****

**Universidad Carlos III de Madrid**

## **Proyecto: AYUDAS A LA PREDICCIÓN EN LA TOMA DE DECISIONES DE CLIENTE (BIG DATA)**

## **Documentación correspondiente al Hito I. Noviembre de 2015.**

**Documento de diseño de los modelos de predicción, caracterización y clasificación**

Carlos Bousoño Calzón

Aníbal R. Figueiras Vidal

Josué Bustarviejo Muñoz

# Introducción

Este documento responde al correspondiente título del Hito I. En él se establecen las consideraciones relativas a los modelos para la caracterización de los experimentos que sirven de base a la posterior inferencia, objeto de futuros entregables.

Comencemos por recordar el objetivo de este proyecto: se trata de explorar la posibilidad de extraer de las trazas de los *usuarios* en *juegos simples* información relevante sobre su comportamiento como *consumidores*. Es por tanto una *prueba de concepto* para la que los materiales desarrollados son simplemente una herramienta hacia el objetivo fundamental.

La idea inicial, tal y como se colige de la *Memoria Técnica* al contrato, consistía en analizar la correspondencia entre los juegos y las ofertas de Telefónica considerando la posibilidad de ofrecer dichas ofertas por los puntos obtenidos en los juegos. La dependencia de este esquema en un alto número de participantes, la limitación procedimental de acceso a los mismos en el colectivo de pruebas de Telefónica, y la dificultad de ofrecer ofertas reales como pagos, motivó modificaciones en la herramienta para adaptarla a este entorno más realista dentro de los recursos del proyecto. Estas modificaciones se convierten así en el principal objetivo de este documento.

El diseño definitivo de la herramienta de captación de datos consiste pues en englobar en un mismo entorno los juegos, siendo aquí fieles a la especificación inicial, con unas *ofertas ficticias* que podrán comprarse con los puntos que se obtengan de juagar. Este diseño define un *meta-juego*, que llamaremos ***Movijuego***, consistente en ganar puntos en los juegos elementales para *comprar* la *oferta* deseada. Para que esta modificación del esquema inicial sea consistente con los objetivos del proyecto, debe satisfacer unos requisitos mínimos como se discute en el Apdo.1. Por otra parte, se puede aprovechar esta circunstancia para diseñar unas *ofertas* *ad hoc* que brinden más información mercadotécnica [1] de la inicialmente prevista, reforzando de esta manera la aplicabilidad de este proyecto y la importancia estratégica silente de sus potenciales extensiones.

# Diseño de incentivos: el paradigma de la revelación de preferencias

Si la idea es inferir comportamientos de consumo en base a las trazas en juegos, es imprescindible la alineación de los mismos. Desde un punto de vista académico, este problema se conoce en el ámbito de la *Teoría de Juegos* como *principio de revelación de preferencias* [2,3]. Según este principio, y expresándolo libremente, si queremos *extraer* información al jugador sobre *consumo*, debemos *premiar o incentivar* al mismo con algo relacionado con *ese consumo* y las *opciones* que escoja al jugar.

La aplicación de este principio al problema que nos ocupa significa si queremos que el jugador *revele* qué ofertas está dispuesto a elegir, debería premiársele con esas mismas ofertas. Ésta era la propuesta original.

Pero si ello no es posible, la segunda versión podría ser que *compre* (de manera virtual, con los puntos obtenidos en el juego) una *oferta* sabiendo que se le va a recompensar con algo relacionado con la misma. Esta versión permite desligar la información que pretendemos extraer del premio que es posible ofrecer. En todo caso, la conexión debería estar clara; es decir, el premio debería ser una oferta real o descuento sobre la misma de parecidas características a la *oferta comprada*. Sería incluso posible, si las limitaciones de recursos así lo exigen, vincularlo a una rifa sobre esa misma oferta o sobre una ventaja adicional en su compra real.

En el caso más restrictivo de que el premio no esté en absoluto vinculado a la *oferta* que revele el jugador (por ejemplo, que sólo pueda darse como recompensa un objeto, similar a un i-Pad, sin vinculación aparente a la *oferta comprada* en el juego), la recomendación es que la información que se ofrece al jugador sea lo menos detallada posible para mantener, en la medida de lo posible, la vinculación entre la *oferta* ficticia y el premio real.

# Diseño de juegos

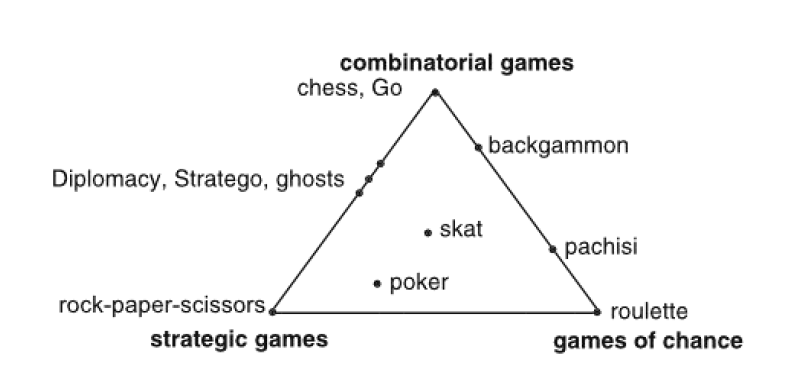
La elección de los juegos básicos sigue el esquema inicialmente propuesto y su justificación la misma que la ofrecida en la Memoria Técnica al contrato: han de ser juegos de sencilla comprensión, para evitar una innecesaria segmentación de los jugadores, con un sistema de recompensas que los haga atractivos a la repetición (sin resultar cansinos) y que revele rasgos no triviales del jugador.

La necesidad de diseñar ofertas ficticias pide un marco de referencia de los aspectos más básicos a fin de que la información que pueda extraerse no se vea limitada *a priori* por un diseño excesivamente limitado. A tal fin se ha buscado un marco conceptual para los juegos y las ofertas lo más abstracto posible para no interferir en los procesos de inferencia de otras fases.

## Un modelo de referencia para la adecuación entre juegos y ofertas

*Comencemos por puntualizar que este marco no pretende en ningún sentido orientar o forzar un modelo para la inferencia: el análisis de los datos no responderá necesariamente al marco aquí establecido, sino que será desarrollado con los objetivos y orientación originalmente propuestos y las restricciones que los datos finalmente obtenidos puedan imponer*.

En ese sentido, y teniendo en cuenta que los juegos básicos ofrecen fundamentalmente una diferencia en la incertidumbre probabilística [4], se propone el siguiente esquema de Bewersdorff [5]:



En este esquema se pueden apreciar tres dimensiones básicas de la incertidumbre:

1. **Incertidumbre debida al azar, IA,** (*games of chance*). Su interpretación es la del simple azar, caracterizado simplemente por una distribución de probabilidad. Un ejemplo simple es el resultado de lanzar un dado.
2. **Incertidumbre estratégica, IE,** (*strategic games*). Esta incertidumbre se asocia a la información que se tiene sobre el juego. Fundamentalmente está aparejada a los juegos estratégicos entre personas, donde la incertidumbre proviene del desconocimiento de lo que el oponente va a hacer, o de la interpretación de las señales que éste pueda ofrecer para una posible colaboración o engaño. Esta incertidumbre es la más relevante en este proyecto.
3. **Incertidumbre combinatoria, IC,** (*combinatorial games*). Corresponde a la incertidumbre debida exclusivamente a la estructura y complejidad del juego. El ejemplo típico es el ajedrez, donde no hay azar (no hay “dados”), no hay incertidumbre estratégica (el oponente no tiene “cartas ocultas”) ya que toda la información posible entre oponentes está en el tablero.

Como nota adicional, se señala que este marco se empleará igualmente para el diseño de la encuesta sobre los juegos que habitualmente usan los jugadores en sus casas.

## Los juegos básicos

Se desarrollan tres juegos: *la caña más larga, adivina el color y elige la caja*. La descripción básica de los mismos (véase también [5]) se expone a continuación junto con sus características según el esquema de Bewersdorff; en el siguiente apartado se estudian sus sistemas de puntuación.

Cada juego tiene dos versiones en las que varía el número de elementos. Esta variación se asocia a un incremento en la incertidumbre combinatoria como se explica en lo que sigue.

### La caña más larga

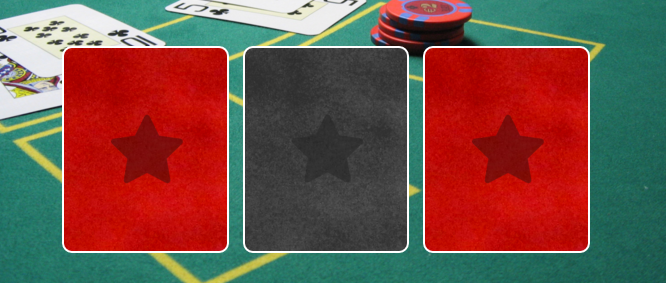


En la versión básica de este juego se muestran tres cañas. Cada una tiene diferente longitud y se gana si se elige la más larga.

En este juego la componente de IE es nula y asumimos que la IC es nula. La única componente de incertidumbre que se muestra claramente es la IA.

La segunda versión añade una caña más. La componente fundamental es la IA igualmente, aunque podemos entender que -de forma absolutamente trivial- se incrementa la IC.

### Adivina el color



En este juego se muestran tres cartas que tienen, respectivamente, dos caras de color rojo; dos caras de color negro; y una cara de color rojo y otra negra. Elegida una carta al azar y enseñando una de sus caras al azar, se pide adivinar el color de la otra cara.

En este juego hay un componente de IA, una componente que podemos asumir nula de IC y una componente esencial de IE. La IE proviene del hecho de que la estructura probabilística del mismo no es trivial y la elección de colores que lleva a una ganancia sistemática no es inmediata, pero se puede aprender de la repetición y observación cuidadosa del mismo.

La segunda versión procede como ésta pero con una carta más. Aquí el incremento de la IC no es trivial y su composición con la IE cambia el carácter del juego.

### Elige la caja



En este tercer juego, se muestran tres cajas de las que sólo una de ellas contiene premio, es decir, conduce a ganar el juego. El jugador empieza eligiendo una de ellas que no se destapa, sino que entre las dos restantes se destapa una que no contiene premio. A continuación, al jugador se le ofrece la posibilidad de destapar la que tiene o cambiar la suya por la que queda sin abrir y destaparla.

Este juego tiene un componte de incertidumbre semejante al de las cartas: una componente IA, asumimos que no tiene IC y su principal componente es la de IE. La estructura de probabilidades en cuantitativamente igual a la de las cartas, este caso de mejor comprensión por parte del jugador, por lo que es posible que al aprendizaje estadístico se una el racional sin necesidad de estudiar estadísticamente sus características.

El incremento en IC en la segunda de las versiones en este caso, es completamente diferente a la de las cartas y su impacto en la dinámica del juego muy distinto.

Nótese que existe, en cuanto a aprendizaje por parte del jugador, una meta-estructura no trivial en el compendio de los tres juegos, ya que el incremento de la componente IC en los tres casos es completamente diferente con lo que el paso de una versión a otra no es extrapolable de un juego a otro y por lo tanto exige, del jugador una consideración conjunta de los tres si quiere extraer el máximo posible de ganancia.

# Diseño de ofertas ficticias

Para el diseño de las ofertas ficticias que el jugador podrá comprar con sus puntos, nos basamos en las mismas dimensiones que para los juegos. El objetivo es proponer el mismo número de ofertas que juegos y que aquéllas tengan una distribución lo más uniforme posible en las tres dimensiones de incertidumbre.

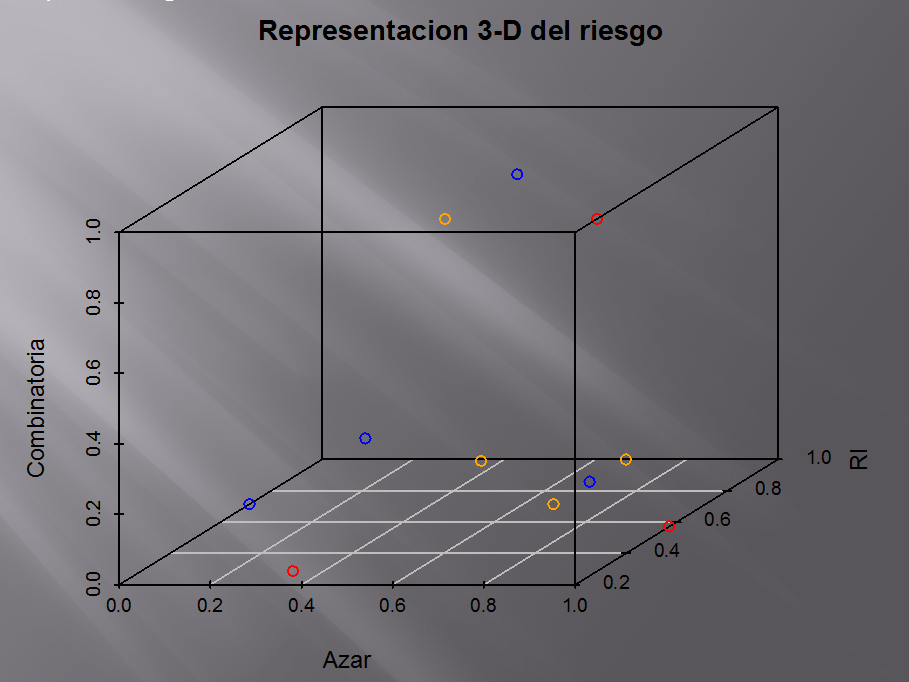
Ya que ahora no se está restringido a las ofertas reales de Telefónica, se aprovecha la ocasión para acercar este proyecto a su interpretación desde el punto de vista del mercado definido por las tres operadoras dominantes: Telefónica, Orange y Vodafone.

**Primero** se exploran las ofertas de estas tres operadoras en la Web, y se clasifican sus características en las dimensiones de incertidumbre del modelo de Bewersdorff, véase apéndice.

**Segundo** se seleccionan aquéllas, independientemente de la compañía, que más se aproximan a una distribución uniforme en el cubo normalizado de las tres dimensiones de incertidumbre.

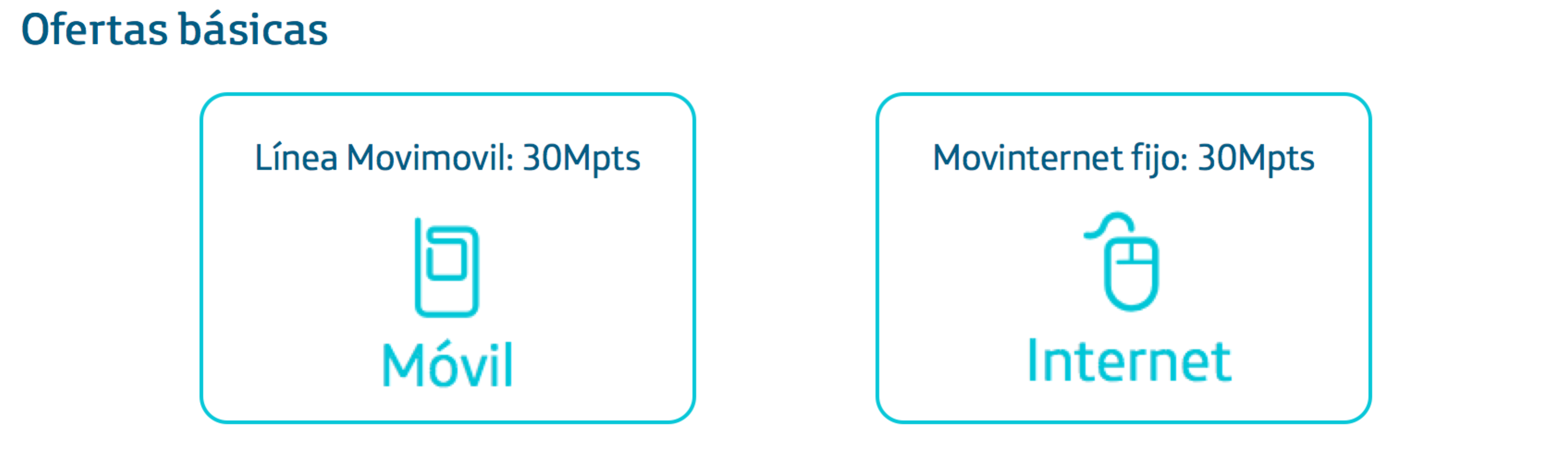
**Tercero** se modifican las ofertas reales en ofertas ficticias añadiendo componentes para que la distribución sea lo más uniforme posible.

El resultado del diseño se puede representar en el cubo como se muestra a continuación.



El listado de ofertas creado a partir de ese análisis es el siguiente:

**Bloque I.**



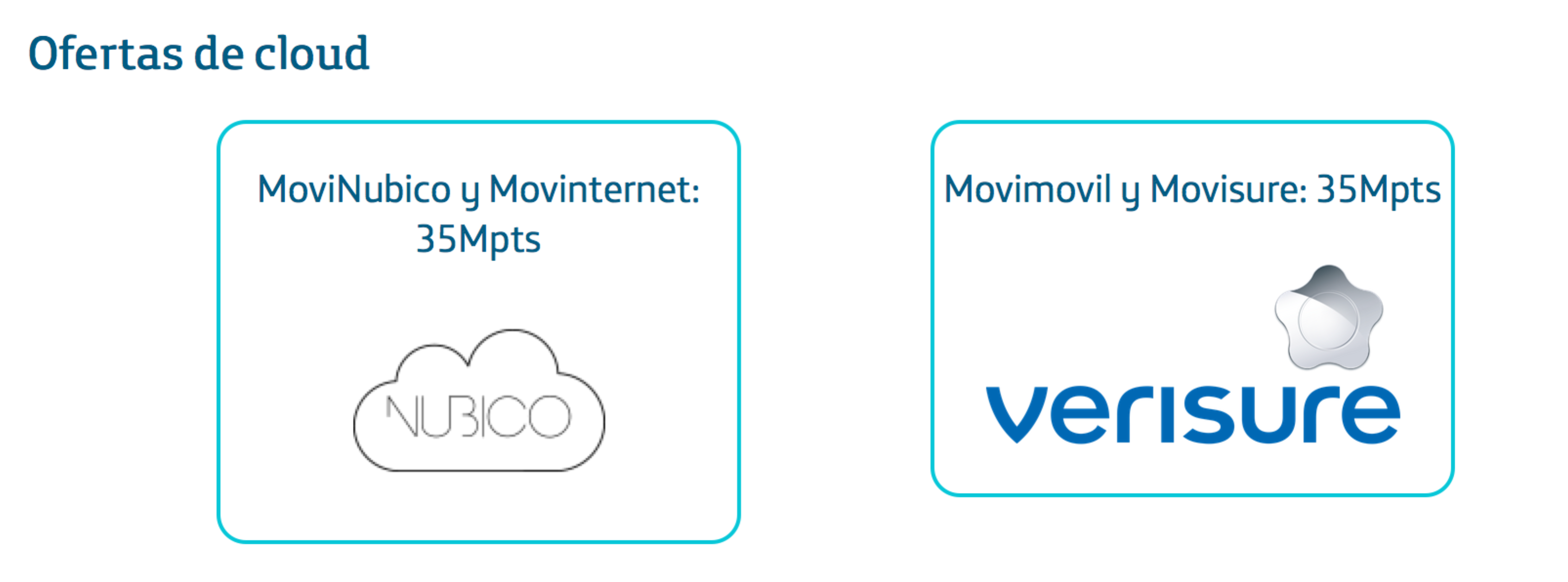
**Línea Movimovil: 30 Movipuntos**

* 200 min.a fijos y móviles nacionales
* 1,5 GB

**Movinternet fijo: 30 Movipuntos**

* Asistencia técnica
* Fibra Óptica 30Mb simétrica (sujeto a cobertura)
* Alta e instalación incluida
* Router Wi-Fi gratis

**Bloque II**



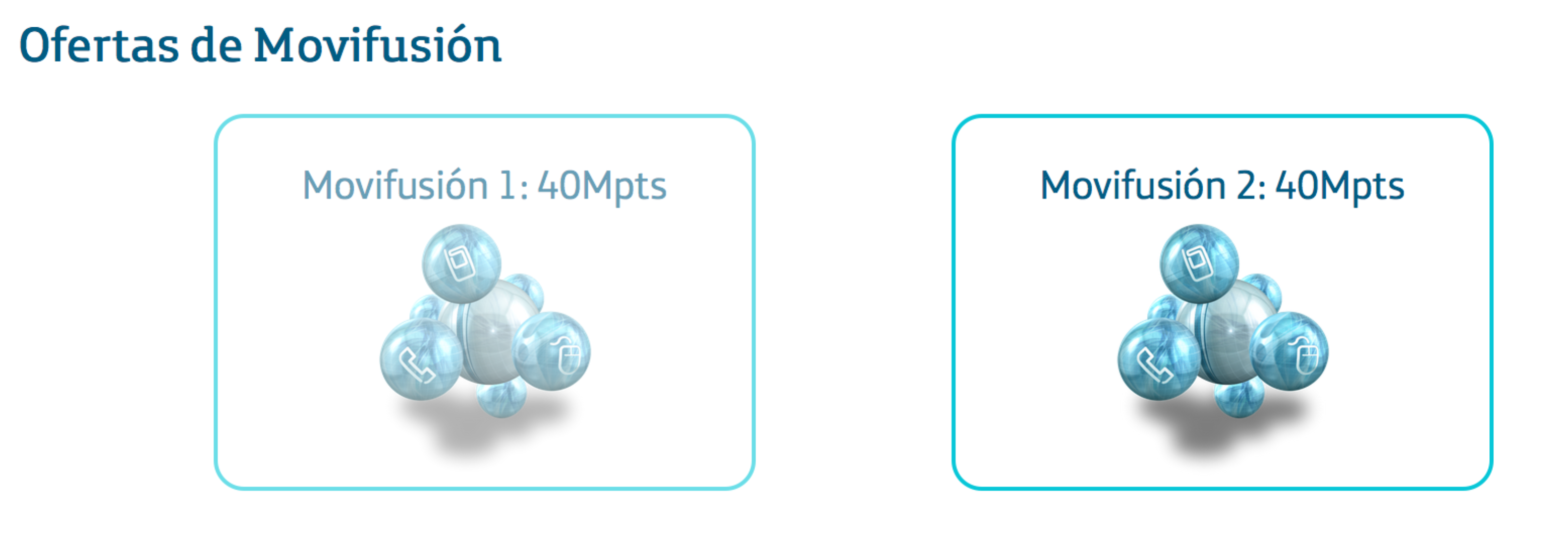
**MoviNubico y Movinternet: 35 Movipuntos**

* Sin compromiso
* Internet 30Mb
* Acceso Cloud a todas las revistas y libros de forma ilimitada
* 5 lectores simultáneos
* Garantía de devolución 20 días

**Movimovil y Movisure: 35 Movipuntos**

* Línea Móvil 4G
* Control de vigilancia en la nube de tu hogar

**Bloque III**

****

**Movifusión 1: 40 Movipuntos**

* Internet 30Mb
* Línea fija
* Nubico
* Garantía de devolución 20 días

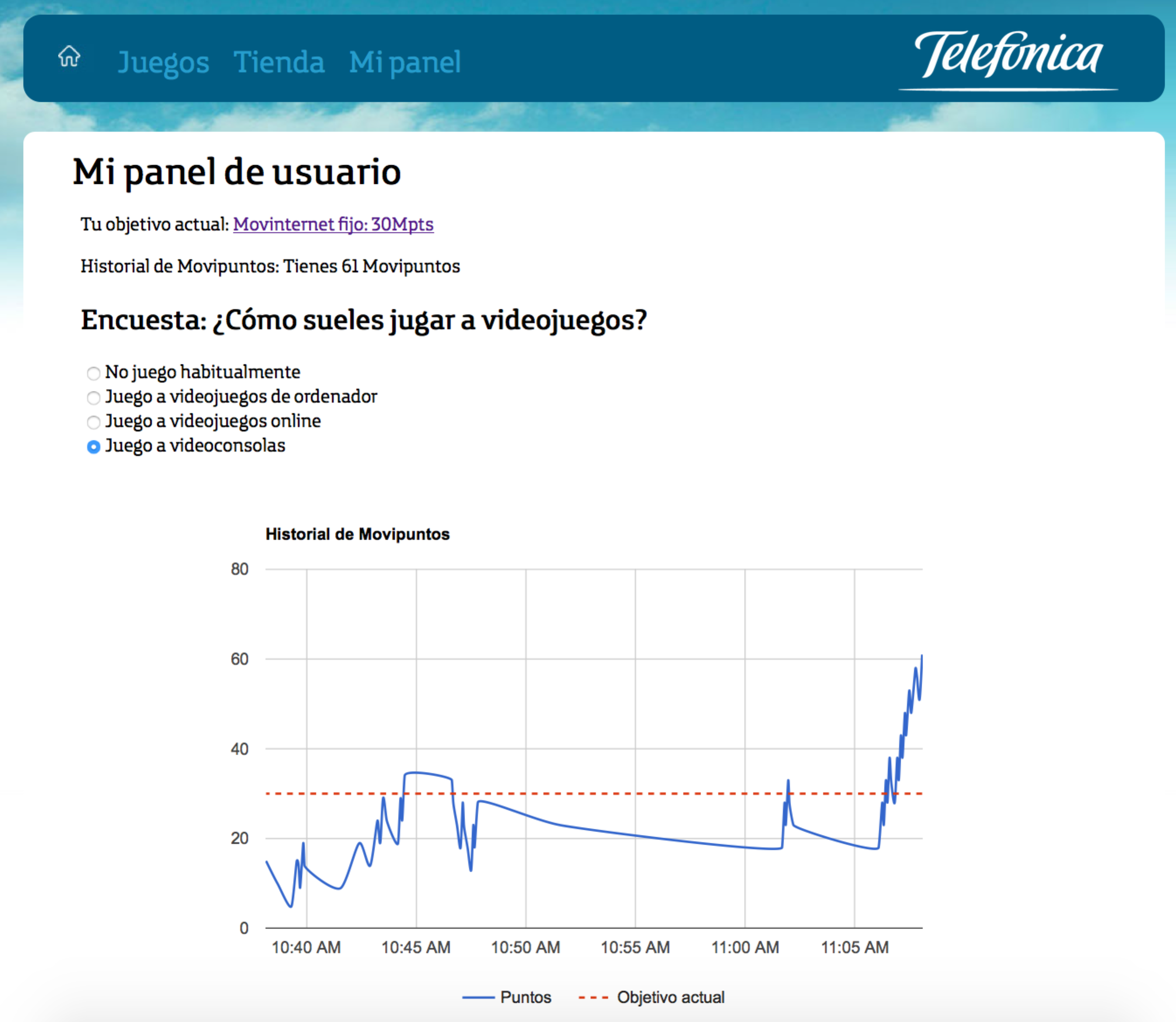
**Movifusión 2: 40 Movipuntos**

* Internet 30Mb
* Línea Móvil 4G
* Verisure

# Sistema de puntuación

El sistema de puntuación comprende el precio de las ofertas ficticias, el dinero virtual (que denominaremos *movipuntos*) que se da a cada jugador en el comienzo del meta-juego y el coste y ganancia asociados a cada uno de los juegos elementales.

Este sistema de puntuación determina en gran parte la dinámica y la capacidad de atracción del jugador, por lo que debe llevar asociado un *cuadro de mandos* (panel de usuario) para que el jugador sepa en cada momento cuál es su situación en el Movijuego, como se muestra en la siguiente figura.



## Criterios de diseño

Los criterios que se exponen a continuación determinan, en gran medida, el diseño completo del sistema de pagos:

1. No debe de ser *pesado*. Si el jugador conoce en profundidad todos los juegos, debe poder comprar la oferta que desee en torno a 10 jugadas con alta probabilidad.
2. Las cantidades de *movipuntos* deben ser fáciles de manejar. Números enteros entre cero y unas pocas decenas.
3. Los costes de los juegos deben de ser parecidos pero no iguales, y no deben dar señales de la estructura probabilística de los mismos, para el que jugador no asocie el más caro al mejor.
4. Las recompensas o premios de los juegos deben reflejar, junto al coste, la estructura de probabilidad: un juego puramente aleatorio no debe permitir ganar nada, por lo que su media debe ser nula. Las varianzas de los mismos deben estar también acotadas en cierta medida, para que sea improbable comprar la oferta más deseada sólo por azar.
5. Los precios de las ofertas deben reflejar que unas son mejores que otras, pero dentro de unos ciertos límites que permitan al jugador revelar sus preferencias aún sin ser capaz de aprender cómo jugar eficazmente. Para compatibilizar este criterio con el anterior, se pide al usuario antes de empezar a jugar una elección preliminar de la oferta que desearía comprar, que podrá actualizar en cualquier momento, pero sobre la que quedarán sus trazas de navegación.

## Costes, ganancias y probabilidades de los juegos básicos

Como base a la aplicación de los criterios anteriores, es necesario conocer la estructura de probabilidad de los juegos. La siguiente tabla muestra los elementos fundamentales: costes y ganancias en movipuntos, probabilidad de ganar si juega bien, probabilidad de ganar si juega mal, la desviación estándar de las ganancias en 10 jugadas y la media de las ganancias –jugando bien- en 10 jugadas. Las ganancias descuentan el coste del resultado.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Juego | Coste | Ganancia | Pr(ganar/bien) | Pr(ganar/mal) | σ (10) | µ (10) |
| Pajita I | 1 | 3 | 1/3 | 1/3 | 5 | 0 |
| Pajita II | 1 | 4 | 1/4 | 1/4 | 4 | 0 |
| Carta I | 5 | 10 | 2/3 | 1/2 | 15 | 16 |
| Carta II | 2 | 4 | 1/2 | 1/2 | 15 | 0 |
| Caja I | 3 | 9 | 2/3 | 1/3 | 13 | 30 |
| Caja II | 2 | 8 | 3/8 | 2/8 | 10 | 10 |

# Movijuego: dinámica global

El Movijuego comienza pidiéndole al usuario identificación. Este requisito supone una modificación de la propuesta original, al estimar que el número final de jugadores -y el sesgo de los mismos- haría extremadamente complicado el análisis de datos sin la identificación *a priori*.

Antes de poder acceder a los juegos, se pide al usuario que rellene la encuesta de juegos y elija una oferta como la más deseable.

El *panel de usuario* y las instrucciones tanto del *Movijuego* como de los juegos básicos está siempre al alcance del usuario.

La asignación inicial es de 10 *movipuntos*. Si el jugador se queda sin puntos se le regalan 8 *movipuntos* las veces que haga falta para que siga jugando.

Por último, se informa al usuario de que el Movijuego sólo estará disponible durante un tiempo limitado –especificando el momento de cierre- durante el que podrá jugar –previa identificación- las veces que desee. Una vez cerrado se procederá a la entrega de premios reales según el procedimiento y criterio establecidos por Telefónica.

# Notas sobre algoritmia de procesado

En este estadio del proyecto no es posible -ni deseable- adelantar los algoritmos ni estrategias que se van a utilizar para el análisis de los datos. Pero sí se pueden adelantar algunos comentarios al respecto.

Primero. Es importante dejar que sean los datos los que *hablen*. En este sentido, no van a proponerse modelos psicológicos, econométricos o para el tratamiento de los datos, sino que van a emplearse algoritmia de A*prendizaje Máquina* para ver las estructuras subyacentes a los mismos.

Segundo. Aunque dada la importancia de la aproximación empleada –basada en el análisis de los datos- se ha tenido gran cuidado en el diseño de la herramienta de captación de los mismos, como muestra este documento, el marco aquí empleado tampoco deber encorsetar el análisis de los mismos.

Tercero. La estructura de procesado prevista sigue el esquema que se muestra abajo:

1. Exploración de los datos y pre-procesado. La gran dimensionalidad de los datos, fruto de la aproximación *greedy* en el diseño de la herramienta, hace indispensable una primera reducción de dimensionalidad guiada por la semántica de los mismos. Por ejemplo, en los ciclos de visualización de un individuo en los juegos o en las ofertas, será necesario extraer parámetros del estilo *recurrencia*, *permanencia*, o similares.
2. Análisis del modelo resultante y dimensionalidad de la representación. Del resultado de la fase anterior y, en función del número de respuestas obtenidas así como de la homogeneidad de las mismas, se tendrá que decidir sobre la posible representación de los datos para su procesado. El análisis en este punto sigue estrategias semejantes a las presentadas en [6].
3. Análisis Estadístico de Datos: Sobre la representación anterior, se aplicarán las técnicas de Minería de Datos [6] que se consideren más adecuadas.

# Referencias

1. M. Solomon: Consumer Behaviour
2. K. Binmore: *Fun and Games. A Text on Game Theory*. D.C. Heath and Co., Lexington, MA., 1992.
3. P. Wakker: Prospect Theory
4. J. Bewersdorff*: Luck, Logic, and White Lies. The Mathematics of Games*. A.K. Peters Ltd., Wellesley, MA, 2005.
5. J. Haigh: “Matemáticas y juegos de azar”
6. Y. Bengio, A. Courville, and P. Vincent: Representation Learning *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013, vol. 35, no 8, p. 1798-1828
7. T. Hastie, R. Tibsirani, J. Friedman: The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2º Ed. Springer, 2009.

# Apéndice. Análisis de Mercado de las ofertas de Movistar, Orange y Vodafone.

Para el análisis se procede a la búsqueda de semejanzas y diferencias de los productos que puedan ser comparables dentro de estas tres compañías. Es decir, no se han tenido en cuenta otros tipos de productos que fueran ofrecidos simultáneamente en las tres operadoras. En los análisis, cada representación gráfica se acompaña de su color corporativo





