

# Marketing de conducción autónoma

# Índice

<b>Prólogo.....</b>	<b>4</b>
<b>Sobre el título .....</b>	<b>5</b>
Nuestras experiencias	
Análisis de datos	
Procesos de conducción autónoma	
¿Big Data, Data Science, Machine Learning?	
<b>Estructura del libro .....</b>	<b>7</b>
Capítulos	
<b>Catálogo y fuerza de ventas.....</b>	<b>8</b>
Punto de partida	
Datos iniciales	
Análisis	
Rutas de navegación	
‘Cestas’ de navegación	
Ejemplos de reglas o cestas de navegación	
Conclusiones	
<b>Comentarios y clientes .....</b>	<b>18</b>
Punto de partida	
Datos iniciales	
Análisis	
Detección de idioma, traducción y extracción de keywords	
Clasificación de comentarios	
Conclusiones	
<b>Promociones y prescriptores.....</b>	<b>24</b>
Punto de partida	
Datos iniciales	
Actores	
Análisis de talleres	
Caracterización	
Modelo predictivo	
Resultados del modelo	
Análisis de flotas	
Caracterización	
Modelo predictivo	
Resultados del modelo	
Conclusiones	
<b>Incentivos y fidelización .....</b>	<b>36</b>
Punto de partida	
Datos iniciales	

Conceptos de facturación	
De la reserva al cliente	
Análisis	
Categorías de clientes	
Incentivos y facturación	
Incentivos y demografía	
Intermediación y facturación	
Conclusiones	
<b>Experimentación .....</b>	<b>50</b>
Punto de partida	
Big data y experimentación	
Metodología	
Hipótesis	
Validación de hipótesis	
Resultados esperados y reales	
Instrumental	
Datos	
Software	
Conclusiones	
<b>Predicción .....</b>	<b>64</b>
Punto de partida	
Predicciones sin experimento	
Respuesta a una campaña	
Predicción de respuestas	
Predicciones con resultados históricos	
Instrumental	
Datos	
Software	
Conclusiones	
<b>Glosario.....</b>	<b>80</b>

# Prólogo

Empezaré con una anécdota.

Hace unos años, un ex-vicepresidente de un banco, de los que llamábamos extranjeros, me invitó a comer. Lo hacía de vez en cuando, en restaurantes muy caros, donde no pasábamos de pedir un primero de legumbres y de compartir media botella de vino: “porque si no, Pedro, nos llenamos”. No eran buenos tiempos.

En ese momento de la comida en que lo único que queda sobre el mantel son copas con dos dedos de vino, le dije: “Algún día tengo que escribir un libro sobre lo que he vivido como profesional; hay cosas interesantes”. Antes de responderme, apartó ligeramente su copa, luego inclinó la cabeza hacia mí, me clavó los ojos y me dijo: “Mira, Pedro, escribir un libro lo escribe cualquiera; lo hacen millones de personas. Hoy en día, lo difícil es leerlo.”

He escrito este libro con la lección aprendida: ¡qué más da lo que a mí me parezca interesante! Lo único que cuenta es lo que pueda interesar a tu lector.

Y ahora, juguemos al dominó; porque de eso va este libro: de una partida de dominó con proveedores, clientes y competidores.

Para ganar al dominó hay que hacer bien dos cosas: recordar las fichas que cada jugador ha puesto sobre la mesa y, basándose en ellas, prever las que puedan poner después y en qué orden. De eso he escrito en este libro: de conocer bien lo que ha ocurrido para prever lo que pueda ocurrir después; de analizar el pasado y predecir, dentro de los márgenes de la probabilidad, el futuro.

Explotar los datos del negocio, lo que vendemos, cuánto, cuándo, dónde, a quién, a qué precio, tiene que tener como objetivo prioritario anticipar el movimiento siguiente de nuestros clientes, de nuestra red de ventas externa, de nuestros competidores. Ese es el enfoque que he intentado mostrar en este libro: analizar, predecir, mover <sup>1</sup>.

Pedro Nevado

---

<sup>1</sup> Algun lector podría pensar “¿y por qué no jugar al ajedrez?” El ajedrez carece del componente aleatorio, de la suerte: los dos jugadores parten con las mismas piezas. En el dominó pueden tocarte seis fichas dobles y a tus rivales ninguna: mala suerte. Los negocios, sin exagerar, también gestionan el azar.

# Sobre el título

En este libro hemos recogido nuestras experiencias en una línea de trabajo: tratar el análisis de datos de negocio como un input intermedio para el diseño y la implantación de procesos de marketing de *conducción autónoma*.

## Nuestras experiencias

Hemos escrito sobre lo que hemos hecho con algunos de nuestros clientes: departamentos de marketing y ventas de medianas empresas, muy enfocados en el día a día, con un poder de decisión reducido en las grandes decisiones tecnológicas que les afectan.

Las experiencias reflejan la presión por obtener primeros resultados en poco tiempo, sin grandes transtornos en sus procedimientos actuales, ni grandes requerimientos de producir nueva información: “esto es lo que hay; a ver qué podemos hacer con ello”, ha sido casi siempre un estribillo en las reuniones.

## Análisis de datos

Hemos trabajado con la información de la que ya disponían nuestros clientes: datos de facturación y de ventas unitarias, principalmente. El objetivo ha sido obtener, a partir de ellos, cambiando las técnicas y los puntos de vista del análisis, una visión diferente, complementaria a la que ya tenían, de sus clientes y de sus políticas de marketing.

También hemos perseguido un tipo de análisis que no termine en sí mismo, en un informe o en un cuadro de mando con los correspondientes gráficos y tablas; que fuera, más bien, el paso previo al diseño de un nuevo proceso (semi)automatizado en la toma de decisiones comerciales.

## Procesos de conducción autónoma

Hay un trecho importante entre tener inventariado cada centímetro del firme, cada señal de tráfico, cada intersección de vías en un tramo de una carretera, y guiar un vehículo sin conductor a través de él. No es tanto la cantidad de información, como la forma en que se dispone de ella: el énfasis no está en saber, sino en decidir qué hacer a cada paso.

Salvando todas las distancias con el símil anterior, hemos intentando establecer procesos de marketing de *conducción autónoma* e indagar en la clase de información que necesitan. En algunos de los casos que contamos, hemos diseñado el proceso y lo hemos implementado en una solución de software. En otros, hemos analizado la información y hemos diseñado cuál podría ser el nuevo proceso.

# ¿Big Data, Data Science, Machine Learning?

En este documento, al igual que en el trabajo con nuestros clientes, hemos puesto muy poco empeño en clasificar lo que hacemos con la nomenclatura *de moda* adecuada. Para empezar, hacerlo con precisión requeriría de nosotros dejar claro primeramente qué entendemos que comprende cada una de esas categorías. No vamos a aburrir al lector haciéndolo.

Responderemos, simplemente, que hemos utilizado métodos y técnicas que podrían encuadrarse, a nuestro juicio, en cada una de las tres categorías que recoge el título de esta sección.

# Estructura del libro

El libro recoge nuestras experiencias en cinco proyectos de consultoría y desarrollo, con cuatro clientes diferentes de LEBEN. El orden de exposición refleja también el orden cronológico de los proyectos: del más antiguo al más reciente.

Las referencias a proyectos concretos nos han servido para dos propósitos:

- Encuadrar los análisis en problemas reales en negocios de variada naturaleza: el sector turístico, los servicios tecnológicos o la fabricación de componentes mecánicos.
- Ilustrar soluciones técnicas diferentes, tanto en el análisis como en la implementación de un nuevo proceso de marketing o ventas.

Para mantener la confidencialidad de la información y para no entrar en un nivel de detalle innecesario para nuestros objetivos, sólo mostramos, en cada caso, los aspectos más relevantes para la línea de argumentación que vertebría el libro: cómo analizar y modelar los datos de negocio para implementar procesos de marketing de *conducción autónoma*.

## Capítulos

En cada capítulo, correspondiente cada uno de ellos a un proyecto, hemos mantenido una secuencia expositiva similar:

- Punto de partida: objetivos del proyecto y los elementos del negocio a estudiar.
- Los datos suministrados por el cliente y cómo los transformamos.
- Las herramientas de análisis y desarrollo que empleamos.
- Las conclusiones que extrajimos.

# Catálogo y fuerza de ventas

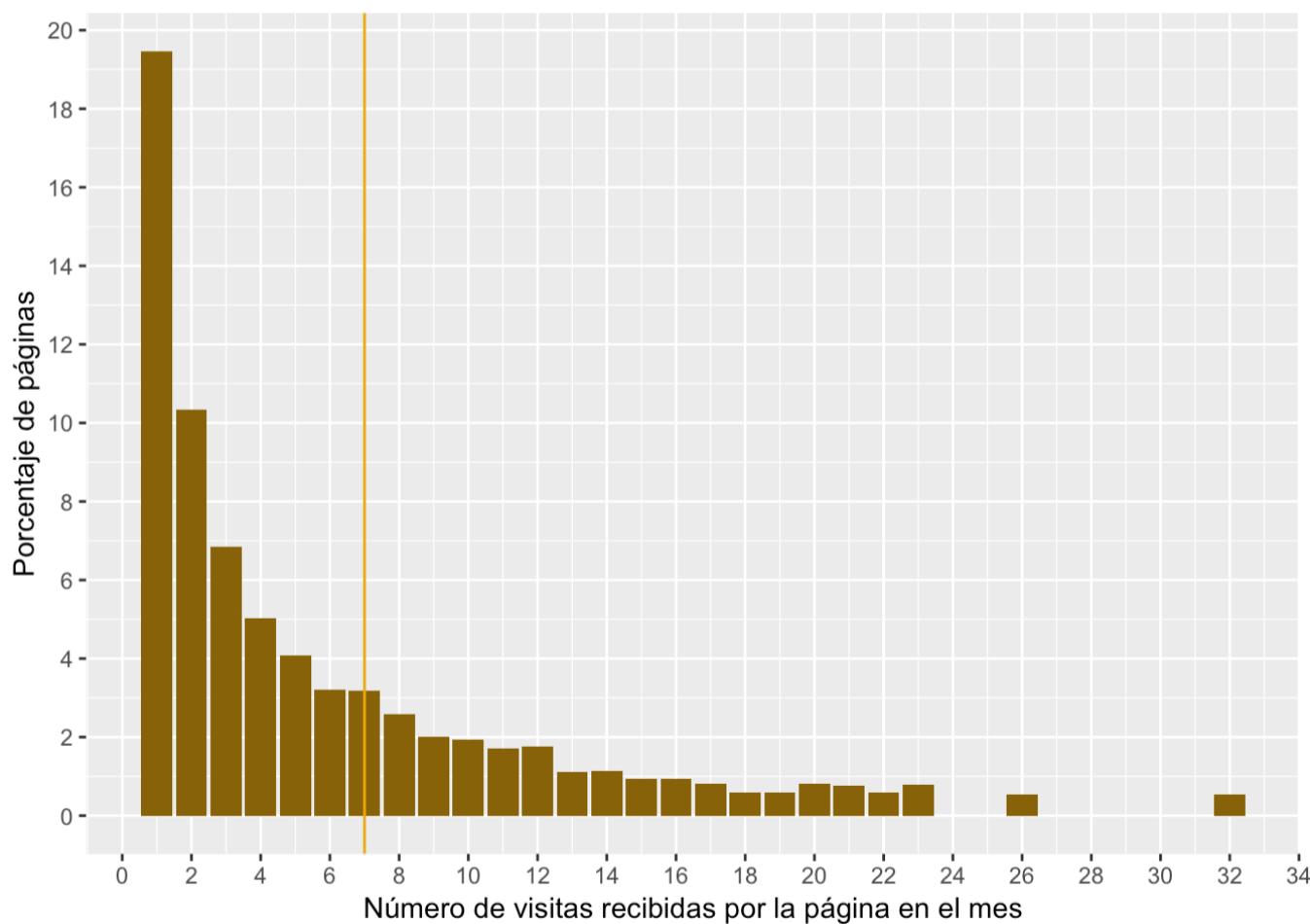
## Punto de partida

¿Cómo mejorar la navegación de un catálogo de productos en Internet, dedicado a una fuerza de ventas extensa y muy diferenciada en sus patrones de consulta de la información?

Para dar una idea de la complejidad del **catálogo** que estudiamos en el proyecto, damos algunas aproximaciones de magnitud:

- Número de páginas: en el orden de unos pocos miles.
- Número de visitas al catálogo por mes: más de medio millón.

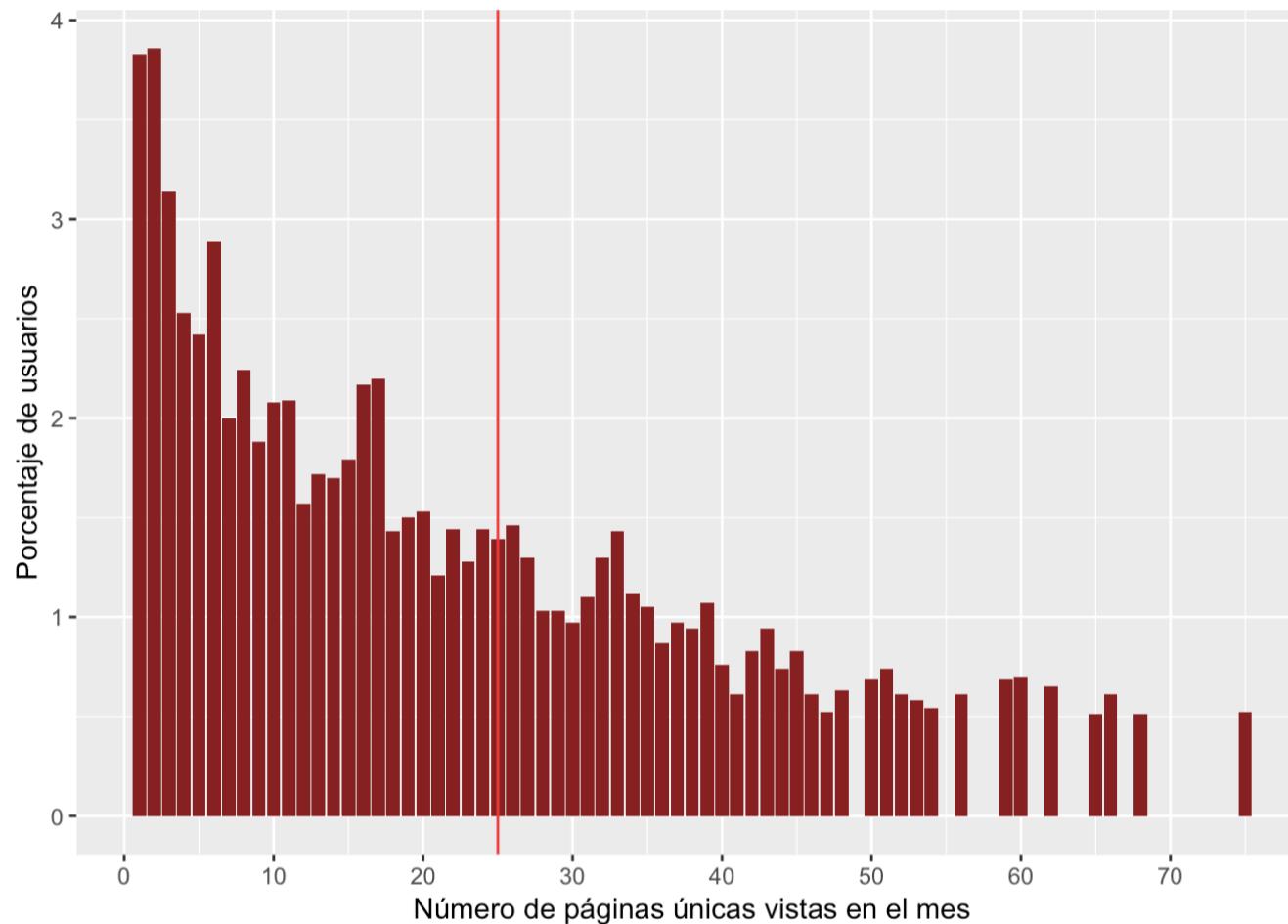
Sin embargo, el 50% de las páginas (la divisoria la marca la línea vertical) recibe, en total, siete visitas o menos en el mes: las visitas se concentran en unas pocas páginas; otras, centenares de ellas, apenas son consultadas.



Desde el punto de vista de los **usuarios**, damos también una idea de las magnitudes:

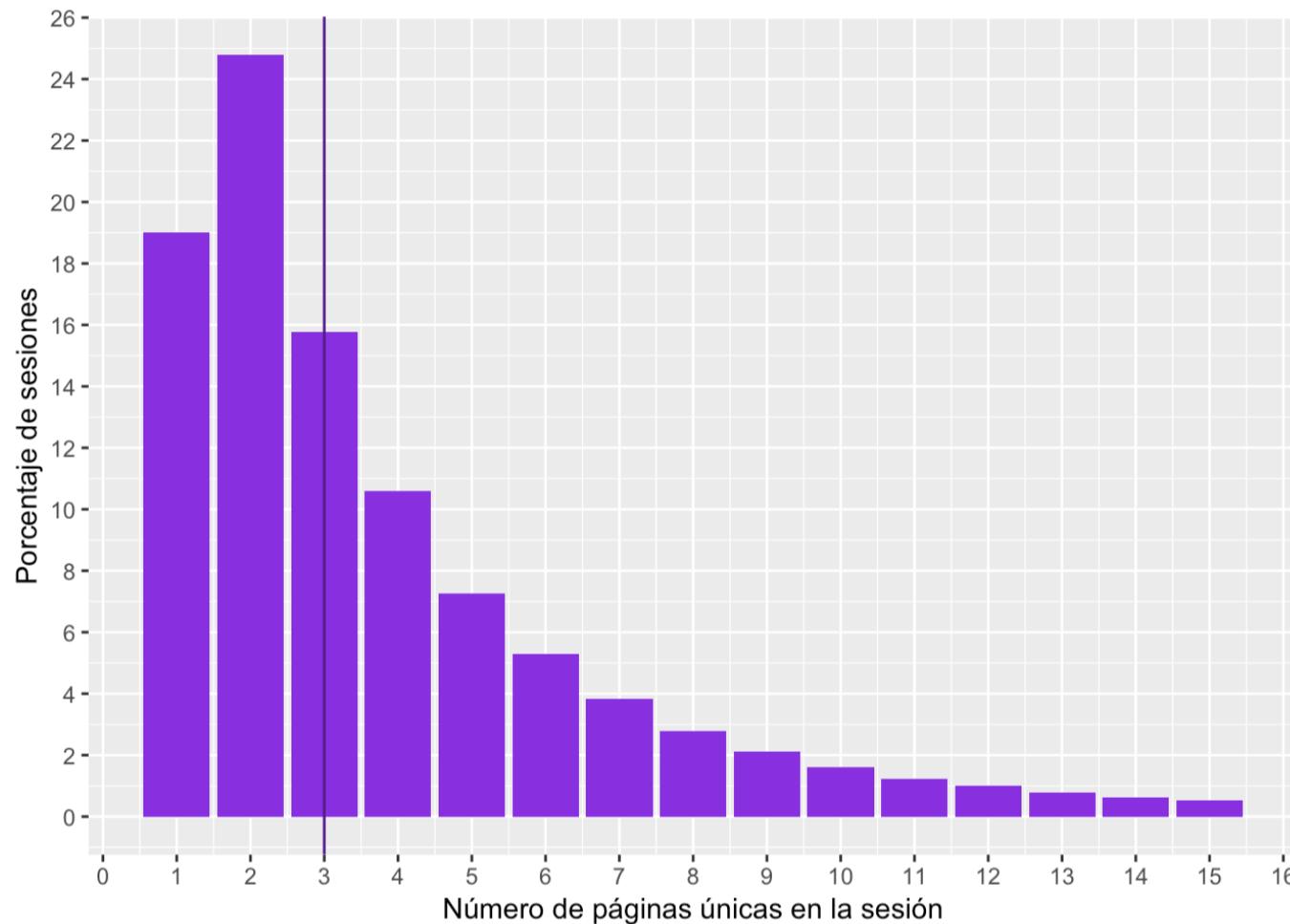
- Número de usuarios registrados: más de mil.
- Tipos de usuario diferenciados por perfil profesional y canal de ventas: más de treinta.

Los usuarios definen pautas de navegación especializadas en función de su perfil: el 50% de ellos (la linea divisoria vertical) ve 25 o menos páginas diferentes en un mes.



Por último, caracterizamos las **sesiones** de navegación. En un catálogo para una fuerza de ventas, las sesiones han de ser previsiblemente de corta duración: el usuario consulta información sobre una promoción o un producto, demandada por un cliente al otro lado de la línea telefónica o del mostrador, y cambia a una aplicación de tramitación de contratos o similar.

Los datos confirman la hipótesis anterior: en más del 50% de las sesiones (línea divisoria vertical) se consultan 3 páginas como máximo.



# Datos iniciales

Los datos para el análisis fueron dos ficheros, con formato CSV, correspondientes a dos meses naturales, con un total aproximado de dos millones y medio de registros. Los registros se compusieron a partir del mecanismo de *tracking* de Google Analytics que incorporaba cada página del catálogo.

Ésta es la información más relevante, para nuestro proyecto, recogida en los ficheros:

- Sesión
- Página actual
- Página anterior
- Página siguiente
- Identificador anónimo del usuario
- Tipo del canal de ventas del usuario
- Perfil funcional del usuario
- Día y hora de la consulta

Procesamos, tabulamos y almacenamos la información en cuatro tablas de una base de datos relacional gestionada por MySQL. Para facilitar el acceso a los diferentes miembros del equipo, utilizamos un servidor remoto en Amazon Relational Database Service.

# Análisis

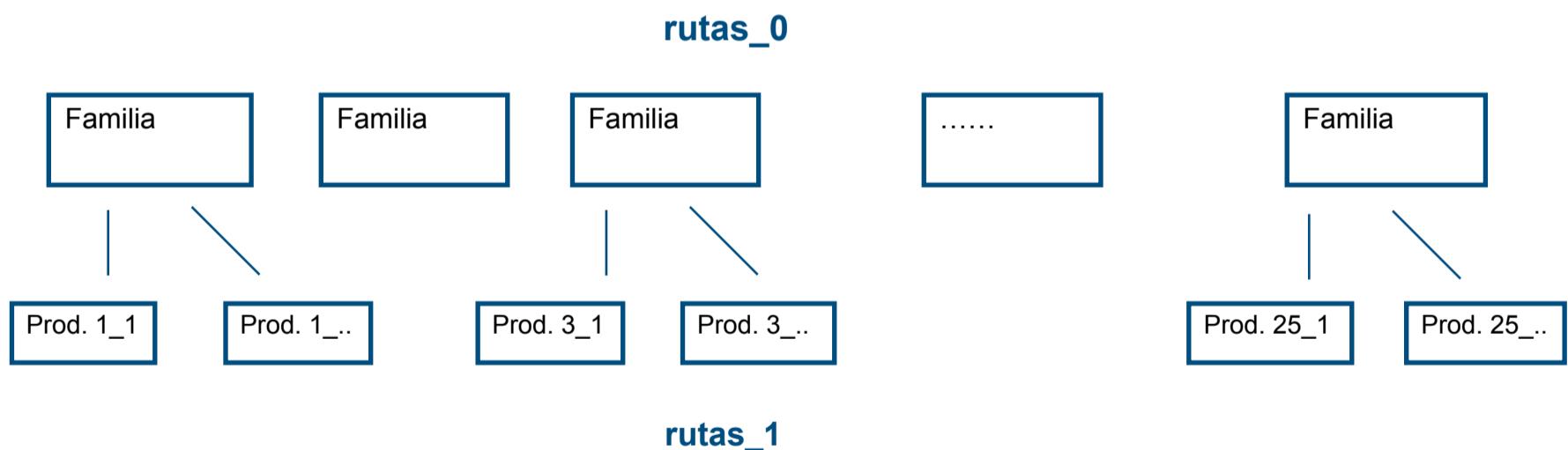
Hay catálogos de productos donde el objetivo de quien lo diseña ha de ser prolongar el tiempo de consulta hasta el máximo posible: cuanto más tiempo pase un usuario viendo títulos de películas, canciones o libros, o imágenes de prendas de vestir, mayor es la probabilidad de que termine comprando alguno de ellos.

En un catálogo de productos para una fuerza de ventas dedicada a atender peticiones de información o de contratación por parte de clientes finales, la motivación ha de ser la contraria: cuanto menos tiempo tarde el usuario en encontrar lo que busca, más clientes atenderá y, probablemente, mayores serán las ventas.

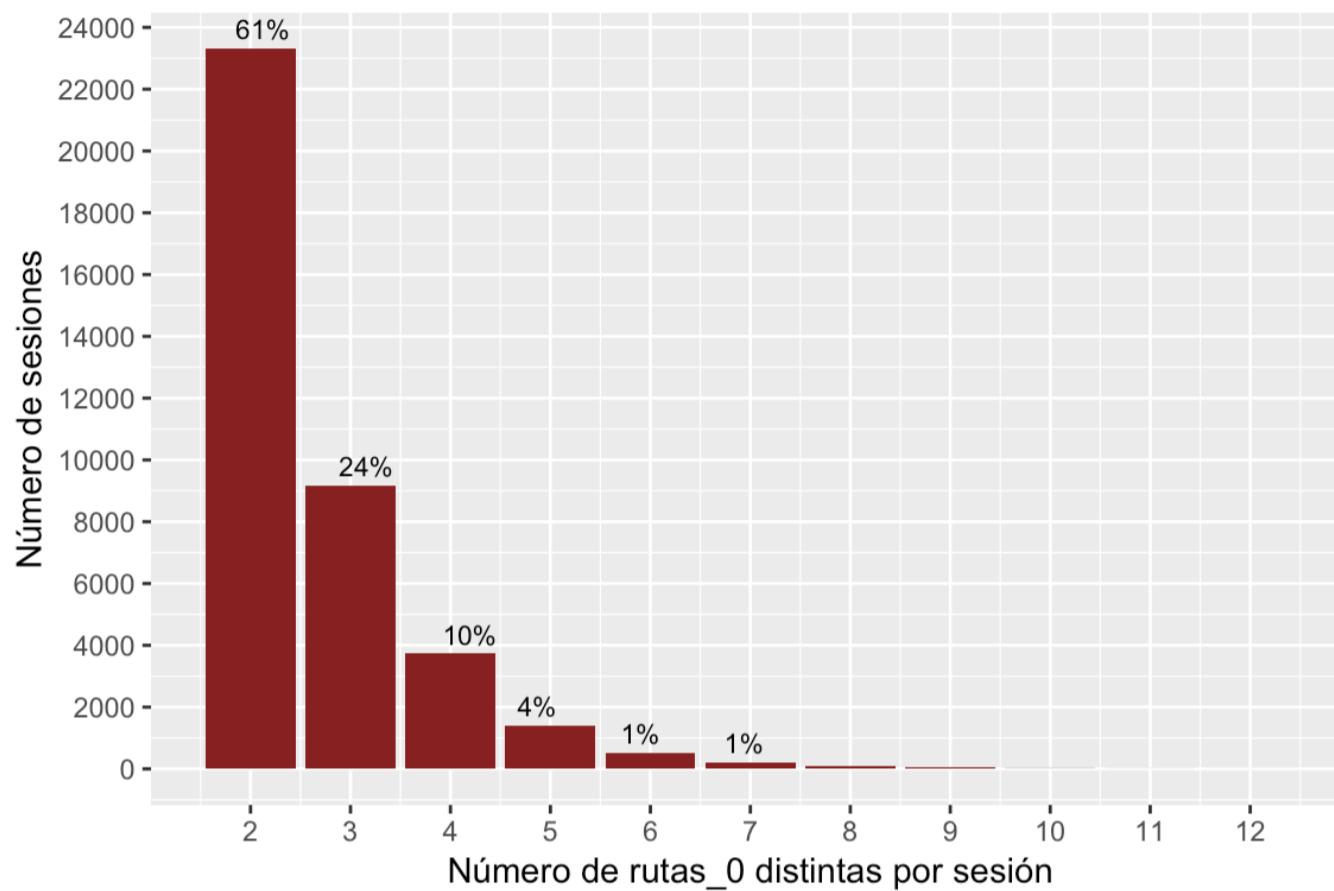
## Rutas de navegación

Nuestro catálogo, como otros muchos, tenía una organización simétrica y jerárquica. Es decir, por un lado, el *real state* de la página es distribuido por igual entre los diferentes menús o enlaces que dan acceso a los productos, con independencia de la importancia relativa de cada uno de ellos para el usuario que navega; por el otro, para acceder a un producto hay que pasar forzosamente por el producto o familia de nivel superior al que pertenece.

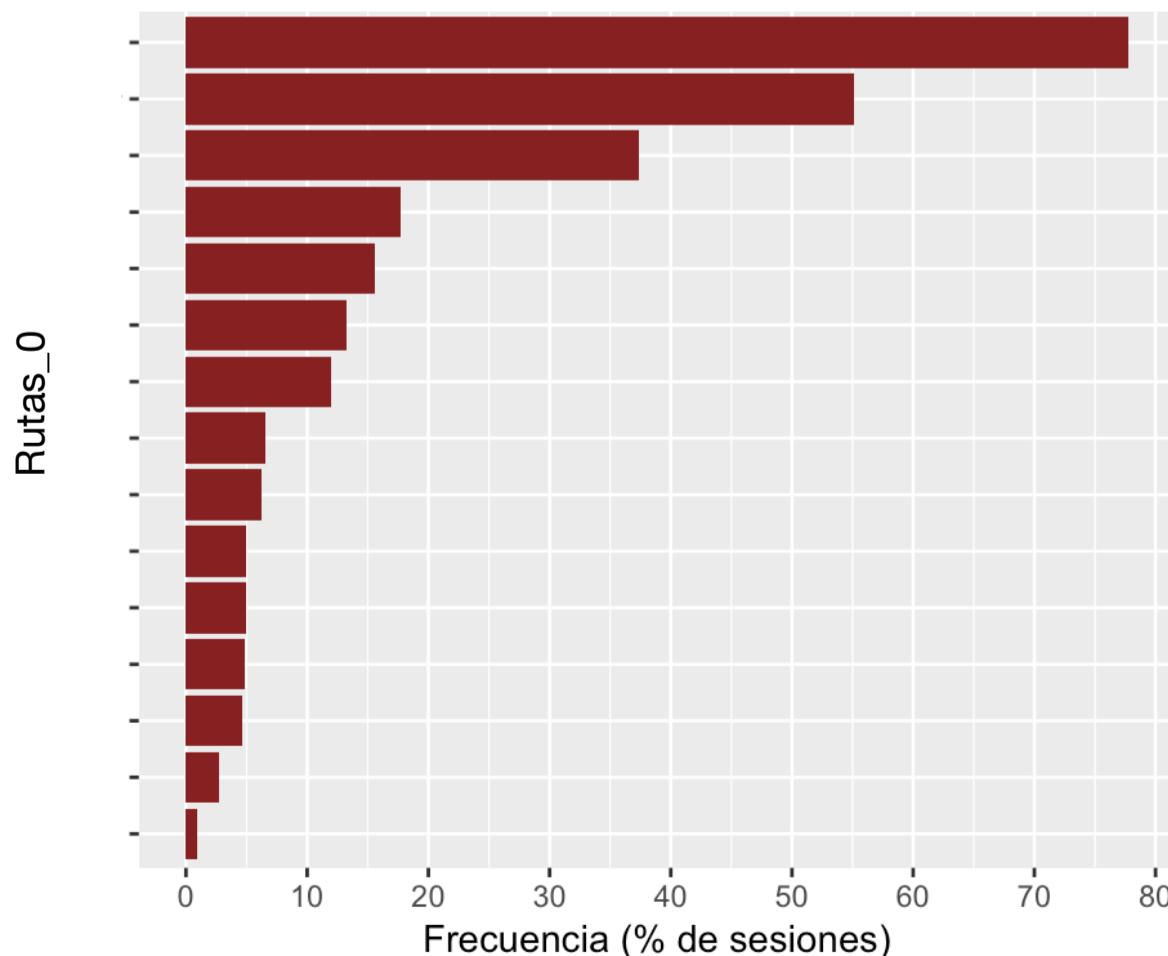
Llamaremos **rutas\_0** a las opciones de menú de primer nivel, aquéllas asociadas a la elección de una familia de productos del catálogo. **rutas\_1** serán las opciones de menú o los enlaces para acceder a los productos de una familia, una vez seleccionada ella.



El usuario, al contrario que el catálogo que consulta, no distribuye simétricamente su tiempo entre los diferentes productos: cuando visita más de una ruta\_0 o familia de productos, en el 61% de los casos sólo visita otra más y en el 85%, no más de otras dos. Y, además, estas pocas rutas\_0 que visita son casi siempre las mismas.

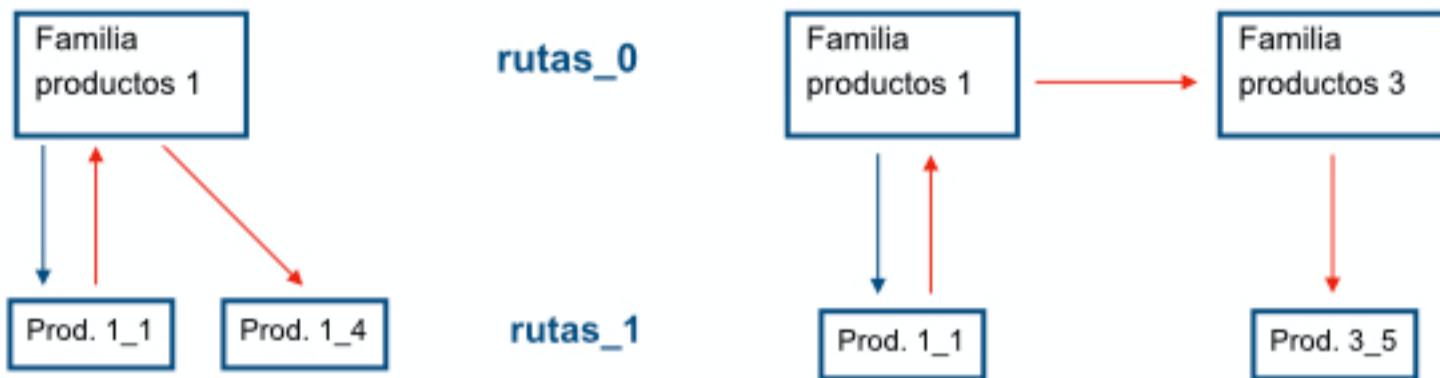


El gráfico siguiente muestra cómo hay dos rutas\_0 que aparecen en el 80% y en el 55% del total de sesiones donde el usuario consulta más de una familia de productos.

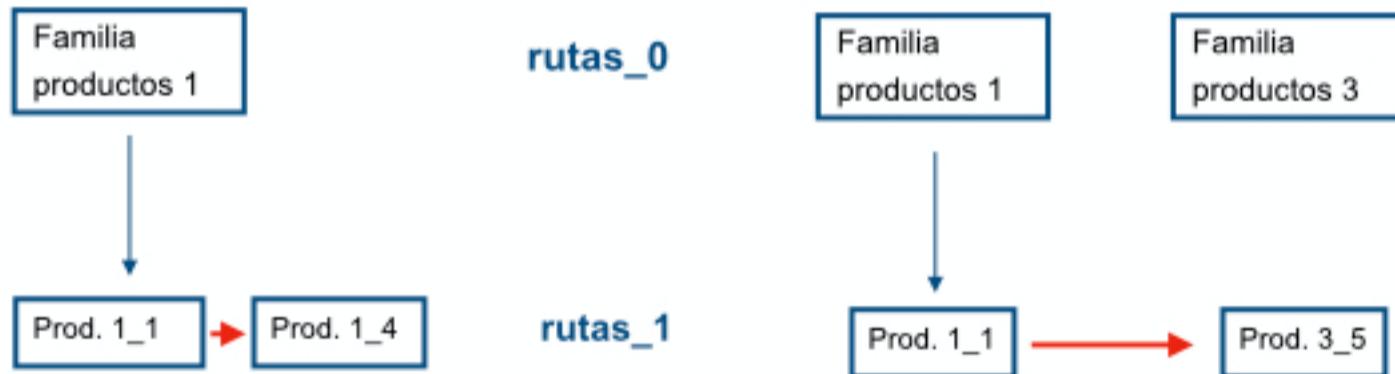


## 'Cestas' de navegación

Además de priorizar en el diseño del catálogo las rutas\_0 más utilizadas, también podríamos mejorar la navegación una vez que el usuario ha entrado en una de ellas. Casos como estos, donde el usuario consulta el producto 1\_4 o el producto 3\_5 tras consultar el producto 1\_1:

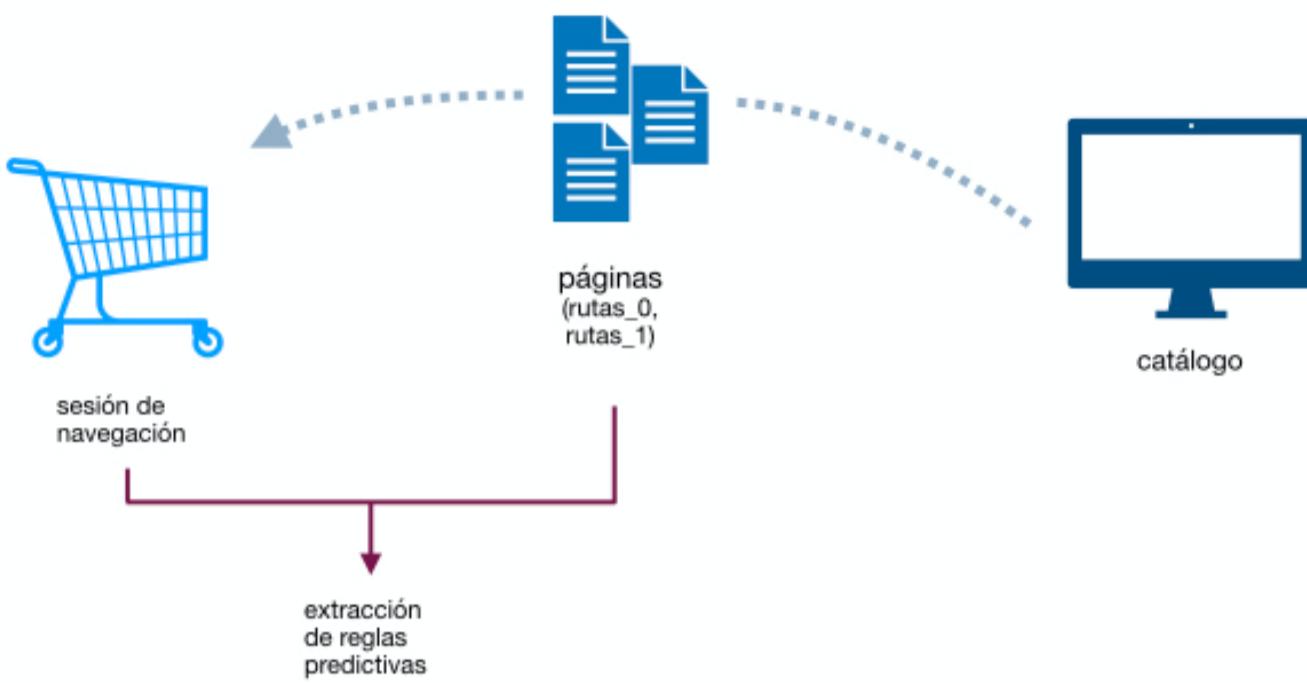


Que podríamos abreviar con una navegación más rápida mediante opciones de navegación cruzada:



Para poder hacerlo, tendríamos que saber qué productos son probables que sean consultados cuando el usuario está en el producto 1\_1, por ejemplo.

Para diseñar opciones de navegación cruzada entre rutas, utilizamos una técnica muy extendida en la venta por Internet: averiguar qué productos *suele* incluir un cliente en una misma cesta de la compra, para hacer venta *cruzada* de ellos al tiempo que va comprando. En nuestro caso, los productos serían las rutas y la cesta de la compra, la sesión de navegación.



Como herramienta estadística para definir nuestras cestas de navegación de rutas del catálogo, utilizamos la librería ARULES implementada en el lenguaje R.

ARULES agrupa las rutas en reglas (cestas) basándose en una medida de la probabilidad (frecuencia relativa) de su aparición conjunta en las sesiones de navegación de los usuarios: cuanto más utilizan los usuarios dos rutas en una misma sesión, más probable es que aparezcan asociadas una regla.

En el análisis consiguiente, obtuvimos *cestas* o reglas de tres tipos<sup>1</sup>: reglas entre familias de productos, reglas entre productos de una misma familia y reglas entre productos de familias diferentes.

---

<sup>1</sup> No detallamos aquí los pormenores técnicos de la obtención de las reglas, ni los criterios que utilizamos para la selección final, basados en conceptos como ‘soporte’ y ‘confianza’.

## Ejemplos de reglas o *cestas* de navegación

Las reglas asociaban una o más rutas, las que figuran a la izquierda del símbolo ‘=>’, con una ruta a su derecha. La asociación se basa en dos medidas de probabilidad<sup>1</sup> de las que, en los cuadros siguientes, sólo presentamos una: la probabilidad de que si el usuario ha consultado las rutas que aparecen a la izquierda consulte, también, la que aparece a la derecha.

Para cada regla, añadimos además el número de sesiones en las que se observó.

Cestas de familias de productos	probabilidad	frec_absoluta
{Fam_3, Fam_5} => {Fam_12}	79,3	16.832

La regla nos dice que en el 79,3% de las sesiones de navegación donde el usuario consulta las familias de productos Fam\_3 y Fam\_5 también consulta Fam\_12. Esta asociación ocurre en 16.382 de las sesiones analizadas en un mes.

Ejemplos similares pueden ilustrar los otros dos tipos de reglas:

Cestas de productos de una misma familia	probabilidad	frec_absoluta
{prod_1_3} => {prod_1_15}	81,5	9.112

Cestas de productos de una misma familia	probabilidad	frec_absoluta
{prod_1_3, prod_2_5} => {prod_1_11}	68,7	6.124

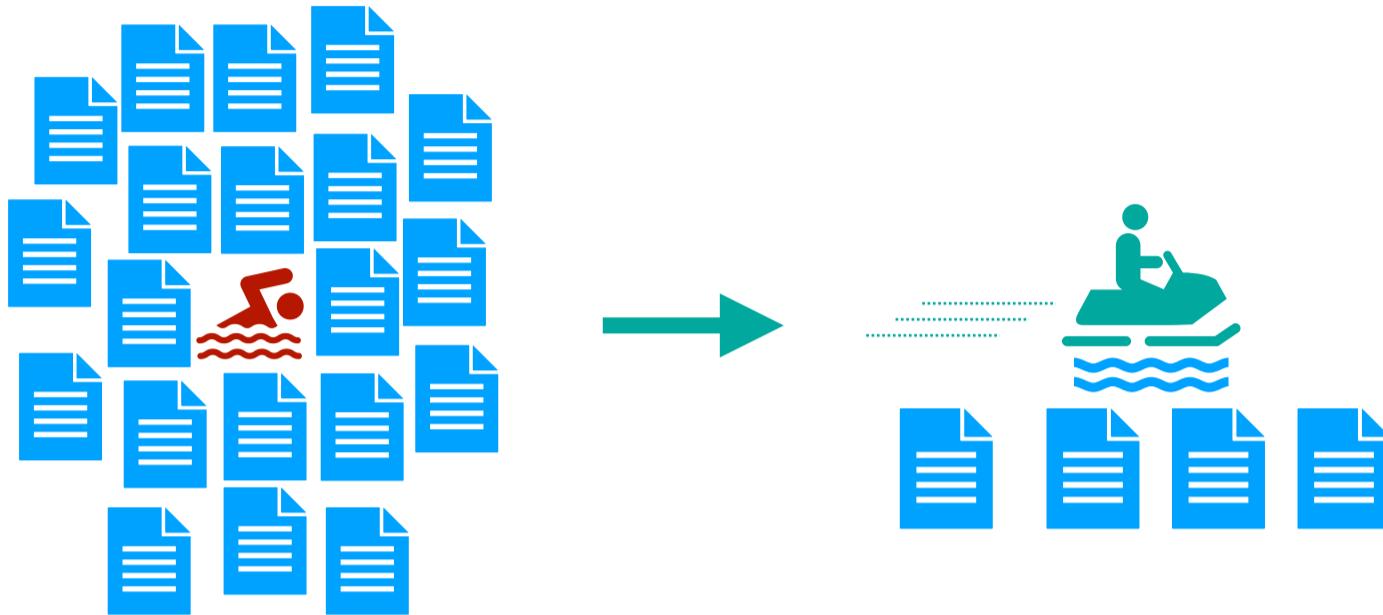
---

<sup>1</sup> La probabilidad, dentro del conjunto de todas las sesiones, de que participen en una misma sesión las rutas incluidas en la regla, simultáneamente.

La probabilidad, restringida al subconjunto de sesiones en las que participan las rutas a la izquierda, de que también participe la ruta a la derecha.

# Conclusiones

Un catálogo con miles de páginas, unos pocos minutos por consulta, sesiones de navegación muy especializadas, con muy pocas páginas por sesión, ...



Nuestra aproximación al problema y nuestra solución:

1. Analizamos patrones de navegación por tipo de canal de venta, perfil funcional del usuario y familias de productos.
2. Definimos rutas de nivel\_0, nivel\_1 y nivel\_2 como unidades de análisis, y estudiamos las consultas entre ellas y dentro de ellas.
3. Establecimos asociaciones entre rutas de diferentes niveles, que nos permitieron establecer atajos de navegación cruzada entre aquellas más probables de aparecer en una misma sesión.

Más allá:

4. Personalizar las reglas/asociaciones entre rutas por tipo de canal y perfil funcional del usuario.

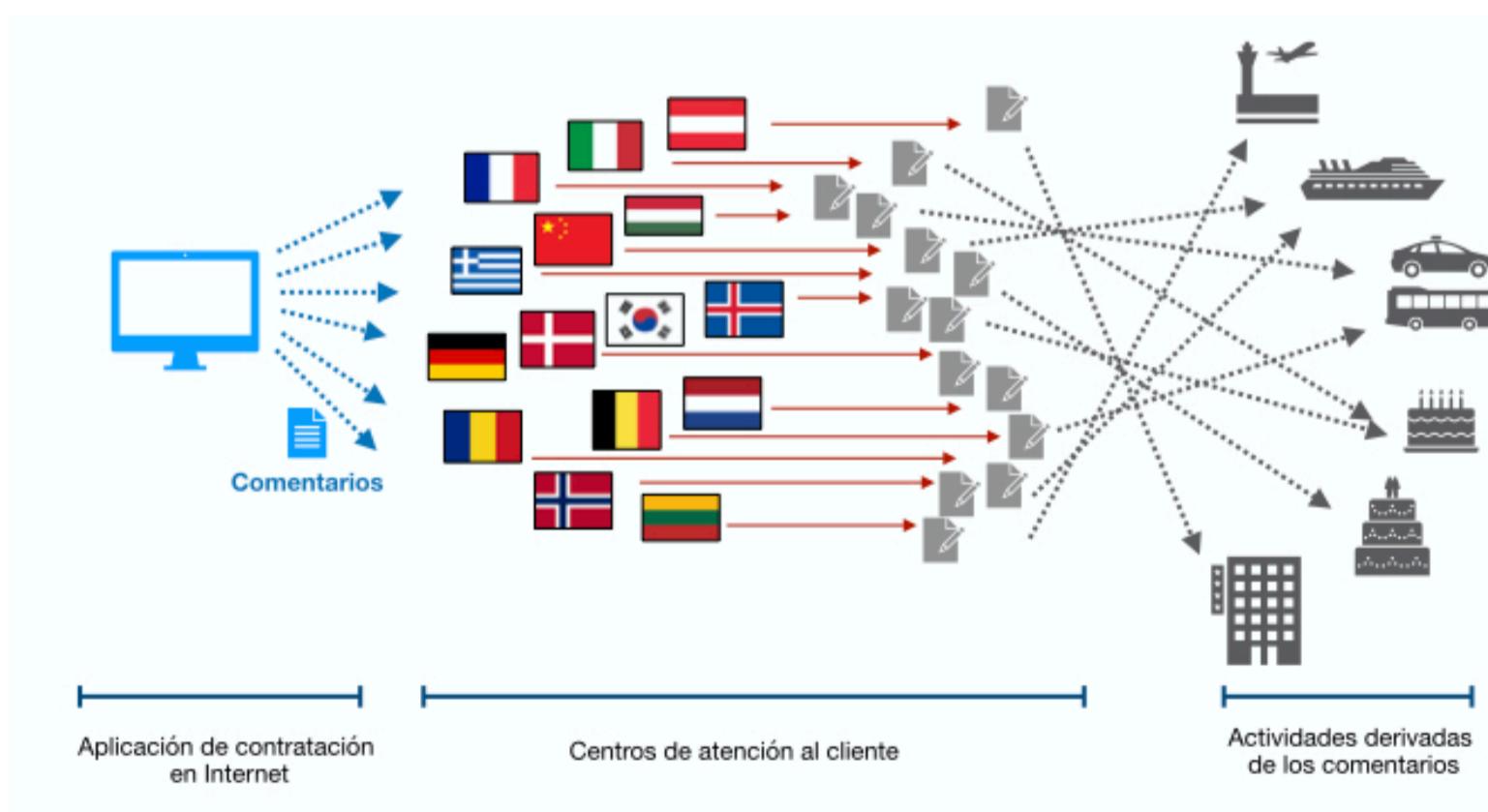
# Comentarios y clientes

## Punto de partida

*¿Cómo procesar la información que contienen los comentarios que dejan los clientes al contratar los servicios de un proveedor?*

Al contratar por Internet los servicios de nuestro cliente, sus clientes tienen la opción de dejar un comentario. En un año, cientos de ellos lo hacen, en casi veinte idiomas diferentes. Algunos detalles relevantes:

- Los comentarios están escritos en un "lenguaje natural": son datos no estructurados.
- Previamente a traducirlos, hay que identificar o confirmar el idioma en que están escritos: no hay una correspondencia unívoca, en todos los casos, entre el idioma del comentario y el idioma del país de procedencia que consigna el usuario en su contrato.
- El tratamiento de los comentarios es descentralizado: lo hace cada uno de los centros donde contrata el cliente. Ellos son responsables de la traducción y de poner en marcha procesos individuales de resolución, cuando así lo requiera el comentario.



# Datos iniciales

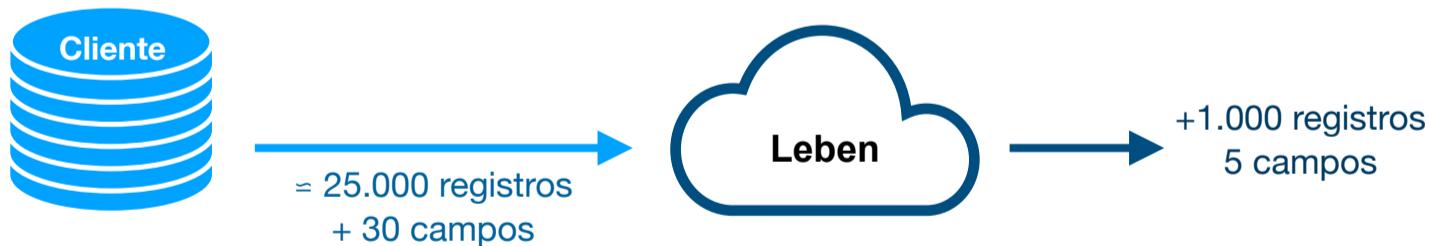
Partimos de un fichero con formato CSV, con cerca de veinticinco mil registros, correspondientes a otros tantos contratos hechos en el período de un año. Cada registro se componía de varias decenas de campos, de los que seleccionamos cinco:

1. Identificador anónimo del cliente (generado por nosotros a partir de direcciones de correo electrónico).
2. El nombre del país de residencia del cliente.
3. El comentario.
4. La fecha de contratación.
5. Identificador anónimo del centro de contratación.

En un segundo ciclo de procesamiento, realizamos las siguientes transformaciones:

- Convertimos los nombres de países en códigos iso6391.
- Filtramos los códigos de países con cuatro o más comentarios y desechamos el resto. Para nuestra experiencia piloto, consideramos este filtro como un umbral mínimo de presencia de un posible idioma.

El número final de comentarios incluidos en el proyecto fue algo superior a los mil.



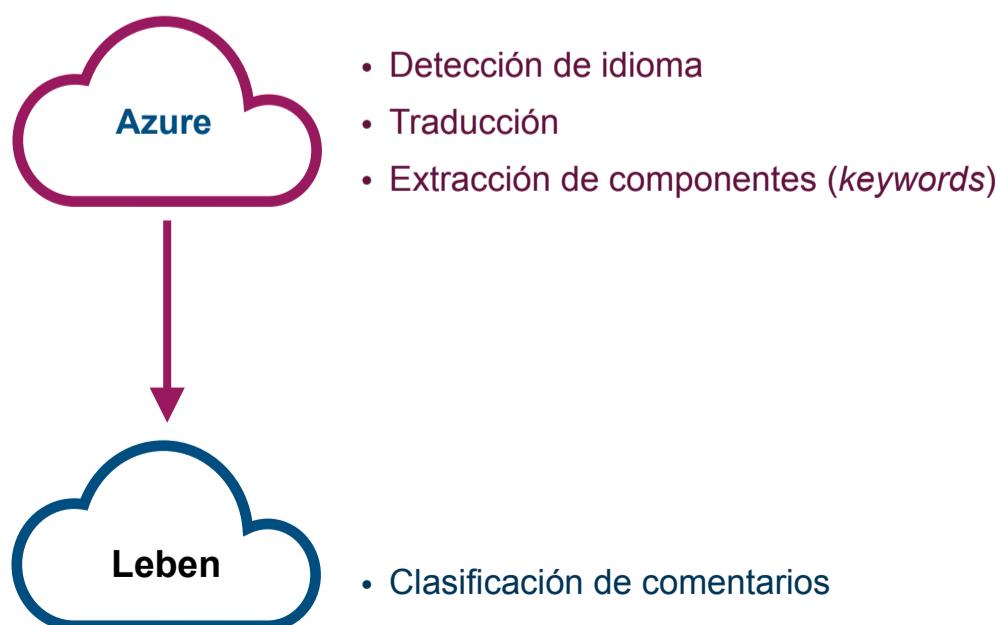
# Análisis

Nuestro objetivo era desarrollar un proceso de conducción semi-autónoma del tratamiento de los comentarios. En el proyecto piloto, las tareas a incluir en el proceso eran:

1. La detección del idioma del comentario.
2. La traducción desde el idioma detectado al español.
3. La extracción, una vez traducido, de los principales componentes semánticos del comentario.
4. La clasificación del comentario, basada en sus principales componentes semánticos, según el tipo de resolución que pudiera requerir.
5. La distribución de los comentarios clasificados, según tipo de resolución, a los centros locales y departamentos involucrados.

En cuanto a las herramientas técnicas a utilizar, optamos por las siguientes:

- Para la detección del idioma, la traducción y la extracción de componentes queríamos contar con herramientas de inteligencia artificial de proveedores externos. Analizamos los servicios de Amazon Web Services y Microsoft-Azure. Seleccionamos finalmente los servicios Translator y Text Analytics de Azure por su facilidad de integración con nuestra aplicación interna.
- Para la clasificación de los comentarios, una vez extraídos sus componentes principales, hicimos un desarrollo propio basándonos en la librería de código abierto TextAnalysis.jl.
- Para la integración de los diferentes procesamientos, tanto internos como con proveedores externos, desarrollamos una aplicación con el lenguaje de programación JULIA, tras barajar la opción de hacerlo con el lenguaje JAVA.



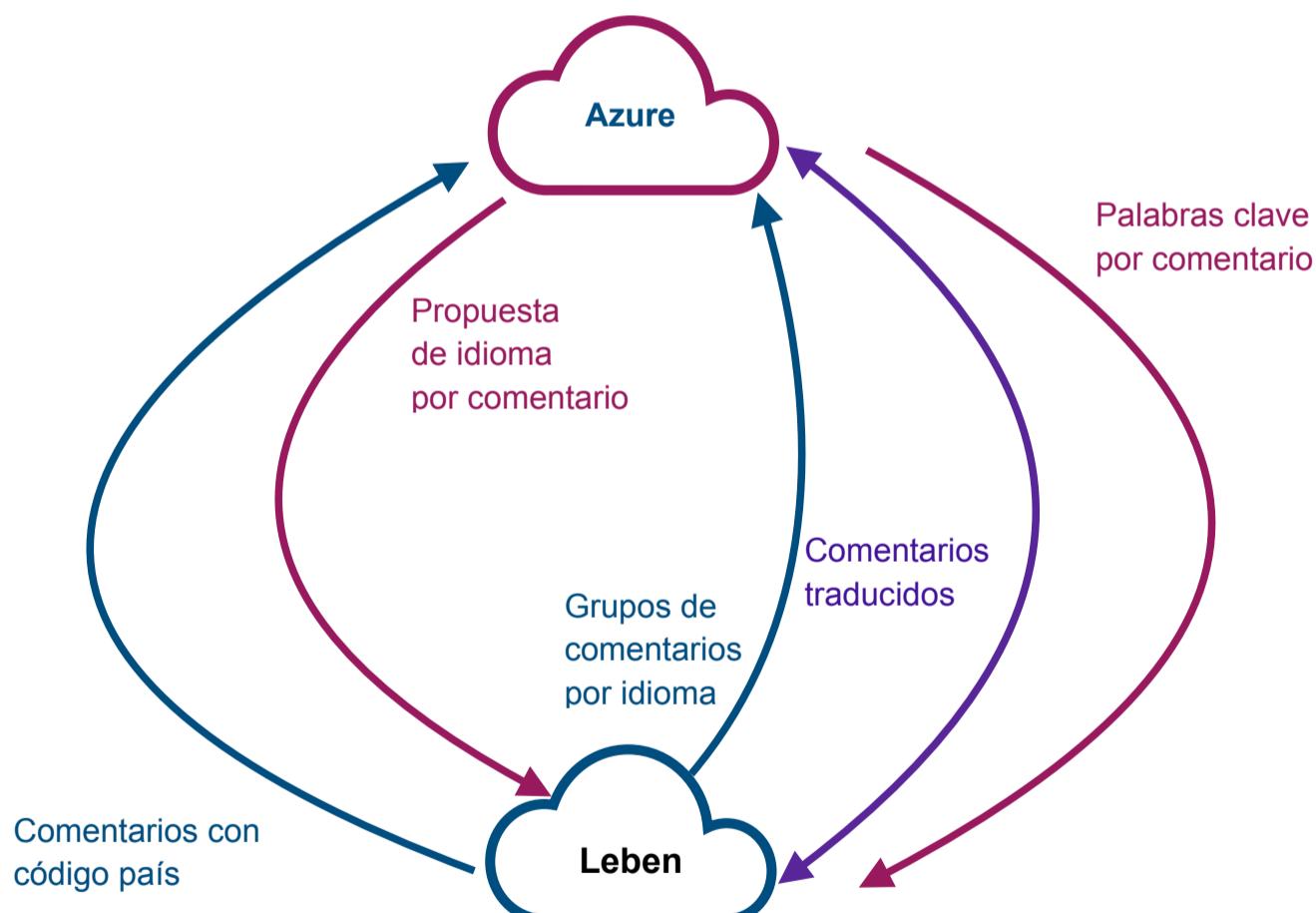
## Detección de idioma, traducción y extracción de *keywords*

Describimos brevemente las operaciones que montamos entre nuestra aplicación en JULIA y los servicios web de Azure.

1. Para la identificación del idioma, enviamos a Azure el conjunto de comentarios etiquetados cada uno de ellos con una *pista*: el código iso6391 del país de procedencia del cliente.

Azure respondía con una *propuesta* de idioma y una estimación de la probabilidad de que el idioma propuesto fuera correcto. Establecimos un umbral mínimo de probabilidad en 0.60: comentarios con un probabilidad inferior quedaban pospuestos para su procesamiento manual (no hubo ninguno).

2. Una vez identificado el idioma del comentario, los agrupamos por idioma y enviamos los diferentes grupos a Azure, que devolvía los textos traducidos al español.
3. El paso final de esta parte del proceso era enviar a Azure, identificando el idioma, los textos traducidos de los comentarios. Azure devolvía una lista de palabras clave por comentario.



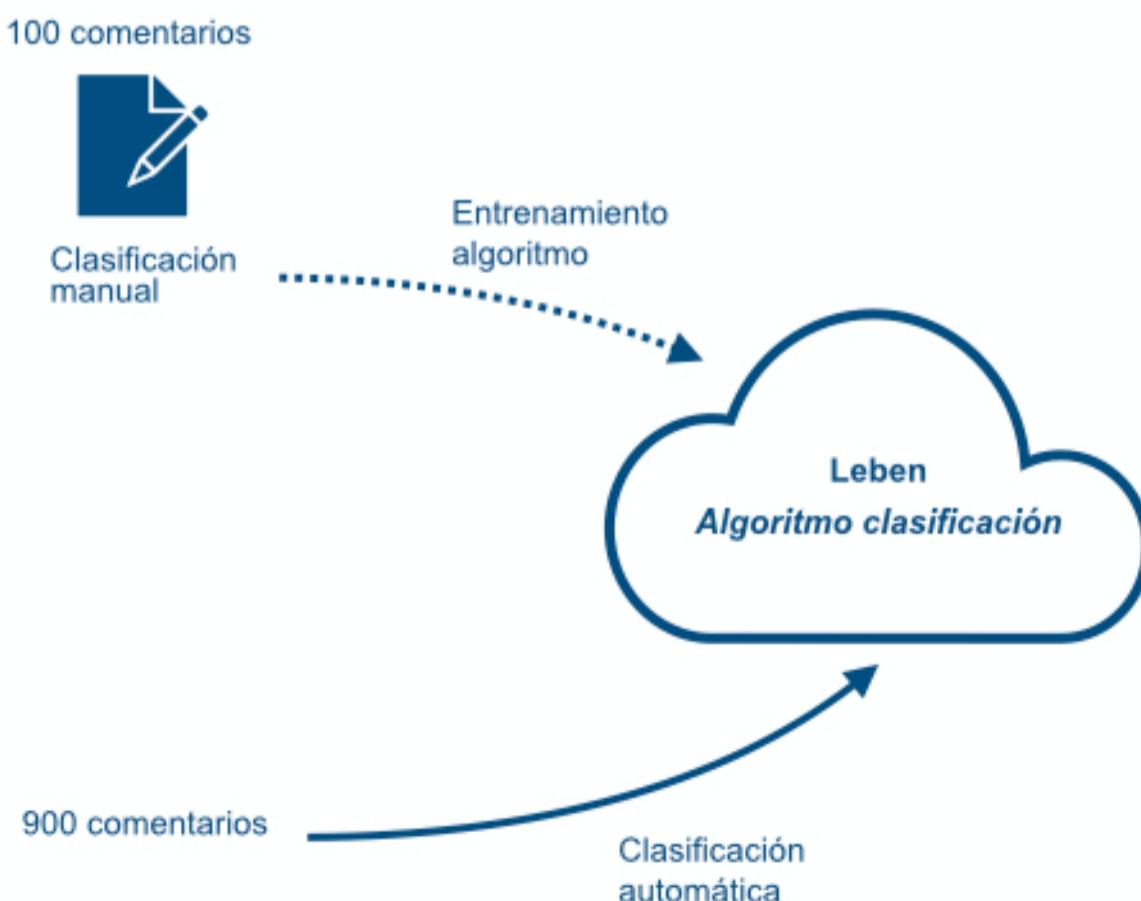
## Clasificación de comentarios

En otro tipo de proyecto, podríamos habernos interesado en un análisis ‘de sentimientos’ de los comentarios: intentar inducir algún tipo de valoración por parte del cliente. En nuestro caso, lo que buscábamos era la automatización del proceso de acciones requeridas por los comentarios, y optamos por una clasificación por tipo de resolución o de acción requerida por cada uno de ellos.

El programa de clasificación lo hicimos con JULIA utilizando las funciones de la librería TextAnalysis.jl para el método *Naive Bayes Classifier*<sup>1</sup>.

El proceso de diseño y desarrollo del programa fue el siguiente:

1. Definimos 8 clases de comentarios, cada clase correspondiente a un tipo de resolución.
2. Seleccionamos 100 comentarios y los clasificamos manualmente en una de las 8 categorías: así *entrenamos* el algoritmo de clasificación. Cada comentario se clasificó atendiendo exclusivamente a la lista de palabras clave asociada a él.
3. A continuación, utilizamos las listas de palabras clave del resto de comentarios para clasificarlos con el algoritmo entrenado en la fase anterior.
4. Una vez clasificado cada comentario, los agrupamos por clase y centro de contratación del cliente, para simplificar el proceso de comunicación final, y control posterior, a la instancia última responsable de la resolución.



<sup>1</sup> Una traducción del nombre podría ser Clasificación Bayesiana Simple. Las características del método quedarán reflejadas en la descripción del proceso.

# Conclusiones

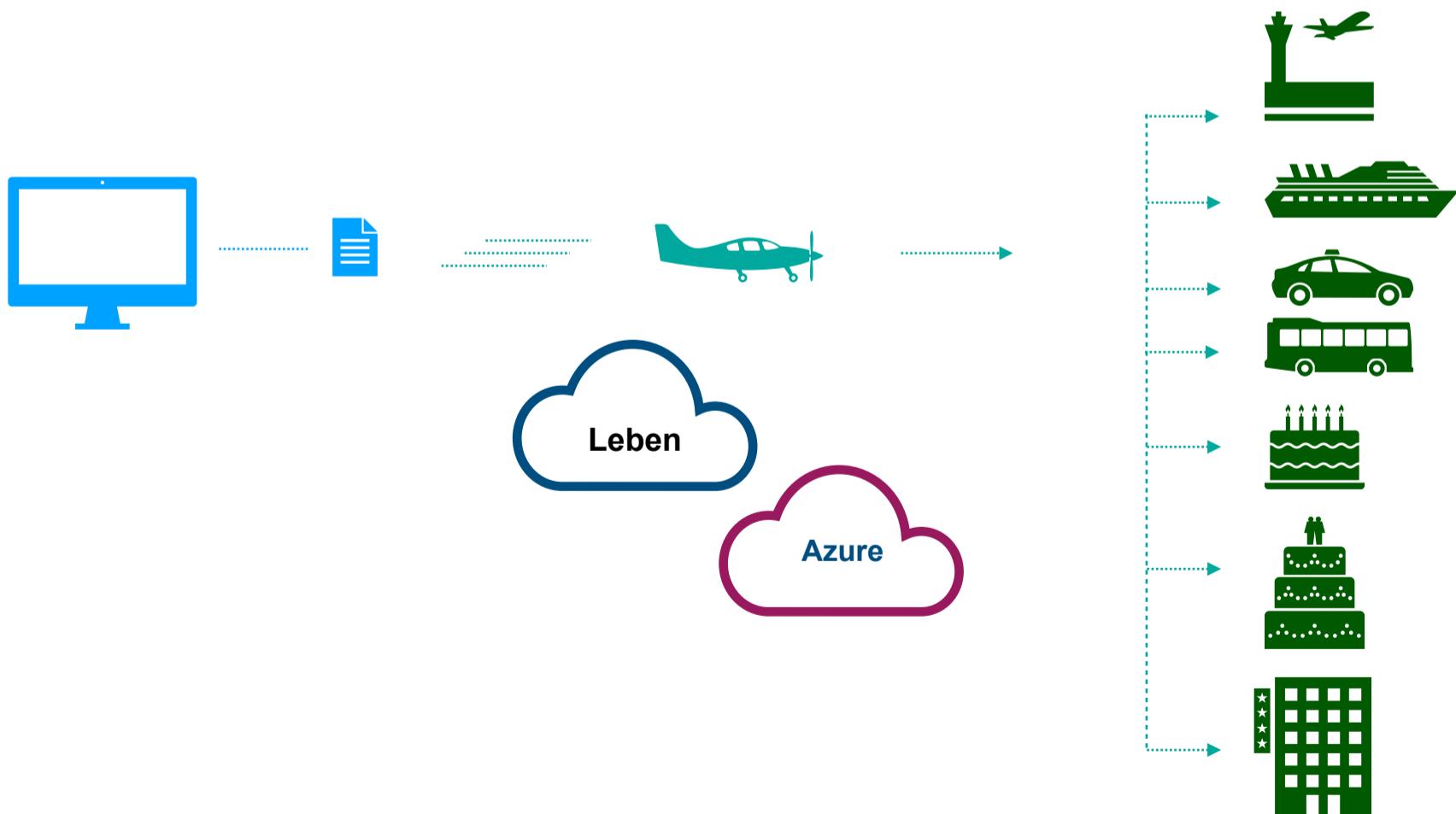
Cientos de comentarios hechos por los clientes en el momento de la contratación por Internet, en idiomas diferentes, con requerimientos de acciones ulteriores precisas, tratados de forma manual y descentralizada en diferentes centros de atención, ...

Nuestra aproximación al problema y nuestra solución:

1. Creamos un proceso de conducción autónoma con identificación de idioma, traducción y extracción de palabras clave.
2. Definimos clases de comentarios por tipo de resolución requerida.
3. Entrenamos y desarrollamos un algoritmo de clasificación de los comentarios.
4. Agrupamos los comentarios por centros y clases para simplificar la comunicación de las acciones de resolución.

Más allá:

5. Integrar dentro de la propia aplicación el control de las acciones requeridas por los comentarios.



# Promociones y prescriptores

## Punto de partida

¿Cómo influye el punto de venta sobre el éxito de mis promociones para clientes finales? ¿Qué otros factores influyen? ¿Cómo cuantificarlos?

Nuestro cliente era un proveedor de componentes para automóviles que se venden casi exclusivamente en talleres mecánicos. Su mercado se compone de dos grandes segmentos:

- B2B: flotas, tanto de camiones o camionetas, como de turismos.
- B2C: propietarios particulares.

El proveedor lanzaba periódicamente promociones (descuentos y reembolsos) para incentivar la compra de sus productos. Las promociones se comunicaban a los clientes a través de dos tipos de canales:

1. Un canal directo: comunicaciones por correo electrónico y un portal en Internet, ambos para usuarios registrados.
2. Un canal indirecto, a través de publicidad disponible en el punto de venta.

El objetivo de nuestro proyecto era identificar y cuantificar, en la medida de lo posible, qué factores intervenían en el éxito de las promociones dirigidas al segmento B2B. Entre ellos, muy especialmente, el papel que desempeñaba el taller como comunicador de la promoción e inductor de la compra.



# Datos iniciales

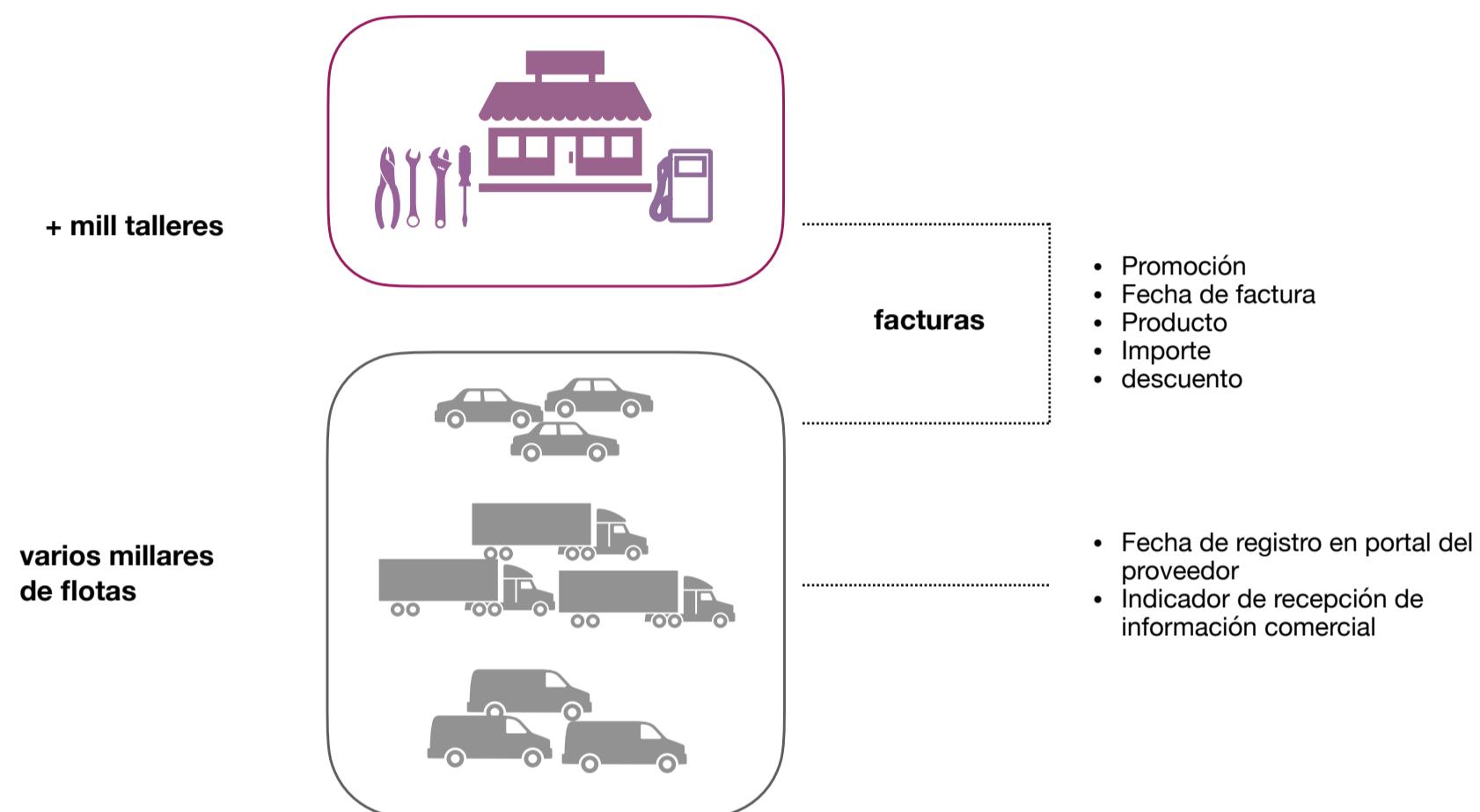
Recibimos del cliente un único fichero, con formato CSV:

- Datos de las facturas emitidas por los talleres vendedores de los productos de nuestro cliente. Estas facturas correspondían, todas ellas, a ventas acreedoras del incentivo promocional. Las características más relevantes del fichero eran:
  - Abarcaba un período de 30 meses.
  - Contenía unos 40.000 registros:
    - ▶ Identificador anónimo único del taller vendedor.
    - ▶ Identificador anónimo de la flota compradora.
    - ▶ Identificador anónimo de usuario asociado a la flota.
    - ▶ Identificador de la promoción.
    - ▶ Fecha de factura.
    - ▶ Fecha de registro en el portal del proveedor.
    - ▶ Indicador de si el usuario deseaba recibir información comercial.
    - ▶ Producto.
    - ▶ Importe y descuento.

## Actores

El primer requerimiento, para construir un modelo de las respuestas de talleres y flotas a las promociones, era extraer unos y otras de los datos de facturación:

- Poco más de mil talleres <sup>1</sup>.
- Varios millares de flotas.



A partir de aquí, a falta de tener otra información que las facturas, decidimos construir un modelo de predicción de la conducta en una campaña ‘actual’ en función de cuál había sido en las campañas anteriores. Nos pareció claro, también, que los modelos habrían de ser diferentes para talleres y flotas.

<sup>1</sup> Por razones de confidencialidad, el lector disculpará nuestra imprecisión en las cifras, meramente aproximadas.

# Análisis de talleres

## Caracterización

La conducta de un taller en una promoción, en cuanto a las ventas que podría hacer en ella, podría caracterizarse por tres aspectos:

1. La participación, como aspecto determinante del resto.
2. La captación de nuevos clientes gracias a la promoción.
3. Las ventas hechas vinculadas al incentivo promocional.

A estos añadimos otro más que nos permitiera estimar el grado de vinculación entre el proveedor y el taller:

4. La antigüedad de la relación entre ambos.

El siguiente paso era definir variables que cuantificaran cada uno de esos cuatro aspectos, utilizando lo que teníamos: facturas. Esas variables definirían la situación previa al arranque de una nueva promoción y las utilizaríamos como predictores de lo que podría ocurrir, a futuro, en ella.

Definimos las siguientes variables:

1. Historial de participación
  - Ratio del número de campañas en las que el taller ha facturado sobre el número de campañas habidas hasta una fecha. Las campañas incluidas en esta ratio van variando en función de la fecha de la campaña ‘actual’, aquella cuya conducta se intenta predecir con el modelo.
2. Historial de captación de nuevos clientes
  - Ratio del número de campañas en las que un taller ha hecho nuevos clientes sobre el número total de campañas anteriores a la ‘actual’.
3. Media (histórica) de nuevos clientes
  - Número medio de nuevos clientes captados por un taller, en las campañas anteriores a la ‘actual’. Un cliente se considera como nuevo en aquella campaña en la que factura por primera vez; obviamente, dentro del intervalo temporal que cubre el fichero de datos.
4. Venta media (histórica)
  - Facturación media hecha por el taller en las campañas anteriores a la ‘actual’.

## 5. Antigüedad (vinculación con el proveedor)

- Número de días transcurridos entre la fecha de la primera factura del taller, en cualquier campaña, y la fecha final de la campaña 'actual'.



- Participación
- Captación
- Ventas



### Variables

- Historial de participación
- Historial de captación de nuevos clientes
- Media de nuevos clientes captados por campaña
- Venta media en campaña
- Antigüedad relación con proveedor

## Modelo predictivo

Llegados a este punto, nuestro objetivo era construir un modelo predictivo de cuál sería la respuesta de un taller a una nueva campaña, sabiendo cuáles eran los valores de las cinco variables anteriores en el momento de iniciarse aquella.

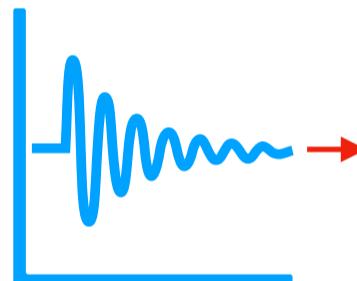
Para hacerlo, teníamos que definir una variable que cuantificara la respuesta del taller. Optamos por la variable dicotómica *alta* o *baja* venta:

- ▶ Un taller tendría una respuesta *alta* cuando el taller terminara la promoción dentro de la mitad de talleres con mayor facturación en la misma; mitad a la que llamamos *superior*<sup>1</sup>.
- ▶ Un taller tendría respuesta *baja* cuando terminara en la *mitad inferior*.

Nuestro modelo clasificaría cada taller en una de las dos categorías de respuesta, antes de iniciarse la campaña y antes de comunicarla. Esta clasificación permitiría, por ejemplo, diferenciar la comunicación de la campaña a los talleres de uno u otro grupo.

### Variables

- Historial de participación
- Historial de captación de nuevos clientes
- Media de nuevos clientes captados por campaña
- Venta media en campaña
- Antigüedad relación con proveedor



Alta venta

Baja venta



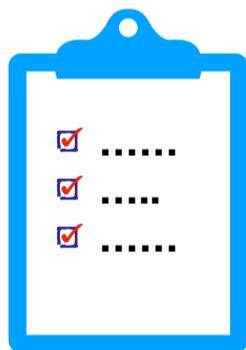
Modelo predictivo

<sup>1</sup> Dicho de otro modo: respuesta *alta* la tendrían los talleres con una facturación igual o superior a la mediana de las facturaciones de todos los talleres; respuesta *baja*, cuando terminaran con una facturación inferior a la mediana.

## Resultados del modelo

Una vez definidas las variables anteriores (predictores y respuesta), estimamos los parámetros del modelo <sup>1</sup>. Estos parámetros nos sirvieron para explicar la conducta promocional de los talleres:

1. Existía una alta correlación <sup>2</sup> entre el historial de participación en campañas de un taller y su historial de captación de clientes: tendían a participar más aquellos talleres que captaban más clientes nuevos en la campaña.
2. Las ventas en campañas anteriores no eran un buen predictor de las ventas en una campaña futura.
3. Las dos variables que más influían en que un taller tuviera una respuesta *alta* eran la media histórica de nuevos clientes captados en una campaña y el historial previo de participación del taller: vendían más aquellos talleres que habían participado más y habían captado más clientes nuevos en campañas anteriores.
4. Los resultados indicaban que los talleres que más vendían eran aquellos que utilizaban las campañas para captar nuevos clientes, de una forma activa o pasiva. Parecía haber poca acción comercial interna dirigida hacia su propia “base de clientes”, si la hubieran tenido (no lo podíamos saber con los datos del proyecto).



- + Quienes más habían participado antes.
- + Quienes más clientes nuevos captaban.

<sup>1</sup> Estimamos un modelo lineal mediante regresión logística. Controlamos la calidad de los resultados dividiendo los datos originales en subconjuntos *training/test* y estimando el modelo repetidas veces, para diferentes subconjuntos. El porcentaje medio de clasificación correcta de los subconjuntos de test estuvo en torno al 82%.

<sup>2</sup> Los valores de las dos variables crecen y decrecen con una pauta similar.

# Análisis de flotas

## Caracterización

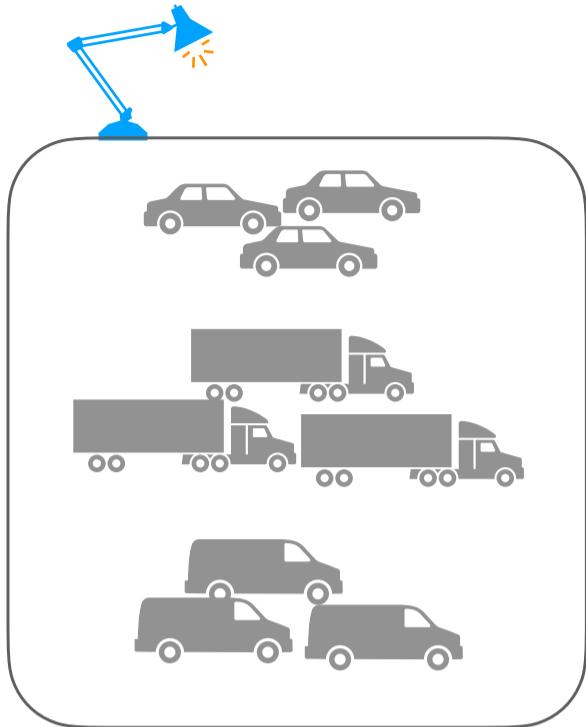
La caracterización de la conducta de las flotas era algo más complicada que la de los talleres. En una flota, además de la relación con el proveedor, es importante caracterizar la relación con el taller: en qué tipo de taller compraban más los clientes finales.

Los aspectos de la conducta que nos interesaba cuantificar eran los siguientes:

1. La participación.
2. El Interés por el incentivo promocional.
3. La vinculación con el proveedor.
4. La relación con el taller.

Como hicimos con los talleres, definimos variables que cuantificaran las conductas:

1. Historial de participación
  - Ratio de participación de la flota en campañas anteriores a la actual.
2. Descuento promocional
  - Descuento obtenido u ‘obtenible’ en una promoción. Hay que interpretar este predictor como el coste de oportunidad de la no participación.
3. Antigüedad (respecto al proveedor)
  - Número de días transcurridos entre la fecha de registro más antigua de un usuario de la flota y la fecha de cierre de la campaña actual.
4. Comunicación proveedor
  - Existe, o no, un usuario registrado asociado a la flota que acepta recibir mensajes publicitarios del proveedor.
5. Historial de captación del taller
  - Número medio de nuevos clientes, en las campañas anteriores a la actual, hechos por el taller donde la flota realiza la mayor facturación en la campaña actual.
6. Historial de participación del taller
  - Historial de participación del taller donde la flota realiza la mayor facturación en la campaña actual.



- **Participación**
- **Incentivo**
- **Vinculación**
- **Relación con taller**



#### Variables

- Historial de participación
- Descuento promocional
- Antigüedad
- Comunicación con proveedor
- Historial de captación del taller
- Historial de participación del taller

## Modelo predictivo

Como respuesta a predecir por nuestras variables, elegimos la participación en la promoción:

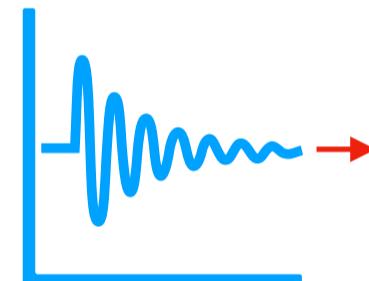
- Una flota participaría en una promoción cuando hubiera, al menos, una factura asociada a ambas.

La respuesta era, pues, una variable dicotómica con los valores *sí* o *no*. Podríamos haber utilizado, como en los talleres, una variable relacionada con la facturación. Pensamos, sin embargo, que para el proveedor la decisión clave, por parte de la flota, era cruzar el umbral de la participación en una campaña.

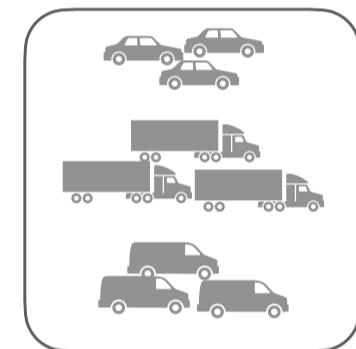
Nuestro modelo clasificaría cada flota asociada a un usuario registrado en el portal como susceptible o no de participar en una promoción futura. Esta clasificación podría utilizarse, por ejemplo, en la estrategia de comunicación de la misma o en el diseño de incentivos.

### Variables

- Historial de participación
- Descuento promocional
- Antigüedad
- Comunicación con proveedor
- Historial de captación del taller
- Historial de participación del taller



Modelo predictivo

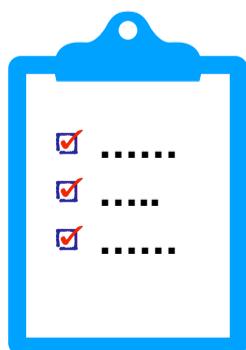
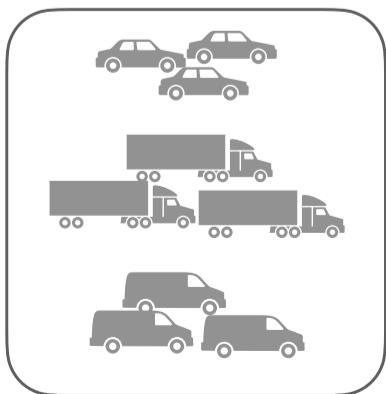


## Resultados del modelo

Con un procedimiento similar al que empleamos con los talleres, estimamos los parámetros del modelo para flotas. Con estos parámetros podíamos asignar, a cada flota, una probabilidad de participación en una próxima campaña.

También nos permitió explicar algunos aspectos de la relación compleja entre campaña, taller y flota:

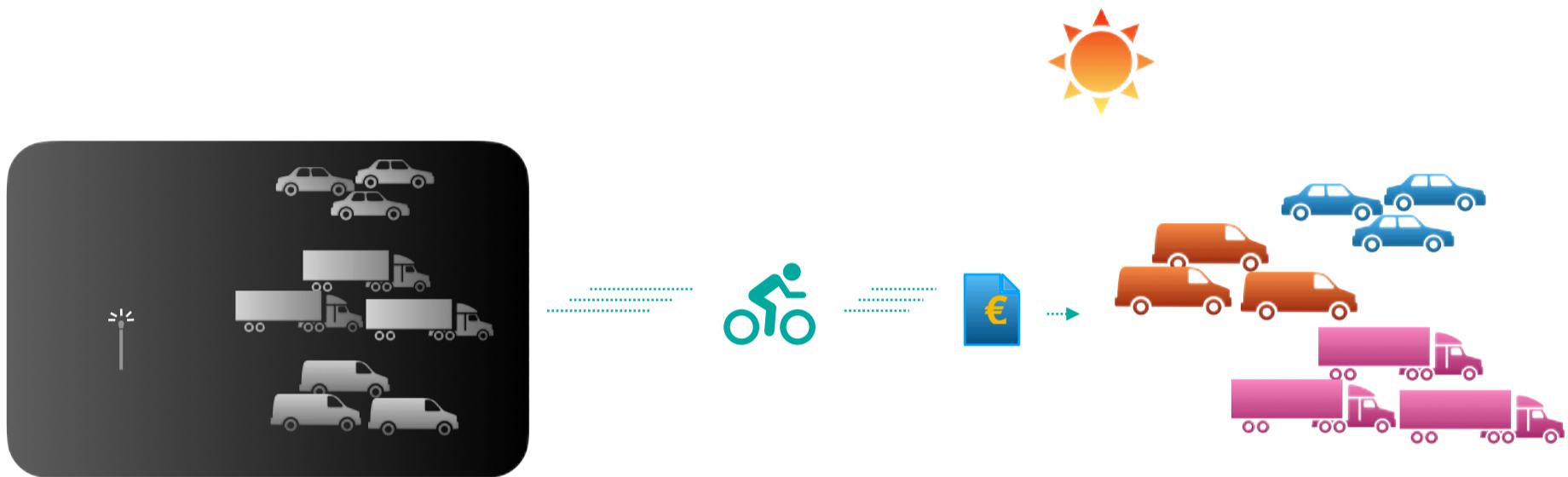
1. Respondiendo la pregunta que nos hacía al inicio del proyecto nuestro cliente, podíamos responder que sí: efectivamente, la variable con mayor poder predictivo en cuanto a si una flota participaría en una campaña, era la inducción prescriptiva del taller.
  - El parámetro del modelo con mayor poder explicativo correspondía a la variable ‘Historial de participación del taller’; es decir, las flotas que más participaban en una campaña eran aquellas que lo hacían en talleres muy activos en campañas anteriores.
2. Es menos probable que compre en una promoción una flota que ha comprado mucho en promociones anteriores. Como razón aventuramos la siguiente: dado que las campañas que estudiamos cubrían un período de dos años y medio, y que los componentes tienen una vida media útil algo inferior a ese plazo, es improbable que una flota repita la compra de muchos de los componentes que compró una primera vez dentro de él.
3. Es más probable que compren flotas con usuarios que desean recibir información comercial del proveedor. Este punto avala la importancia de mantener una comunicación directa con los clientes.
4. La antigüedad como usuario registrado tiene poco poder predictivo. Lo importante no es el registro, que es obligatorio para poder disfrutar de los incentivos de una promoción, sino ese interés suplementario que implica el querer recibir, además, comunicaciones comerciales.
5. Tampoco era relevante el incentivo promocional para explicar la participación. Quizás porque las diferencias entre los descuentos de diferentes promociones eran muy bajas a lo largo del período que estudiamos.



- + Quienes compran en talleres históricamente más activos.
- + Quienes reciben información comercial.
- Quienes han comprado en promociones anteriores.

# Conclusiones

Un doble canal de comunicación con mis clientes (directo, indirecto), la opacidad de los mensajes que circulan por uno de ellos, la dificultad consiguiente para estimar la eficacia por separado de cada uno de los canales en la comunicación de promociones, .....



Nuestra aproximación al problema y nuestra solución:

1. Estimamos un modelo predictivo de la conducta de los talleres en una campaña, basado en la probabilidad de obtener un nivel alto/bajo de ventas.
2. Estimamos un modelo predictivo de la conducta de las flotas en una campaña, basado en la probabilidad de participación.
3. Cuantificamos la influencia del poder prescriptor de la comunicación del taller a la flota, y concluimos que esta era la variable que mejor explicaba la participación de una flota en una campaña.
4. Cuantificamos, por último, la importancia de los canales directos de comunicación entre proveedor y cliente.

Más allá:

5. Modular la política de incentivos y su comunicación según las estimaciones individuales de participación y ventas, de talleres y flotas, como un proceso de marketing de conducción autónoma.

# Incentivos y fidelización

## Punto de partida

¿Cómo son los clientes que repiten, los que vienen una temporada y la siguiente? ¿Cómo establecer comunicación directa con ellos? ¿Cómo incentivarlos a repetir?

Nuestro cliente era una cadena hotelera con unos pocos establecimientos <sup>1</sup>, ubicados casi únicamente en puntos geográficos de sol y playa. Disponía de sus propios canales de venta (centro de atención telefónica, portal en internet y los propios hoteles), pero cerca del 80% de las reservas se hacían a través de canales indirectos: operadores, agencias y portales en Internet de agregadores y buscadores.

Simplificando la argumentación y concentrándonos sólo en los aspectos que nos interesan en este capítulo, diremos que la cadena perseguía aumentar las ventas directas por tres razones:

- Evitar el coste de los canales indirectos, percibido como muy alto.
- Ampliar la facturación por servicios extraordinarios, y superar las limitaciones que imponía un modelo de negocio basado en la venta con descuento de grandes volúmenes de estancias a precio cerrado.
- Controlar la comunicación con el cliente, hasta ahora más vinculado con quien le vendía la reserva que con el hotel donde la disfrutaba.

El medio considerado para primar la venta directa y mejorar la comunicación con los clientes finales era un programa de fidelización que incentivara la repetición y la facturación de servicios extras.



<sup>1</sup> Nuevamente, justificamos nuestro oscurantismo por la necesidad de mantener la confidencialidad de nuestro cliente.

# Datos iniciales

Recibimos del cliente un fichero con formato CSV conteniendo unos 180.000 registros, correspondientes a las facturadas presentadas al cobro, a clientes finales, en un período de unos tres años y medio. No había una relación de uno a uno entre facturas y registro: dos o tres registros, representando diferentes conceptos de facturación, podían estar asociados a la misma factura.

Cada registro tenía 40 campos, de los que extraímos 15 que agrupamos en 3 grupos:

1. De mera identificación, sin valor informativo alguno:
  - Identificador anónimo del titular de la factura
  - Identificador de reserva
2. De caracterización del titular de la factura:
  - Fecha de nacimiento
  - País de nacimiento
  - Sexo
  - Indicador V/F de suscripción a información comercial
3. De caracterización de la reserva y de la factura asociada:
  - Fecha de entrada
  - Fecha de salida
  - Número de noches
  - Número de adultos
  - Número de niños
  - Canal de ventas
  - Régimen de alojamiento
  - Concepto de facturación
  - Importe facturado

## Conceptos de facturación

Dado el papel central que iba a desempeñar el campo del concepto de facturación, realizamos con él las siguientes transformaciones: agrupamos los 54 conceptos de facturación en 14, de los que finalmente incluimos 9 en nuestro análisis:

- ▶ Alojamiento y régimen (desayuno, media pensión, pensión completa)
- ▶ Servicios extras de alquileres
- ▶ Servicios extras de actividades
- ▶ Servicios extras de restauración

- ▶ Servicios extras en tiendas
- ▶ Servicios extras de telecomunicaciones
- ▶ Facturación por paquetes ‘todo incluido’ (alojamiento + régimen + servicios extras)
- ▶ Facturación por mejoras de las condiciones contratadas originalmente
- ▶ Descuentos promocionales aplicados a la factura

## De la reserva al cliente

El paso siguiente era convertir los datos de caracterización de las reservas en datos de caracterización de los clientes que las hacían. Al final del camino, queríamos tener únicamente clientes definidos por su conducta: cuántas noches pernoctaban, con qué régimen de alojamiento, con qué tipo de acompañantes, qué actividades realizaban, con qué equipamiento adicional, qué canales de contratación preferían, ....

A partir de los 9 campos de caracterización de la reserva, obtuvimos 18 variables adicionales para caracterizar a los clientes, además de las 4 que ya teníamos. Esta es la lista final:

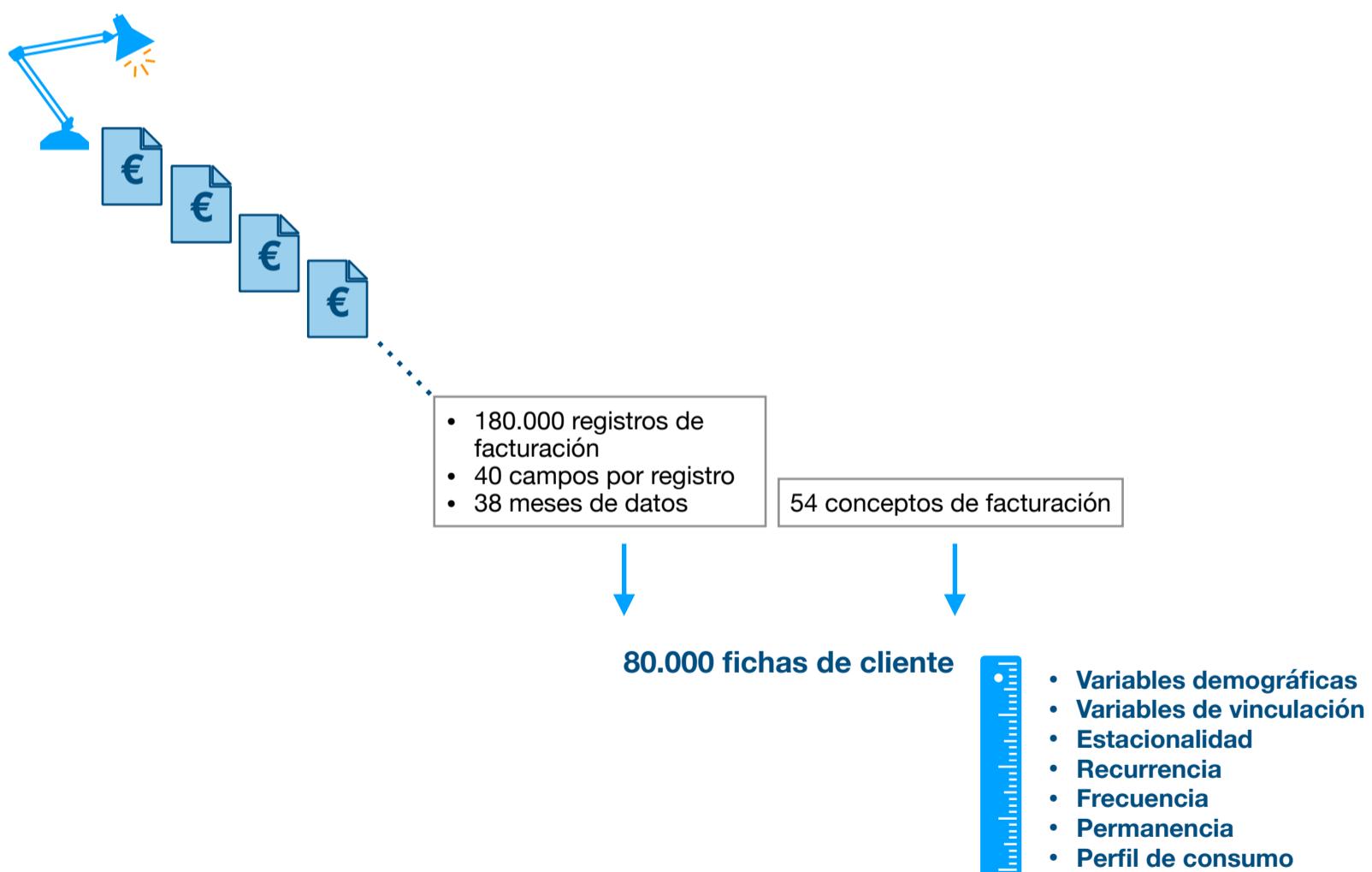
- Variables demográficas
  1. Sexo
  2. País de nacimiento
  3. Edad media en el período de estudio
  4. Número medio de adultos en las reservas que hizo el cliente
  5. Número medio de niños en las reservas que hizo el cliente
- Variables de vinculación
  6. Indicador de suscripción a información comercial
  7. Preferencia por un canal de ventas (último utilizado)
  8. Número medio de *titulares* (de facturación) diferentes agrupados en *lotes de reservas* en las que se incluían las de este cliente, durante el período de estudio.
- Estacionalidad
  9. Trimestre del año donde el cliente hizo la reserva con mayor número de noches
- Recurrencia
  10. Número de años en los que el cliente hizo una reserva
- Frecuencia
  11. Media anual de reservas en los años con reserva
- Permanencia

## 12. Media anual de noches reservadas en los años con reserva

- Perfil de gasto o consumo

13. Gasto medio anual en alojamiento y régimen
14. Gasto medio anual en extras de restauración
15. Gasto medio anual en paquetes 'todo-incluido'
16. Gasto medio anual en mejora de condiciones
17. Gasto medio anual en actividades extras
18. Gasto medio anual en alquileres de materiales o equipamiento extras
19. Gasto anual en extras de telecomunicaciones
20. Gasto medio anual en tiendas
21. Facturación media anual del cliente
22. Descuento medio anual por participación en promociones

Tras esta proceso de extracción y transformación, obtuvimos unos 80.000 clientes caracterizados por 22 variables.



# Análisis

En este apartado, como hemos hecho en los anteriores, extraeremos únicamente los componentes del análisis más relevantes para el objetivo de este capítulo: mostrar la adecuación de un tipo de análisis a un objetivo de negocio específico, que es, en este caso, el diseño de un programa de fidelización <sup>1</sup>.

Por diversas razones previas al proyecto, manejábamos tres categorías de clientes como escala de partida para el programa de fidelización. Teníamos, pues, que definir, primero, y caracterizar, después, a estas tres categorías.

## Categorías de clientes

Para definir las categorías, utilizamos el agrupamiento en *clusters* en torno a *centroides* o medias centrales <sup>2</sup>. Esta técnica arroja grupos con elementos muy similares dentro de ellos y muy dispares respecto a los elementos en otros grupos: alta homogeneidad interna y alta heterogeneidad externa. Como variables a tener en cuenta para valorar la similitud entre clientes, utilizamos aquellas relacionadas con su perfil de consumo (descritas más arriba); por dos razones:

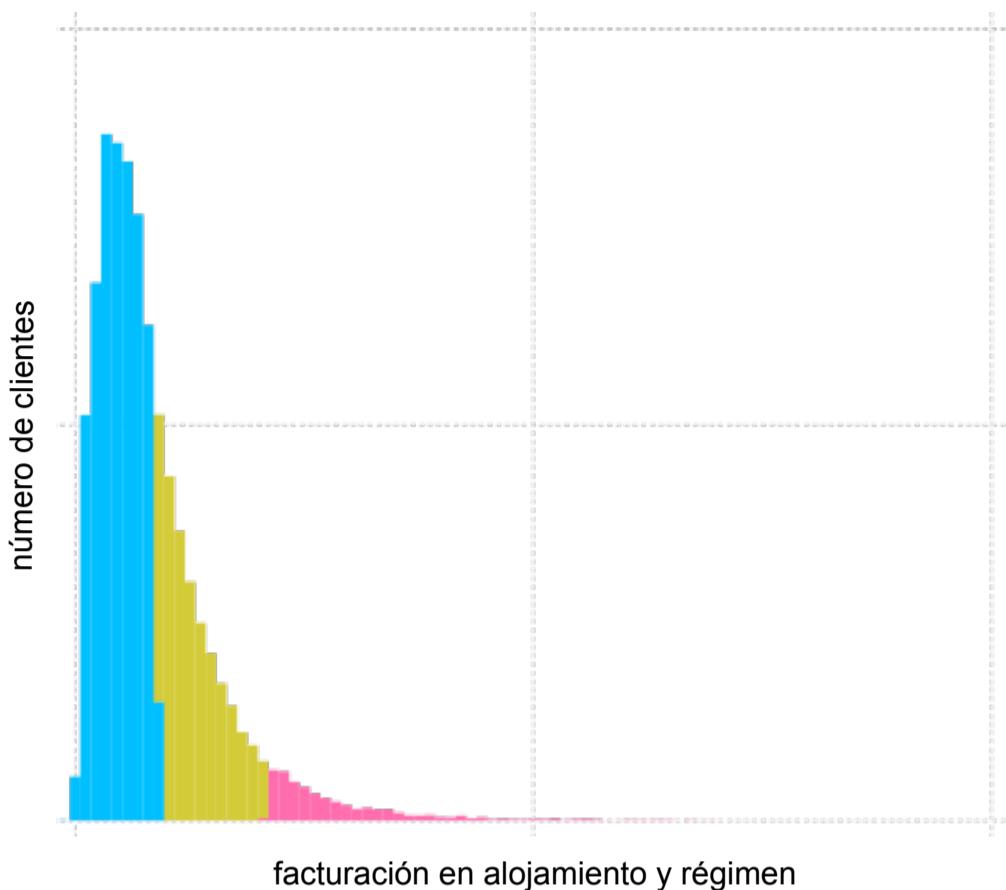
1. El programa iba a premiar, sobre todo, la facturación de los clientes.
2. Los incentivos del programa iban a estar ligados, a su vez, a conceptos de facturación: regalo de estancias, o de actividades, o de otras opciones de consumo ofrecidas por la propia cadena hotelera.

---

<sup>1</sup> Mantendremos también una cierta vaguedad en la presentación de los resultados, obligados por la confidencialidad que debemos a nuestro antiguo cliente.

<sup>2</sup> Utilizamos el algoritmo K-means implementado en una librería de JULIA. Todas las tareas de análisis, así como las de carga y transformación de los datos originales, se programaron con el lenguaje JULIA. Por simplificar, no utilizamos un gestor de base de datos: los resultados intermedios los almacenábamos en ficheros en formato CSV.

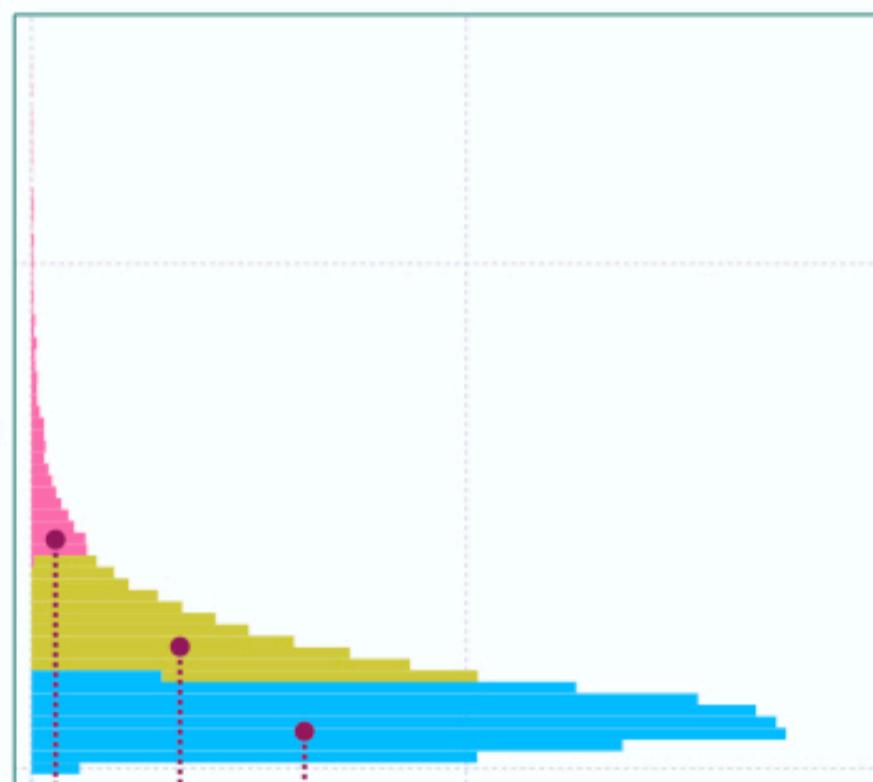
El siguiente gráfico muestra los tres *clusters* que obtuvimos.



De izquierda a derecha, del azul al fucsia, los *clusters* comprenden clientes de menor a mayor gasto en alojamiento y régimen. Reseñamos varios hechos:

- Cuanto más gasto (más a la derecha en el eje horizontal) en un *cluster*, menos clientes (altura de la barra vertical) pertenecen a él.
- La relación anterior, sin embargo, no se cumple en el *cluster azul*, el de menor gasto: hay más clientes en el centro de cluster que al inicio de él.
- El gráfico, al girarlo, nos ofrecía una pirámide de partida muy similar a la que necesitaríamos en un programa de fidelización: podíamos utilizar los *clusters* para encuadrar a los clientes en los tres niveles de membresía del programa.

Programa  
de incentivos  
para aumentar  
la facturación



Clasificación en los tres niveles  
de membresía del programa

## Incentivos y facturación

Al igual que el número de categorías, también era una decisión previa al proyecto que los incentivos serían servicios ofrecidos por la propia cadena.

Como hipótesis de partida, establecimos que los mejores servicios para ofrecer como incentivos serían aquellos más populares dentro de cada *cluster*, entendiendo como tales no los de mayor consumo ‘medio’, sino aquellos con un consumo más ‘homogéneo’: los servicios con mayor número de consumidores dentro de cada cluster<sup>1</sup>. Por ejemplo, en un cluster determinado, la actividad recreativa más practicada podía ser el mini-golf y en otro, los baños en aguas termales.

Los resultados de nuestro análisis<sup>2</sup> pueden ilustrarse en tablas y gráficos como los dos siguientes.

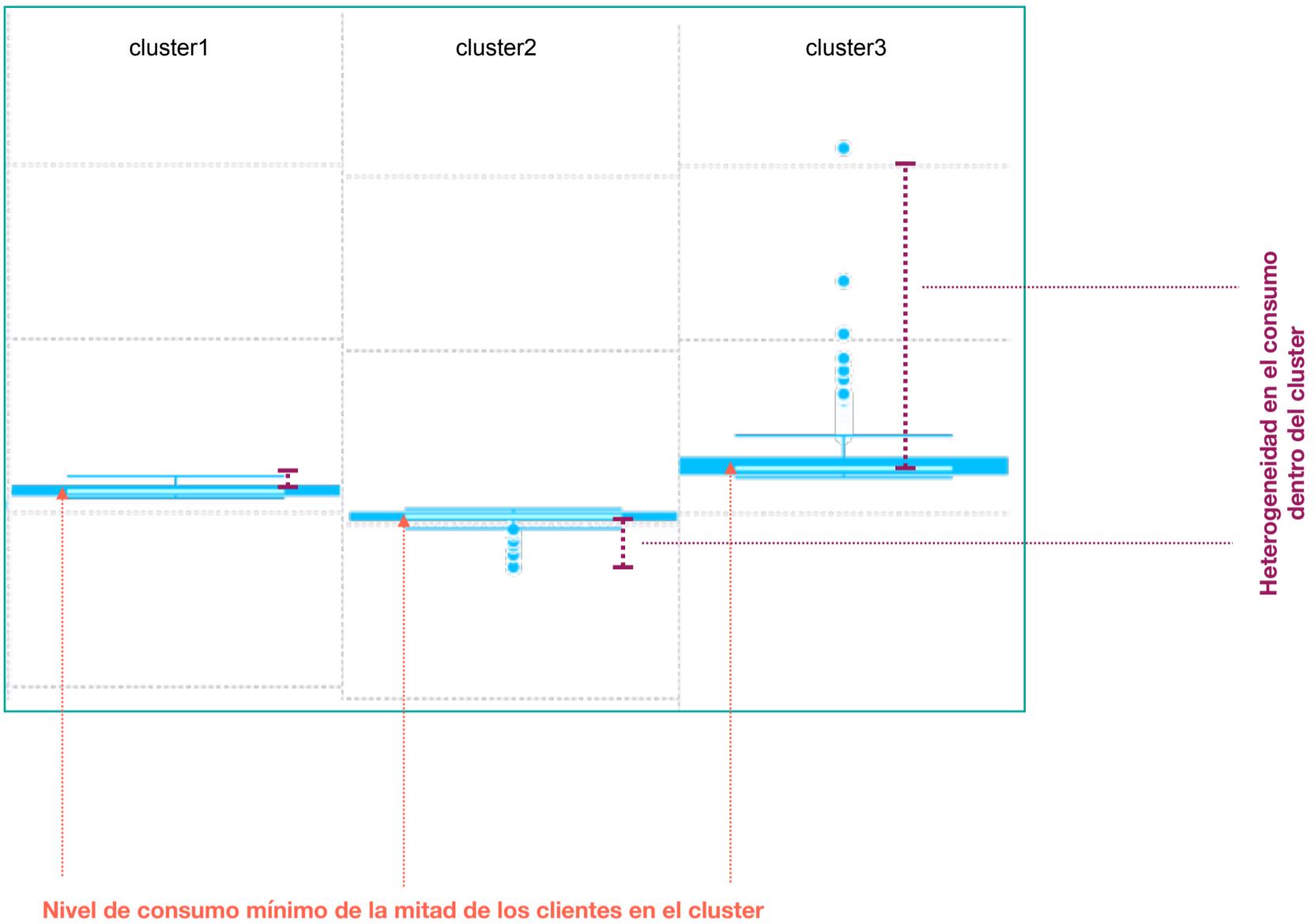
decil	cluster1	cluster2	cluster3
1.0	94.49	16.5	218.48
2.0	100.72	26.54	226.64
3.0	107.28	33.84	236.52
4.0	114.14	40.85	246.95
5.0	123.42	47.54	261.58
6.0	133.39	54.26	276.74
7.0	145.8	62.31	301.04
8.0	159.82	69.85	336.69
9.0	179.06	78.52	399.13
10.0	210.76	89.2	2102.8

La tabla muestra la facturación media anual por cliente en una actividad recreativa, en cada *cluster* y en cada grupo resultante de dividir los clientes del *cluster* en diez grupos de igual tamaño (decil).

- En el cluster1, el cliente que más factura dobla la facturación del que factura menos.
- En el cluster2, la facturación mayor multiplica por cinco la menor.
- En el cluster3, la mayor facturación multiplica la menor por diez.
- Esta actividad extra tiene un consumo muy homogéneo en el cluster1 y en los nueve primeros deciles (el 90% de los clientes del cluster) del cluster3.

<sup>1</sup> El concepto estadístico que utilizamos fue la distribución de la facturación de cada servicio dentro de los clusters, medida agregadamente por la mediana e individualizada por decil de clientes.

<sup>2</sup> Los datos de las tablas están ligeramente alterados respecto a los datos reales que obtuvimos, para mantener la confidencialidad.



Nivel de consumo mínimo de la mitad de los clientes en el cluster

El gráfico llega a conclusiones similares, desde otro punto de vista:

- Muestra el nivel mínimo de consumo de la actividad en cada *cluster*, medido por el punto medio del rectángulo central. Vemos nuevamente que el nivel de consumo más bajo lo tienen los clientes del cluster2.
- La heterogeneidad dentro de cada *cluster* la mide la longitud de las ‘patillas’ y de las filas de puntos que salen de las rectángulos centrales que agrupan el grueso de los clientes. La mayor heterogeneidad la presentan los clientes del cluster3 y la menor, los del cluster1.
- El cluster1 presenta el perfil idóneo para ofrecer esta actividad como incentivo en un programa de fidelización: tiene un nivel alto de consumo y es valorada por todos los clientes del *cluster* de una forma similar.

Este enfoque analítico nos daba una primera solución al problema de la selección de incentivos para cada uno de los tres niveles de membresía del programa de fidelización.

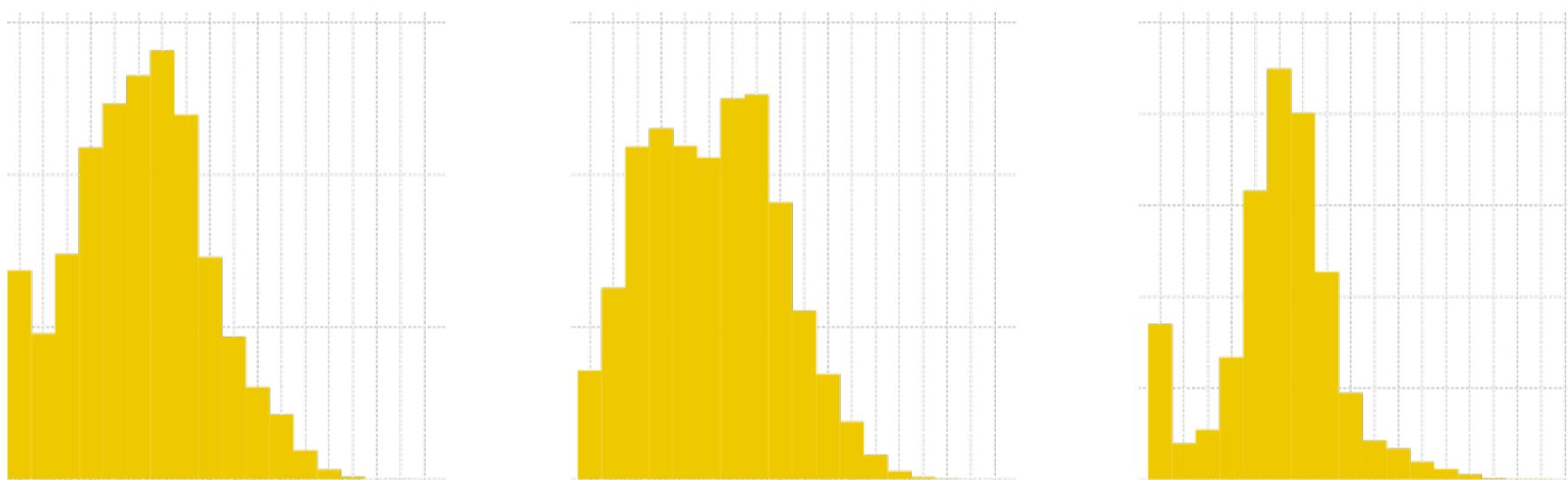
## Incentivos y demografía

La política de incentivos por nivel de membresía habría de tener en cuenta, además, otros tipos de factores:

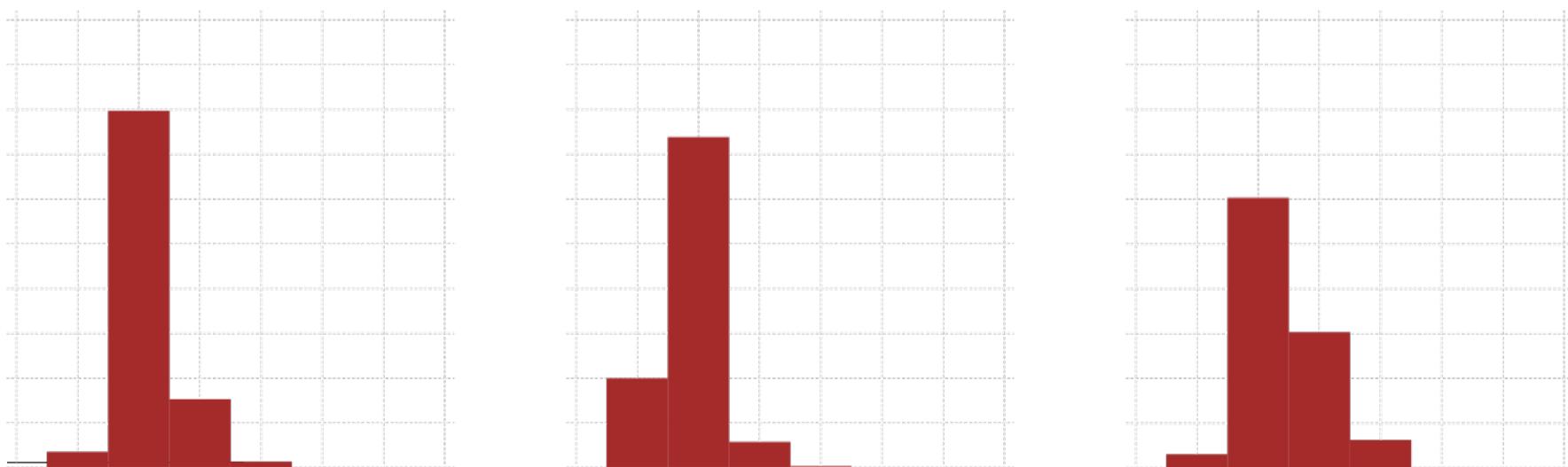
1. El coste marginal para los hoteles de la provisión del incentivo y el posible aumento de facturación derivado.
2. La caracterización demográfica de cada *cluster*: no tiene sentido, por ejemplo, ofrecer cama supletoria gratis, como incentivo, si mayoritariamente solo hay reservas con 1 ó 2 adultos en un *cluster*.

En el primer tipo de consideraciones no entraremos, pero sí ilustraremos brevemente, sin poder hacerlo con detalle, el segundo aspecto <sup>1</sup>.

**Porcentaje de clientes (vertical) en cada *cluster* por tramo de edad**

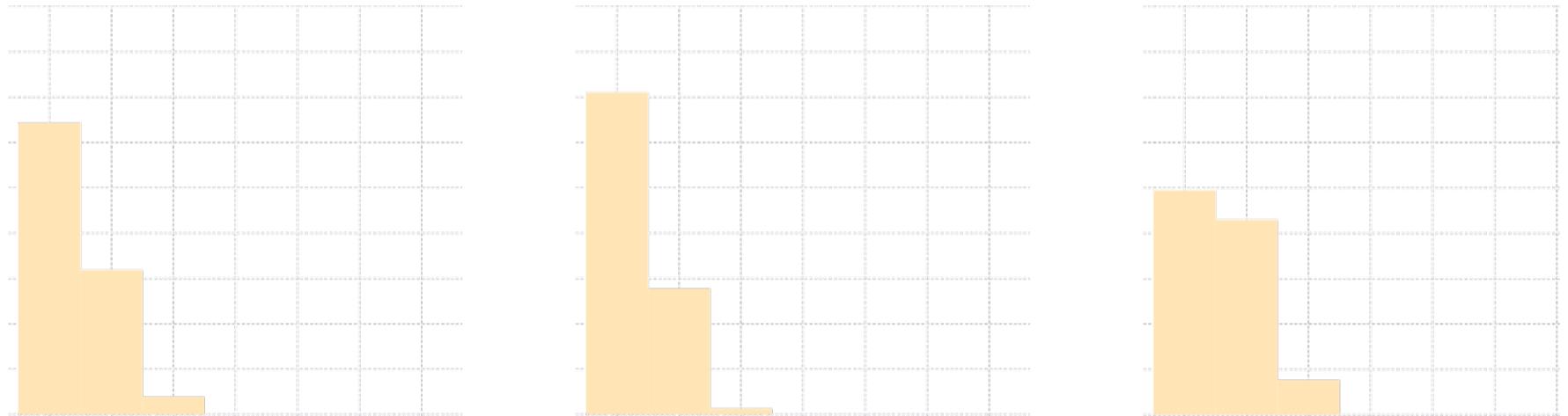


**Porcentaje de clientes (vertical) en cada *cluster* según el número medio de adultos (1-4) en sus reservas**



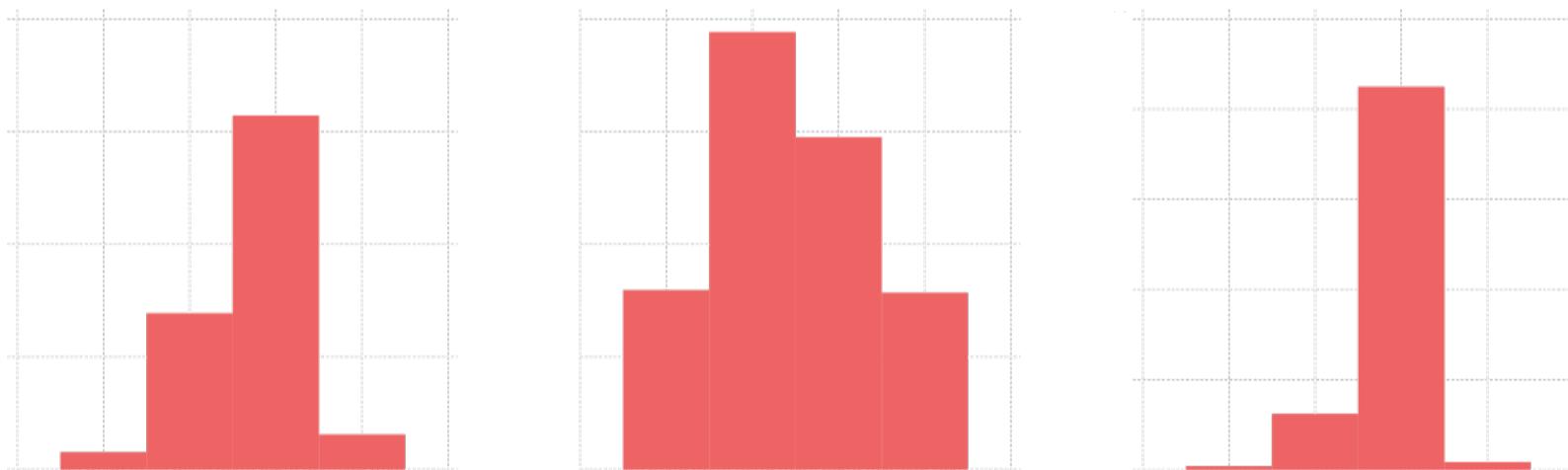
<sup>1</sup> No identificaremos los gráficos con su número de *cluster*, ni aparecerán siempre los *clusters* en el mismo orden, por confidencialidad de los datos.

**Porcentaje de clientes (vertical) en cada *cluster* según el número medio de niños (1-3) en sus reservas**



Incluimos también un gráfico sobre la estacionalidad predominante en cada *cluster*, en la medida en que también delimita qué incentivos ofrecer.

**Porcentaje de clientes (vertical) en cada *cluster* según el trimestre más frecuente en sus reservas**



Todos los gráficos anteriores muestran la heterogeneidad demográfica y estacional de los diferentes *clusters* y apuntan a la conveniencia de tenerlo en cuenta en el diseño de la política de incentivos.

## Intermediación y facturación

Retomamos en este apartado una de las motivaciones originales del proyecto: ganar mayor control de la relación con el cliente, a través de los canales directos de ventas y de un programa de fidelización dirigido a él.

La hipótesis de partida era, pues, que ‘desintermediar’ las ventas podía aumentar la facturación. ¿Apoyaban los datos de las reservas esta asunción?

Respondimos esta pregunta analizando los datos desde dos puntos de vista:

1. ¿Facturan más los clientes que prefieren canales directos de venta? O en términos de nuestras unidades de análisis: ¿qué canales utilizan más los clientes del *cluster* con mayor facturación?
2. Nos habíamos tomado la molestia de definir una variable con un nombre muy largo, ‘Número medio de *titulares* (de facturación) diferentes agrupados en *lotes de reservas* en las que se incluían las de este cliente’, porque pensábamos que podía ser el mejor indicador del grado de intermediación de una venta: cuando un cliente reserva habitación en la misma reserva que lo hacen otros cincuenta clientes es muy improbable que la reserva la esté haciendo él directamente. ¿Cómo se comportaban los *clusters* atendiendo a esta variable?

Desde el primer punto de vista, los resultados<sup>1</sup> confirmaban, sin ser rotundos, nuestra hipótesis: los clientes del *cluster* con mayor facturación utilizaban menos los canales de venta de operadores y agencias, y más otros canales como el call-center o los portales de internet.

Desde el segundo punto de vista, los resultados eran más nítidos:

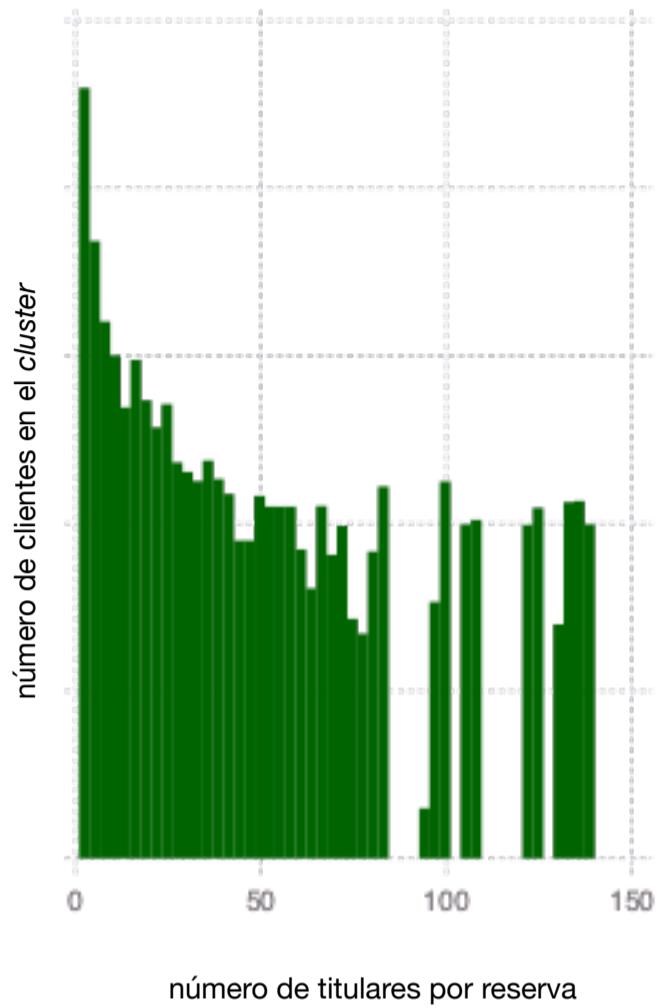
- Los clientes del cluster de mayor facturación se concentraban principalmente en reservas con 5 titulares (de facturación) o menos, mientras que los clientes del cluster de menor facturación se concentraban en el grupo de 60 titulares o menos.

Los gráficos siguientes muestran, para los clusters de mayor y menor facturación, el número de clientes que pertenecen, en cada cluster, a cada tramo de número de titulares por reserva.

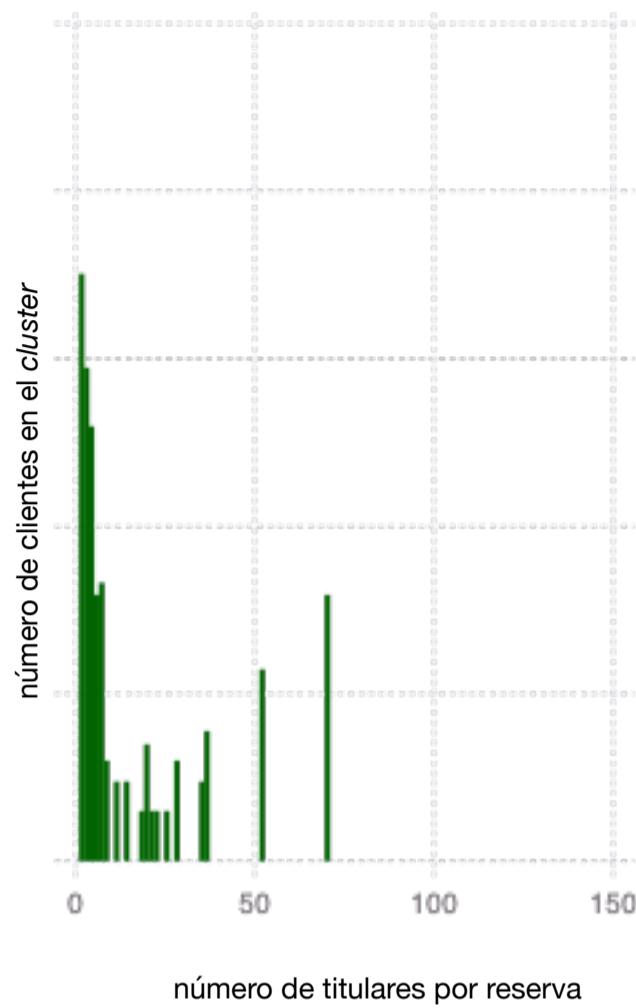
---

<sup>1</sup> No podemos presentar el detalle de los resultados por razones de confidencialidad.

**cluster menor facturación**



**cluster mayor facturación**



# Conclusiones

Clientes más vinculados con los canales externos de venta que con el proveedor del servicio final, alto coste por la intermediación de agencias y operadores, un programa de fidelización que aumente la vinculación con el cliente final y prime el canal de ventas directo...



Nuestra aproximación al problema y nuestra solución:

1. Compusimos 80.000 fichas de clientes a partir de datos de facturación exclusivamente.
2. Clasificamos a los clientes en tres *clusters* según su perfil de consumo y traspusimos los *clusters* en tres niveles de membresía de un programa de fidelización.
3. Diseñamos una política de incentivos basada en el perfil demográfico de los clientes en cada *cluster* y de los servicios más consumidos por ellos.
4. Verificamos las asunciones sobre la relación entre intermediación en la venta y facturación del cliente final.

Más allá:

5. Actualización automática de los niveles de membresía y de la política de incentivos del programa de fidelización, en función de la evolución de los datos de facturación individual.

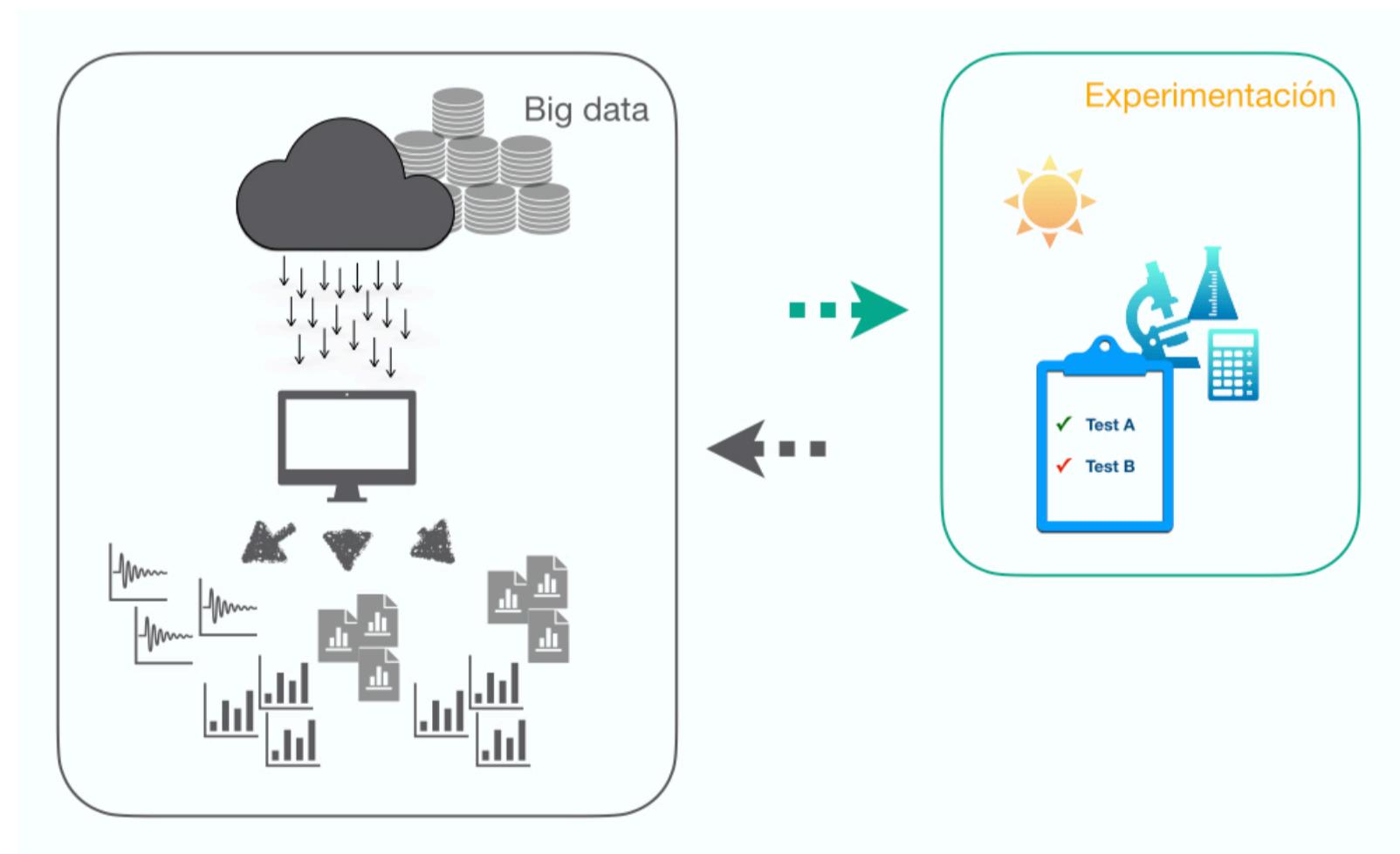
# Experimentación

## Punto de partida

¿Qué volumen, qué tipo de datos hacen falta para implantar técnicas sofisticadas de análisis de datos? ¿Qué metodología? ¿Qué infraestructura tecnológica? ¿Qué inversión?

Estas eran algunas de las preguntas que surgían en las conversaciones con nuestros clientes, a las que fuimos dando respuestas parciales como las que hemos narrado en los capítulos anteriores, y a las que hemos dado una respuesta más global en una plataforma multi-cliente que hemos denominado LebenLab.

Aunque la tentación comercial era irresistible, desechamos desde el principio construir LebenLab sobre un concepto tan elefantiásico como el de *big data*. La idea debía ser otra más enfocada al resultado final, en lugar de al pedregoso camino; algo así, por seguir en inglés, como *just knowledge*. Y nos pusimos en marcha.



# Big data y experimentación

Simplificando mucho, distinguiremos entre dos enfoques para enfrentarnos a los datos:

## Big data

Aprovechar *toda* la información disponible, aquella que surge de una manera *natural* de la actividad de la empresa: datos del ‘día a día’ del negocio, principalmente. Este es el enfoque que ha dado lugar, junto a otros factores, a lo que denominamos *big data*. Requiere una serie de actividades, entre las que podemos mencionar:

- Capturar información; cuanta más, mejor.
- Almacenarla en repositorios.
- Estructurarla en unidades lógicas (variables y la relaciones entre ellas).
- Analizarla de mil y una maneras, hasta encontrar alguna conclusión *significativa* para el negocio.
- Publicar y dar acceso a los resultados, preferiblemente a través de facilidades de auto-servicio construidas con *software* de *business intelligence*.

## Experimentación

Buscar y crear *artificialmente* datos de los que ahora no disponemos, que podrían clarificar la toma de ciertas decisiones. A este atajo para aprender es lo que en este capítulo llamamos experimentar. Para hacerlo basta con establecer una hipótesis y montar un dispositivo (un experimento) para validarla.

Desvíémonos por unos minutos e imaginemos un científico que quiera estudiar los efectos de una vacuna. Démole nuestros dos enfoques:

## Big data

Olvidándonos de todas las trabas legales con que se encontraría, nuestro científico podría vacunar a *toda* la población en riesgo y luego hacer un seguimiento individual y muy pormenorizado de cada uno de los vacunados.

Tras unos meses, podría comparar las cifras de infección actuales con las cifras históricas y ver si había una disminución significativa. Si la hubiera, podría deberse al carácter fuertemente estacional de la infección, a un cambio de hábitos sociales o al efecto de la vacuna. Podría esperar entonces un año completo y comparar la evolución de la infección en esos doce meses, controlando el factor estacional, pero olvidando que en ese año podrían haber cambiado otros factores.

En paralelo, haría un estudio exhaustivo de los millones de historiales médicos de las personas vacunadas, buscando patrones de reacción adversa, de completa ineficacia o de

adecuada inmunización, según características como la edad, el sexo, la localización, la tensión arterial, el tiempo que dedican a ejercitarse, etc...

Por fin, tras dos o tres años de estudio, podría concluir que a un grupo de vacunados hubiera sido mejor no vacunarlos; a otros, hubiera dado lo mismo, y para un tercer grupo, vacunarlos era imperativo.

## Experimentación

En lugar de todo lo descrito más arriba, nuestra científica, antes de vacunar a toda la población, podría haber hecho un experimento con dos pequeñas muestras aleatorias, bien seleccionadas desde el punto de vista de las variables epidemiológicas a estudiar, administrar la vacuna a una muestra y, por último, hacer un seguimiento de la evolución de los dos grupos durante el mínimo tiempo necesario para establecer una conclusión.

En el ejemplo anterior, bastaría sustituir ‘vacunación de toda la población’ por ‘campaña a toda la base de clientes’ para que viéramos inmediatamente la utilidad de experimentar, también, en el ámbito del marketing promocional. Esto es lo que hicimos en Leben: desarrollar un laboratorio para que nuestros clientes pudieran experimentar en él, sin tener que montar cada uno de ellos el suyo propio.

Para avanzar en nuestro proyecto necesitábamos una metodología, hipótesis e instrumental.

# Metodología

Como metodología, en Leben elegimos los tests A/B:

1. Se definen dos propuestas de valor, para los clientes que participarán en el experimento, que difieren en uno de sus elementos.
2. Se extraen dos muestras aleatorias estratificadas del conjunto de clientes destinatarios de la acción promocional.
3. Se pone en marcha la promoción entre los participantes seleccionados en las muestras, asignándole a cada uno de ellos una de las dos propuestas (tests).
4. Se miden los resultados en una y otra muestra, y se valora la significación estadística de la diferencia entre ellos.
5. Se define la propuesta final común para todos los clientes.

The screenshot shows the Leben software interface for creating a new experiment. The top navigation bar includes 'EXPERIMENTOS' (selected), 'SIMULACIONES', and 'FICHEROS'. Sub-options 'NUEVO' and 'LISTA' are visible under 'EXPERIMENTOS'. The main section is titled 'NUEVO EXPERIMENTO' with a red information icon. Below it is a 'DATOS EXPERIMENTO' section. The 'Nombre' field contains 'exp2\_febrero\_2021'. The 'Fecha de Inicio' is '27-02-2021' and 'Fecha de Fin' is '08-03-2021'. Under 'Mercados', there is a dropdown menu with 'España', 'Portugal', and 'Andorra' selected. Under 'Conceptos', there is a dropdown menu with 'Red\_ventas\_1', 'Red\_ventas\_2', 'Red\_ventas\_3', and 'Red\_ventas\_4'. A note at the bottom left says 'El conjunto actual de participantes en base de datos comprende 150 participantes. Para utilizar un conjunto diferente al actual, clique en la casilla y seleccione el fichero que los contiene.' A 'Choose File' button shows 'no file selected'. A red callout arrow points from this button to the text 'Conjunto total de clientes susceptibles de participar en una promoción'. Another red callout arrow points from the 'Mercados' and 'Conceptos' sections to the text 'Variables a tener en cuenta en la extracción de muestras: selección de mercados y redes de venta de los participantes.' and 'Fechas del experimento' respectively.

# Hipótesis

Un experimento vale tanto como la hipótesis que se somete a experimentación.

En nuestro laboratorio, las hipótesis son el fruto del *know-how* acumulado durante años por los equipos de marketing y ventas de nuestros clientes. Frente a otros enfoques de análisis de datos, la experimentación se apalanca en el capital invertido en la experiencia del equipo humano de una empresa; nuevamente: *knowledge* frente a *big data*.

DATOS DE CAMPAÑAS

**CAMPAÑA A**

Nombre  
 Entre 5 y 25 caracteres; letras, números, guión bajo; sin espacios.

Productos  
Selección hasta 3  
Sistemas de frenado  
Iluminación  
Encendido

Incentivo  
Descuento en metálico

Comunicación  
sms

Texto comunicación  
Texto del mensaje SMS.  
Máximo 500 caracteres.

**CAMPAÑA B**

Nombre  
 Entre 5 y 25 caracteres; letras, números, guión bajo; sin espacios.

Productos  
Selección hasta 3  
Sistemas de frenado  
Iluminación  
Encendido

Incentivo  
Tarjetas regalo

Comunicación  
email

Texto comunicación  
Texto del mensaje para correo electrónico.  
Máximo 500 caracteres.

**OPCIONES DE PARA EXPERIMENTAR**

Venta cruzada de productos

Tipo de incentivo

Canal de comunicación

Mensaje a comunicar

**OBTENER MUESTRAS**

Muestra test A

Muestra test B

En la versión actual, a la que pertenece la pantalla parcialmente mostrada más arriba, podemos experimentar con los siguientes elementos del diseño de una promoción:

- Opciones de venta cruzada de productos
- Tipo de incentivo promocional
- Canal de comunicación
- Mensaje enviado por el canal de comunicación

Estos son algunos ejemplos de los tipos de hipótesis que pueden experimentarse:

- Las ventas de ‘Sistemas de frenado’ son mayores cuando el incentivo promocional incluye también la venta cruzada de ‘Iluminación’ y ‘Encendido’.
- En el mercado español, es más atractivo el incentivo del ‘Descuento en metálico’ que las ‘Tarjetas de regalo’ por el mismo importe. En el mercado portugués, ocurre lo contrario.
- En la Red\_ventas\_1, tienen más éxito las promociones comunicadas con SMS que con correo electrónico.
- El texto de tipo A, para comunicaciones por correo electrónico, está asociado a mayores ventas en las promociones de los productos X, V y Z, que las que se registran cuando se utiliza el texto de tipo B.

# Validación de hipótesis

Durante el período de ejecución de un experimento, pueden consultarse los resultados parciales obtenidos para nuestra hipótesis:

- La venta media por participante en cada uno de los tests o campañas.
- Si la comunicación es por correo electrónico, número de aperturas hasta la fecha.
- La probabilidad de que las diferencias entre las ventas de los dos tests NO se deban a la hipótesis que estamos experimentando: valores superiores a 0.05 (5%) se muestran en rojo e indican que las diferencias NO son estadísticamente significativas; es decir, que nuestra hipótesis no es válida <sup>1</sup>.

**leben**

EXPERIMENTOS ▾ SIMULACIONES FICHEROS ▾

**RESULTADO DEL EXPERIMENTO**

**exp1\_febrero\_2021**  
2021-02-19 2021-02-28

Mercados Conceptos

Campaña A		Campaña B	
exp1_febrero_2021_A	dto_dinero	Campaña	exp1_febrero_2021_B
Número participantes - TEST A	75	Incentivo	tarjeta_regalo
email		Participantes	75
Texto promocional para comunicación por correo electrónico, TIPO A.		Comunicación	email
0		Texto	Texto promocional para comunicación por correo electrónico, TIPO B.
		Aperturas(%)	0

**PG1\_2**  
6,08 Venta media 4,88  
Probabilidad de que las diferencias en ventas no sean significativas: **0,127**

**PG1\_1**  
4,07 Venta media 4,63  
Probabilidad de que las diferencias en ventas no sean significativas: **0,467**

Venta media por participante en el producto 2 - TEST A  
↓  
Venta media por participante en el producto 1 - TEST A

Venta media por participante en el producto 2 - TEST B  
↓  
Venta media por participante en el producto 1 - TEST B

Significación estadística de las diferencias entre las ventas medias del producto 2 en los dos tests  
↓

Significación estadística de las diferencias entre las ventas medias del producto 1 en los dos tests  
↓

The screenshot shows the Leben software interface with the title 'RESULTADO DEL EXPERIMENTO'. Below it is the experiment name 'exp1\_febrero\_2021' and dates '2021-02-19 2021-02-28'. There are two tabs: 'Mercados' (Markets) and 'Conceptos' (Concepts). The 'Mercados' tab is active, displaying a table comparing 'Campaña A' and 'Campaña B'. Both campaigns have 75 participants. The table includes columns for campaign names, incentive types, participant counts, communication types, text descriptions, and opening percentages. Below the table are two bar charts labeled 'PG1\_2' and 'PG1\_1'. Each chart has two bars: '6,08' and '4,88' for PG1\_2, and '4,07' and '4,63' for PG1\_1. Below each bar is a red arrow pointing down to a text label: 'Venta media por participante en el producto 2 - TEST A' for the first bar and 'Venta media por participante en el producto 1 - TEST A' for the second. Red arrows also point from the text labels to the bottom of the charts. At the bottom of each chart is a red text box containing a probability value: '0,127' for PG1\_2 and '0,467' for PG1\_1. Red arrows point from these text boxes to the bottom of the charts. The bottom of the charts also contains text about statistical significance: 'Significación estadística de las diferencias entre las ventas medias del producto 2 en los dos tests' and 'Significación estadística de las diferencias entre las ventas medias del producto 1 en los dos tests'.

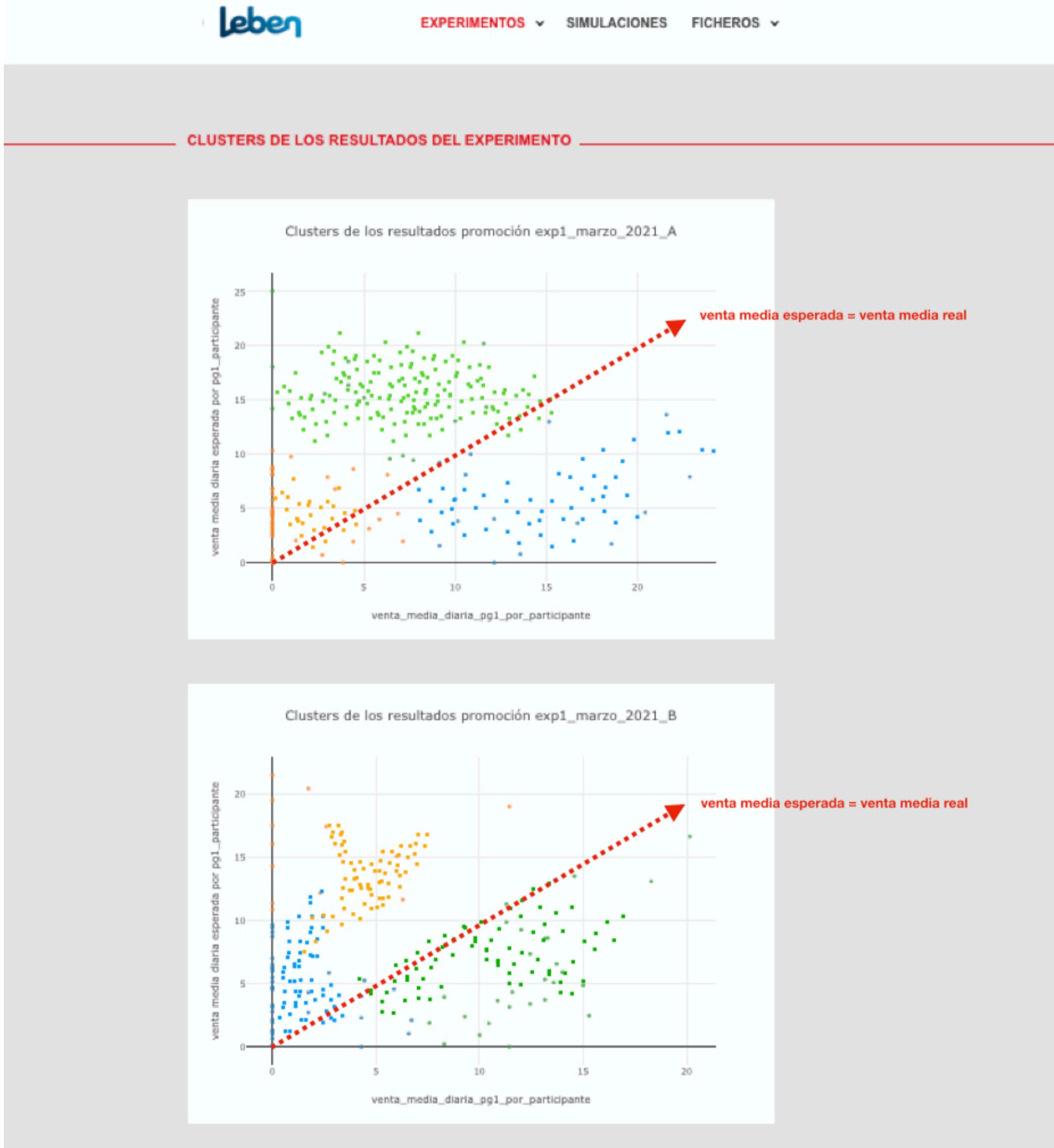
<sup>1</sup> Las muestras de participantes se estratifican por mercado y por tipo de red de venta, para asegurar su homogeneidad respecto a estas dos variables. A las ventas por participante obtenidas en cada test, se les aplica un test de diferencias de medias basado en una distribución de t-Student, cuyo resultado se consigna en el informe de seguimiento.

En el gráfico que hemos puesto más arriba, hemos experimentado con la hipótesis siguiente: en todos los mercados y en todas las redes, para promocionar las ventas de los productos 2 y 1 es más eficaz un incentivo de 'descuento en metálico' que otro de 'tarjeta de regalo'. Los resultados que arroja el experimento son:

- En el producto 2, las ventas medias por participante en el test A son superiores a las del test B: 6,08 y 4,88, respectivamente.
- En el producto 1, la diferencia es menor: 4,07 frente a 4,63, en el test A y en el test B, respectivamente.
- En ambos tests, la probabilidad de que la diferencia NO sea significativa estadísticamente es superior al 5%: 12,7% en el producto 2 y 46,7% en el producto 1. Con estas cifras, no parece claro que el 'descuento en metálico' sea un incentivo más atractivo que la 'tarjeta regalo', especialmente en el caso del producto 1.

# Resultados esperados y reales

¿Cuál es la *perturbación* que introduce una campaña en la conducta previsible de los participantes?



La conducta previsible de un participante en una campaña es el dato que en el gráfico denominamos ‘venta media esperada’, que aparece representado en el eje vertical. En LebenLab, calculamos el resultado esperado multiplicando dos variables históricas:

- La venta media diaria hecha por un participante en las promociones anteriores
- La probabilidad de que un participante venda en la campaña actual, estimada a partir del porcentaje de promociones anteriores en las que el participante efectivamente vendió.

En el eje horizontal, hemos representado la ‘venta media real’ hecha por un participante hasta el momento de la consulta del informe de resultados.

Cada gráfico ilustra los resultados de un test. Cada punto de los gráficos es un participante representado por un par de números: venta media esperada, que es su coordenada en el eje vertical, y venta media real, que es su coordenada en el eje horizontal.

Las flechas rojas que cruzan diagonalmente los gráficos representan aquellos participantes cuyas ventas reales son exactamente las mismas que sus ventas esperadas: cuanto más cerca esté un punto de la flecha, más se ajusta la venta real a la venta que habíamos previsto para un participante.

Los puntos por encima de las flechas rojas representan participantes para los que la campaña ha producido una perturbación negativa: sus ventas reales están por debajo de las ventas previstas o esperadas. Por el contrario, los puntos por debajo de las flechas corresponden a participantes con una perturbación positiva: las ventas reales superan las ventas esperadas. Un buen concepto de campaña es aquel que coloca la mayoría de los participantes por debajo de la flecha diagonal.

Además de la información individual que recoge cada punto, el gráfico agrupa los participantes en tres grupos o *clusters*<sup>1</sup>, que podrían servirnos ulteriormente para modular el diseño de la promoción en cada uno de ellos.

Entresacamos algunos ejemplos del tipo de conclusiones que ilustran los gráficos:

1. El diseño promocional del test A (gráfico superior) afecta positivamente a las ventas esperadas de los participantes en el grupo azul. Este grupo, sin embargo, responde desfavorablemente al diseño del test B (gráfico inferior): casi todos los participantes del grupo están por encima de la flecha diagonal.
2. En el grupo verde, ocurre exactamente lo contrario que en el grupo azul: es más favorable la respuesta al test B que al test A.
3. El grupo naranja no parece atraído por ninguna de las dos propuestas, aunque es ligeramente más favorable su respuesta en el test A.

---

<sup>1</sup> Los tres clusters son el resultado de aplicar el algoritmo K-means al conjunto de resultados esperados y reales de todos los participantes que intervienen en un experimento.

4. Ni el test A, ni el test B producen una perturbación homogénea, positiva o negativa, en el conjunto de los participantes en el experimento.

En el capítulo siguiente, ‘Predicción’, veremos cómo utilizar estas conclusiones para extrapolarlas a quienes no participaron o para predecir los resultados de promociones similares.

# Instrumental

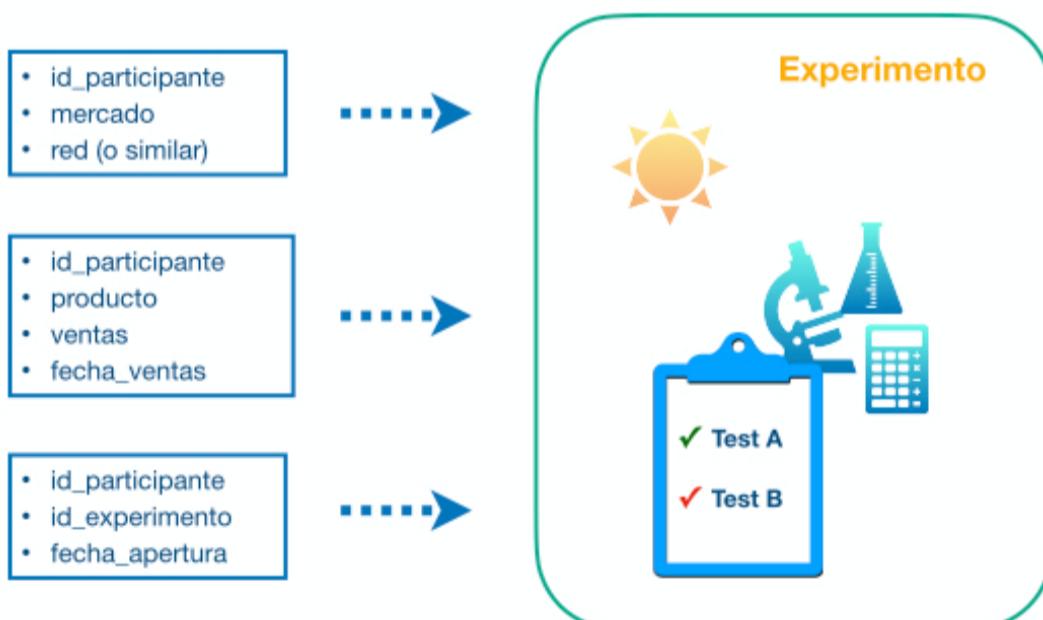
El instrumental del LebenLab lo componen las herramientas informáticas de carga, transformación, almacenamiento, cálculo, acceso y presentación de los datos que utilizamos en un experimento.

## Datos

Para que cualquiera de nuestros clientes, por pequeño que sea, pueda utilizar nuestro laboratorio cuando desee hacerlo, el primer requerimiento ha de ser que pueda experimentar con un conjunto de datos que ya tenga disponible. Dicho de otra manera: experimentar no tiene que implicar para el cliente un coste de fabricación de datos a la medida del experimento. Y no lo tiene.

En LebenLab, un experimento requiere tres tipos de información que, en total, comprenden diez datos:

- Datos de participantes
- Datos de ventas unitarias de los productos en campaña.
- Opcionalmente, datos de aperturas cuando la comunicación sea por correo electrónico.



## Software

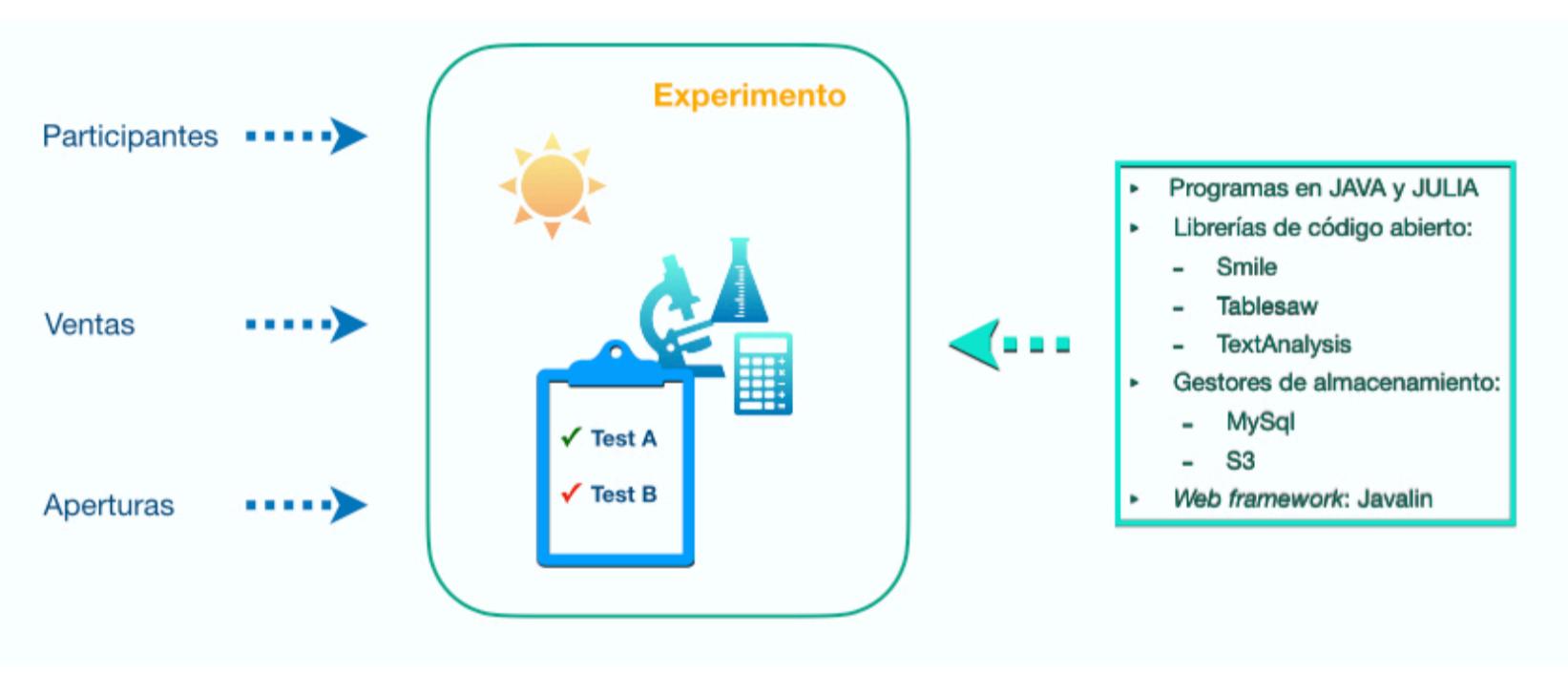
Para los procesos de carga y transformación de datos, y de obtención de muestras de participantes, desarrollamos nuestros propios programas en JAVA, principalmente, y en JULIA.

Para las estimaciones estadísticas y la *clusterización* de los resultados de los experimentos, utilizamos la librería de código abierto SMILE (Statistical Machine Intelligence and Learning Engine). Los gráficos están programados con la librería de código abierto Tablesaw.

Para *entrenar* nuestro algoritmo de clasificación de mensajes de comunicación, utilizamos la librería de código abierto TextAnalysis.jl. La clasificación posterior de los nuevos mensajes la hicimos con un programa que desarrollamos en JAVA y en JULIA, siguiendo las especificaciones del algoritmo *Naive Bayes Classifier*<sup>1</sup>.

El almacenamiento de datos lo hacemos en MySql y S3-AWS<sup>2</sup>.

Para dar acceso a los usuarios a nuestros servicios a través de Internet, desarrollamos una aplicación con el *web framework* de código abierto Javalin.



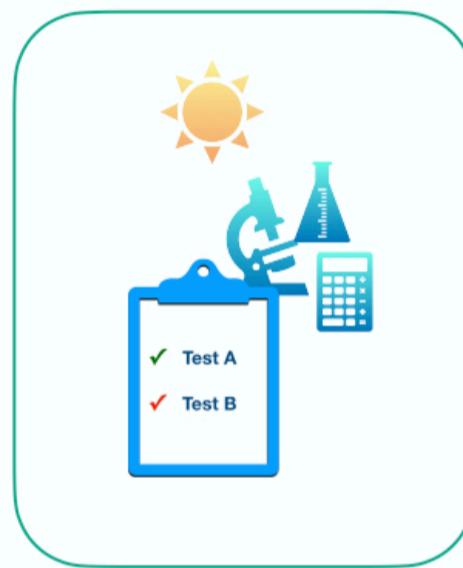
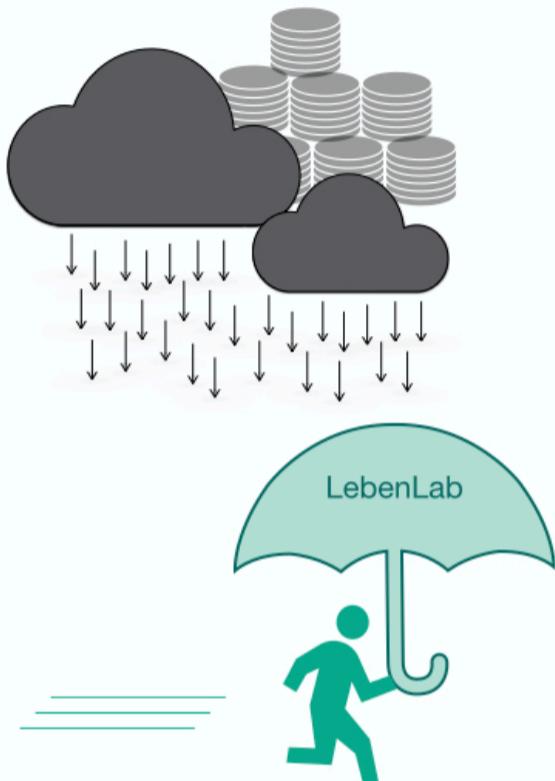
<sup>1</sup> Nos basamos, especialmente, en la exposición que se hace en Daniel Jurafsky & James H. Martin: *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition 3rd edition draft*.

<sup>2</sup> LebenLab instala un entorno de producción propio por cada cliente del laboratorio: ni los recursos de cálculo ni de almacenamiento de datos son compartidos entre clientes. Los entornos actuales están instalados en Amazon Web Services, regiones de París, Frankfurt e Irlanda, sometidas todas ellas a la legislación de la UE sobre protección de datos.

No obstante, pueden instalarse entornos, en las mismas condiciones anteriores, en España, en la propia red de servidores de Leben, para aquellos clientes que así lo prefieran.

# Conclusiones

Una solución de análisis de datos más allá del tradicional software de *business intelligence*, sin inversión inicial, sin requerimientos complejos de extracción de datos, a la medida de cualquier empresa, escalable a un número creciente de clientes, ....



Nuestra aproximación al problema y nuestra solución:

1. Nos enfocamos en extraer conocimiento, no información: *just knowledge* frente a *big data*.
2. Minimizamos los requerimientos de datos para el cliente final, hasta reducirlos a una decena de variables.
3. Utilizamos el método experimental como inspiración.
4. Estudiamos, seleccionamos y utilizamos software de código abierto, con gran implantación en el mercado, como la base informática sobre la que construir nuestro propio laboratorio de *marketing*.

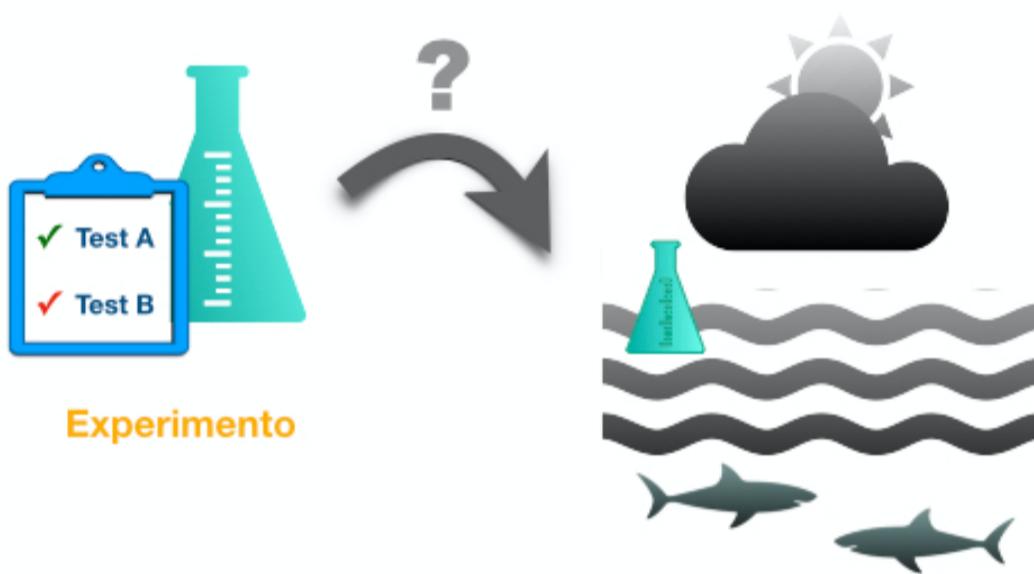
Más allá:

5. Aumentar el rango de hipótesis sobre el que experimentar y la panoplia de técnicas analíticas con que explotar los resultados.

# Predicción

## Punto de partida

¿Podemos predecir resultados de promociones futuras basándonos en los resultados obtenidos en los experimentos? ¿Qué tipo de predicciones podríamos hacer? ¿Qué técnicas de estimación nos convendría utilizar?



Resumamos qué hemos obtenido con nuestro experimento, asumiendo, para simplificar, que sólo hay un producto en promoción:

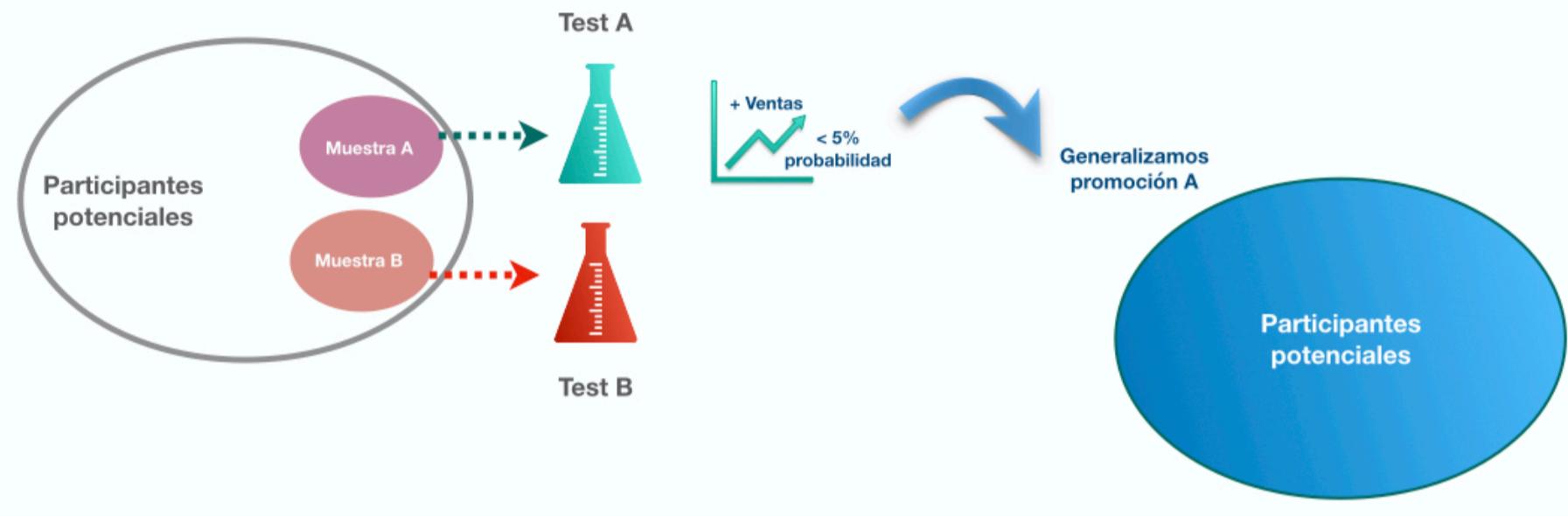
1. Una estimación de venta media por participante en la muestra en cada test.
2. La probabilidad de que la diferencia entre las ventas medias NO se deba a la hipótesis objeto de experimentación, entendiendo por 'hipótesis' la existencia de una relación causa-efecto entre un elemento del diseño de la campaña y las ventas resultantes en ella.

Veamos en un ejemplo qué tipo de predicción podríamos hacer con esos datos:

1. El test A tiene unas ventas superiores a las del test B.
2. La probabilidad de que la diferencia, a favor de la promoción A, *no* se deba a la hipótesis experimentada es del 5%.
3. Podríamos predecir, entonces, que si repitiéramos las dos promociones 100 veces, en 95 de ellas las ventas serían superiores en la promoción A.

4. Con algo menos de rigor, podríamos predecir, también, que si generalizamos el diseño de la promoción A al conjunto de posibles participantes, obtendríamos una cifra de ventas mayor que si generalizamos el diseño de la promoción B.

Este primer análisis nos muestra el primer tipo de predicción que podemos hacer basándonos en nuestro experimento: de los resultados de dos promociones *efectivamente* realizadas podemos inferir cuál de ellas, si la extrapolamos al conjunto de participantes potenciales, tendría mayores ventas.



Este método de predicción requiere siempre la realización previa de aquellos conceptos de campaña (hipótesis) que son susceptibles de generalizarse al conjunto de nuestros clientes (potenciales participantes).

¿Y si pudiéramos hacer predicciones, con fiabilidad, sin pagar el peaje previo de hacer *siempre* un experimento?

# Predicciones sin experimento

Imaginemos que hemos realizado unos 10 experimentos, la mitad de los cuales han registrado un diferencia en las ventas estadísticamente significativa entre sus tests A y B, y que hemos generalizado posteriormente al conjunto de los participantes la hipótesis ganadora: tendríamos un historial de  $(10 \times 2) + 5 = 25$  promociones con sus correspondientes resultados.

Hagamos unas cuentas más sobre los participantes que pueden haber intervenido en esas 25 promociones:

- En cada uno de los 20 tests, digamos que han intervenido, por redondear, unos 100 participantes; en total, tenemos 2.000 datos de ventas correspondientes a unos 20 diseños diferentes de campaña.
- En cada una de las 5 promociones generales, cifremos en 1.500 el número de participantes. En total, tendríamos unos 7.500 datos de venta correspondientes a todas ellas.
- Sumando unos datos y otros, tras 10 experimentos y 5 promociones globales, tendríamos unos 9.500 datos de venta individuales hechas por los participantes en las diferentes acciones.

Analicemos, ahora, con detalle, qué es exactamente lo que tenemos dentro de eso que hemos llamado ‘datos de venta individuales’ y cómo lo podríamos conceptualizar en una ‘respuesta a la campaña’.

## Respuesta a una campaña

Siguiendo con el símil del test de vacunación (capítulo anterior), en el que cada participante inoculado da una respuesta inmunológica, que es registrada como resultado del test, en nuestras promociones podríamos entender que cada participante da también una *respuesta*, comercial en este caso.

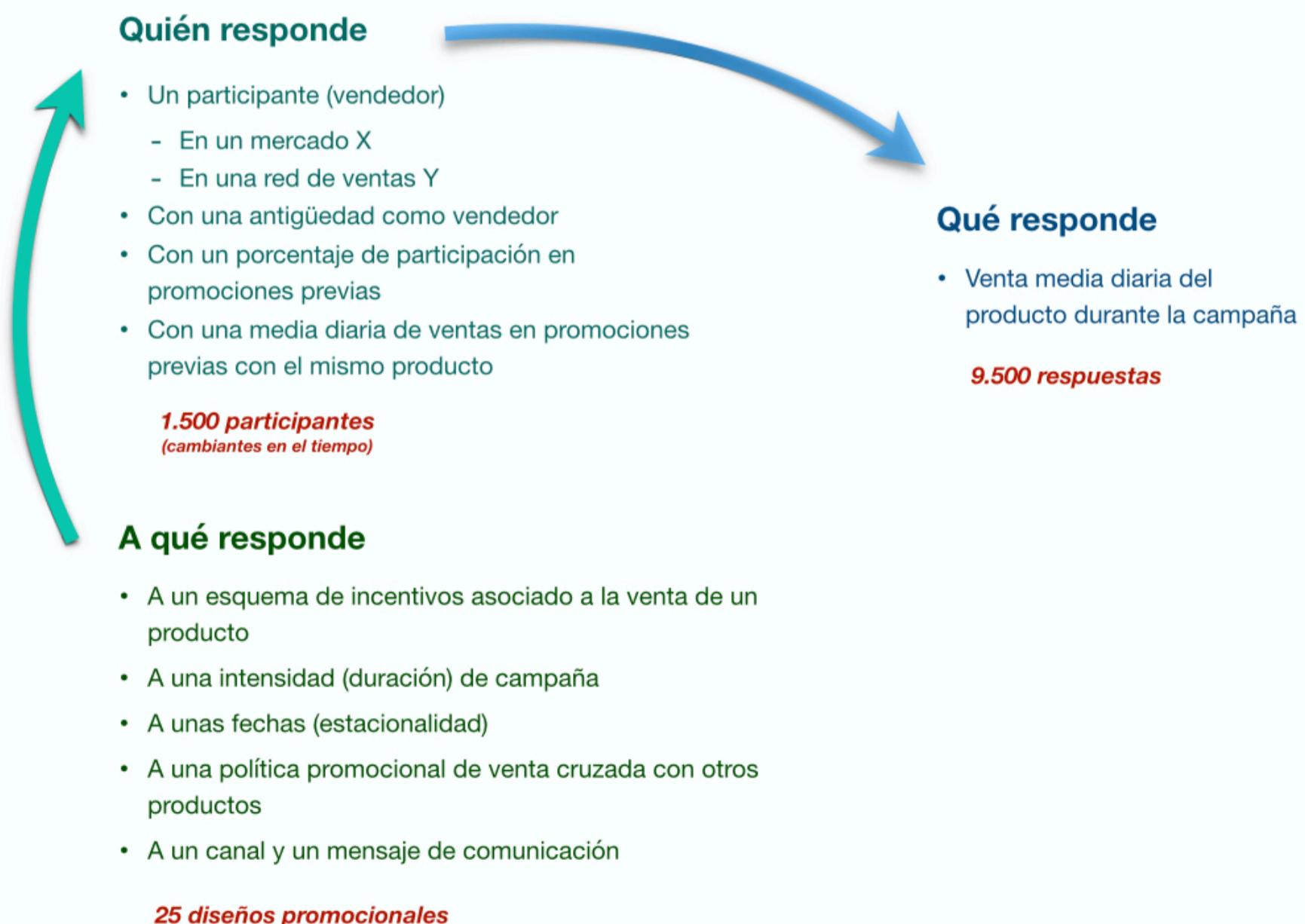
En un test de vacunación, podemos asumir que nos interesaría saber qué compuesto se administra, en cuántas dosis, durante qué intervalo de tiempo, en qué fechas, qué otro medicación tomaba concurrentemente el participante, la edad de este, sus antecedentes médicos y, finalmente, su respuesta al fármaco en un intervalo de tiempo prefijado. Estos datos los podríamos agrupar en tres categorías:

1. Qué fármaco y en qué condiciones se administra al paciente: A QUÉ RESPONDE.
2. A quién se administra la vacuna: QUIÉN RESPONDE.
3. Qué respuesta da el paciente: QUÉ RESPONDE.

Pasemos al otro lado del eje imaginario de la comparación: ¿cuáles serían las categorías equivalentes a las tres anteriores en nuestras promociones? Cambiando la nomenclatura, estas tres:

1. A QUÉ RESPONDE: qué productos en campaña, con qué esquema de incentivos, con qué duración, en qué fechas, con qué canal y mensaje de comunicación.
2. QUIÉN RESPONDE: qué participantes se seleccionan, con qué antigüedad, con qué historial en promociones previas.
3. QUÉ RESPONDE: cuál es la venta media diaria por participante y producto en campaña.

Esquematizando y recuperando los cálculos que hicimos más arriba, este sería nuestro conocimiento tras 10 experimentos y 5 promociones generales *efectivamente realizados*:

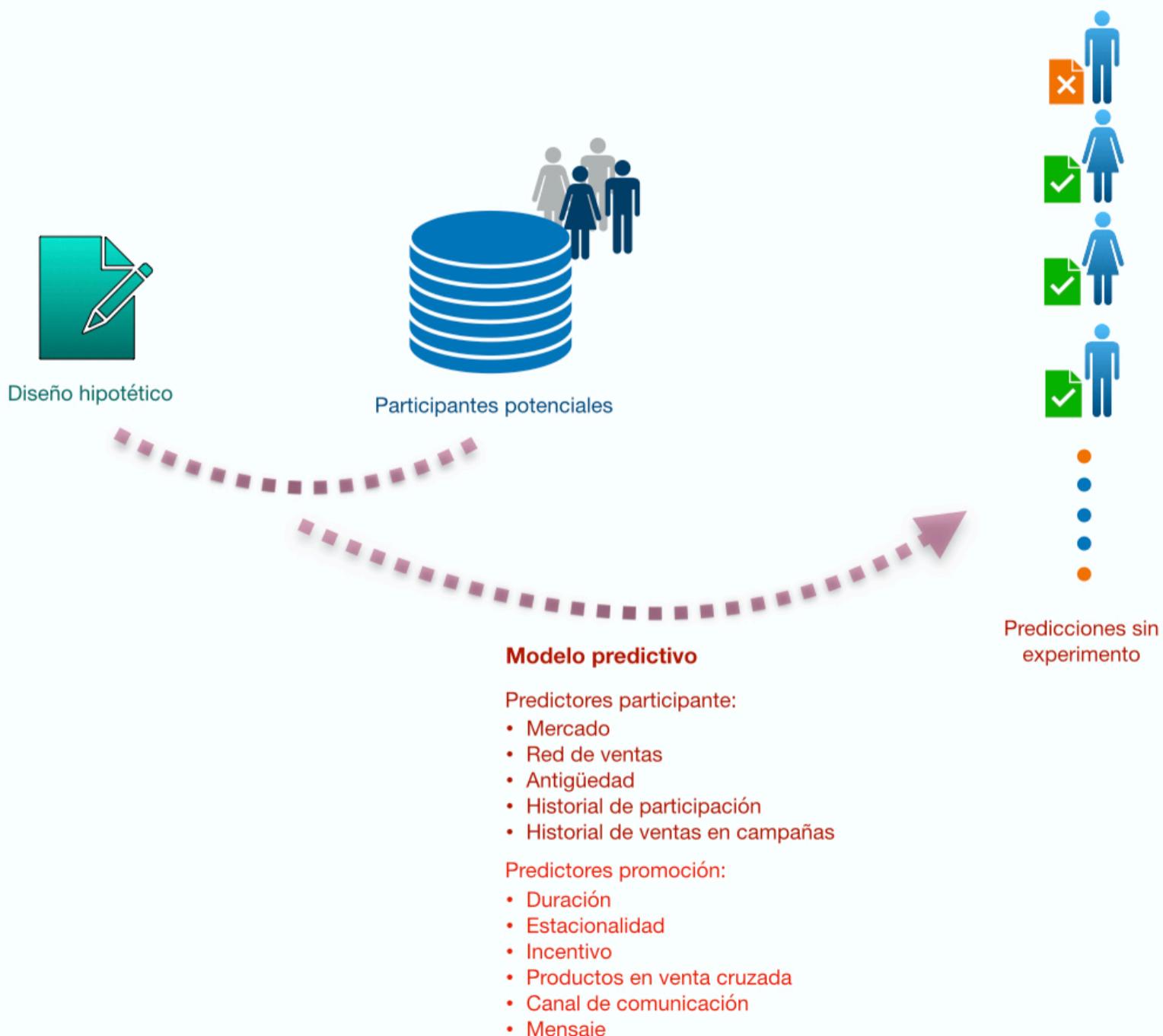


# Predicción de respuestas

¿En qué punto nos encontramos de nuestra indagación?

- Tenemos un método de predicción de las *respuestas* a una campaña, basado en la experimentación previa de dos posibles diseños promocionales que difieren en alguno de sus elementos.
- Buscamos, además, un método para predecir *respuestas*, basado en nuestra experiencia histórica, sin tener que realizar necesariamente un experimento previo.
- Siguiendo nuestro esquema a-qué-responde -> quién responde -> qué responde, el nuevo método se basaría en diseñar *sobre el papel* una nueva promoción y predecir qué responden quienes se encuentren en ese momento, con sus momentáneas características, en nuestro repositorio *real* de participantes.

Ese es el enfoque que implementamos en LebenLab.



Mostramos a continuación tres pantallas de la versión actual de LebenLab. La primera contiene un formulario para definir una promoción *hipotética* que no va a experimentarse realmente; tan sólo van a estimarse sus resultados probables.

**SIMULACIÓN DE RESULTADOS DE UNA CAMPAÑA**



**DATOS SIMULACIÓN**

Nombre  
promo\_simulada\_marzo\_1  
Entre 5 y 25 caracteres: letras, números, guión bajo; sin espacios.

Fecha de Inicio  
01-03-2021  
Formato de fecha dd-mm-aaaa

Fecha de Fin  
10-03-2021  
Formato de fecha dd-mm-aaaa

Mercados  
Selección al menos uno  
España  
Portugal  
Andorra

Conceptos  
Red\_ventas\_1  
Red\_ventas\_2  
Red\_ventas\_3  
Red\_ventas\_4

Productos  
Selección hasta 3  
Sistemas de frenado  
Iluminación  
Encendido

Incentivo  
Descuento en metálico

Comunicación  
email

Texto comunicación  
Texto para mensaje de comunicación por correo electrónico, promoción simulada.  
Máximo 500 caracteres.

**Selección de participantes**  
• Mercado  
• Red de ventas

**Incentivo**

**Productos en campaña**

**Canal de comunicación**

**Mensaje promocional**

**SIMULAR**

La segunda pantalla muestra los resultados estimados para la promoción: venta media diaria por producto y participante.

Reseñamos algunos aspectos de ella:

1. Los resultados se estiman para *todos* los participantes potenciales que entran en la selección de mercados y redes de venta, y que, además, tienen un historial previo de ventas en el producto cuyo resultado se estima.
2. A diferencia de lo que hacemos en un experimento, la predicción no se hace para una muestra, sino para el conjunto de participantes que se ajustan a las dos condiciones descritas en el punto anterior.
3. Si no hubiera historial previo de ventas para, al menos, un participante en alguno de los productos, no estimaríamos respuesta para ese producto.

**Leben** EXPERIMENTOS ▾ SIMULACIONES FICHEROS ▾

**RESULTADOS GLOBALES POR PRODUCTO EN CAMPAÑA**

**promo\_simulada\_marzo\_1**  
2021-03-01 - 2021-03-10  
Incentivo: **dto\_dinero**  
Duración: **10 días**  
Comunicación: **email**  
Texto:  
*Descuento en metálico del 25% en compras de equipos iluminación delantera y anti-niebla.*

Nº participantes seleccionados: **1000** → Número de participantes potenciales que se ajustan a la selección de la campaña

Mercados Conceptos

PG1\_1

Número de respuestas simuladas: **1000** → Número de respuestas estimadas - producto\_1  
Venta media diaria: **10,91** → Venta media por participante - producto\_1

PG1\_2

Número de respuestas simuladas: **1000** → Número de respuestas estimadas - producto\_2  
Venta media diaria: **16,19** → Venta media por participante - producto\_2

La tercera pantalla realiza un análisis muy similar al que hacíamos en los experimentos, pero esta vez comparando la respuesta *esperada* a la respuesta *estimada*, en lugar de hacerlo con la respuesta real o experimentada. Recordamos ambos conceptos:

1. La respuesta esperada la calculamos multiplicando dos variables históricas:

- La venta media diaria hecha por un participante en las promociones anteriores
- La probabilidad de que un participante venda en la campaña actual, estimada a partir del porcentaje de promociones anteriores en las que el participante efectivamente vendió.

2. La respuesta estimada la calculamos con el algoritmo *random forest* para *regression trees* (árboles con nodos con variables continuas). Véase, más abajo, la sección ‘Instrumental’.

Los participantes por debajo de la flecha diagonal son aquellos cuyas ventas estimadas serían superiores a las ventas esperadas según su historial en campañas anteriores; es decir, aquellos que encuentran esta posible propuesta de valor promocional más atractiva que la ‘media’ de las propuestas de campañas previas. En este ejemplo, nuestra predicción nos dice que son los participantes en el cluster 2, en azul, quienes claramente valoran más la nueva propuesta respecto a otras que se les haya hecho anteriormente.



# Predicciones con resultados históricos

Imaginemos que una empresa, con un solape parcial o total en las fechas, monta dos promociones independientes de un mismo producto, e incluye en las dos a un mismo participante. ¿A cuál de las dos campañas imputaríamos el aumento observado en sus ventas?

Una posible respuesta sería atribuir las ventas al efecto combinado de las dos campañas: a la combinación de dos incentivos, de dos duraciones, de dos cestas de productos en venta cruzada, de dos canales de comunicación, de dos mensajes, ... Es fácil ver cómo la nitidez de establecer una relación causa -> efecto en un experimento, controlando los valores de diseño de las dos promociones que lo definen, se empieza a emborronar. Tampoco es fácil inferir consecuencias operativas del análisis de los resultados: si quiero repetir una de las campañas, ¿qué puedo esperar: la mitad de lo obtenido por el efecto combinado de las dos campañas? ¿Cuál de las dos comunicaciones empleo?

Para evitar situaciones como la descrita anteriormente, el instrumental de LebenLab impone unas restricciones en la definición de un experimento:

1. No puede haber otro experimento en vigor con el mismo producto y con un mismo participante.
2. Las dos promociones de un mismo experimento han de compartir forzosamente, al menos, un mismo producto, de lo contrario no hay hipótesis que experimentar: el objeto del experimento es cuantificar el efecto de un diseño de campaña, frente a otro alternativo, sobre las ventas de unos *mismos* productos en un mismo período de tiempo.
3. Sólo puede experimentarse con productos, mercados, redes de venta, incentivos, canales de comunicación y tipos de mensaje que el instrumental predictivo *entiende*, es decir, aquellos que tiene previamente definidos como variables predictoras.

Gracias a estas restricciones, es posible predecir las respuestas de un diseño hipotético de campaña, en las condiciones en que lo hacemos en LebenLab. Si intentáramos hacer lo mismo con resultados de campañas históricas, para que nuestras predicciones tuvieran la misma calidad necesitaríamos que esas campañas, en su día, se hubieran montado con las tres restricciones que imponemos en los experimentos, lo cual es improbable que ocurriera.

No obstante, lo que sí puede hacerse, como hemos mencionado en apartados anteriores, es utilizar las respuestas de campañas generales, no necesariamente experimentos, definidas *dentro* de LebenLab, como historial para realizar predicciones<sup>1</sup>.

---

<sup>1</sup> En la versión actualmente en desarrollo de LebenLab, incluimos la posibilidad de definir una campaña general, no encuadrada en un experimento, y de incorporar sus respuestas a los datos históricos del modelo predictivo.

# Instrumental

El instrumental que necesitamos para predecir se compone, como ocurría en la experimentación, de datos y del software necesario para tratarlos.

## Datos

Los datos que utilizamos para las predicciones son los que genera endógenamente la actividad de experimentar, tanto en los propios experimentos, como en las posteriores generalizaciones que podamos hacer de ellos.

Diferenciamos los dos casos que hemos tratado en este capítulo.

1. La predicción es la mera extrapolación del resultado de uno de los dos tests de un experimento.

Para extraer el resultado basta con multiplicar únicamente dos datos: la venta media por participante obtenida en el test y el número total de participantes potenciales que se ajustan a la selección utilizada en él.

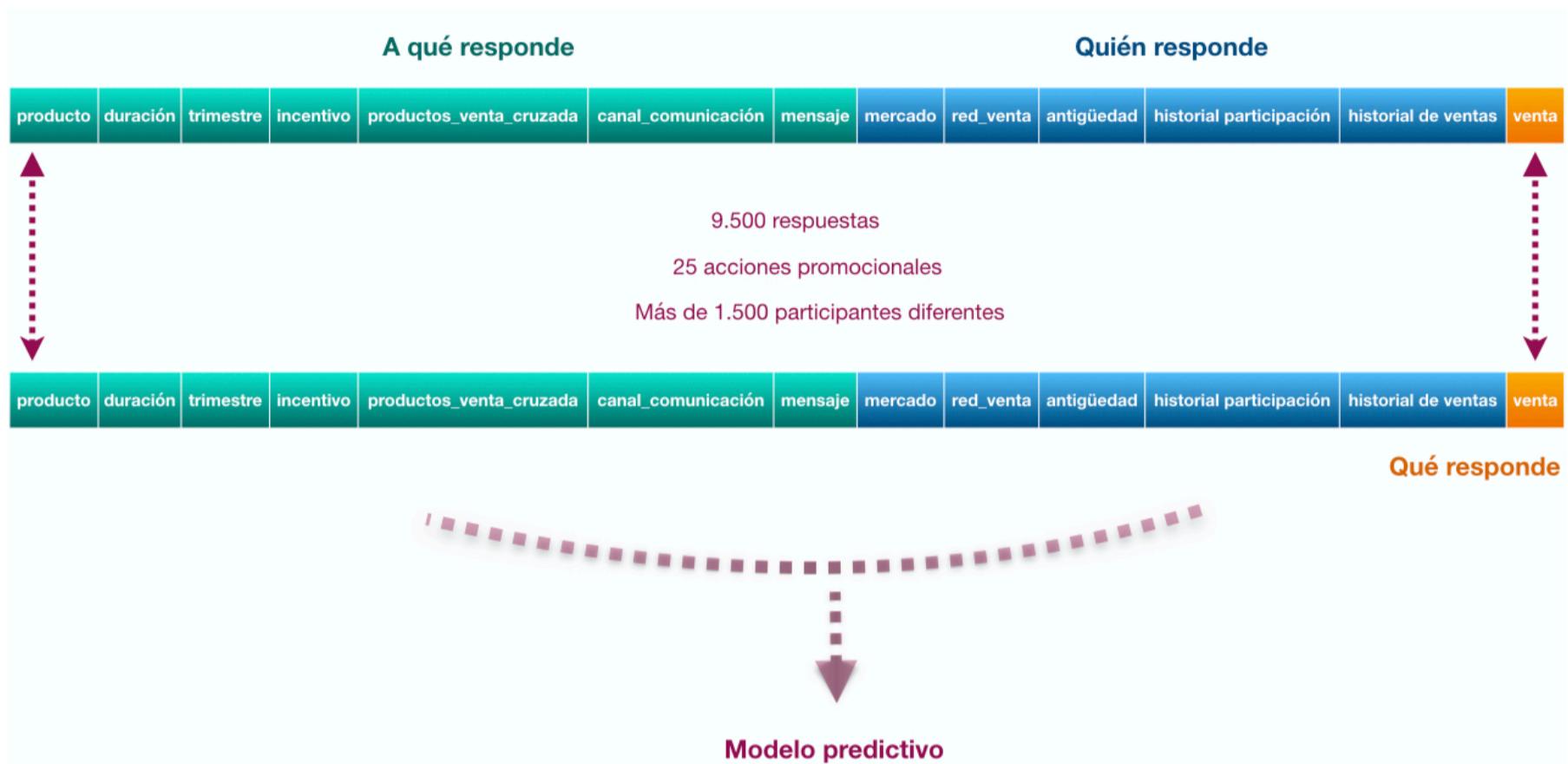
Por ejemplo: realizamos un test B con una muestra de 75 participantes del mercado M\_1 y de la red de ventas RV\_7, donde obtenemos una venta media de 15 unidades. Si el número de participantes potenciales en el mercado M\_1 pertenecientes a la red de venta RV\_7 es de 1.100, la predicción sería de  $1.100 \times 15$  unidades para una campaña general con las mismas características que el test B.

2. La predicción se basa en resultados históricos de experimentos y campañas generales.

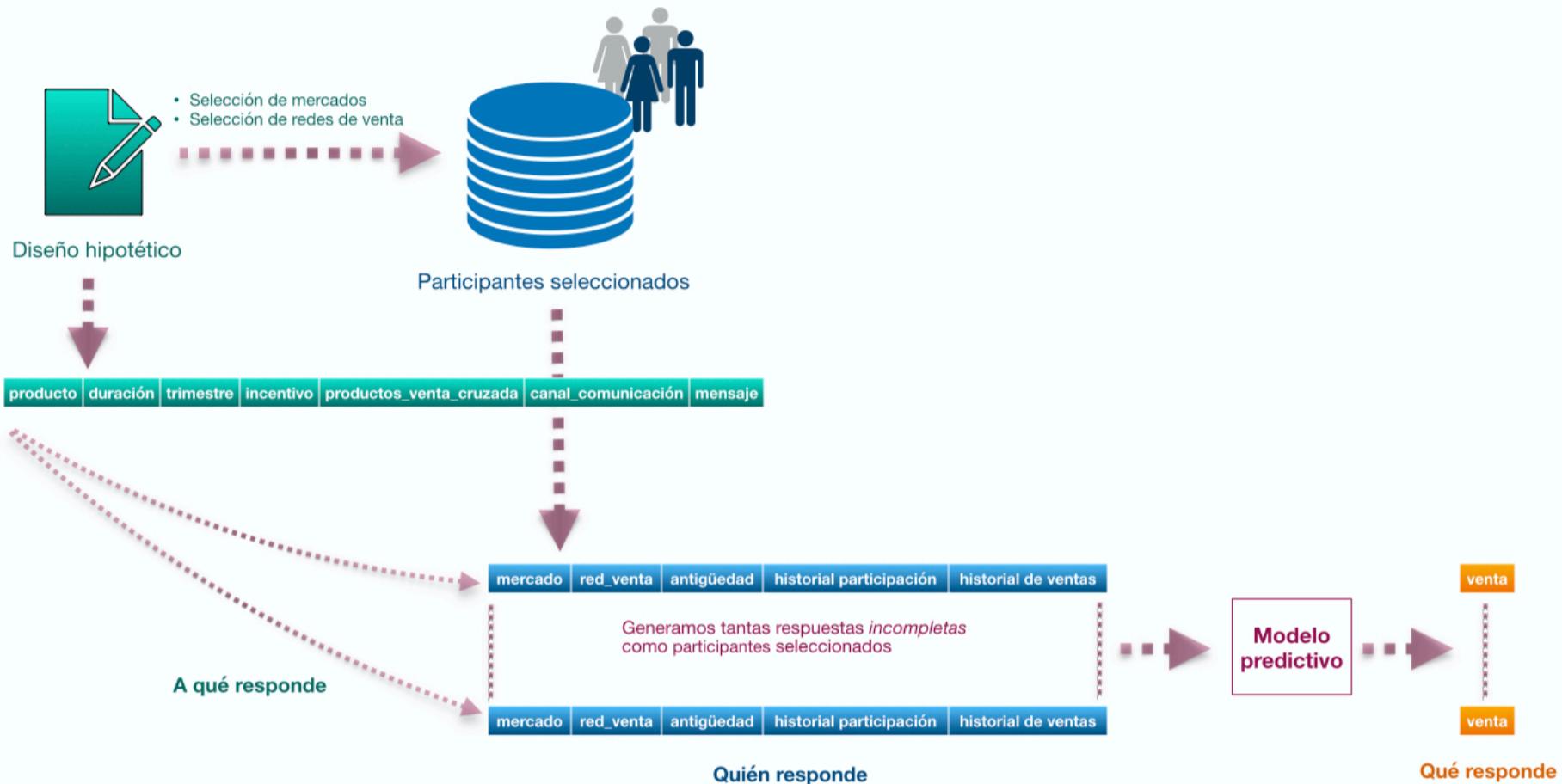
En este segundo caso, necesitamos dos conjuntos de datos:

- Los datos necesarios para construir el modelo predictivo, que sólo necesitaremos cuando construyamos por vez primera el modelo o cuando lo actualicemos, no cada vez que hagamos una simulación.
- Los datos con que alimentar el modelo, una vez construido, y realizar la predicción posterior para ese preciso conjunto de datos.

Reutilizando como ejemplo los cálculos que hemos ido haciendo a lo largo de este capítulo, para construir el modelo predictivo necesitamos el historial promocional compuesto por las 9.500 respuestas obtenidas en las 25 campañas, en las que han participado 1.500 participantes (no necesariamente los mismos).



Para realizar un predicción para un diseño promocional hipotético, necesitamos fabricar, primero, tantas *respuestas incompletas* como participantes seleccionados, y suministrárselas después al modelo, para que las complete con la venta media diaria por producto y participante en campaña. Estos trabajos los realiza LebenLab a partir de la información recogida en el formulario ‘Simulación de datos de una campaña’ (más arriba).

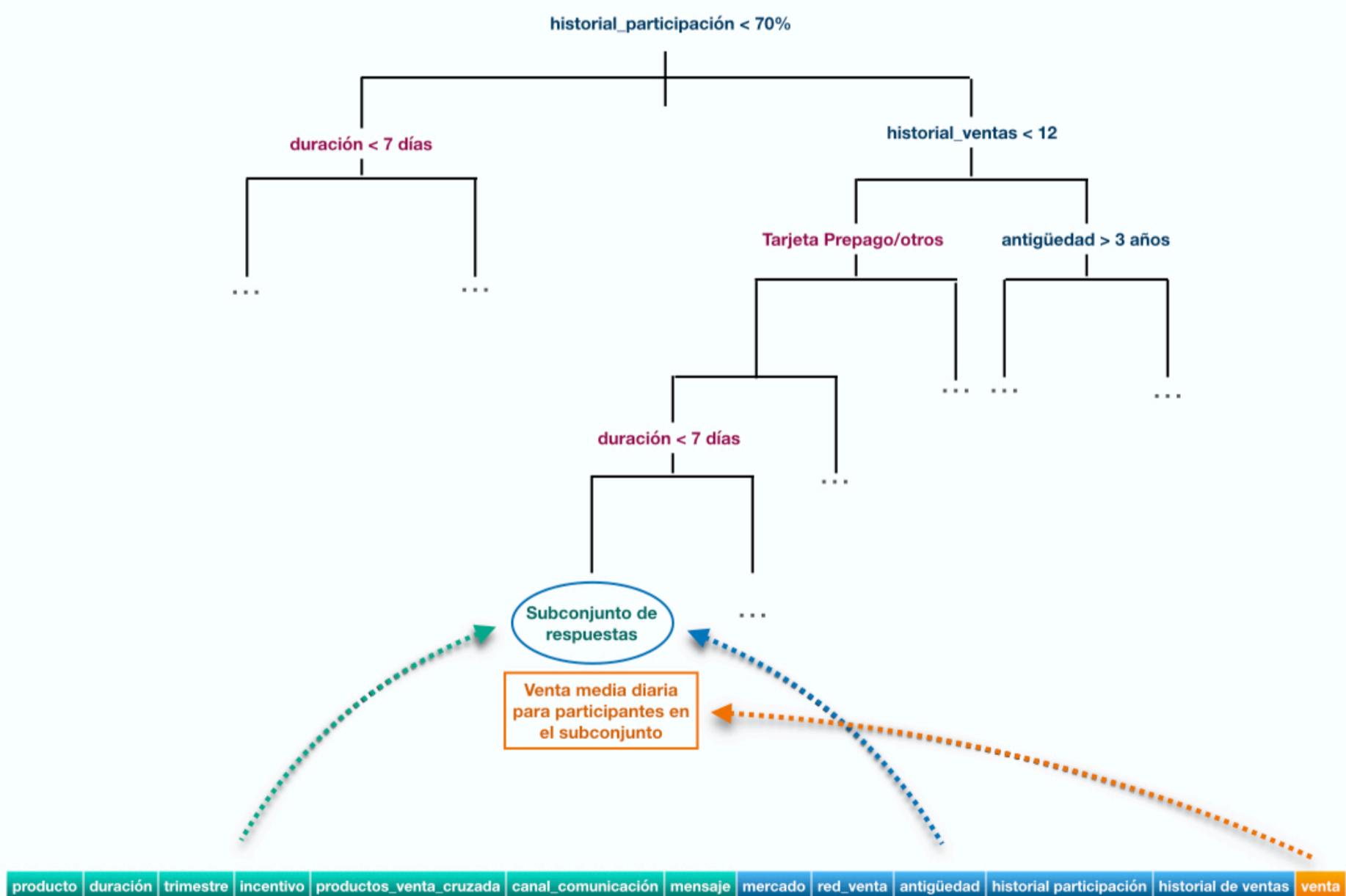


# Software

Además del software que utilizamos en la experimentación, descrito más arriba, necesitábamos un componente adicional para la construcción del modelo predictivo y las posteriores estimaciones.

Para el cálculo del modelo y de las predicciones que con él se realizan, decidimos utilizar la implementación del algoritmo *random forest* en la librería SMILE (Statistical Machine Intelligence and Learning Engine). No vamos a hacer aquí una descripción detallada de él<sup>1</sup>; tan sólo, con la ayuda de un par de gráficos, esbozaremos cómo encaja con las ideas que hemos ido desgranando a lo largo del capítulo.

Modelo predictivo basado en respuestas históricas



<sup>1</sup> Nos ha sido muy útil la descripción en Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani: *An Introduction to Statistical Learning*. Springer, 2017

Los *random forests* se componen de árboles (*regression trees*), como el que hemos dibujado más arriba. Un árbol clasifica las respuestas extendidas (a-qué-responde/quién-responde/qué-responde) de nuestro historial de campañas en nodos, que son las terminaciones del árbol. Cada nodo queda definido por el subconjunto de respuestas que contiene. Haciendo un promedio del valor *qué-responde* de todas las respuestas del nodo obtenemos la venta media diaria por participante en el nodo.

Por ejemplo, en el único nodo que hemos dibujado en el gráfico, están contenidas las respuestas con las siguientes características:

- A qué responde: a una campaña con una duración inferior a siete días, con una tarjeta prepago como incentivo.
- Quién responde: un participante que ha intervenido en un 70% de las promociones anteriores, con un historial de venta media diaria inferior a 12 unidades, perteneciente al mercado ES (España).
- Qué responde: el promedio simple de las cifras de ventas medias diarias por participante consignadas en las respuestas clasificadas dentro del nodo. Por ejemplo, si hubiera 5 respuestas en el nodo, con las cifras de ventas 12, 11, 10, 9 y 8, el nodo tendría asignado, como única venta media por participante, el valor 10.

Todas las respuestas históricas estarían clasificadas en uno y sólo uno de los nodos.

Hagamos ahora una predicción para un diseño promocional hipotético:

1. Cumplimentamos el formulario ‘Simulación de datos de una campaña’ y lo procesamos.
2. LebenLab genera tantas respuestas incompletas (sin estimación de ventas) como participantes se hayan seleccionado.
3. Cada respuesta generada es ubicada en un nodo del árbol. En nuestro ejemplo, si la respuesta tiene una duración inferior a 7 días, una tarjeta prepago como incentivo y un participante del mercado español que ha intervenido en un 70% de las promociones anteriores, con una venta media diaria inferior a 12, entonces es ubicada en el nodo del gráfico que hemos dibujado más abajo.
4. Completamos las respuestas. Para todas aquellas ubicadas en el mismo nodo estimamos la misma venta diaria: el promedio calculado en el modelo. En nuestro ejemplo, a todas las respuestas con los mismos valores que los descritos en el párrafo anterior, se les asignaría un valor de 10.
5. Una vez estimadas las respuestas individuales de cada participante podríamos estimar la respuesta para el conjunto de la campaña.

## Predicción

producto	duración	trimestre	incentivo	productos_venta_cruzada	canal_comunicación	mensaje	mercado	red_venta	antigüedad	historial_participación	historial_de_ventas
----------	----------	-----------	-----------	-------------------------	--------------------	---------	---------	-----------	------------	-------------------------	---------------------

A qué responde

Quién responde

¿A qué nodo pertenece esta respuesta *incompleta*?

historial\_participación < 70%

historial\_ventas < 12

Tarjeta Prepago/otros

Mercado ES

duración < 7 días

Subconjunto de respuestas

Venta media diaria para participantes en el subconjunto

Predicción

producto	duración	trimestre	incentivo	productos_venta_cruzada	canal_comunicación	mensaje	mercado	red_venta	antigüedad	historial_participación	historial_de_ventas
----------	----------	-----------	-----------	-------------------------	--------------------	---------	---------	-----------	------------	-------------------------	---------------------

venta

# Conclusiones

Experimentos promocionales con muestras de participantes, campañas generales, una metodología capaz de unificar todos los resultados, todas las *respuestas*, en un único modelo predictivo fácilmente interpretable en términos del negocio, sin requerir necesariamente un experimento previo ...



Nuestra aproximación al problema y nuestra solución:

1. Definimos como unidad de análisis la respuesta extendida, aprovechando su similitud con la reacción a un tratamiento médico experimental: a-qué-responde -> quién-responde -> qué responde.
2. Construimos nuestro modelo sobre un algoritmo fácilmente interpretable: la clasificación sucesiva de las repuestas en grupos binarios, que se subdividen a su vez, según los valores de las variables de negocio que más diferencian las respuestas que van quedando en los grupos resultantes.
3. Utilizamos el modelo construido sobre respuestas históricas reales para predecir las respuestas individuales de participantes potenciales, en diseños promocionales hipotéticos.
4. Aplicamos la misma herramienta que en la experimentación real, para valorar la perturbación de un diseño hipotético sobre tres grandes *clusters* de participantes.

Más allá:

5. Ampliar la versatilidad del concepto actual de respuesta extendida, extendiéndolo a otros tipos de configuraciones promocionales.
6. Explorar otras construcciones del modelo con diferentes algoritmos.

# Glosario

- **Campaña general:** es una promoción simple que no forma parte de un experimento. La utilidad de definir una campaña general dentro de LebenLab radica en que se le aplican las mismas restricciones que a las promociones dentro de un experimento y que, además, sus resultados se capturan al mismo tiempo, sin ningún trámite adicional, que los de estas otras. Esta captura permite tener en cuenta el historial de respuestas de las campañas generales en la predicción de resultados de diseños promocionales hipotéticos.
- **Experimento:** es la realización simultánea de dos promociones en un test A/B. Las respuestas a ambas se incorporan al historial de respuestas del modelo predictivo.
- **Respuesta esperada:** es una estimación de la venta<sup>1</sup> por participante, que calculamos a partir de dos datos históricos:
  1. La venta media diaria de un producto hecha por un participante en las promociones anteriores en las que participó.
  2. El porcentaje de promociones anteriores en las que el participante efectivamente vendió.

La multiplicación de ambos datos arroja lo que podríamos entender como la respuesta probable de un participante, en una nueva campaña, a la vista únicamente de su conducta pasada.

- **Respuesta estimada:** es una estimación de la venta por participante, que calculamos a partir de su respuesta extendida. Difiere de la respuesta esperada en dos aspectos:
  1. Las variables que se tienen en cuenta para hacer la estimación, además del historial de ventas y de participación que se utilizan igualmente en la respuesta esperada.
  2. La utilización del historial de ventas y de participación de otros participantes en la estimación individual para cada uno de ellos. Es posible, incluso, calcular una respuesta estimada para un participante que no entró nunca en campaña, basándonos en el historial de otros que sí lo han hecho.

---

<sup>1</sup> LebenLab sirve tanto para campañas en las que el participante vende, como para aquellas en las que el objetivo es que compre. Aunque todos los ejemplos que hemos incluido en este libro se refieren a participantes que actúan como vendedores de quien realiza el experimento, bastaría con modificar la nomenclatura para que puedan realizarse experimentos en situaciones donde el participante es comprador.

- **Respuesta extendida:** es el conjunto de variables que definen un contexto (diseño promocional + participante) y la cifra individual de ventas que en él se obtiene.

En las páginas precedentes, la respuesta extendida la hemos definido con tres grupos de variables:

1. A qué se responde (el contexto promocional).
2. Quién responde (el contexto individual).
3. Qué responde (la cifra individual de ventas).

- **Test A/B:** es el conjunto de los siguiente elementos:

1. Una selección de participantes definida por los valores de dos variables (mercado, red de ventas o similar).
2. Dos muestras aleatorias sin ningún elemento en común, extraídas de la selección de participantes.
3. Dos conjuntos de productos con intersección.
4. Un plazo definido por dos fechas.
5. Dos diseños promocionales que difieren en un elemento.
6. Un resultado asociado a cada muestra. En nuestros experimentos, la venta media diaria por participante.
7. Un test de significación estadística. En nuestro caso, un test de diferencia de medias con una distribución de t-student.