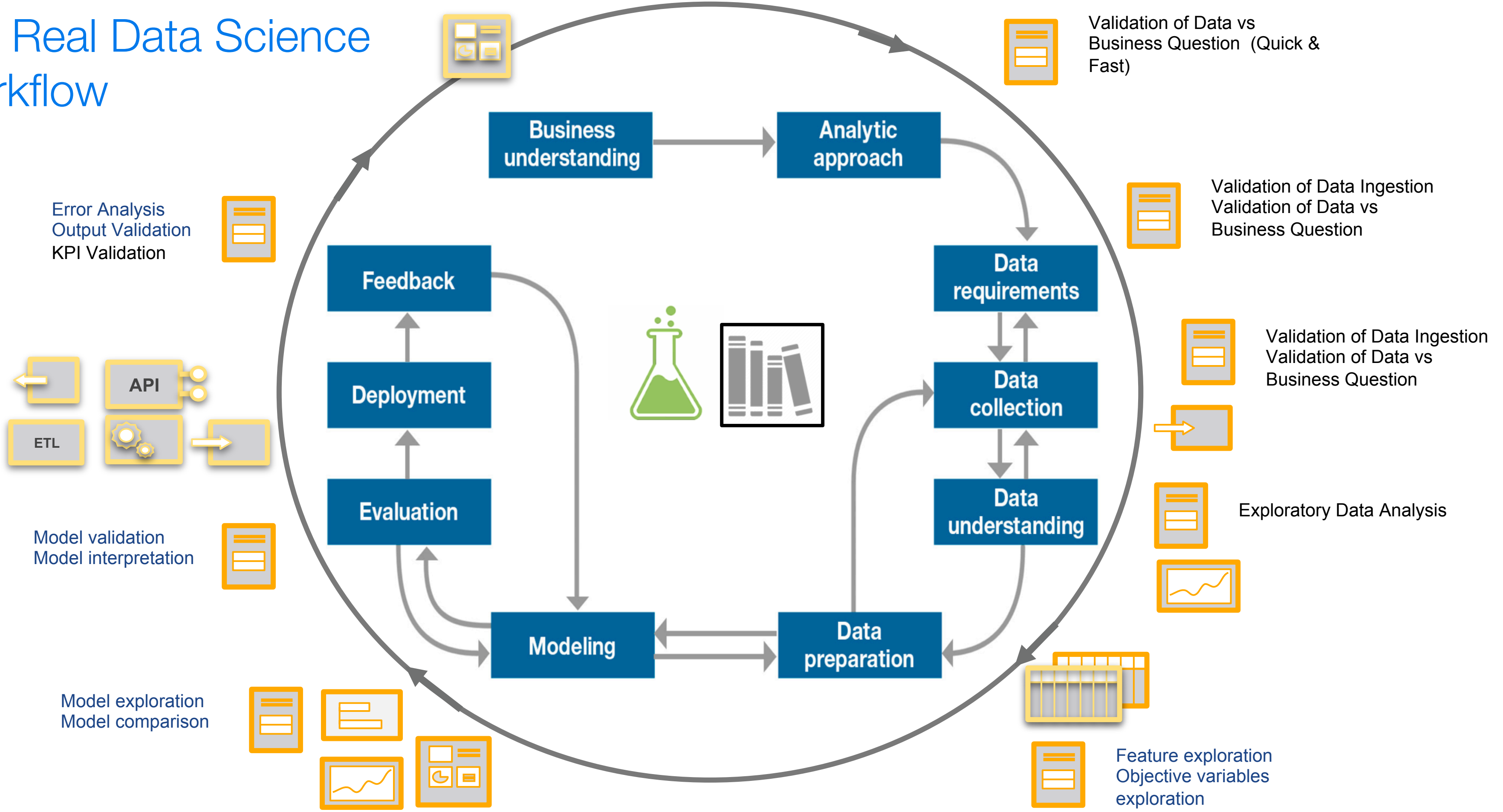
Gil Press Contributor ⓘ

I write about technology, entrepreneurs and innovation.

TWEET THIS

- data scientists found that they spend most of their time massaging rather than mining or modeling data.
- 76% of data scientists view data preparation as the least enjoyable part of their work

We did document The Real Data Science Workflow



Validation of Data vs Business Question (Quick & Fast)

Validation of Data Ingestion
Validation of Data vs
Business Question

Validation of Data Ingestion
Validation of Data vs
Business Question

Exploratory Data Analysis

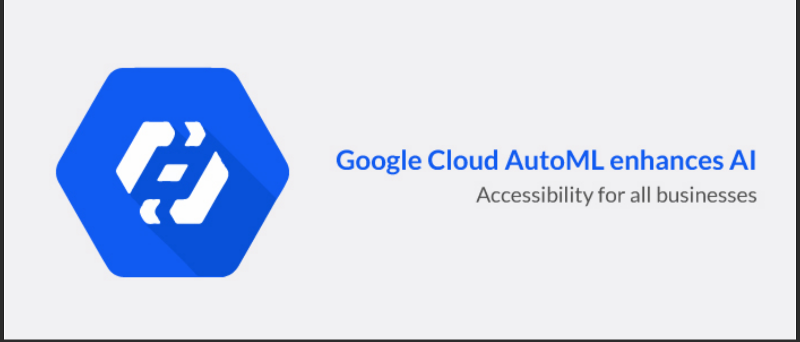
Feature exploration
Objective variables exploration

Model exploration
Model comparison

Model validation
Model interpretation

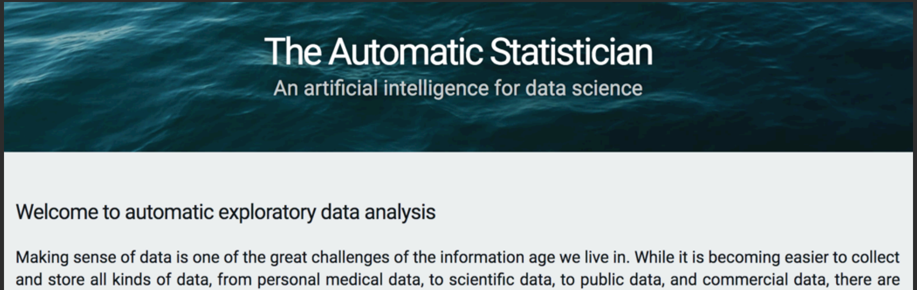
Error Analysis
Output Validation
KPI Validation

There do exist attempts to automate those workflows



Google Cloud AutoML enhances AI

Accessibility for all businesses



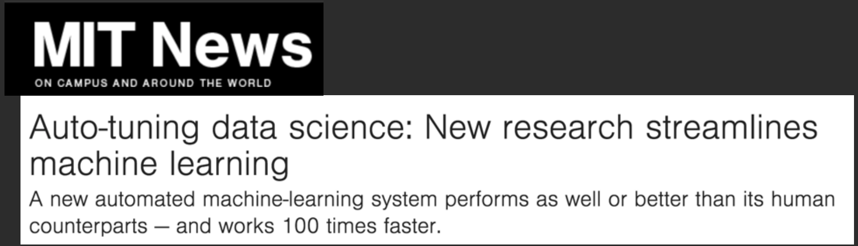
The Automatic Statistician

An artificial intelligence for data science

Welcome to automatic exploratory data analysis

Making sense of data is one of the great challenges of the information age we live in. While it is becoming easier to collect and store all kinds of data, from personal medical data, to scientific data, to public data, and commercial data, there are

Consider TPOT your **Data Science Assistant**. TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming.



MIT News

ON CAMPUS AND AROUND THE WORLD

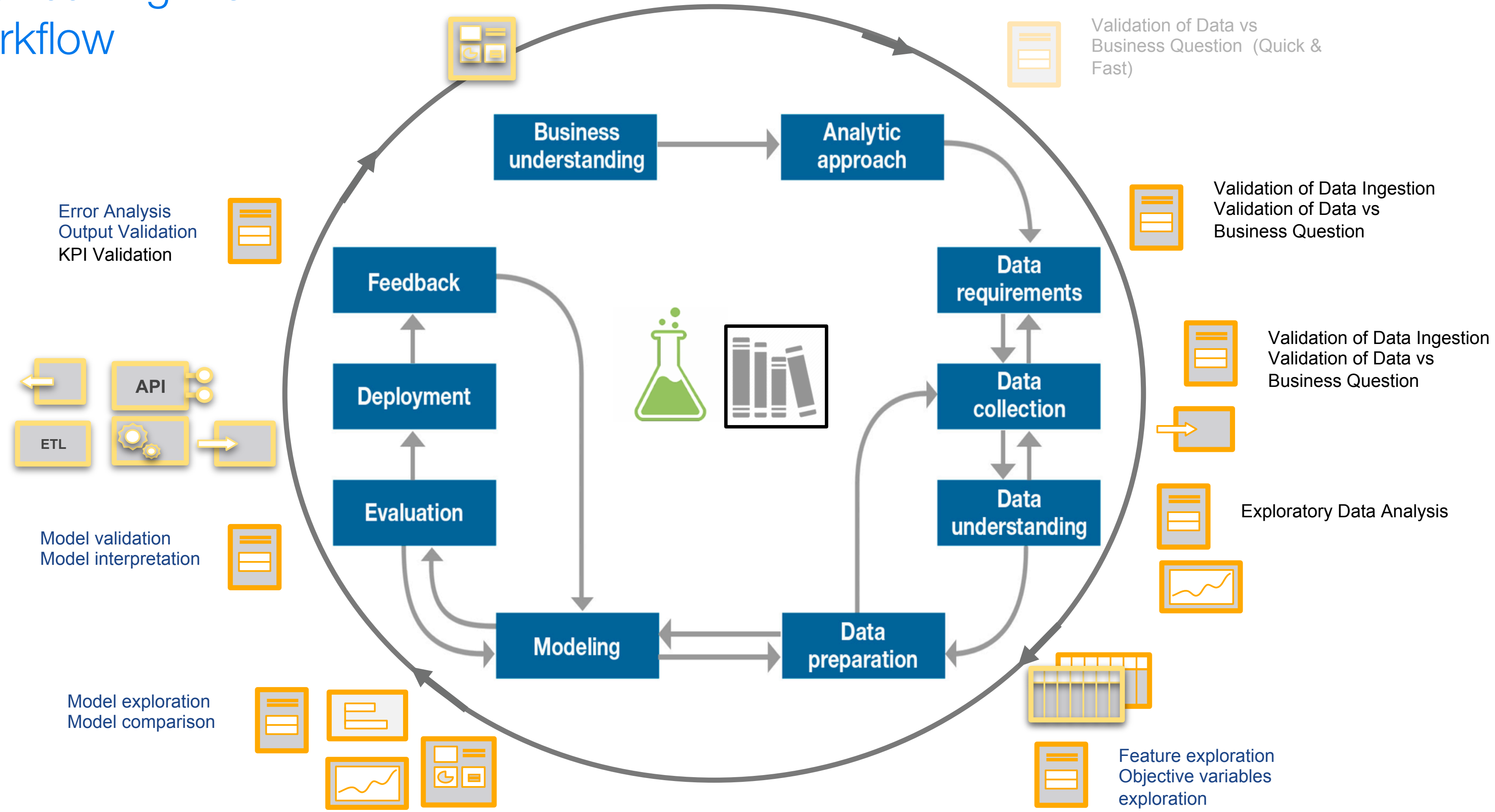
Auto-tuning data science: New research streamlines machine learning

A new automated machine-learning system performs as well or better than its human counterparts — and works 100 times faster.

Towards Increased Data Science Efficiency at BBVA Data & Analytics

client2vec:
Systematic Baselines

Shortcutting the Workflow



Validation of Data vs Business Question (Quick & Fast)

Validation of Data Ingestion
Validation of Data vs
Business Question

Validation of Data Ingestion
Validation of Data vs
Business Question

Exploratory Data Analysis

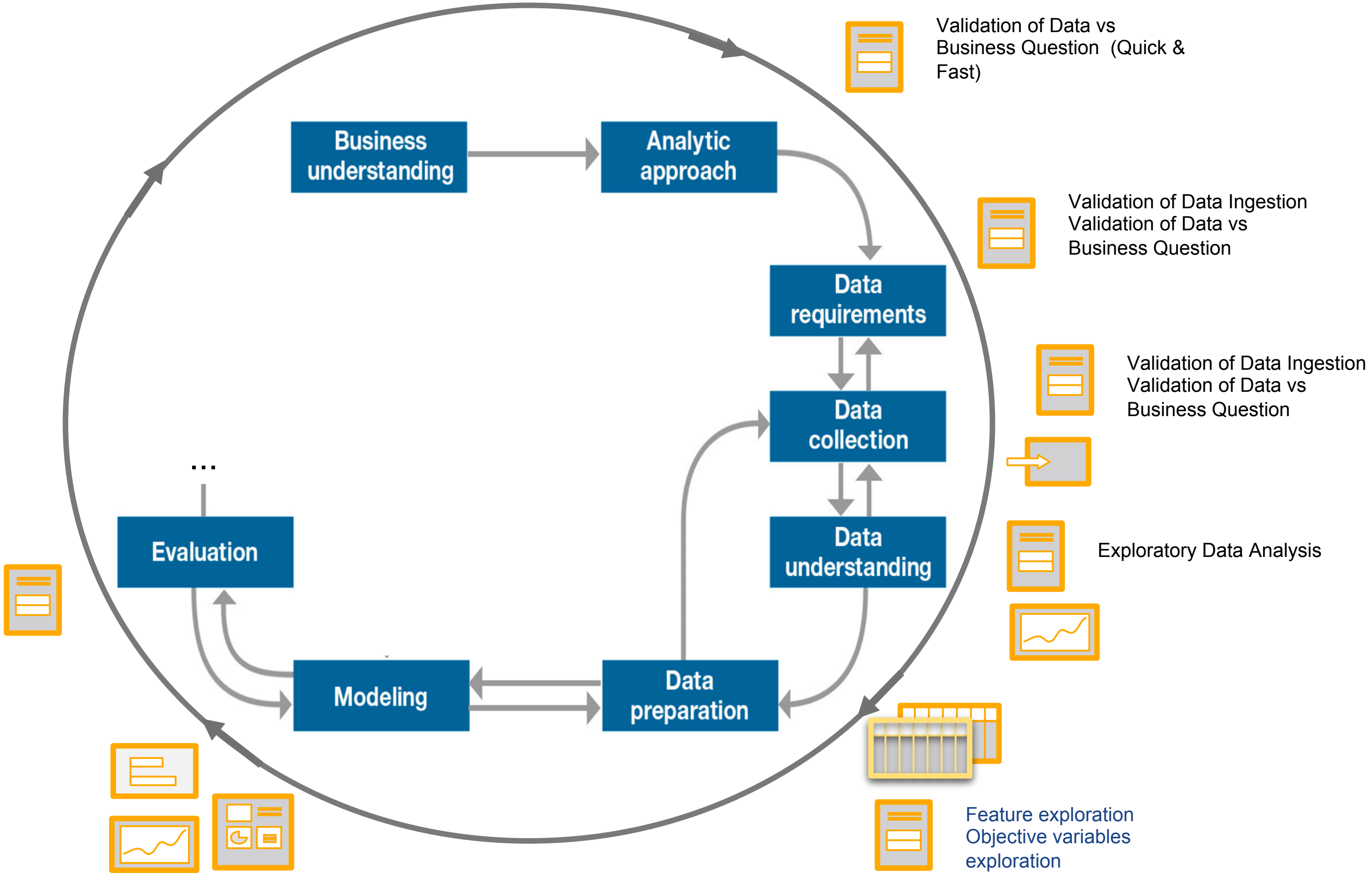
Feature exploration
Objective variables exploration

Model exploration
Model comparison

Model validation
Model interpretation

Error Analysis
Output Validation
KPI Validation

Shortcutting the Workflow



Business understanding

Analytic approach

Data requirements

Data collection

Data understanding

Data preparation

Modeling

Evaluation

...

Validation of Data vs Business Question (Quick & Fast)

Validation of Data Ingestion
Validation of Data vs Business Question

Validation of Data Ingestion
Validation of Data vs Business Question

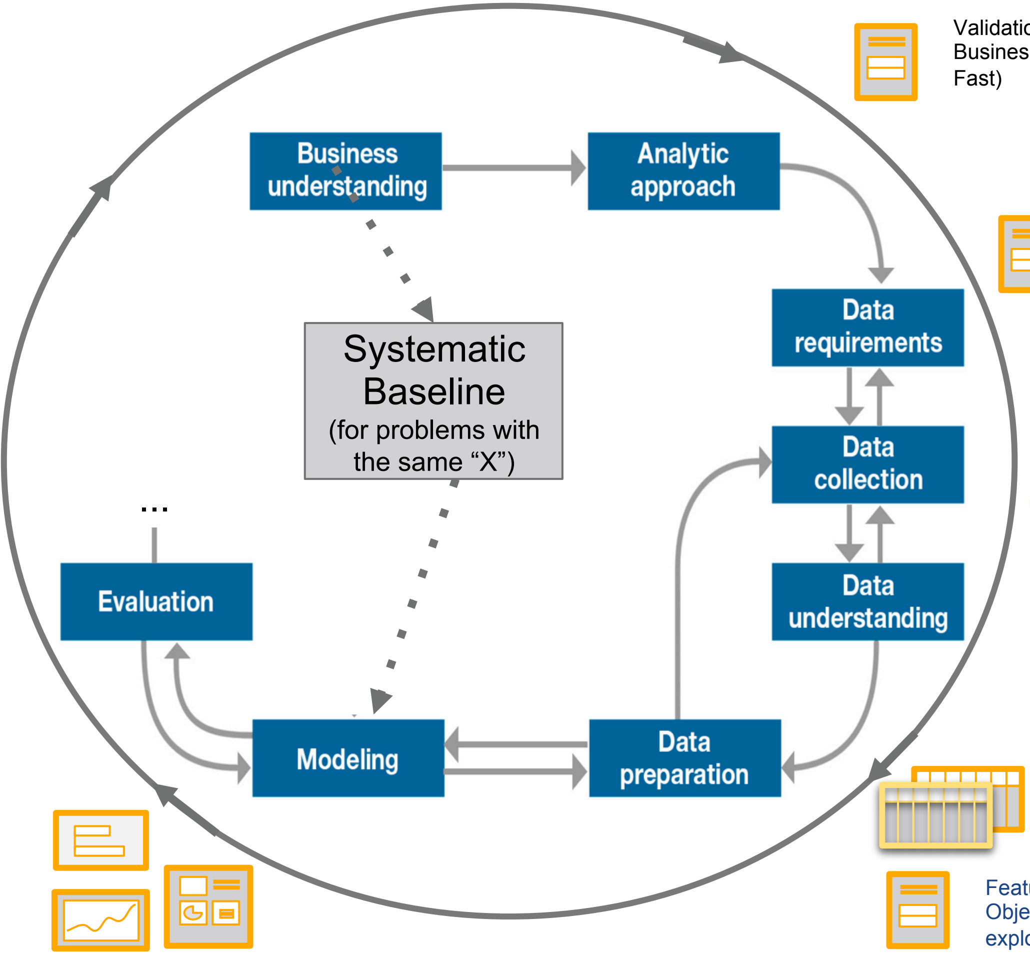
Exploratory Data Analysis

Feature exploration
Objective variables exploration

Model validation
Model interpretation

Model exploration
Model comparison

Shortcutting the Workflow…



Validation of Data vs Business Question (Quick & Fast)

Validation of Data Ingestion
Validation of Data vs
Business Question

Validation of Data Ingestion
Validation of Data vs
Business Question

Exploratory Data Analysis

Feature exploration
Objective variables exploration

Model validation
Model interpretation

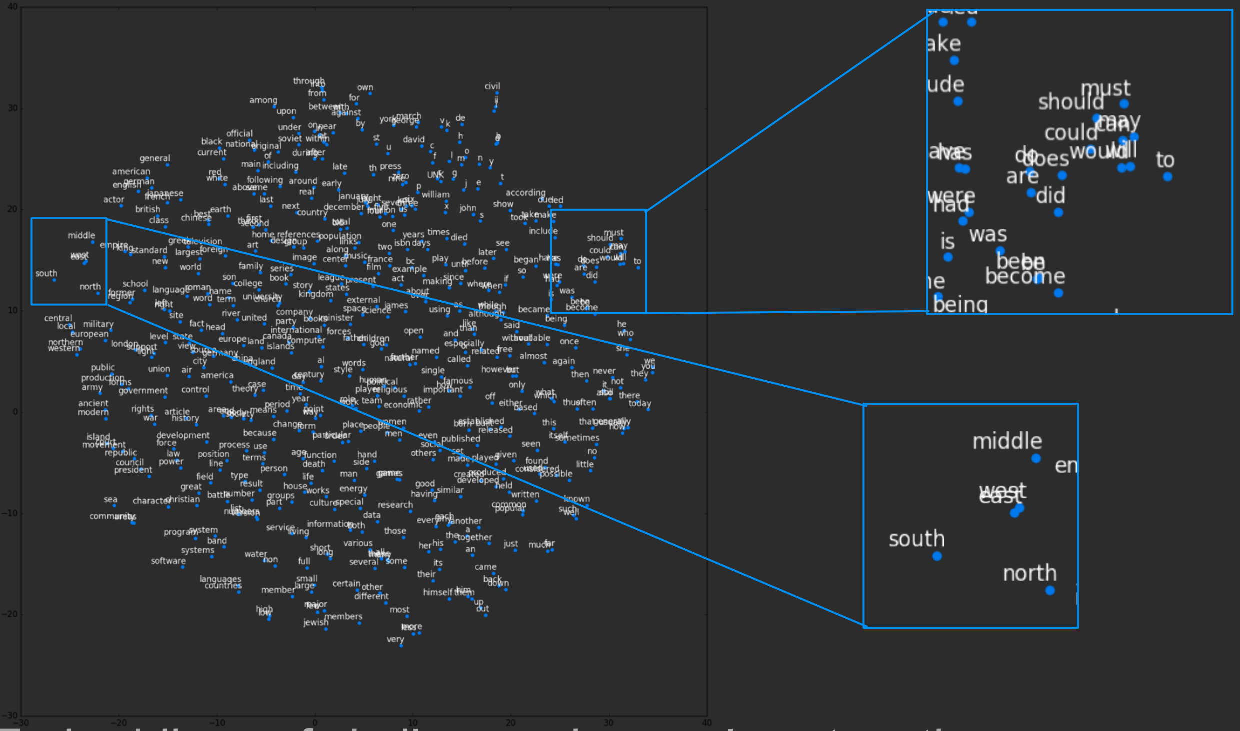
Model exploration
Model comparison

Systematic Baseline =

Generic entity representation
+ flexible & simple model

Do Generic Attributes Exist?

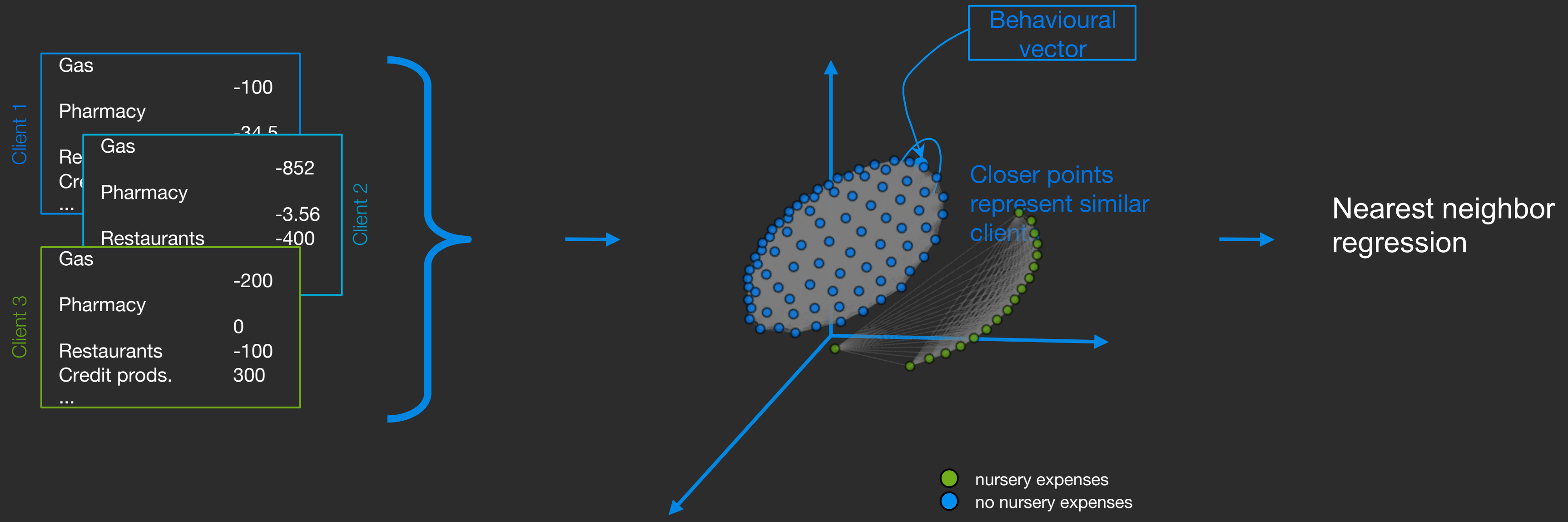
DATA & ANALYTICS



word2vec: Embeddings of similar words are close together

Our Systematic Baseline

BBVA
DATA & ANALYTICS



Client 1

Gas -100
Pharmacy -34.5
Re
Cre
...

Client 2

Gas -852
Pharmacy -3.56
Restaurants -400

Client 3

Gas -200
Pharmacy 0
Restaurants -100
Credit prods. 300
...

Behavioural vector

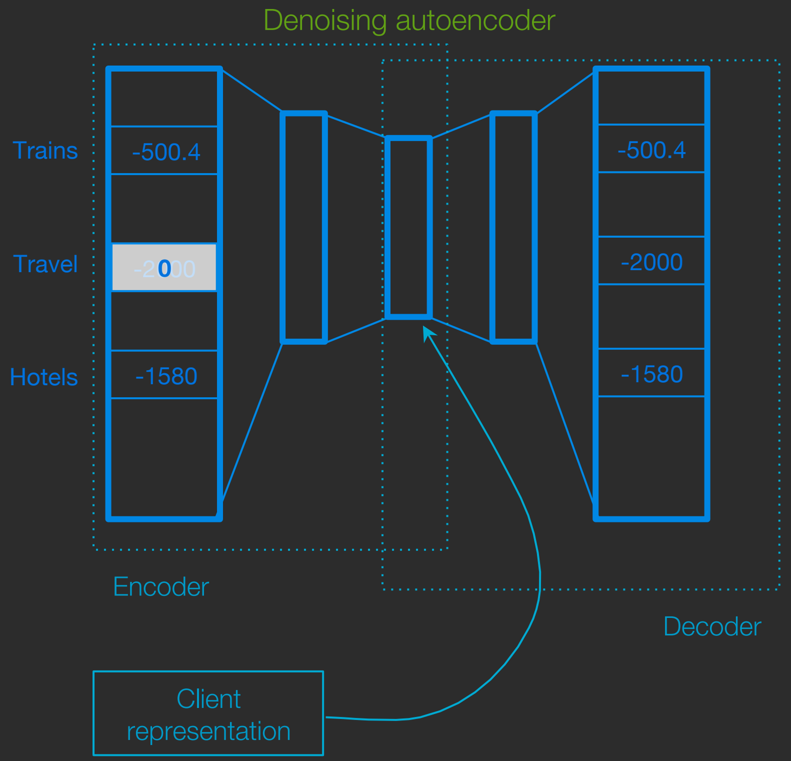
Closer points represent similar client

Nearest neighbor regression

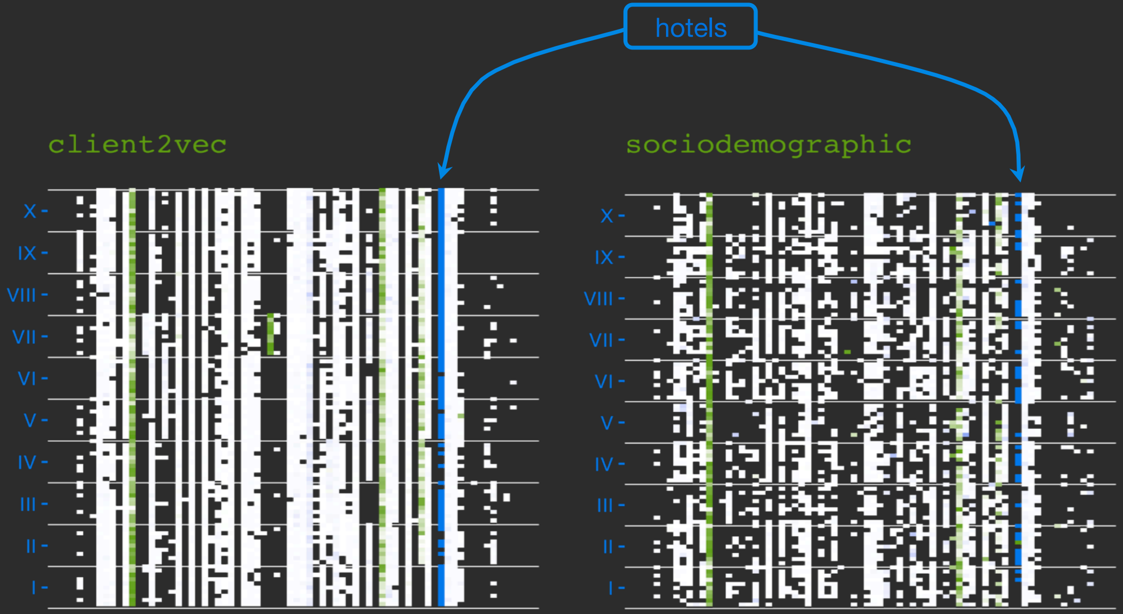
Computing the generic representation

BBVA

DATA & ANALYTICS



- Learn to reconstruct corrupted data
- Reconstruction ≈ similarity
- Marginalized stacked denoising autoencoders (Chen et al, ICML 2012)



Sociodemographic variables don’t capture typical behaviour

BBVA

DATA & ANALYTICS

Improvement on two use cases

+61.6%

Client clustering

Group similar clients and compare their expenses in a target category

+76.1%

Category prediction

By looking at similar clients, guess whether a client had an expense in a target category

Baldassini et al., client2vec: Towards Systematic Baselines for Banking Applications, arXiv, 2018

Our bet

Algorithmic research

Explore, evaluate and generate state-of-the-art algorithms

- State-of-the-art algorithm
- Best performance in our use cases

Implementation

Deliver algorithmic solutions as a software package

- Lightweight training
- Works on bank's infrastructure

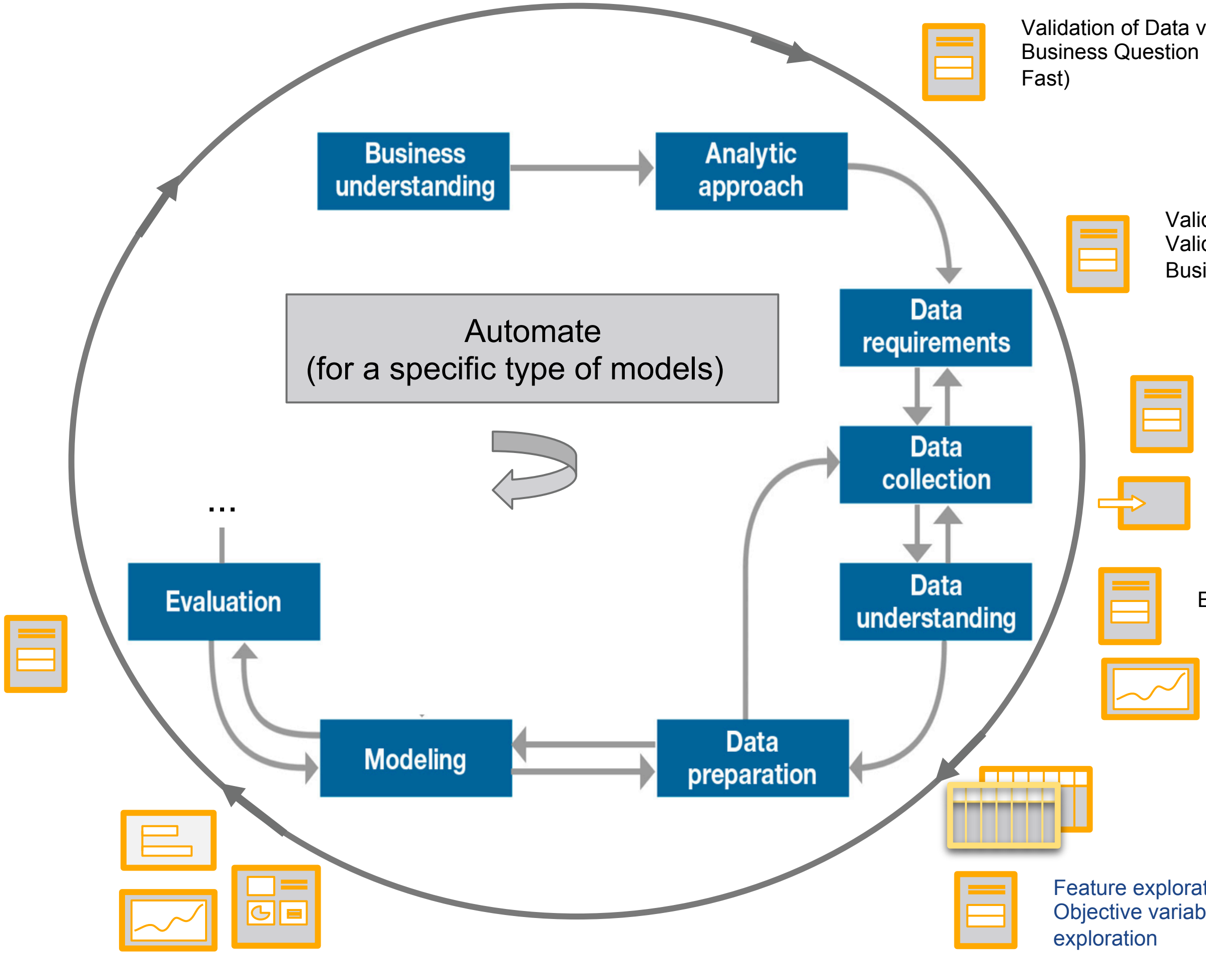
Product enablement

Generate capabilities to accelerate product development

- Better method to compare clients
- Tools to evaluate embedding methods

Automatic Modeller:
Searching end-to-end linear, interpretable models

Shortcutting the Workflow…



Business understanding

Analytic approach

Data requirements

Data collection

Data understanding

Data preparation

Modeling

Evaluation

…

Automate
(for a specific type of models)

Validation of Data vs Business Question (Quick & Fast)

Validation of Data Ingestion
Validation of Data vs Business Question

Validation of Data Ingestion
Validation of Data vs Business Question

Exploratory Data Analysis

Feature exploration
Objective variables exploration

Model validation
Model interpretation

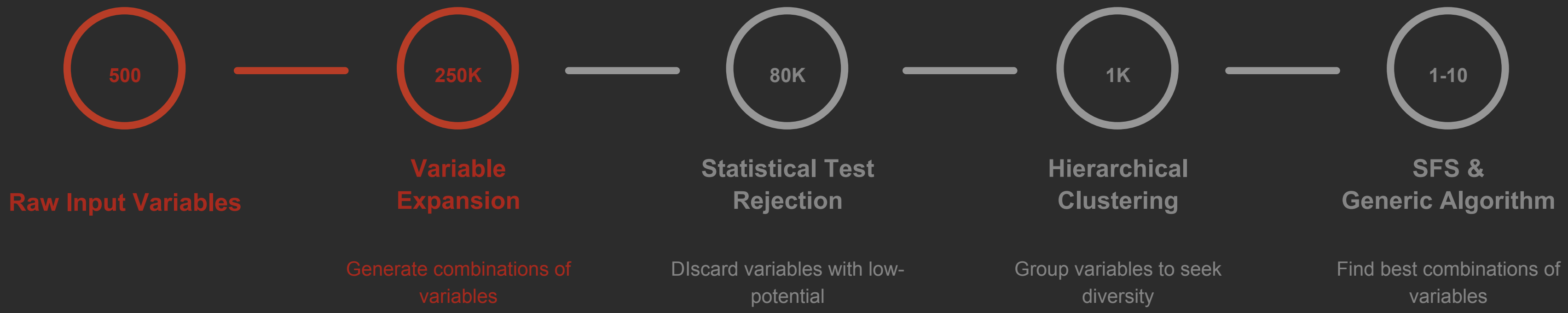
Model exploration
Model comparison

Automatic Modeller

Pipeline to seek

- Linear Models
- “Interpretable” variable meanings
- Across multiple metrics (e.g. model quality vs number of features)

(Number of variables in circles)



500

Raw Input Variables

250K

Variable Expansion

Generate combinations of variables

80K

Statistical Test Rejection

DIscard variables with low-potential

1K

Hierarchical Clustering

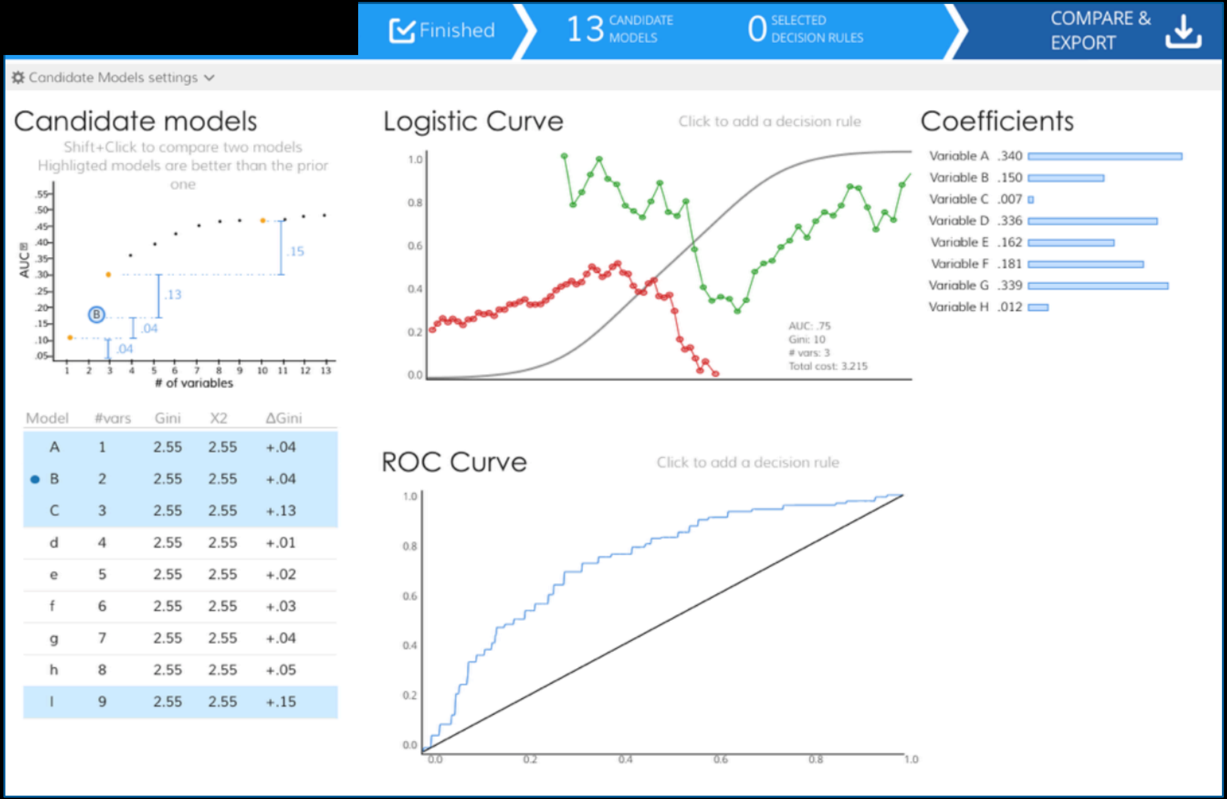
Group variables to seek diversity

1-10

SFS & Generic Algorithm

Find best combinations of variables

DATA & ANALYTICS



Candidate models

Shift+Click to compare two models
Highligted models are better than the prior one

Model	#vars	Gini	X2	ΔGini
A	1	2.55	2.55	+.04
● B	2	2.55	2.55	+.04
C	3	2.55	2.55	+.13
d	4	2.55	2.55	+.01
e	5	2.55	2.55	+.02
f	6	2.55	2.55	+.03
g	7	2.55	2.55	+.04
h	8	2.55	2.55	+.05
I	9	2.55	2.55	+.15

Logistic Curve

Click to add a decision rule

ROC Curve

Click to add a decision rule

Coefficients

Variable A	.340
Variable B	.150
Variable C	.007
Variable D	.336
Variable E	.162
Variable F	.181
Variable G	.339
Variable H	.012

Learnings

- Tools to make data science projects more efficient
  - Generic Customer Attributes
  - Linear, Interpretable Model Construction

- Real tools available to Data Scientists & Data Engineers
- The philosophy somewhat experimental
  - The “experiment” is ongoing and subject to important checks: e.g. real reusability

Take-Home Message:
A Buy-vs-Make Learning

- Commoditization of ML algorithms is a reality
- Speed-up for “mainstream” problems (image classification, text classification)
- Still a long tail of problems need very specific domain knowledge and are not addressed by these tools (e.g. classification of text in bank transactions, pricing, etc)
- Still room for “commoditizing” internally

Acknowledgements

Many thanks for providing insights & part of the contents:

- Leonardo Baldassini
- Roberto Maestre
- Alberto Rubio
- César de Pablo
- Juan Arévalo
- Axel Brando
- Javier Lopez

Thanks!

BBVA
DATA & ANALYTICS

Questions?

Get in touch at “Office Hours” @ DataEngConf:

2:15 PM - 3:00 PM

Or visit **bbvatada.com**