

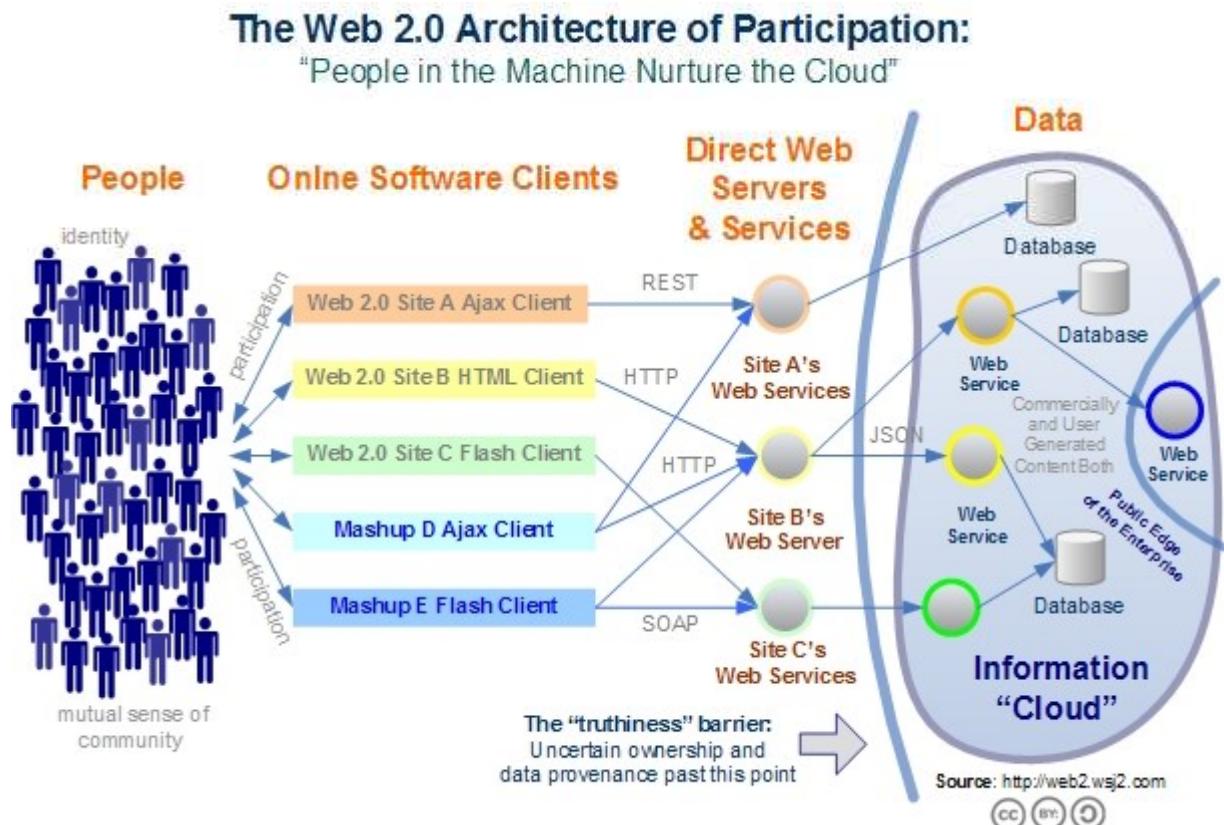
El origen del Big Data

O'Reilly Conference: 2004

| Web1.0 | Web2.0 |
|-------------------------|-------------------------|
| DoubleClick | GoogleAdsense |
| Ofoto | Flickr |
| Akamai | BitTorrent |
| Mp3.com | Napster |
| Enciclopedia británica | Wikipedia |
| Webs personales | Blogs |
| Páginas vistas | Coste por click |
| Publicar | Participación |
| Gestión de contenidos | Wiki |
| Directorios (taxonomía) | Etiquetas (folksonomía) |

El origen del Big Data

► Elementos de la Web 2.0



Boom del Social Media

Social Media Landscape



Obtención de datos en la Web

Bedding (3)

CHOOSE BY CUSTOMER RATING

| |
|-----------|
| ★★★★★ (2) |
| ★★★★ (2) |
| ★★★★★ (2) |
| ★★★★ (1) |
| ★★★ (2) |

CHOOSE BY PRICE RANGE

| |
|----------------|
| Under 5 (1) |
| 5 to 10 (2) |
| 10 to 20 (4) |
| 20 to 40 (7) |
| 40 to 60 (2) |
| 60 to 100 (2) |
| 100 to 250 (2) |

FREE DELIVERY

heather combz
★★★★ (20 Ratings) **SALE** RRP 15.99 ADD TO CART

Redkite Cot Twink - Pretty Flower Compare

Red Kite Comfy Carry Micro Fleece Baby Sling (Medium)
"My wife loves it (I would use it more, but the medium size is too small for me)"
JB - Oxford
★★★★ (1 Rating) RRP 13.99 ADD TO CART

Red Kite Comfy Carry Micro Fleece Baby Sling (Large)
★★★★ (1 Rating) RRP 14.99 ADD TO CART

Redkite Musical Mobile - Woodland
"My little boy is fascinated by this and has been smiling every day since we got it!"
Jade - London
Compare



Etiquetas

tripadvisor.es

Hotel estupendo , con excelente conexión al aeropuerto

★★★★★ MIAMOBILIAO BILBAO 1 sep 2009 | Tipo de viaje: En familia

Nos hospedamos tres noches en agosto ya que teníamos una conexión en un vuelo internacional. Eramos dos adultos y dos niños de 5 y 2 años. Cuando llegamos nos dijeron que aunque teníamos reservada la habitación con antelación, no tenían disponibles, y que nos daban una suite. En la planta cuarta, nos dieron la habitación con dos baños, salón-comedor, y... más

Excelete Relacion Precio -Calidad

★★★★ sebastorrob Buenos Aires, Argentina 31 ago 2009 | Tipo de viaje: En pareja

Lo he reservado desde Argentina por internet sin saber la ubicación exacta, y me llevé una grata sorpresa, desde el bus que te lleva desde el Aeropuerto de Barajas al Hotel en forma gratuita y viceversa, hasta el espléndido desayuno, las cenás, el personal del hotel. Obviamente está alejado del centro de Madrid, pero está a pasos de la estación... más

Fotos y videos de viajeros

Best Western Atlantic ****

★★★★ 1 de 467 hoteles en Madrid 842 opiniones

excelete 17 dic 2009

Muy buen hotel 19 nov 2009

VER PRECIOS 146 € Precios media

Hotel Preciosos ****

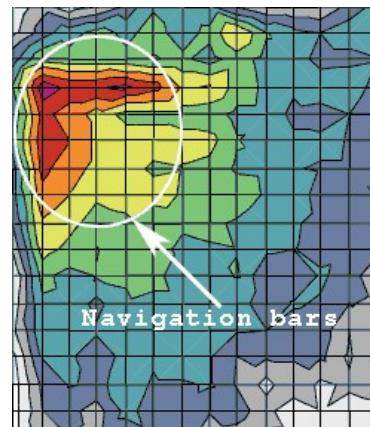
★★★★ 2 de 467 hoteles en Madrid 325 opiniones

Casi perfecto 7 5 ene 2010

Muy buena ubicación y muy buen trato 26 dic 2009

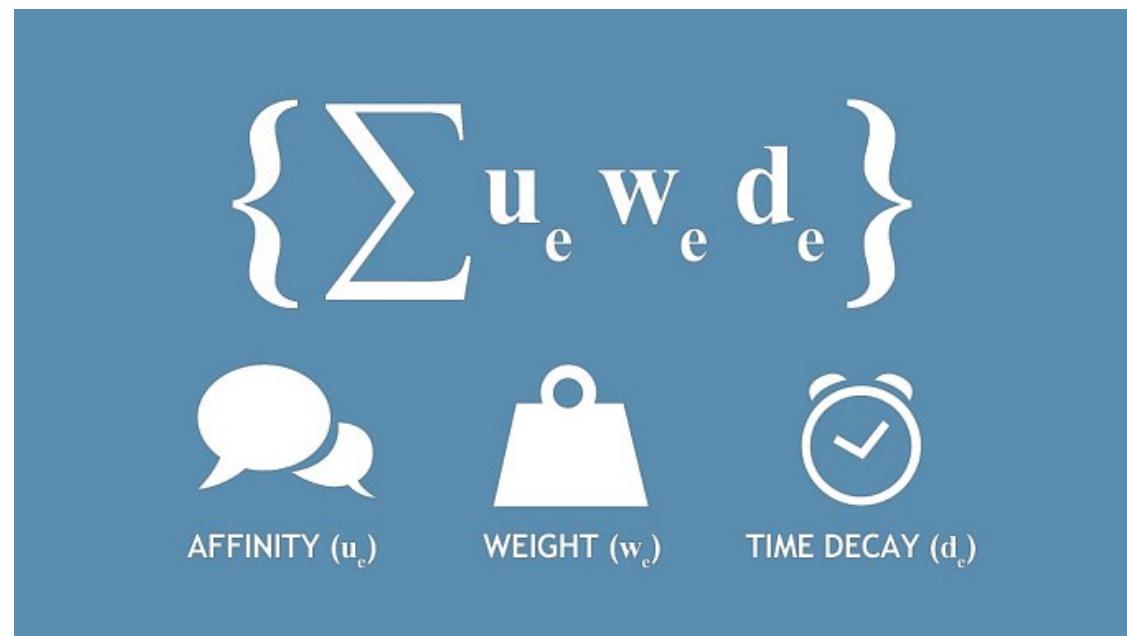
VER PRECIOS 141 € Precios media

Evaluaciones



Click-Stream

Comentarios



Google



[Shopping](#) [Gmail](#) [more ▾](#)

Google

[Make Google your homepage](#) [Advertising Programs](#) [Business Solutions](#) [About](#)
- Go to Google India

©2008 - Privacy





Ambientes inteligentes



Source: Fraunhofer FOKUS und Mikromedia

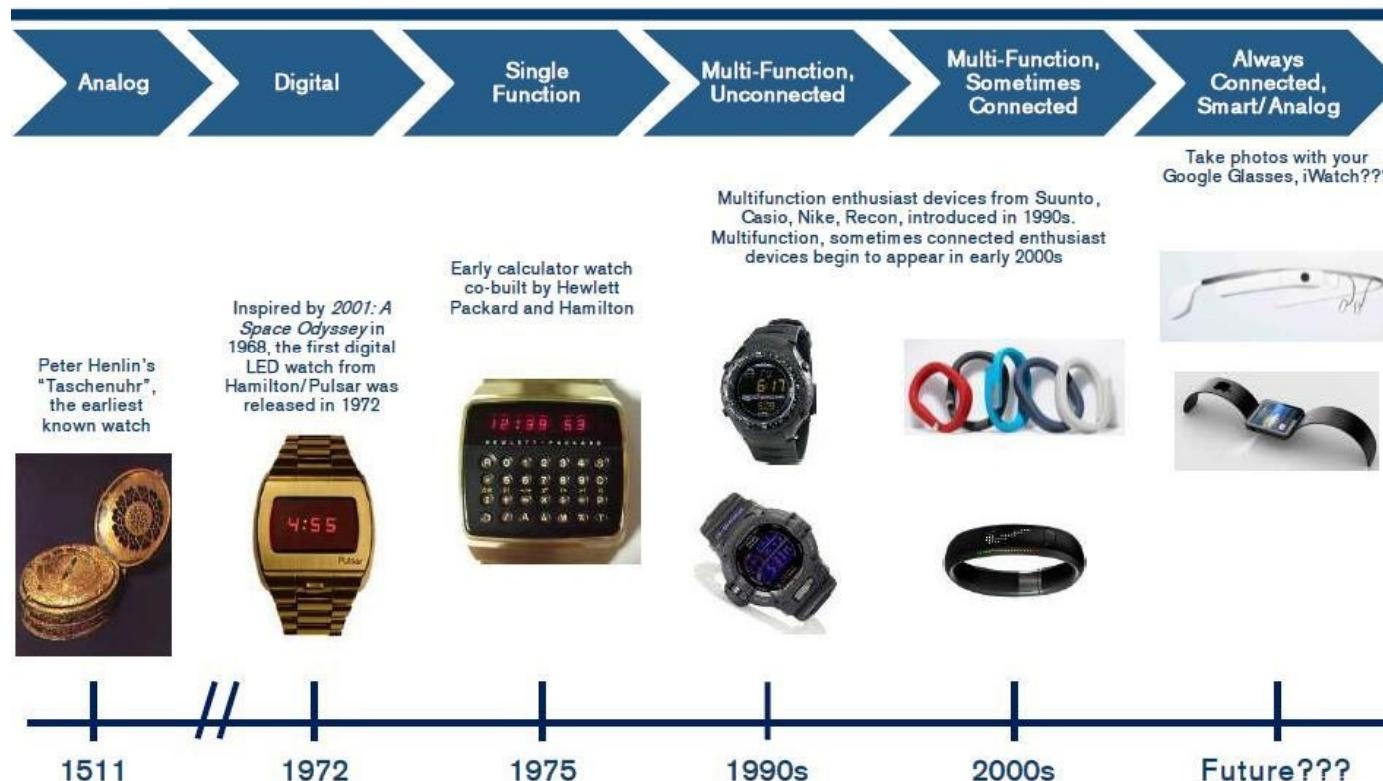
Infraestructura: Computación empotrada e Internet de las Cosas

Smart Cities

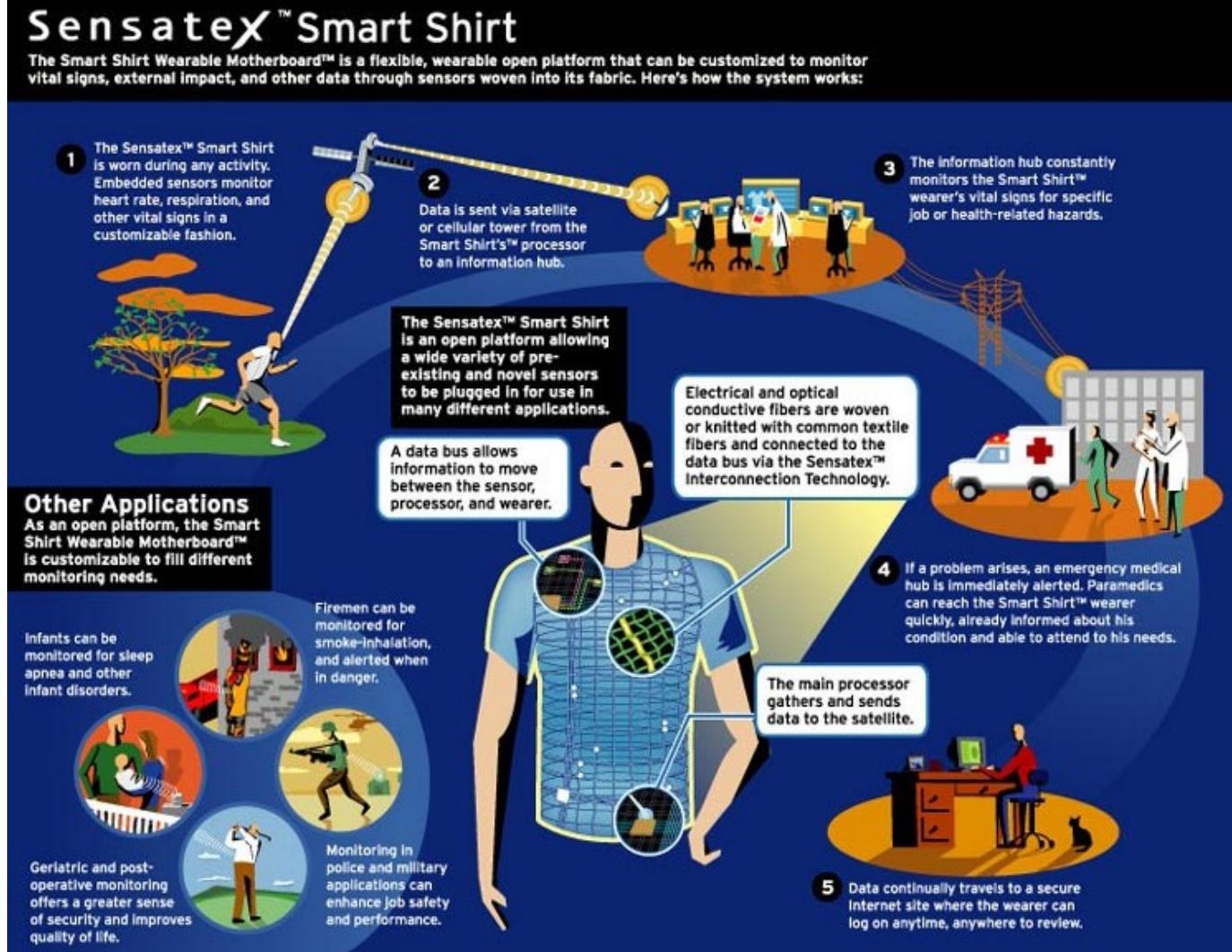


Smart Clothes y Wearables

The Evolution of Wearables



Smart Clothes y Wearables

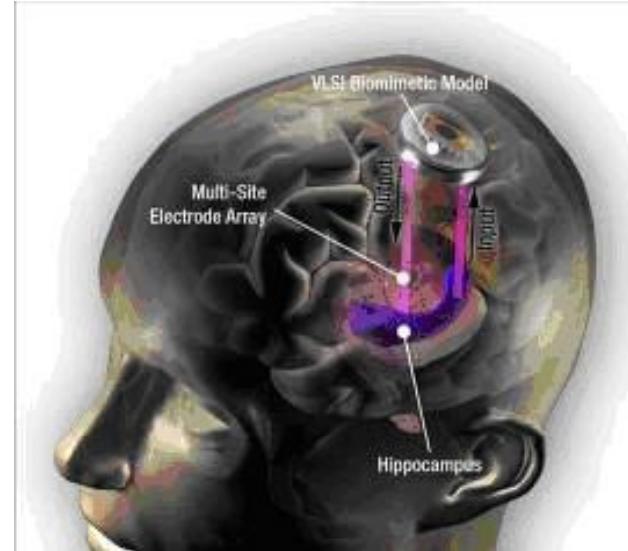
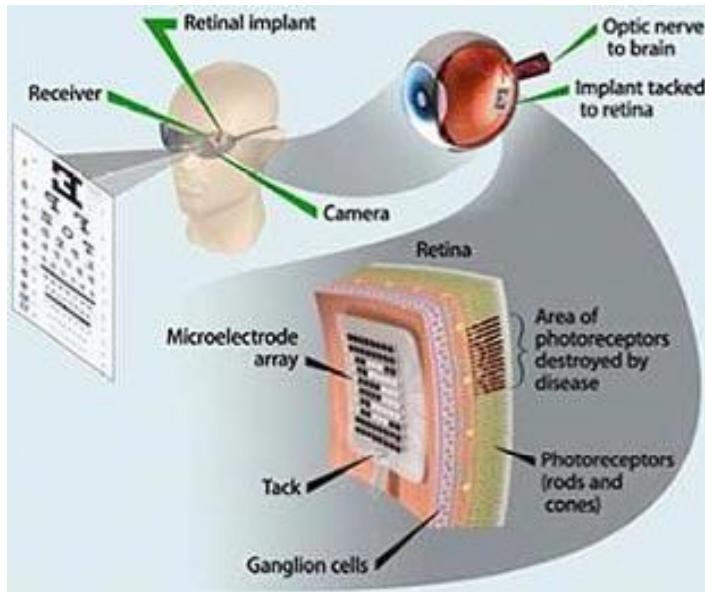


Asistentes Inteligentes



Siri, Alexa y Google Assistant

Prótesis



(Enero, 2024)



Sociedad

EDUCACIÓN · MEDIO AMBIENTE · IGUALDAD · SANIDAD · CONSUMO · LAICISMO · COMUNICACIÓN · ÚLTIMAS NOTICIAS

MENORES >

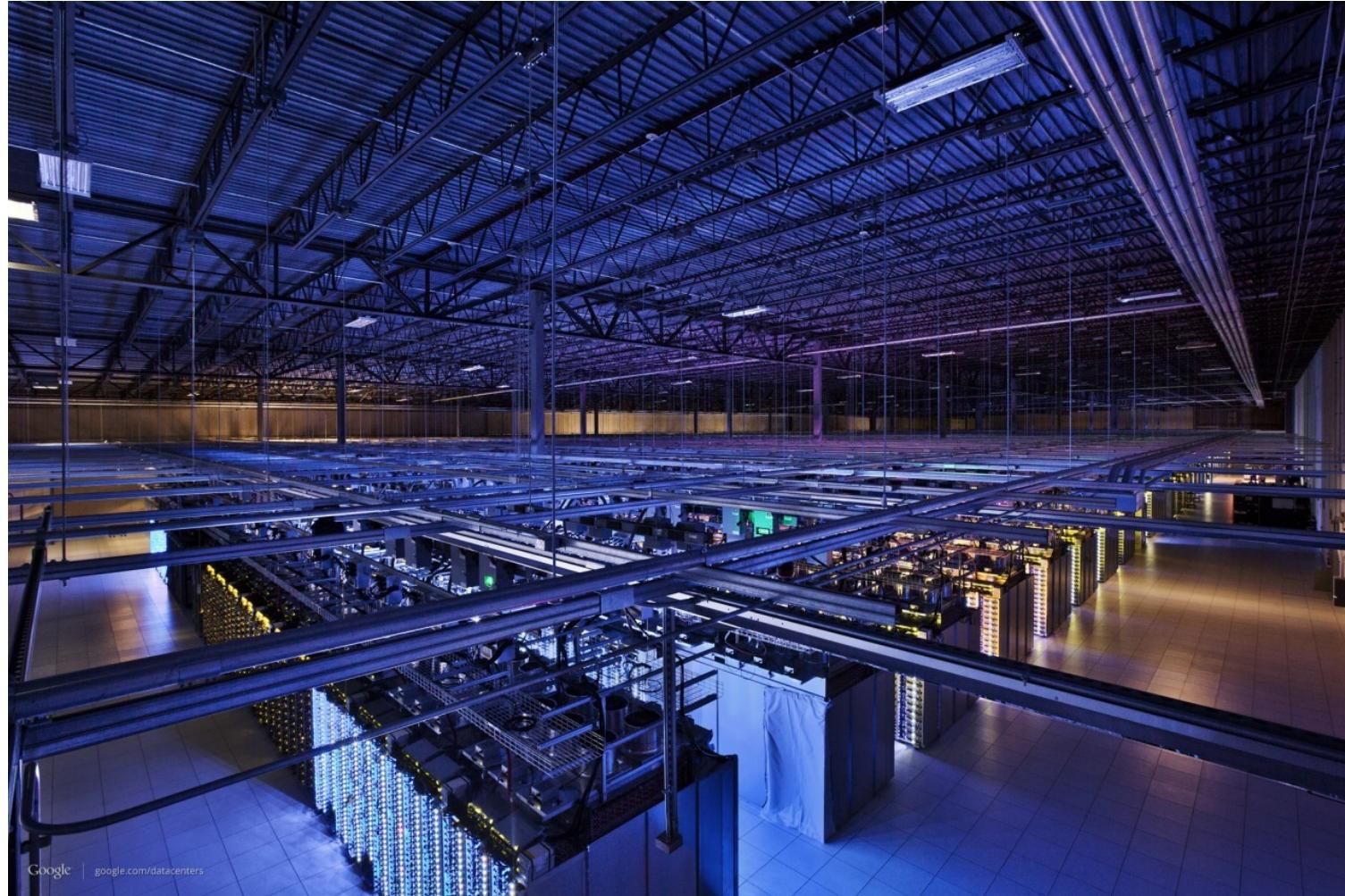
Mar España: “Hasta los seis años los niños no deberían recibir enseñanza digital en el colegio”

La directora de la Agencia Española de Protección de Datos advierte de que la sociedad está en alerta roja por el consumo abusivo y precoz de contenidos inadecuados online y anuncia que trabajará para que la nueva ley de protección de los menores en internet regule sus derechos digitales

<https://elpais.com/sociedad/2024-01-29/mar-espana-hasta-los-seis-anos-los-ninos-no-deberian-recibir-ensenanza-digital-en-el-colegio.html>

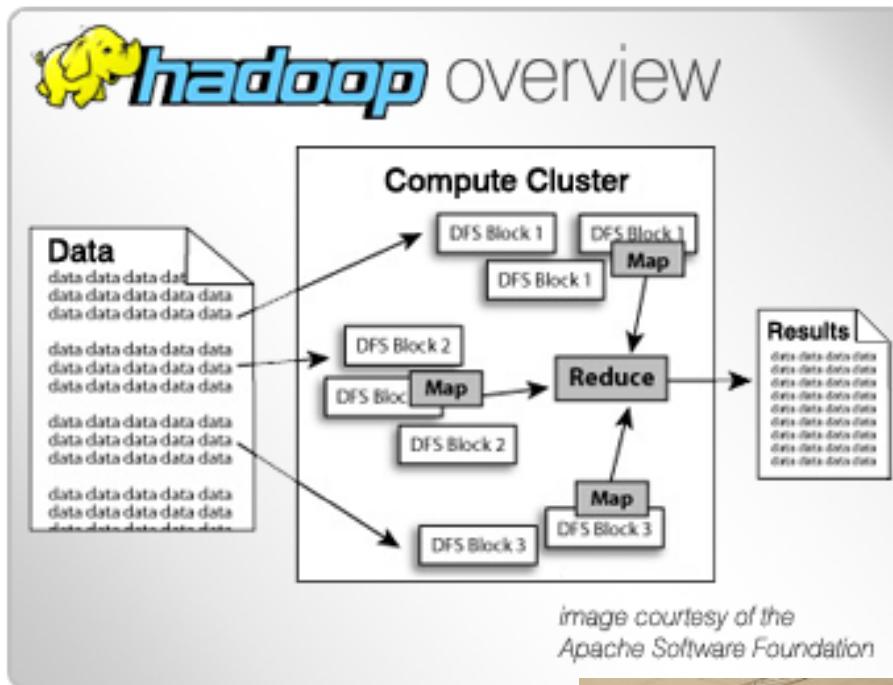
¿Qué opinas al respecto? ¿Qué datos crees que habría que proteger?

Big Data 1.0: Guardar y procesar datos



Google Data Center

Big Data 1.0: Guardar y procesar datos



Big Data 2.0: Analizar datos

DATA



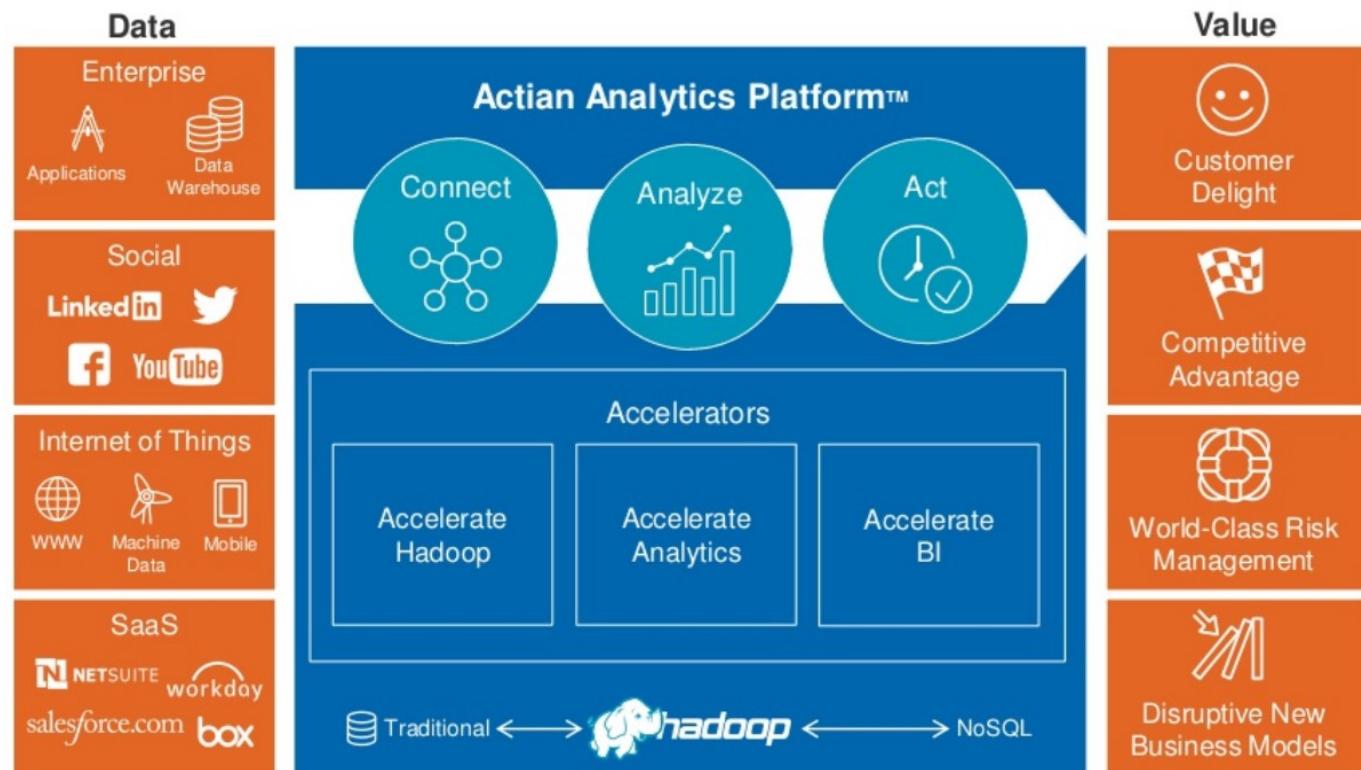
+

MATH



= ANALÍTICA

Actian Accelerates Big Data 2.0 Across the Entire Analytics Value Chain



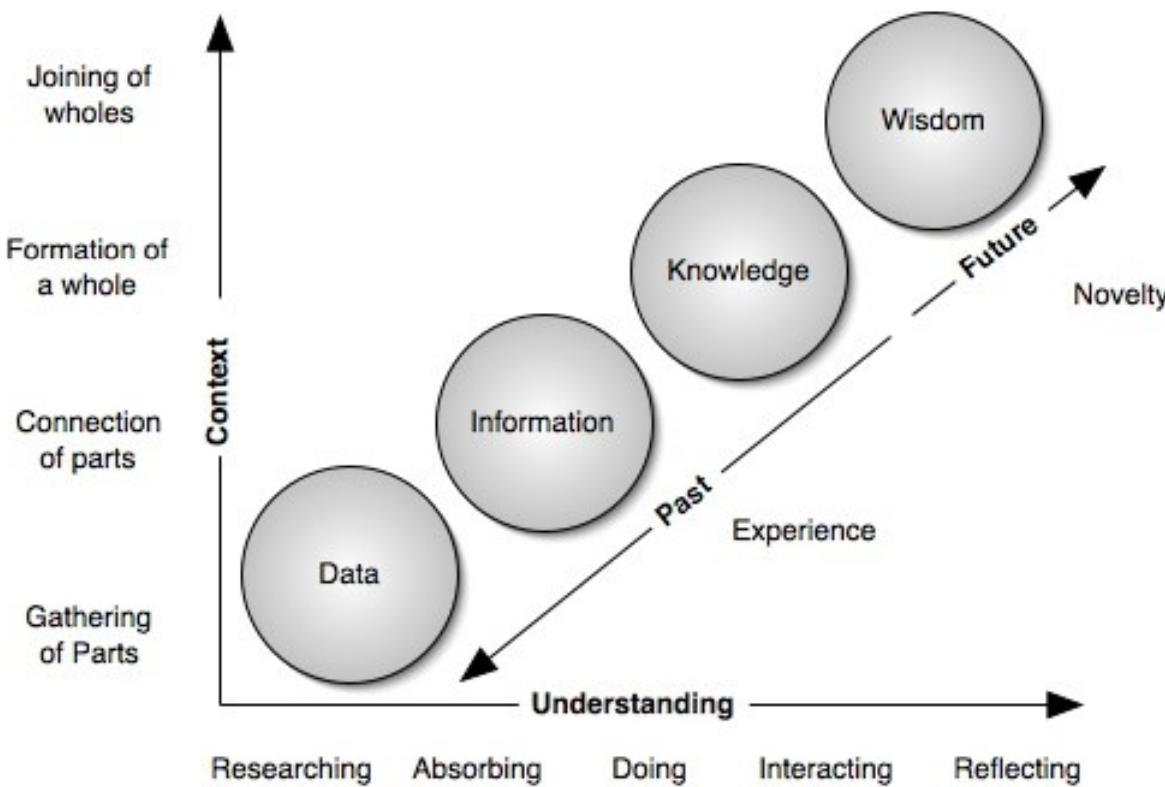
TAREAS DE ANALÍTICA



Realiza la Tarea 3 del boletín de tareas del Tema 1:

Tema 1: Tareas de Analítica

DIKW



DATOS

Good Morning 0 1 1 0 0 0 0 1 0 1 0 0 1 1 0

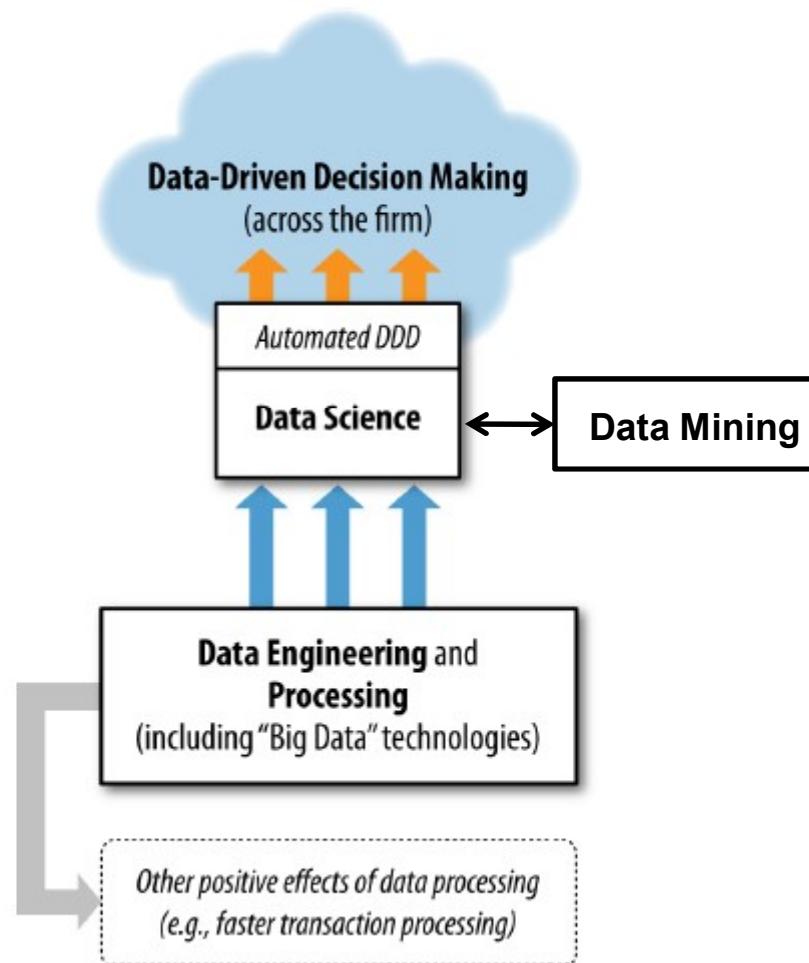
INFORMACIÓN

| Película | Drama | Aventura | |
|--------------|-------|----------|------|
| Good Morning | 1 | 0 | |

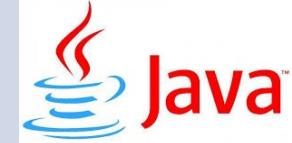
CONOCIMIENTO

Agrupamiento: Película A es similar a Película B
A los usuarios que les ha gustado la Película A
también les ha gustado la Película C
El 75% de los fans de la Película A son menores de
30 años
Si te gusta el drama te gustará la Película B

Big Data 3.0: Decisiones basada en datos



Big Data: conocimientos y tecnologías

| | Entorno de Pruebas/Aprendizaje | Entorno de Producción |
|--------------|--|---|
| Conocimiento |   |   |
| Tecnologías |    |   |

Empresas dirigidas por datos: Amazon.com

Amazon killed Borders and RadioShack

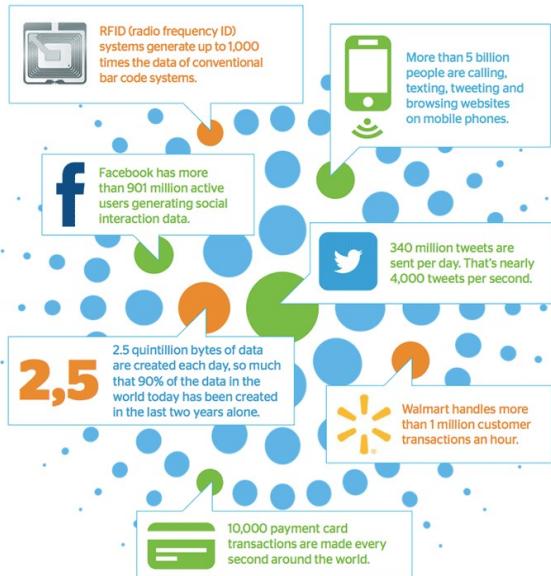


Empresas dirigidas por datos: Netflix.com

Netflix killed Blockbuster



Requisitos para Data Science/Big Data

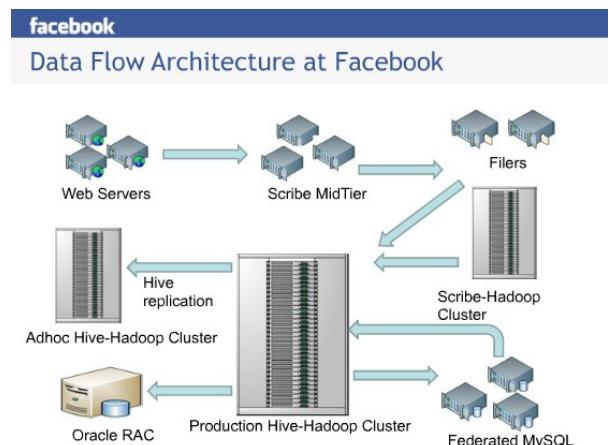


Data

Data
Scientists



Infrastructure



Vuestro trabajo

Explore data:
Look for patterns



Predict the future:
Build predictive models

PENSAMIENTO ANALÍTICO BASADO EN DATOS



Grupo de Sistemas *inteligentes*



Eduardo M. Sánchez Vila
eduardo.sanchez.vila@usc.es

Grupo de Sistemas Inteligentes
Universidad de Santiago de Compostela

Experimenta con Amazon

¿Qué recomendaciones te presenta Amazon?



Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.



[The Little Big Thing: 163 Ways to Pursue Excellence](#) [Fascinate: Your 7 Triggers to Persuasion and Captivation](#)



[Sherlock Holmes \[Blu-ray\]](#)



[Alice in Wonderland \[Blu-ray\]](#)

Experimenta con TripAdvisor



¿Qué recomendaciones te presenta TripAdvisor?

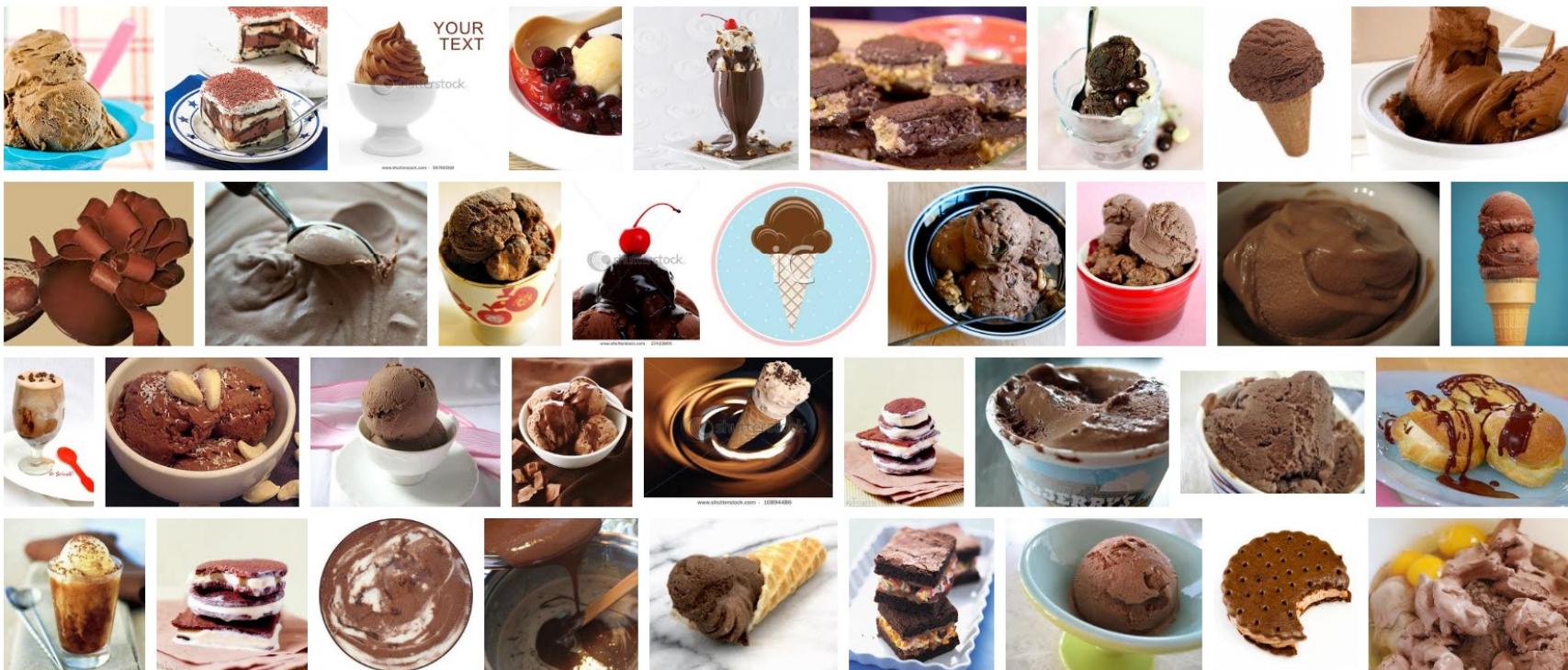


Tripadvisor.

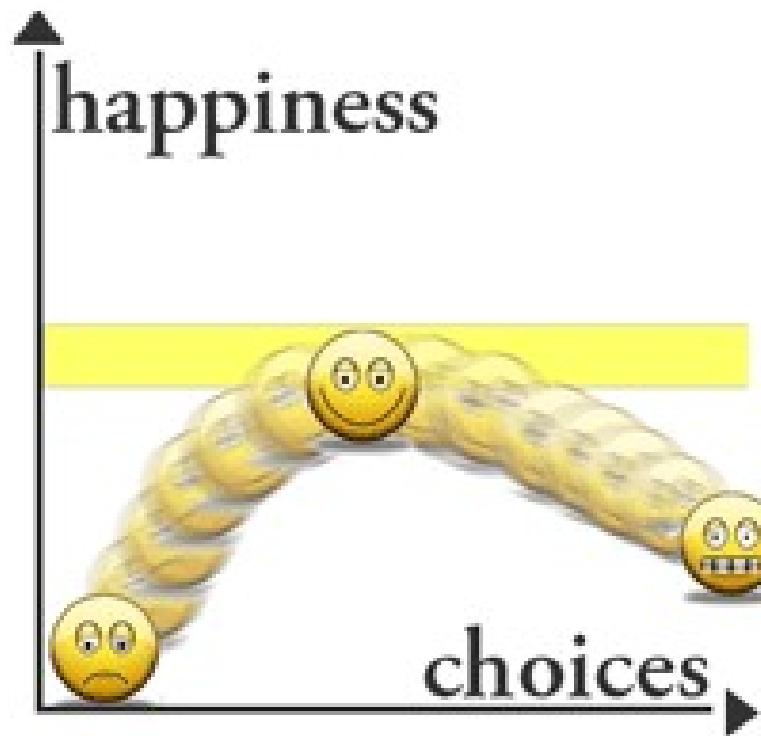
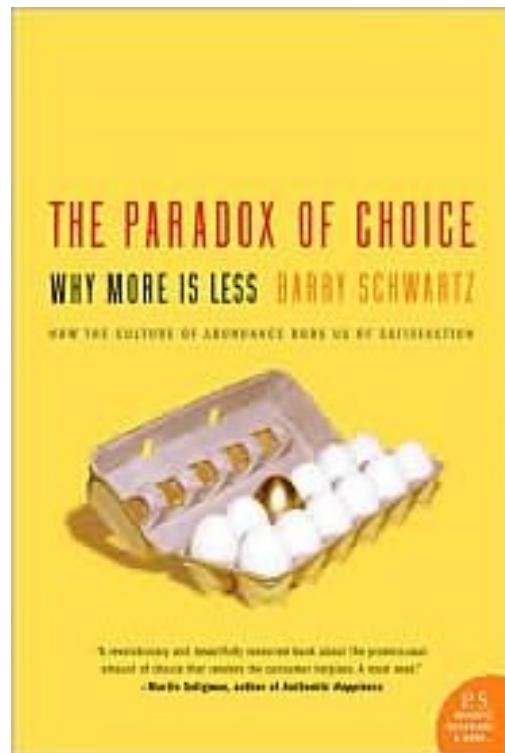
El problema de la elección



El problema de la elección



El problema de la elección



¿Qué son las recomendaciones?

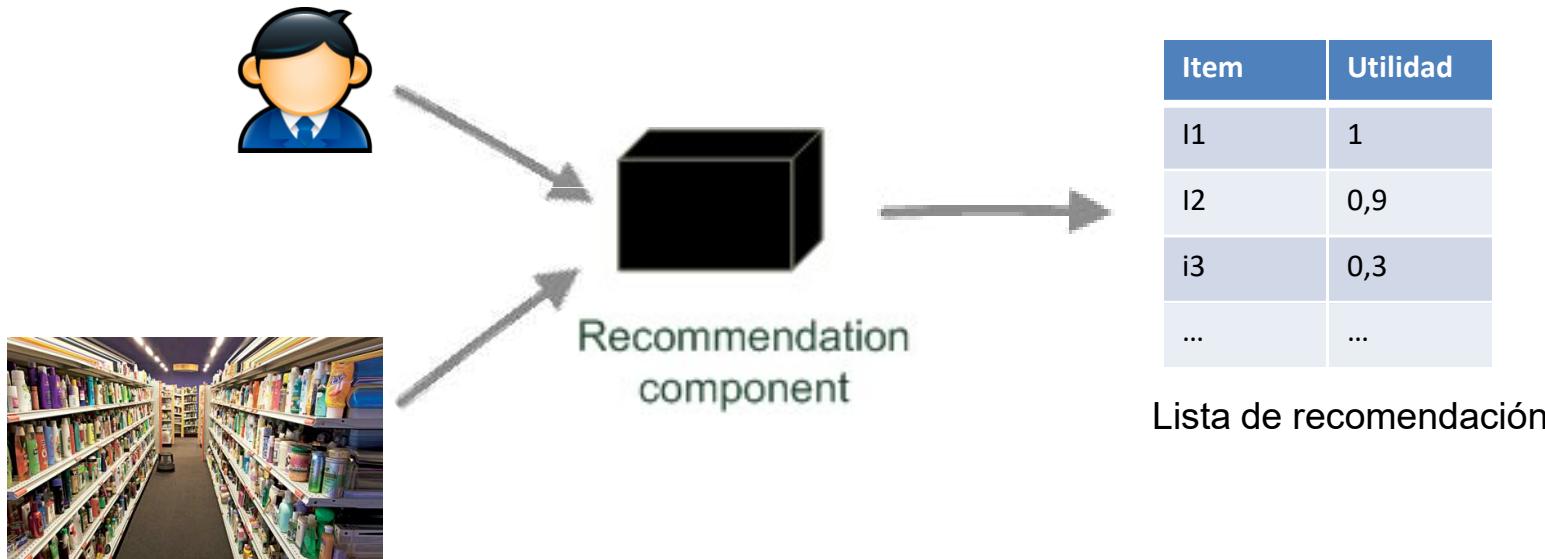


¿Qué son las recomendaciones?



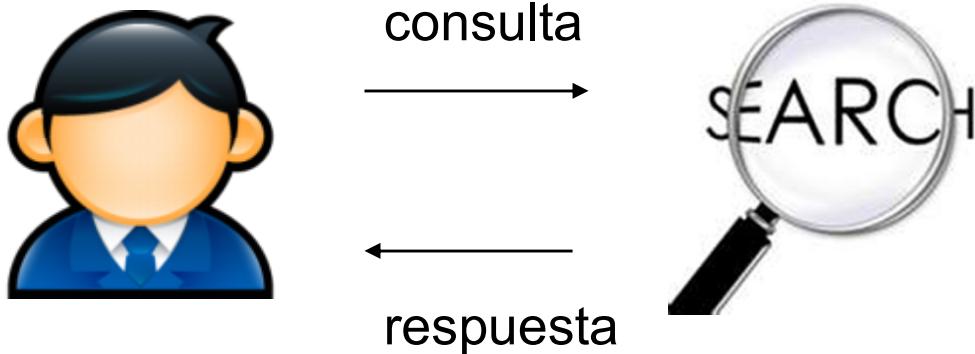
¿Qué son las recomendaciones?

Problema: Predecir la utilidad de un ítem para un usuario!!!

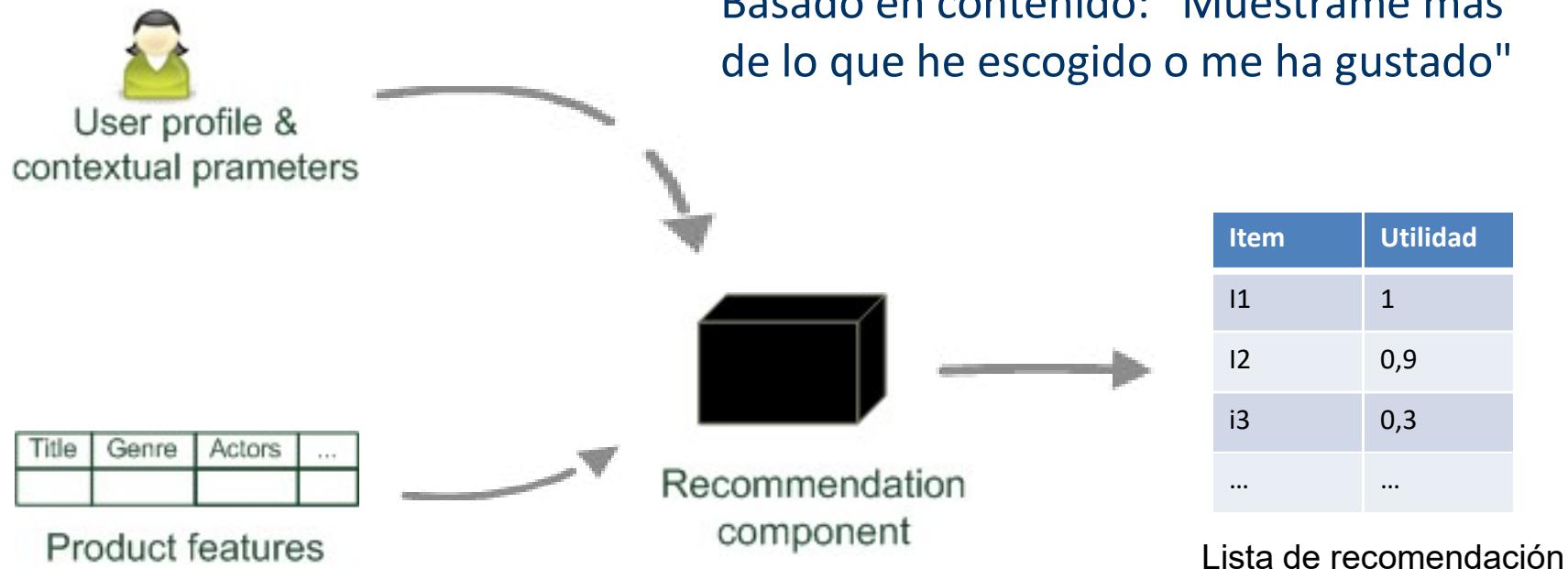


Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Recomendación Vs Búsqueda

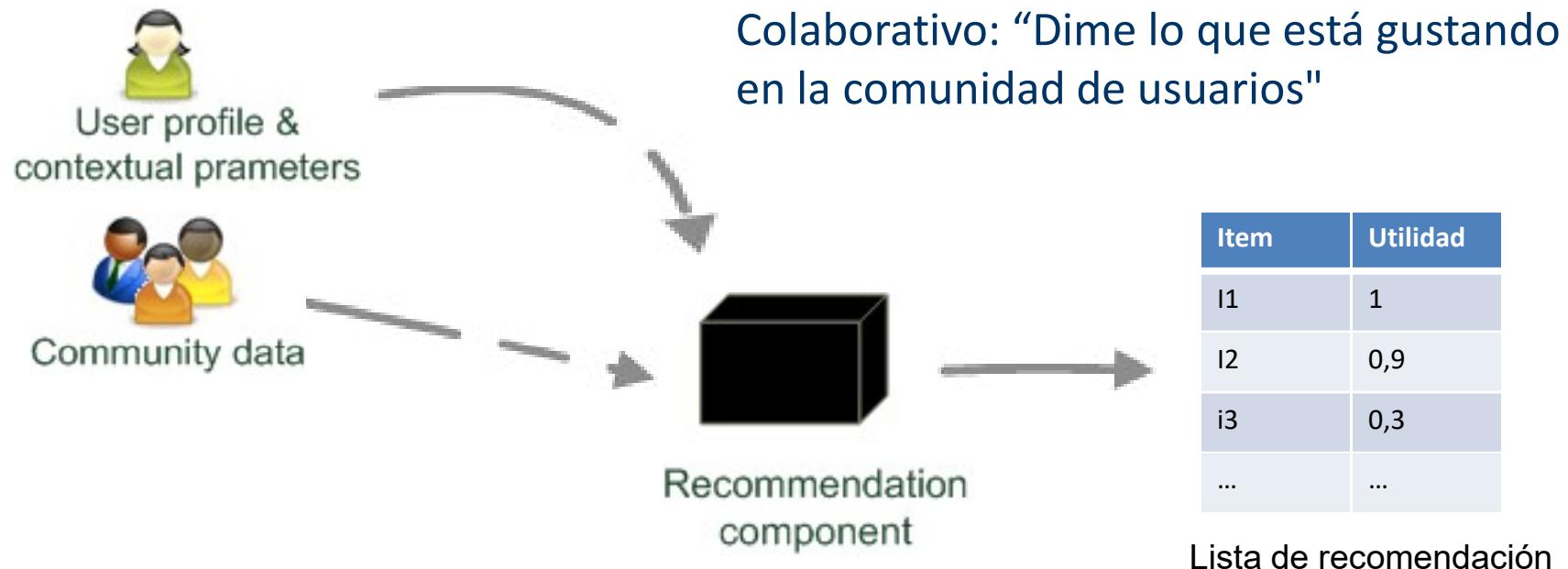


Estrategias de recomendación



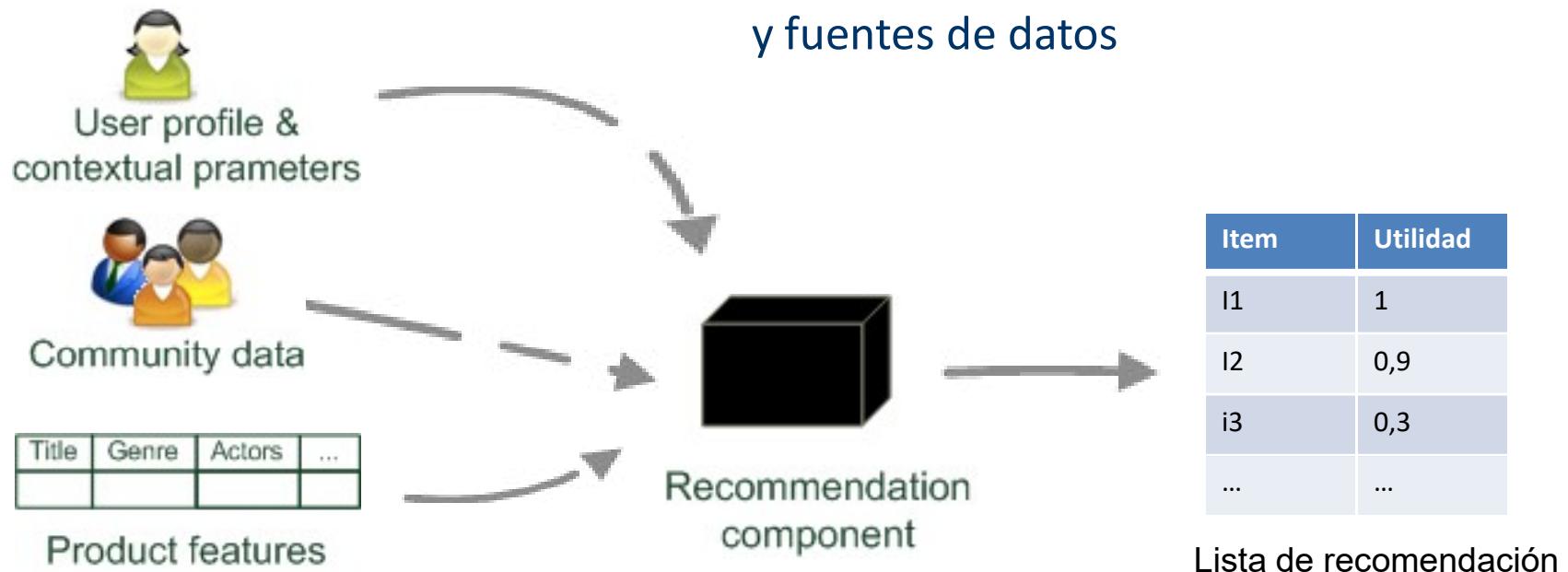
Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Estrategias de recomendación



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Estrategias de recomendación



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Recomendación basada en contenido

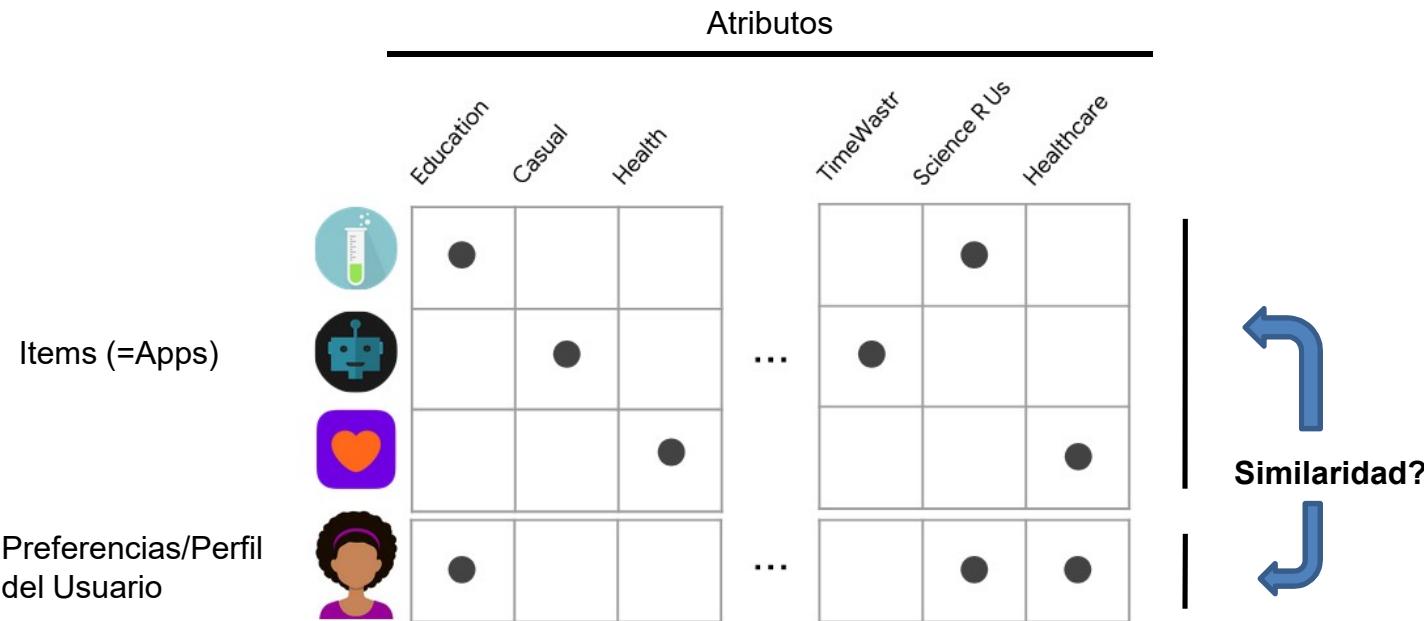
- Objetivo:

Recomendar items al usuario-objetivo similares a aquellos items que el usuario ha escogido, comprado o mostrado algún tipo de preferencia en el pasado.

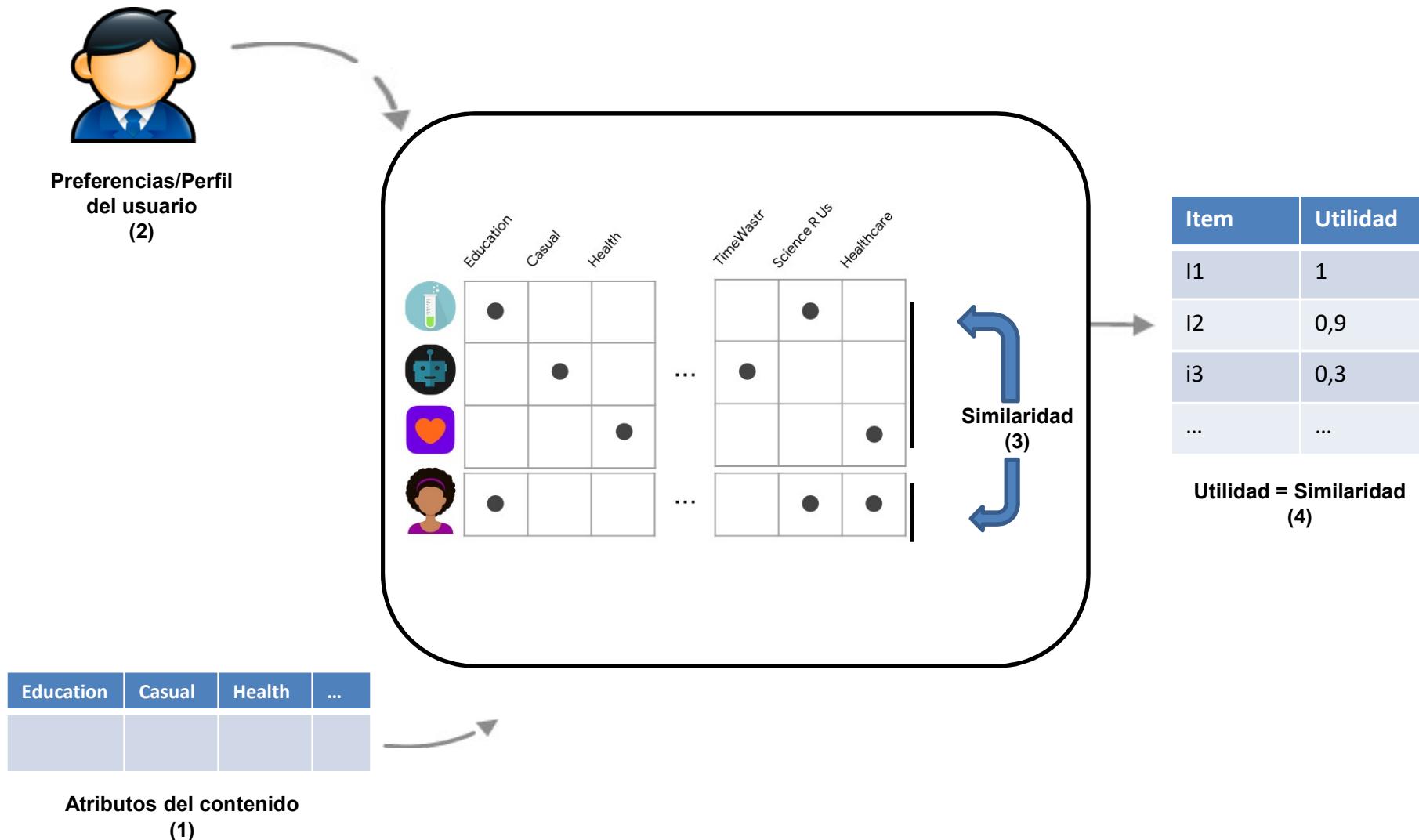
- ¿Qué necesitamos?:

1. **Caracterizar el contenido** (items/alternativas): Información sobre los *atributos del contenido*. Ejemplo: el género de las películas.
2. **Aprender/estimar las preferencias del usuario.** Las preferencias representan la importancia/relevancia que tiene para el usuario cada uno de los atributos del contenido. El conjunto de preferencias constituye el *perfil del usuario*.
3. **Calcular la similaridad.** Hay que calcular la similaridad entre el perfil del usuario y cada item o alternativa que esté disponible para recomendar.
4. **Asumir utilidad = similaridad.** Vamos a aceptar el supuesto de que la utilidad de un item para un usuario viene dada por la similaridad entre el item y el perfil del usuario

Recomendación basada en contenido



Recomendación basada en contenido



Caracterizando el contenido

- Contenidos. Ejemplo: Libros
- Atributos de un libro:
 1. Precio (variable numérica) - Valores: entre 0 y N euros
 2. Género (variable discreta) - Valores: {ficción, histórico, novela...}
- Binarización de los atributos:
 1. Precio: 0-15 euros -> “Precio Bajo” – Valores: {0,1}
 2. Precio: >15 euros -> “Precio Alto” – Valores: {0,1}
 3. Genero: ficción -> “Género ficción” – Valores: {0,1}
 4. Genero: histórico -> “Género histórico” – Valores: {0,1}

Caracterizando el contenido

Atributos

| Items | Producto | Precio | Género |
|-------|----------|--------|-----------|
| | Libro1 | 15 | Ficción |
| | Libro2 | 50 | Histórico |
| | | | |
| | LibroN | 10 | Ficción |

Atributos

| Items | Producto | Precio Bajo | Precio Alto | Género ficción | Género histórico |
|-------|----------|-------------|-------------|----------------|------------------|
| | Libro1 | 1 | 0 | 1 | 0 |
| | Libro2 | 0 | 1 | 0 | 1 |
| | | | | | |
| | LibroN | 1 | 0 | 1 | 0 |

Caracterizando el contenido

| Items | Atributos | | | | |
|--------|-----------|-------------|-------------|----------------|------------------|
| | Producto | Precio Bajo | Precio Alto | Género ficción | Género histórico |
| Libro1 | 1 | 0 | 1 | 0 | 0 |
| Libro2 | 0 | 1 | 0 | 1 | 0 |
| | | | | | |
| LibroN | 1 | 0 | 1 | 0 | 0 |



Representación como vector

Vector Libro1 = (1, 0, 1, 0)

Formalizando para todo item:

$$\text{Vector del Item} = (x_{11}, x_{21}, \dots, x_{1K}, x_{2K}, \dots)$$

donde x es una variable binaria y k indica el número del atributo

Aprendiendo las preferencias/perfil del usuario



Conjunto de datos del individuo c (dataset)

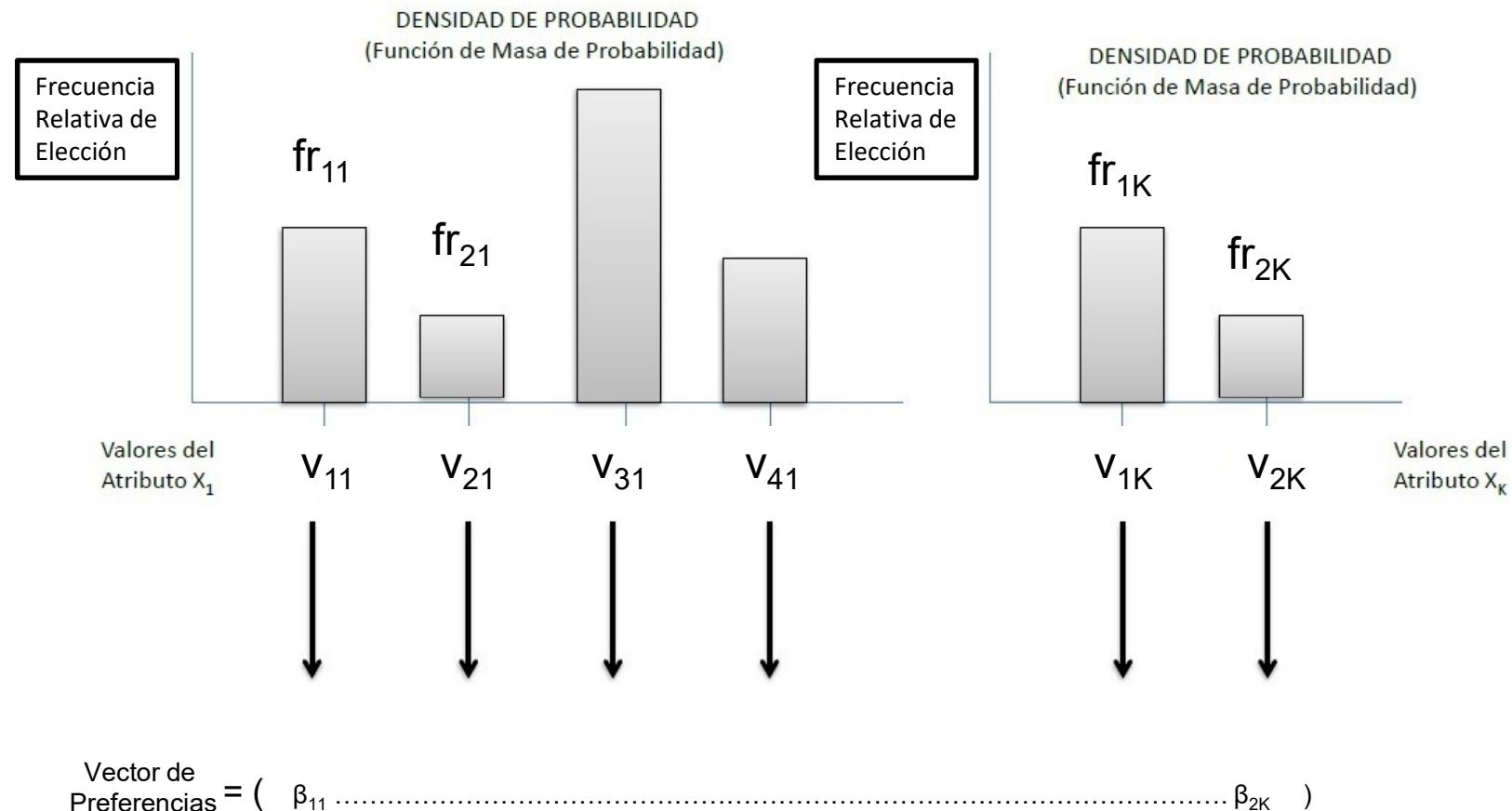
| Atributos | | | | | | |
|-----------|----------|-------|-------|--------|-----------|----------------|
| Items | Producto | Autor | Fecha | Precio | Género | Acción usuario |
| | Libro1 | --- | --- | 15 | Ficción | Comprado |
| | Libro2 | --- | --- | 50 | Histórico | Comprado |
| | | | | | | |
| | LibroN | --- | --- | 10 | Ficción | Comprado |

¿Cuáles son las preferencias del individuo c?

Frecuencia de Elección de cada valor = Veces que un valor ha sido “Comprado”

$$\text{Frecuencia Relativa de Elección de cada valor} = \frac{\text{Frecuencia de Elección de cada valor}}{\text{Número total de compras}}$$

Aprendiendo las preferencias/perfil del usuario



Aprendiendo las preferencias/perfil del usuario

Conjunto de datos del usuario c (dataset)

Atributos

| Producto | Autor | Fecha | Precio | Género | Acción usuario |
|----------|-------|-------|--------|-----------|----------------|
| Libro1 | --- | --- | 10 | Ficción | Comprado |
| Libro2 | --- | --- | 50 | Histórico | Comprado |
| | | | | | |
| LibroN | --- | --- | 50 | Ficción | Visto |

Items



| Producto | Autor | Fecha | Precio | Género | Utilidad |
|----------|-------|-------|--------|-----------|----------|
| Libro1 | --- | --- | 10 | Ficción | 10 |
| Libro2 | --- | --- | 50 | Histórico | 10 |
| | | | | | |
| LibroN | --- | --- | 50 | Ficción | 5 |

Aprendiendo las preferencias/perfil del usuario



Conjunto de datos del usuario c (dataset)

Atributos

| Producto | Autor | Fecha | Precio | Género | Utilidad |
|----------|-------|-------|--------|-----------|----------|
| Libro1 | --- | --- | 10 | Ficción | 10 |
| Libro2 | --- | --- | 50 | Histórico | 10 |
| | | | | | |
| LibroN | --- | --- | 50 | Ficción | 5 |

Items

¿Cuáles son las preferencias del individuo c?

1

$$\text{Utilidad media de cada valor} = \frac{\sum \text{utilidad de la compra con ese valor}}{\text{Número de compras con ese valor}}$$

Aprendiendo las preferencias/perfil del usuario

Conjunto de datos del usuario c (dataset)

Atributos

| Producto | Autor | Fecha | Precio | Género | Utilidad |
|----------|-------|-------|--------|-----------|----------|
| Libro1 | --- | --- | 10 | Ficción | 10 |
| Libro2 | --- | --- | 50 | Histórico | 9 |
| Libro3 | --- | --- | --- | --- | |
| LibroN | --- | --- | 30 | Novela | 3 |



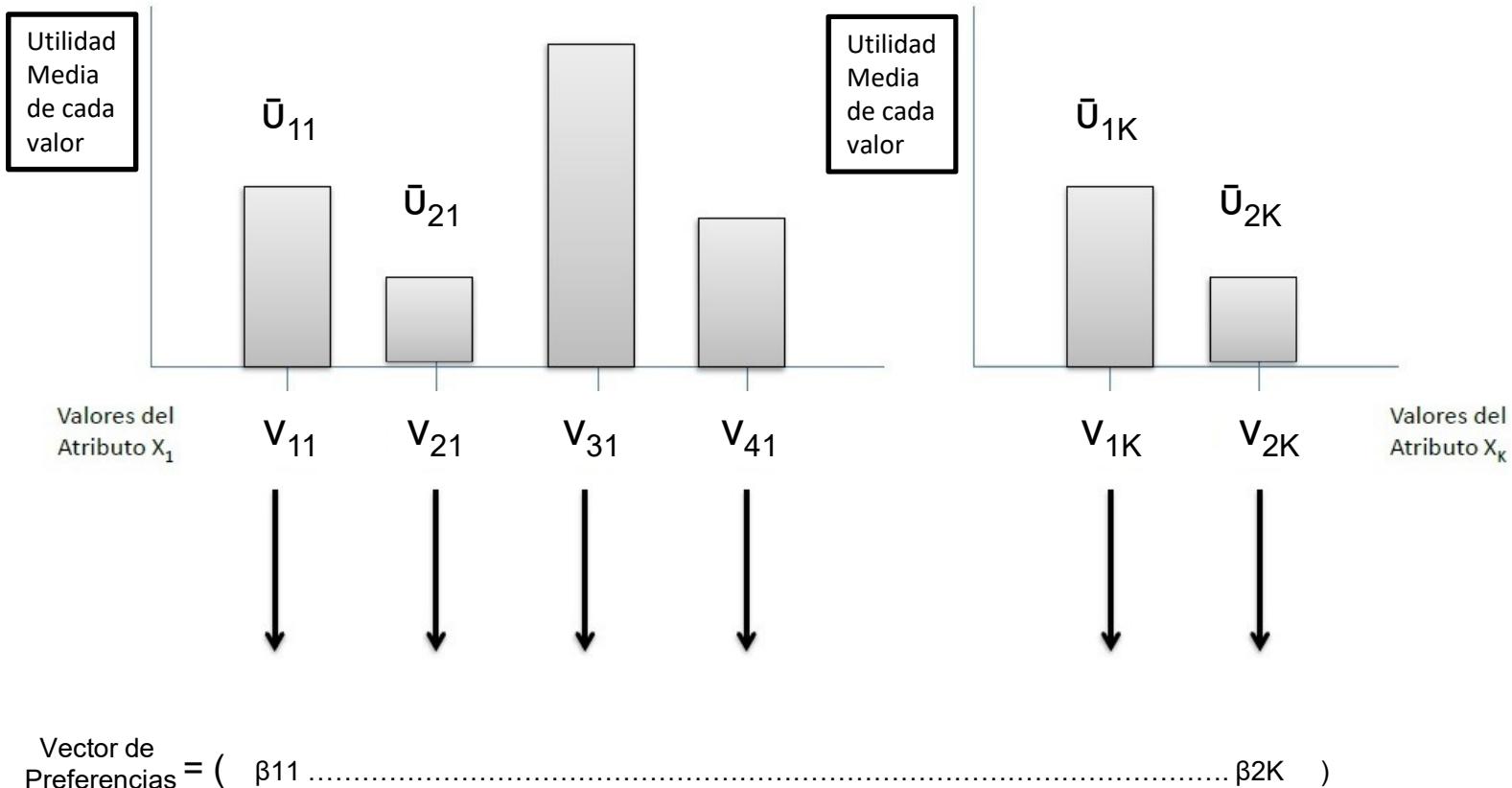
Cuestionario
para conocer
la utilidad

Técnica: Conjoint Analysis

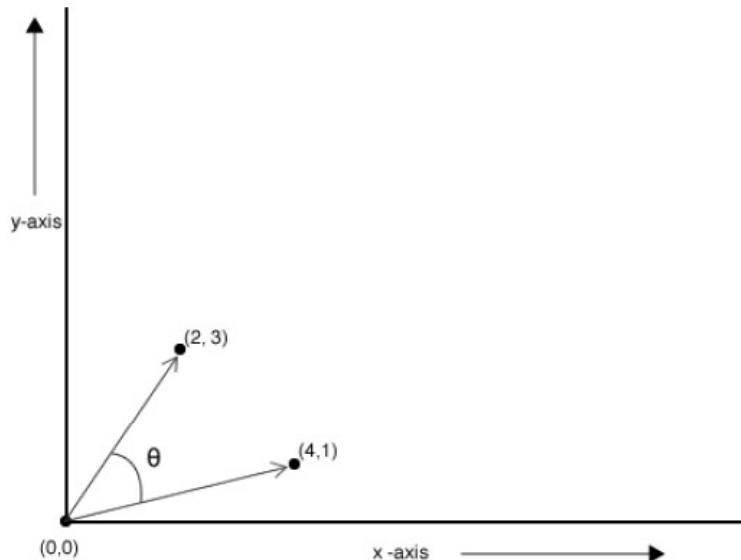
1

$$\text{Utilidad media de cada valor} = \frac{\sum \text{utilidad de la compra con ese valor}}{\text{Número de compras con ese valor}}$$

Aprendiendo las preferencias/perfil del usuario



Calculando la similaridad

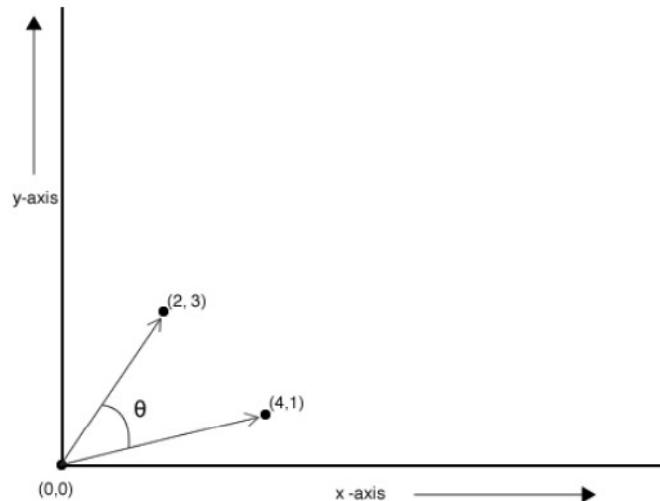


Medida del coseno:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

Similaridad(Perfil Usuario, Item) = Coseno(Vector Preferencias, Vector Item)

Calculando la similaridad



Distancia Euclídea entre dos vectores:

$$\delta(a, b) = \|a - b\| = \sqrt{(a - b)^T(a - b)} = \sqrt{\sum_{i=1}^m (a_i - b_i)^2}$$

Similaridad(Perfil Usuario, Item) = $1 / (1 + \text{distancia(Vector Preferencias, Vector Item)})$

Creando el ranking de los items

1. Asumimos: Utilidad = Similaridad
2. Creamos un ranking de los items ordenándolos de mayor a menor valor de utilidad

| Item | Utilidad |
|------|----------|
| I1 | 1 |
| I2 | 0,9 |
| i3 | 0,3 |
| ... | ... |

3. Generamos las recomendaciones: Seleccionamos los N primeros items (Top-N recommendations) .

Variaciones: contenido descrito con Keywords

- Most CB-recommendation techniques were applied to recommending text documents.
 - Like web pages or newsgroup messages for example.
- Content of items can also be represented as text documents.
 - With textual descriptions of their basic characteristics.
 - Structured: Each item is described by the same set of attributes



| Title | Genre | Author | Type | Price | Keywords |
|----------------------|-------------------|-------------------|-----------|-------|--|
| The Night of the Gun | Memoir | David Carr | Paperback | 29.90 | Press and journalism, drug addiction, personal memoirs, New York |
| The Lace Reader | Fiction, Mystery | Brunonia Barry | Hardcover | 49.90 | American contemporary fiction, detective, historical |
| Into the Fire | Romance, Suspense | Suzanne Brockmann | Hardcover | 45.90 | American fiction, murder, neo-Nazism |

- Unstructured: free-text description.

Variaciones: contenido descrito con Keywords

- Item representation

| Title | Genre | Author | Type | Price | Keywords |
|----------------------|-------------------|-------------------|-----------|-------|--|
| The Night of the Gun | Memoir | David Carr | Paperback | 29.90 | Press and journalism, drug addiction, personal memoirs, New York |
| The Lace Reader | Fiction, Mystery | Brunonia Barry | Hardcover | 49.90 | American contemporary fiction, detective, historical |
| Into the Fire | Romance, Suspense | Suzanne Brockmann | Hardcover | 45.90 | American fiction, murder, neo-Nazism |

- User profile

| Title | Genre | Author | Type | Price | Keywords |
|-------|---------|------------------------------|-----------|-------|-----------------------------|
| ... | Fiction | Brunonia, Barry, Ken Follett | Paperback | 25.65 | Detective, murder, New York |

- Simple approach

- Compute the similarity of an unseen item with the user profile based on the keyword overlap (e.g. using the Dice coefficient)



$$\frac{2 \times |\text{keywords}(b_i) \cap \text{keywords}(b_j)|}{|\text{keywords}(b_i)| + |\text{keywords}(b_j)|}$$

- Or use and combine multiple metrics

$\text{keywords}(b_j)$
describes Book b_j
with a set of
keywords



Conexión con teorías de toma de decisiones

- ▶ En la teoría de la elección racional, la utilidad es una función lineal sobre los valores de los atributos. Dado individuo c y un ítem a:

$$u(c, a) = \sum_k \beta_{c,k} x_{a,k}$$

Con x indicando los valores del atributo k del ítem a

Con b indicando la preferencia del individuo c sobre x

Con K indicando el cto de todos los valores de los atributos de a

- ▶ En la estrategia basada en contenido, la utilidad se puede calcular a través de la similaridad, y ésta a través de la medida del coseno:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

- ▶ Por tanto:

utilidad(estrategia_basada_contenido) = utilidad(eleccion_racional) normalizada

- ▶ ¿Qué significa esta conexión?

La estrategia de recomendación basada en contenido asume que la toma de decisiones de los humanos es de tipo racional

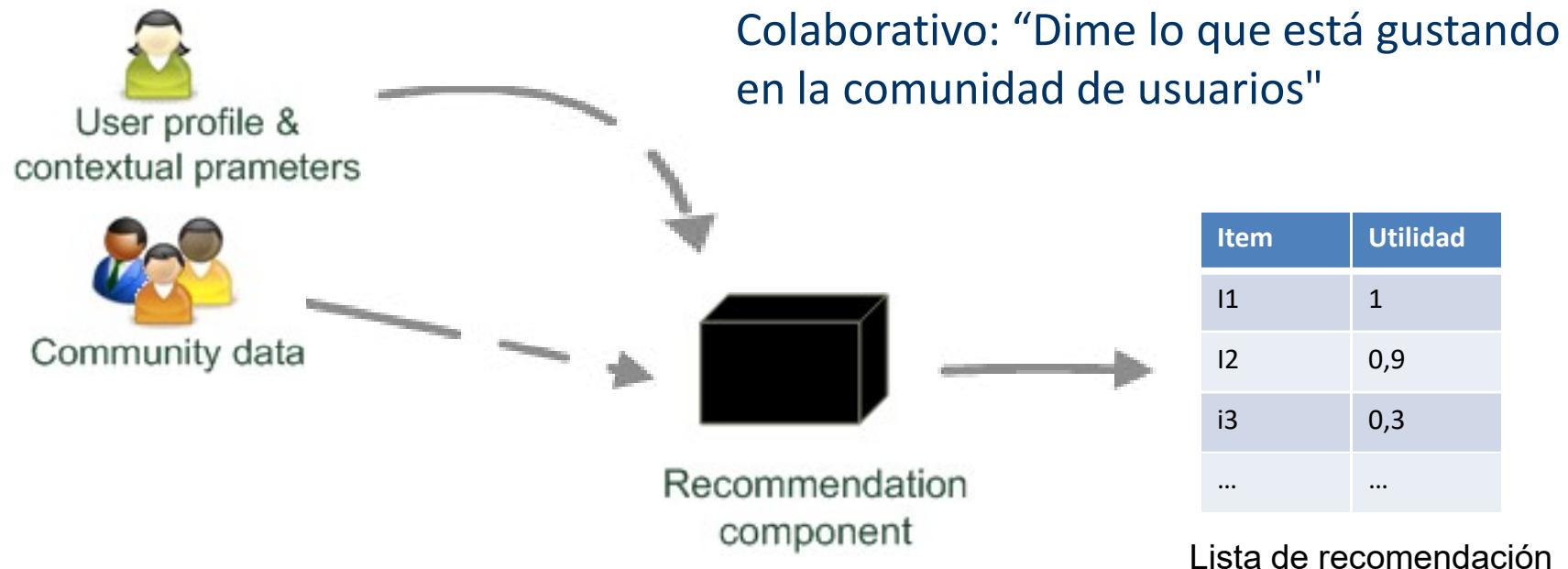
Recomendación basada en contenido

- Resumen de la estrategia:
 1. Representar los items en formato vector.
 2. Representar las preferencias del usuario en formato vector
 3. Calcular la similaridad.
 4. Asumir utilidad = similaridad.
 5. Crear un ranking de items basados en el valor de utilidad
 6. Recomendar los N primeros.

Limitaciones:

1. **Para el usuario:** Poca originalidad de las recomendaciones. Las recomendaciones suelen ser productos muy similares a los ya consumidos por el usuario.
2. **Para el ingeniero/científico:** Necesidad de conocer en detalle el dominio de la aplicación: productos, atributos y valores
3. **En la satisfacción con la recomendación:** Los valores de los atributos no aportan información acerca de la calidad del producto. Estimar la utilidad de un producto solamente a través de sus atributos no garantiza la satisfacción de la recomendación.
4. **En la hipótesis/supuestos en los que se basa la estrategia:** Asume que la toma de decisiones de los humanos se basa en una estrategia racional. ¿Es esto correcto?

Recomendación colaborativa



Recomendación colaborativa

- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

Recomendación colaborativa

- Objetivo:

Recomendar items al usuario-objetivo en base a las valoraciones/ratings de otros usuarios sobre los items disponibles.

- ¿Qué necesitamos?:

1. **Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
2. **Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
3. **Asumir utilidad = valoración predicha.**

Recomendación colaborativa

- Es la estrategia de recomendación más popular
 - Utilizada por grandes plataformas comerciales de e-commerce
 - Aplicable en cualquier ámbito ya que no depende del contenido (libros, películas, música ..)
- Aproximación
 - Utiliza la "wisdom of the crowds" para predecir las valoraciones
- Supuestos
 - Los usuarios mantienen sus gustos constantes.
 - Los usuarios que valoran de forma similar tienen gustos similares



Basada en usuarios

- Objetivo:

Recomendar items al usuario-objetivo en base a las valoraciones aportadas por usuarios similares a él.

- ¿Qué necesitamos?:

1. **Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
2. **Estimar la similaridad entre usuarios.** La similaridad entre el usuario-objetivo y otro usuario que ha valorado el item/alternativa se calcula comparando si las valoraciones realizadas por ambos sobre los mismos items son similares o no. *Se asume que dos usuarios tienen gustos similares si sus valoraciones también son similares.*
3. **Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
4. **Asumir utilidad = valoración predicha.**

Basada en usuarios

- Datos que se necesitan
 - Una matriz de valoraciones usuario-item

| | | Items | | | | | |
|-------|-----|-------|---|-----|-----|-----|-----|
| | | 1 | 2 | ... | i | ... | m |
| Users | 1 | 5 | 3 | | 1 | 2 | |
| | 2 | | 2 | | | | 4 |
| | : | | | 5 | | | |
| | u | 3 | 4 | | 2 | 1 | |
| | : | | | | | 4 | |
| | n | | | 3 | 2 | | |

 a

| | | | | | |
|---|---|--|---|---|--|
| 3 | 5 | | ? | 1 | |
|---|---|--|---|---|--|

Basada en usuarios

- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

Basada en usuarios

- Calculando la similaridad entre usuarios

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

- A popular similarity measure in user-based CF: Pearson correlation

a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

- Possible similarity values between -1 and 1

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

- Ejercicio: Calcula la similaridad utilizando la correlación de Pearson y los datos anteriores

Basada en usuarios

- A popular similarity measure in user-based CF: Pearson correlation

a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

- Possible similarity values between -1 and 1

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

sim = 0,85

sim = 0,70

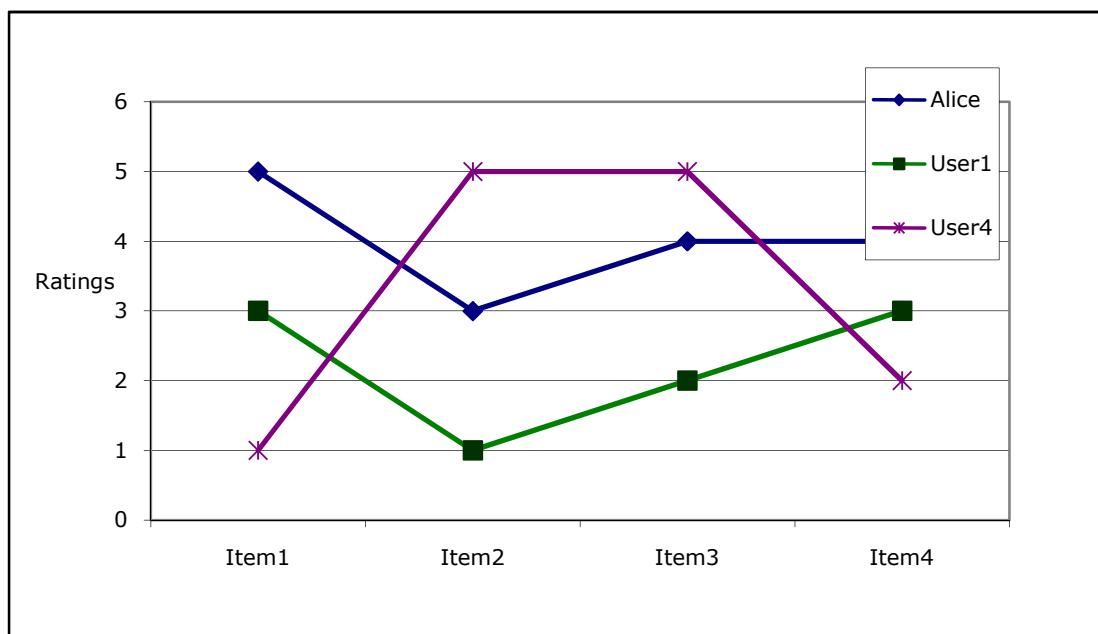
sim = 0

sim = -0,79



Basada en usuarios

- Ejercicio: ¿Otras formas de calcular la similaridad?
- Compara los resultados con la similaridad obtenida por correlación de Pearson



- Predicción de la valoración con una media ponderada

$$p_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

con $p_{a,i}$ la predicción del usuario-objetivo i sobre el ítem a

con $r_{a,j}$ la valoración del usuario j sobre el ítem a

y con $w_{i,j}$ la similaridad entre el usuario-objetivo i y el usuario j

- A common prediction function:

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$



- Calculate, whether the neighbors' ratings for the unseen item i are higher or lower than their average
- Combine the rating differences – use the similarity with a as a weight
- Add/subtract the neighbors' bias from the active user's average and use this as a prediction

- No todas las valoraciones aportan la misma información
 - Coincidir en la valoración de items con opiniones diversas es más informativo que coincidir en la valoración de items que gustan a todos.
 - **Possible solución:** Dar más peso en la predicción a los items con mayor varianza en sus valoraciones
- Valorar el número de items valorados conjuntamente entre dos usuarios
 - El peso de un usuario en la predicción debe estar correlacionado con el número de items valorados conjuntamente.
- Selección del vecindario (número de usuarios similares)
 - Utilizar un umbral mínimo de similaridad para seleccionar el número de usuarios similares

Basada en items

- Objetivo:

Recomendar items al usuario-objetivo en base a las valoraciones realizadas sobre otros items similares a los items a recomendar.

- ¿Qué necesitamos?:

1. **Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
2. **Estimar la similaridad entre items.** La similaridad entre el item-objetivo y otro item valorado por los usuarios se calcula comparando si las valoraciones realizadas por ambos usuarios son similares o no. *Se asume que dos items tienen características similares si sus valoraciones también son similares.*
3. **Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
4. **Asumir utilidad = valoración predicha.**

Basada en items

- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

- A popular similarity measure in user-based CF: Pearson correlation

a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

- Possible similarity values between -1 and 1

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

- Produces better results in item-to-item filtering
- Ratings are seen as vector in n-dimensional space
- Similarity is calculated based on the angle between the vectors

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$



- Adjusted cosine similarity
 - take average user ratings into account, transform the original ratings
 - U : set of users who have rated both items a and b

- Predicción de la valoración con una media ponderada

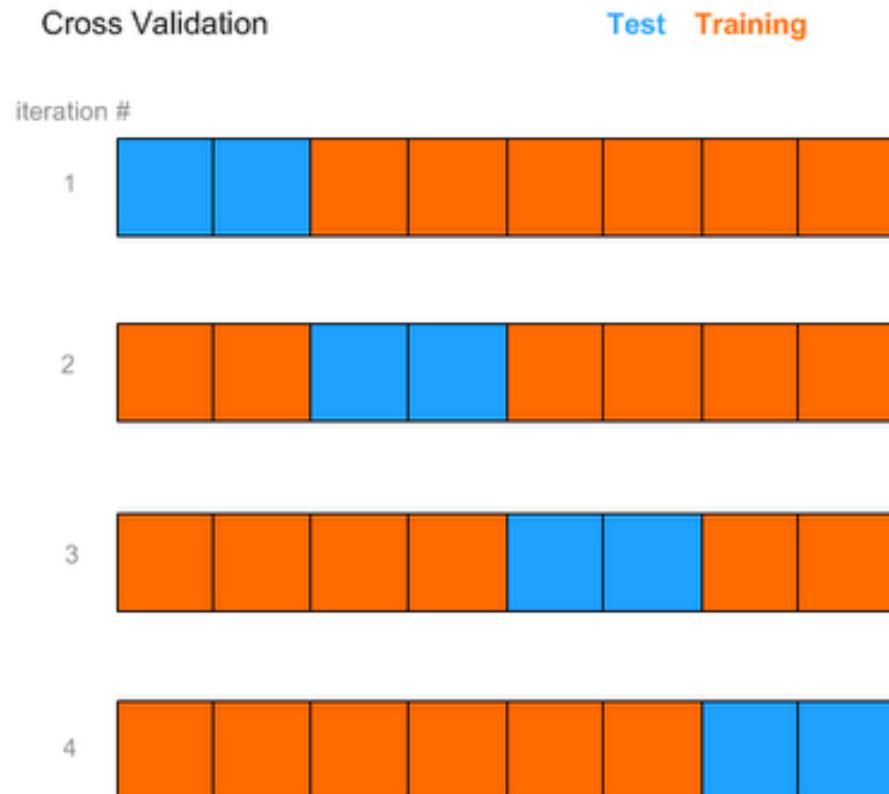
$$p_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

con $p_{a,i}$ la predicción del usuario-objetivo i sobre el item a

con $r_{a,j}$ la valoración del usuario j sobre el item a

y con $w_{i,j}$ la similaridad entre el usuario-objetivo i y el usuario j

Validación de algoritmos



1. Validación cruzada con k subgrupos
2. Validación cruzada con 2 subgrupos
3. Validación dejando una instancia fuera (Leave one out)

Validación de algoritmos

MEAN SQUARED ERROR
(Error cuadrático medio)

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$$

MEAN ABSOLUTE ERROR
(Error absoluto medio)

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|.$$

Evaluación de recomendaciones Top-N



| Items Recomendados | Items que realmente han gustado (o que han sido comprados) por el usuario |
|--------------------|---|
| i2 | i2 |
| i5 | i5 |
| i8 | i8 |
| i10 | i7 |

Comparación entre predicción y realidad en un catálogo de 10 items

Evaluación de recomendaciones Top-N

Table 2: 2x2 confusion matrix

| actual / predicted | negative | positive |
|--------------------|----------|----------|
| negative | a | b |
| positive | c | d |

Matriz de confusión para el ejemplo

| Real/Predicción | Negativo | Positivo |
|-----------------|----------|----------|
| Negativo | 5 | 1 |
| Positivo | 1 | 3 |

Evaluación de recomendaciones Top-N

$$Accuracy = \frac{\text{correct recommendations}}{\text{total possible recommendations}} = \frac{a + d}{a + b + c + d}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\epsilon_i| = \frac{b + c}{a + b + c + d},$$

$$Precision = \frac{\text{correctly recommended items}}{\text{total recommended items}} = \frac{d}{b + d}$$

$$Recall = \frac{\text{correctly recommended items}}{\text{total useful recommendations}} = \frac{d}{c + d}$$

Métricas de rendimiento para el ejemplo

| Accuracy | MAE | Precision | Recall |
|-----------|----------|-----------|----------|
| 8/10= 0,8 | 2/10=0,2 | 3/4=0,75 | 3/4=0,75 |

Problemas: la calidad de los datos

| Training Set | | | |
|--------------|----------|-----|-------|
| Id | Status | Age | Class |
| 1 | Single | 20 | Bad |
| 2 | Single | 30 | Good |
| 3 | Single | 50 | Bad |
| 4 | Single | 60 | Good |
| 5 | Married | 20 | Good |
| 6 | Married | 30 | Good |
| 7 | Married | 40 | Good |
| 8 | Married | 50 | Good |
| 9 | Divorced | 40 | Bad |
| 10 | Divorced | 60 | Good |

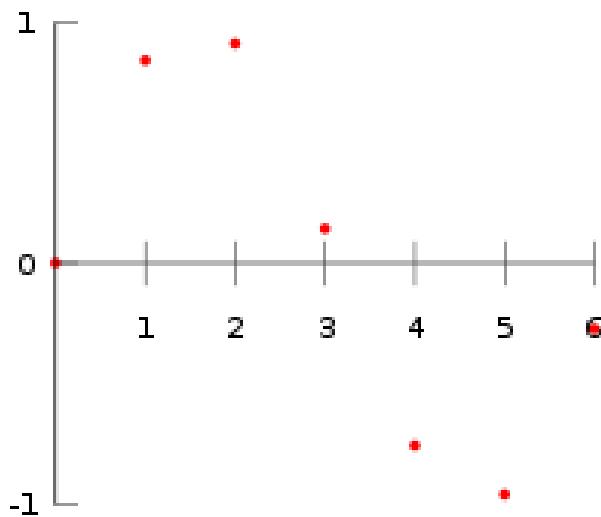
| Testing Set | | | |
|-------------|----------|----|--------|
| 11 | Single | 40 | (Bad) |
| 12 | Married | 60 | (Good) |
| 13 | Divorced | 20 | (Bad) |
| 14 | Divorced | 30 | (Bad) |
| 15 | Divorced | 50 | (Good) |

1. Exactitud de los datos?

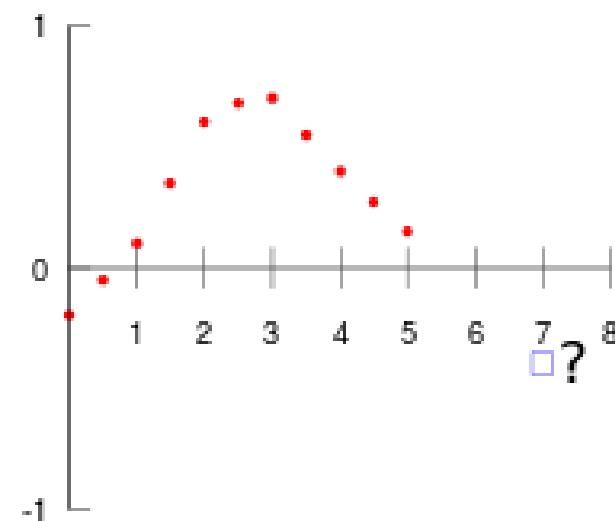
2. Imparcialidad del Testing Set Respecto al Training Set?

Problemas: La calidad de los datos

3. Completitud de los datos



Interpolación:
Situación deseable



Extrapolación:
Problema complicado

Sistemas predictivos: Sistemas de Recomendación



Grupo de Sistemas *inteligentes*



Eduardo M. Sánchez Vila
eduardo.sanchez.vila@usc.es

CITIUS
Grupo de Sistemas Inteligentes
Universidad de Santiago de Compostela

Recomendaciones: Todo es un Ensemble



Netflix prize was won by an ensemble of 107 models (2007):

- Matrix factorization
- Restricted Boltzman Machines
- k-NN
- Regression models

The BellKor solution to the Netflix Prize

Robert M. Bell, Yehuda Koren and Chris Volinsky
AT&T Labs – Research
BellKor@research.att.com

Our final solution ($\text{RMSE}=0.8712$) consists of blending 107 individual results. Since many of these results are close variants, we first describe the main approaches behind them. Then, we will move to describing each individual result.

The core components of the solution are published in our ICDM'2007 paper [1] (or, KDD-Cup'2007 paper [2]), and also in the earlier KDD'2007 paper [3]. We assume that the reader is familiar with these works and our terminology there.

Problemas con las Recomendaciones

The screenshot shows the TripAdvisor website for Santiago de Compostela, Spain. The top navigation bar includes links for Opinion, Unirse, Iniciar Sesión, and currency conversion. The main search bar shows "Encuentra: Qué hacer" and "Cerca de: Santiago de Compostel ...". Below the search bar, the breadcrumb navigation shows: Europa > España > Galicia > Provincia de A Coruña > Santiago de Compostela > Qué hacer en Santiago de Compostela.

Cosas que hacer en Santiago de Compostela

Ordenar por: Clasificación Reservar en línea

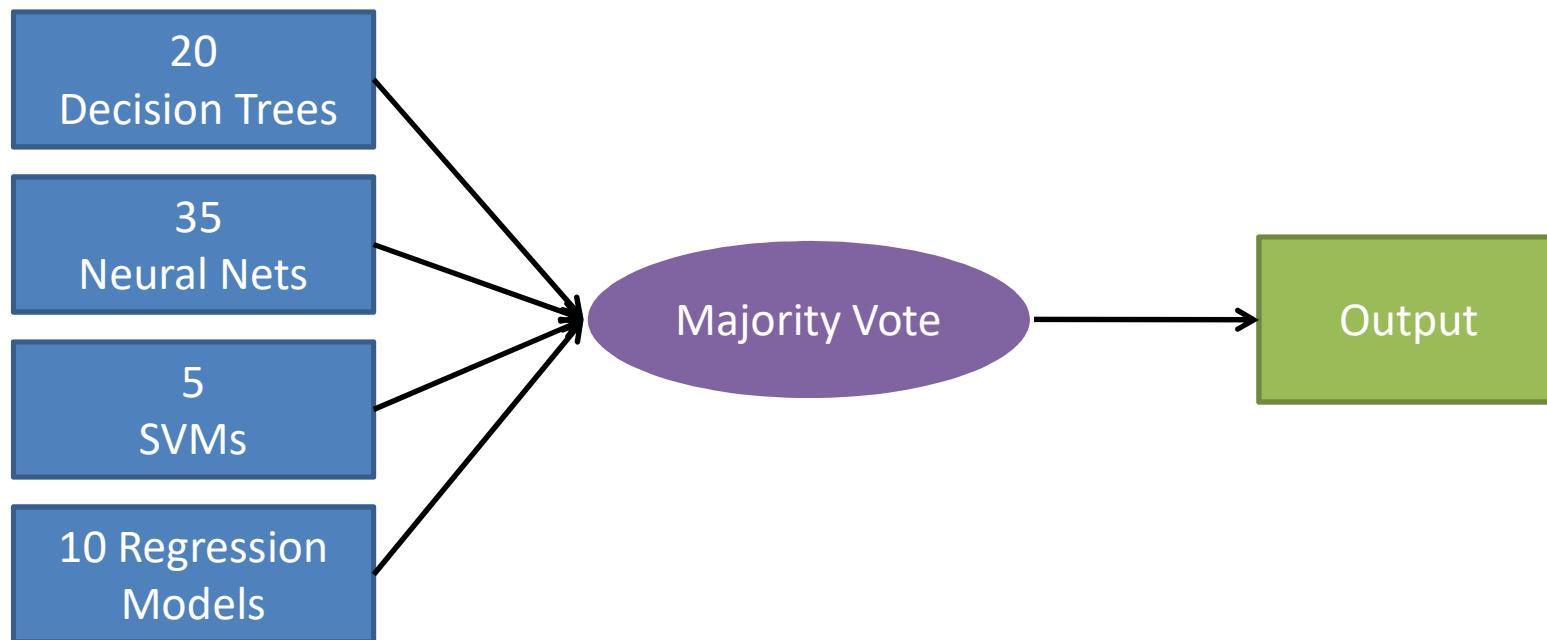
| Atracción | Opiniones | Precio | Opciones |
|---|-----------------|---------------------|--------------------------|
| Catedral de Santiago de Compostela | 6.250 opiniones | desde EUR 42,09 €* | Detalles |
| Botafumeiro | 1.620 opiniones | desde EUR 560,00 €* | Detalles |
| Casco histórico de Santiago de Compostela | 1.153 opiniones | desde EUR 380,00 €* | Detalles |

TIPO DE ATRACCIÓN:

- Monumentos y puntos de interés (39)
- Visitas guiadas (26)
- Museos (10)
- Naturaleza y parques (7)
- Compras (12)
- Vida nocturna (54)
- Conciertos y espectáculos (1)

“Users do not trust explicit recommendations”
(2016)

Problemas con las Recomendaciones



¡¡Cómo lo explicamos!!

Fundamentos teóricos

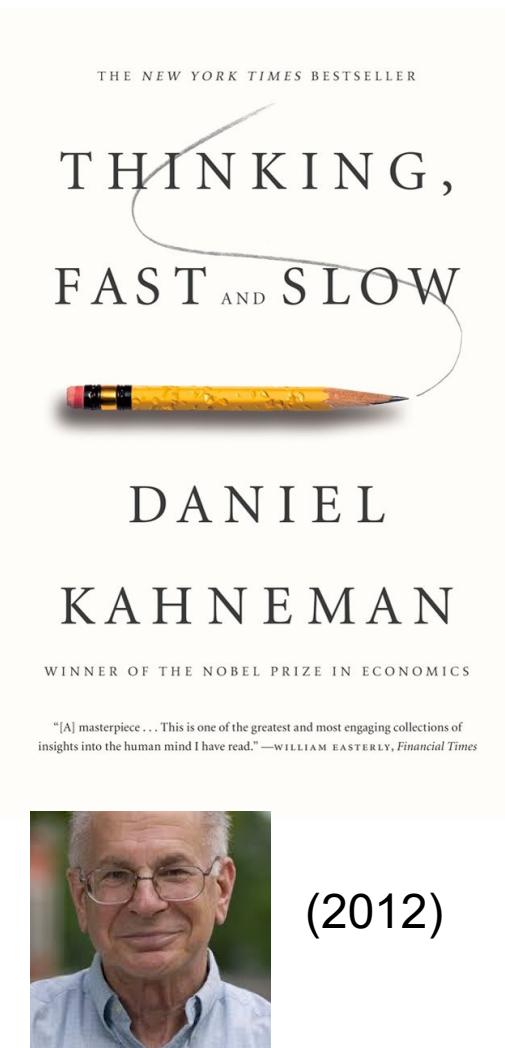
Más fácil con una teoría

Nuestra mente tiene dos sistemas:

- Sistema I: Rápido, automático, poco esfuerzo, sin control consciente.
- Sistema II: Lento, racional, costoso, con control consciente.

Psicólogos ante la teoría económica:

- Econos Vs. Humanos.
- El poder de la multidisciplinariedad



Tomando decisiones



Referencia: Brain Games

| Economist.com | SUBSCRIPTIONS |
|----------------------|---------------|
| OPINION | |
| WORLD | |
| BUSINESS | |
| FINANCE & ECONOMICS | |
| SCIENCE & TECHNOLOGY | |
| PEOPLE | |
| BOOKS & ARTS | |
| MARKETS & DATA | |
| DIVERSIONS | |

Welcome to
The Economist Subscription Centre

Pick the type of subscription you want to buy or renew.

Economist.com subscription - US \$59.00
One-year subscription to Economist.com.
Includes online access to all articles from *The Economist* since 1997.

Print subscription - US \$125.00
One-year subscription to the print edition of *The Economist*.

Print & web subscription - US \$125.00
One-year subscription to the print edition of *The Economist* and online access to all articles from *The Economist* since 1997.

Referencia: (Ariely, 2010)

¿Qué es la toma de decisiones?

Definición 1:

The thought process of selecting a logical choice from the available options.

When trying to make a good decision, a person must weight the positives and negatives of each option, and consider all the alternatives. For effective decision making, a person must be able to forecast the outcome of each option as well, and based on all these items, determine which option is the best for that particular situation

Fuente: <http://www.businessdictionary.com/>

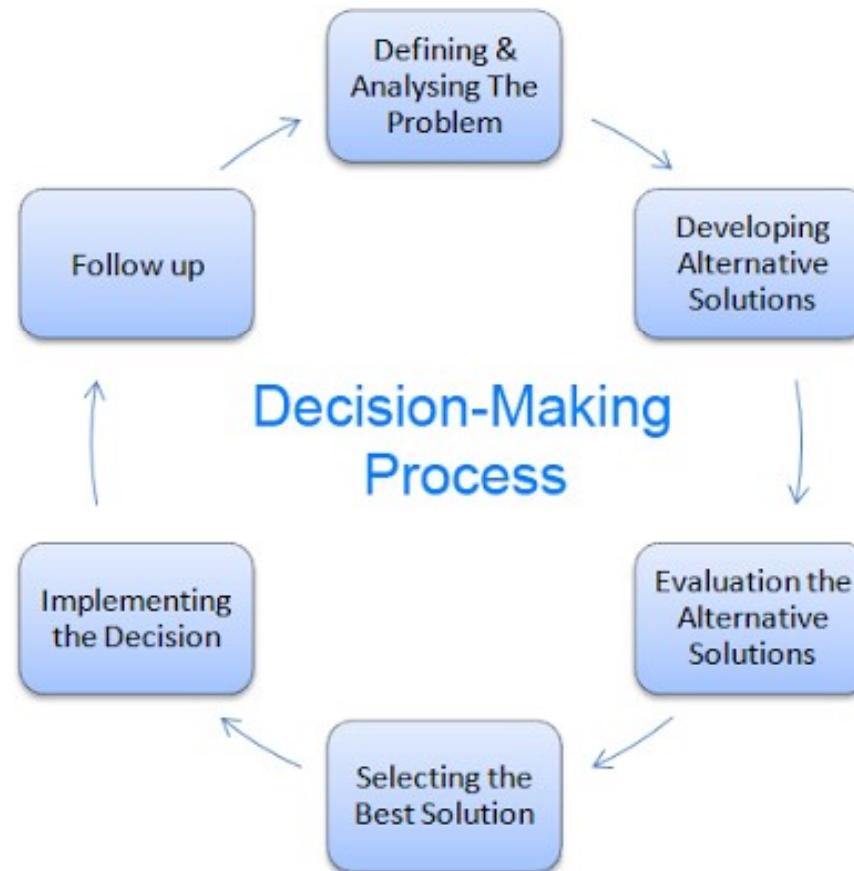
Definición 2:

Decision-making can be regarded as a problem-solving activity terminated by a solution deemed to be optimal, or at least satisfactory. It is therefore a process which can be more or less rational or irrational and can be based on explicit or tacit knowledge and beliefs.

Fuente: <https://en.wikipedia.org/wiki/Decision-making>

Decisiones racionales

Teoría de elección racional: el proceso



Fase I: Definiendo el problema



Tipo de problema: Informal

Tipo de problema: Serio



Ejemplo: ¿Tipo de problema?



2 euros

3.25 euros

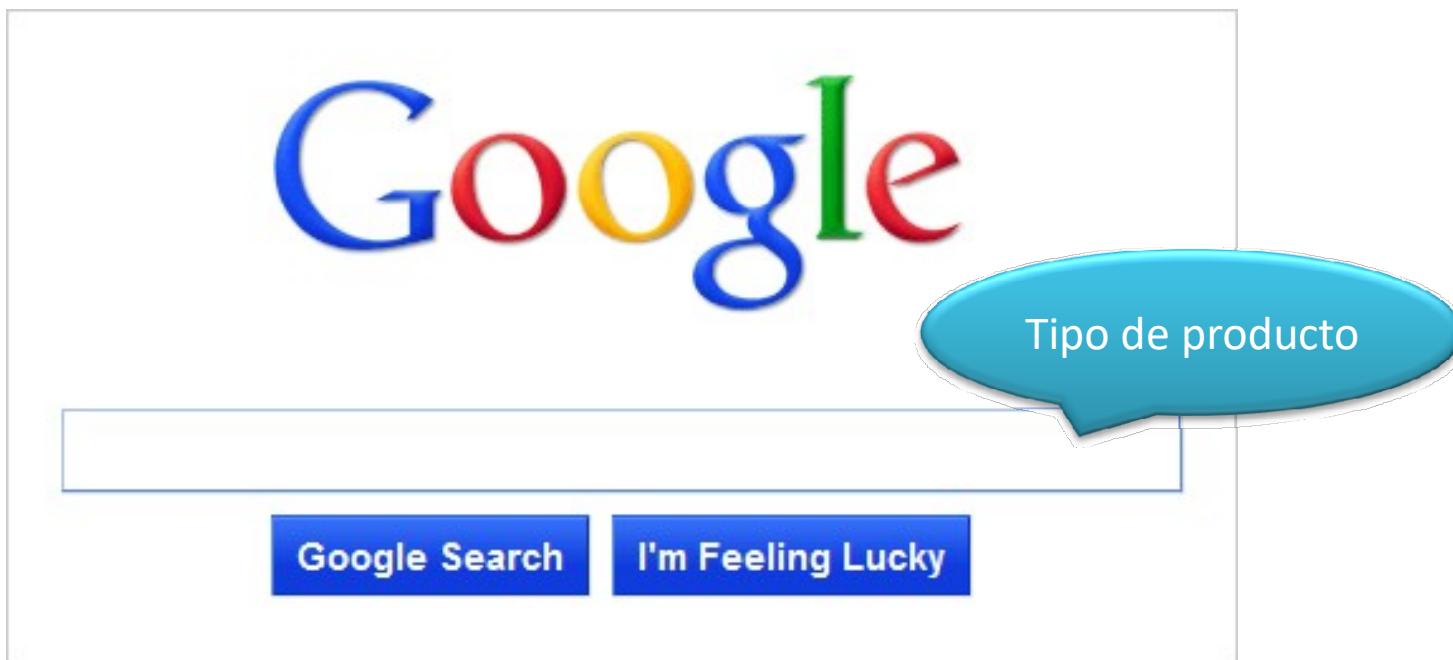
3.30 euros

Teoría de elección racional: Fase II

Fuentes de información:
Mi memoria



Fuentes de información: Buscadores



Fuentes de información: Tiendas online o tiendas físicas



Fuentes de información: Mi red social

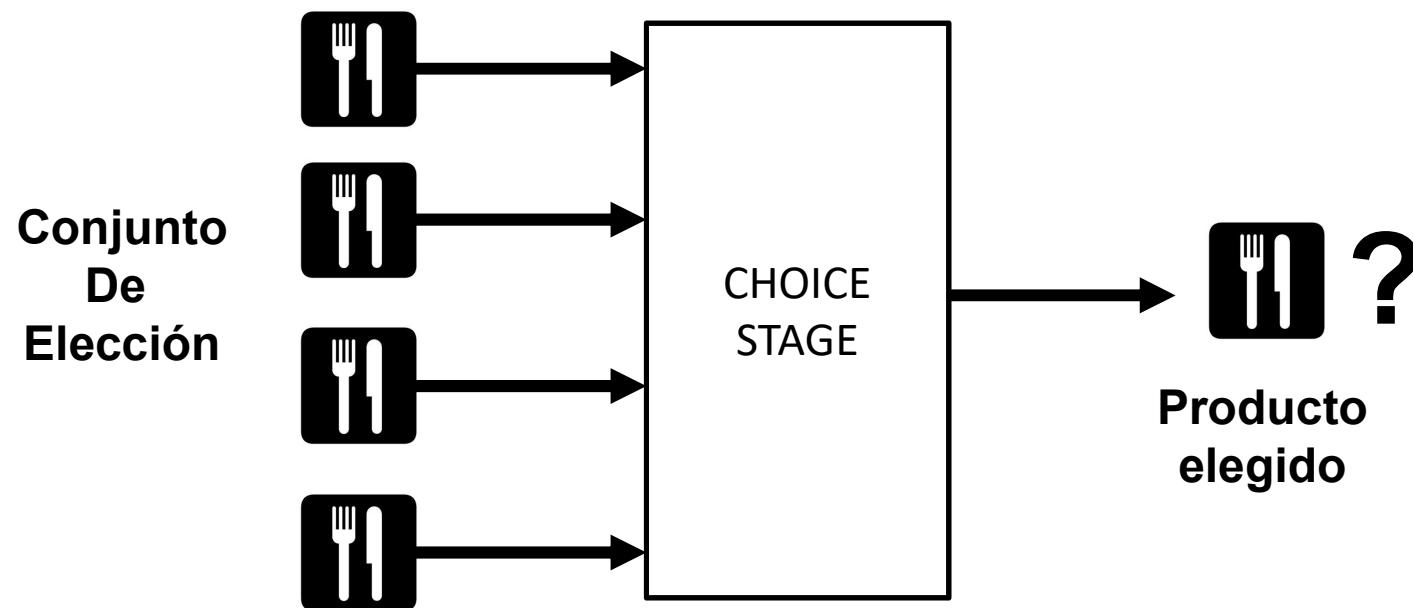


Teoría de elección racional: Fase III

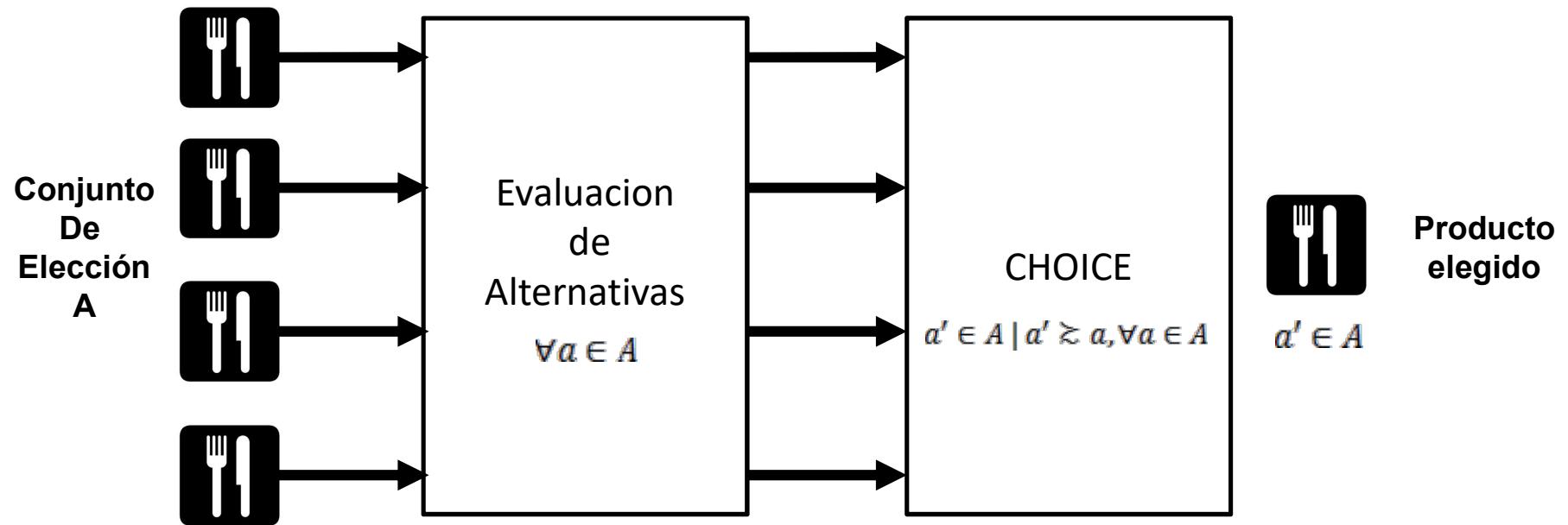
El Conjunto de Elección



Fase III y IV: El problema de la Elección



Fase III y IV: El problema de la Elección



Regla de Elección con Preferencias

$$RE(A, \geq) = \{a' \in A \mid a' \geq a, \forall a \in A\}$$

Regla de Elección con Preferencias

$$RE(A, \succsim) = \{a' \in A \mid a' \succsim a, \forall a \in A\}$$

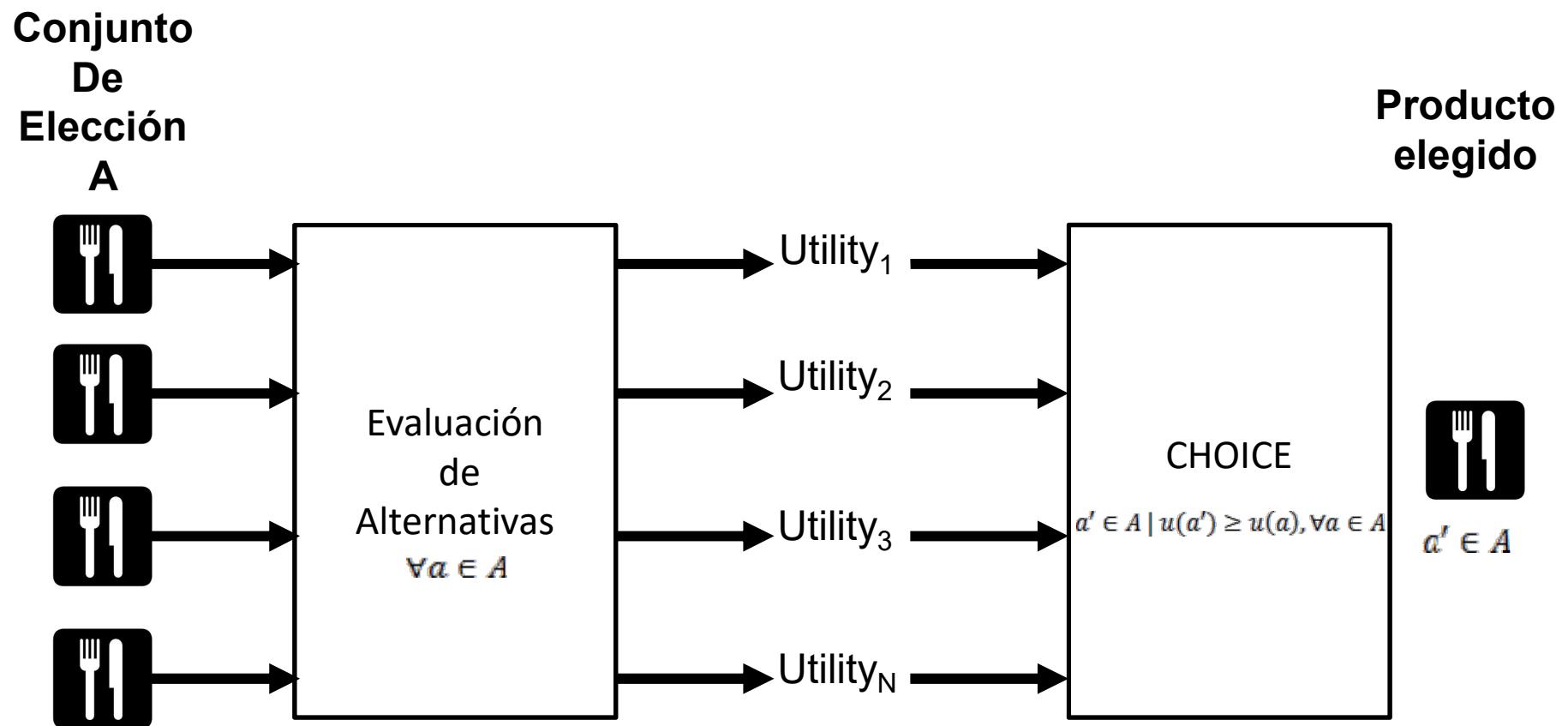
Axioma de la Teoría de la Utilidad

$$a \succsim b \Leftrightarrow u(a) \geq u(b)$$

Regla de Elección con Utilidades

$$RE(A, \geq) = \{a' \in A \mid u(a') \geq u(a), \forall a \in A\}$$

Teoría de la Utilidad



- ▶ ¿Qué es la Utilidad?:
 1. Perspectiva de Marshall:
 - Funcionalidad de una alternativa.
 - Se estima comparando los atributos de una alternativa con las preferencias del sujeto.
 - Concepto de Utilidad estándar.
 2. Perspectiva de Bentham:
 - Satisfacción que se espera obtener de una alternativa.
 - Concepto original de Utilidad.

- ▶ ¿Cómo construimos la función de utilidad?:
 1. Identificar los atributos relevantes de las alternativas
 2. Identificar los valores de los atributos anteriores
 3. Estimar las preferencias del individuo sobre los valores de los atributos.
 4. Matemáticamente: Función lineal sobre los valores de los atributos. Dado individuo c y alternativa a:

$$u(c,a) = \sum_k \beta_{c,k} x_{a,k}$$

Con K indicando el cto de valores de todos los atributos de a.

Teoría de la Utilidad

Ejercicio: ¿Cómo comparas entre estas alternativas?



Centro ciudad
20 mins de playa
100 m²
3 habitaciones
210.000 euros

10 min coche del centro
15 mins de playa
160 m²
4 habitaciones
250.000 euros

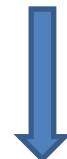
20 min coche del centro
Pie de playa
90 m²
3 habitaciones
150.000 euros

Teoría de la Utilidad

Estimate utilities

| Alternatives | Short Distance | Medium Distance | Long Distance | Small Size | Big Size | Low Price | High Price |
|----------------|----------------|-----------------|---------------|------------|----------|-----------|------------|
| $A_1 - X(A_1)$ | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $A_2 - X(A_2)$ | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| $A_3 - X(A_3)$ | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

| Preferences | Short Distance | Medium Distance | Long Distance | Small Size | Big Size | Low Price | High Price |
|--------------|----------------|-----------------|---------------|------------|----------|-----------|------------|
| My prefs - B | 6 | 3 | 1 | 2 | 8 | 6 | 4 |



$$u(c, a) = \sum_k \beta_{c,k} x_{a,k}$$

Teoría de la Utilidad



Building a table

| Alternatives | Utility | MU | Chosen Alternative |
|----------------|---------|----|--------------------|
| A ₁ | 12 | | |
| A ₂ | 15 | 15 | A ₂ |
| A ₃ | 9 | | |

Conexión con estrategias de recomendación

- ▶ En la teoría de la elección racional, la utilidad es una función lineal sobre los valores de los atributos. Dado individuo c y un ítem a:

$$u(c, a) = \sum_k \beta_{c,k} x_{a,k}$$

Con x indicando los valores del atributo k del ítem a

Con b indicando la preferencia del individuo c sobre x

Con K indicando el cto de todos los valores de los atributos de a

- ▶ En la estrategia basada en contenido, la utilidad se puede calcular a través de la similaridad, y ésta a través de la medida del coseno:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

- ▶ Por tanto:

utilidad(estrategia_basada_contenido) = utilidad(eleccion_racional) normalizada

- ▶ ¿Qué significa esta conexión?

La estrategia de recomendación basada en contenido asume que la toma de decisiones de los humanos es de tipo racional

Teoría de elección racional: Fase V

Fase V: Experiencia con el producto



Teoría de elección racional: Fase V

Fase V: Experiencia con el producto

ACME Restaurant Customer Survey

Thank you in advance for taking the ACME Restaurant survey. Please cross the box which is most relevant to your experience.

How satisfied are you with your experience at the ACME restaurant today?

| | Very dissatisfied 1 | - | - | - | - | Very satisfied 10 | | | | |
|-----------------------------------|---------------------------|---|---|---|---|-------------------------|---|---|---|----|
| The cleanliness of the restaurant | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Ease of booking a table | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The decor of the restaurant | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The breadth of the food menu | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The breadth of the wine menu | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Catering for special diets | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Quality of your eating experience | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Quality of service from staff | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The speed of service | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

How important or unimportant are the following requirements for you when visiting the ACME Restaurant?

| | Totally unimportant 1 | - | - | - | - | Very important 10 | | | | |
|-----------------------------------|-----------------------------|---|---|---|---|-------------------------|---|---|---|----|
| The cleanliness of the restaurant | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Ease of booking a table | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The decor of the restaurant | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The breadth of the food menu | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The breadth of the wine menu | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Catering for special diets | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Quality of your eating experience | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Quality of service from staff | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| The speed of service | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Predicción: el problema del científico

Ejercicio: ¿Con qué atributos representarías estas alternativas?
¿Cómo harías la predicción considerando TUS preferencias?



El problema de la incertidumbre

Problema: ¿Hay completa seguridad de predecir el resultado/utilidad una vez tomada una decisión? ¿Es posible que no siempre se obtenga el mismo resultado/utilidad?

Ejemplo de decisión con incertidumbre: ¿Qué alternativa prefieres?

Opción A:

70% probabilidad de ganar 1.000 euros
30% probabilidad de no ganar nada

Opción B:

50% probabilidad de ganar 500 euros
50% probabilidad de ganar 200 euros.

Opción A:

33% probabilidad de ganar 2.500 euros
67% probabilidad de no ganar nada

Opción B:

34% probabilidad de ganar 2.400 euros
66% probabilidad de no ganar nada.

Teoría de la Utilidad Esperada

Problema: ¿Hay completa seguridad de predecir el resultado/utilidad una vez tomada una decisión? ¿Es posible que no siempre se obtenga el mismo resultado/utilidad?

Solución: *Teoría de la Utilidad Esperada*

Principio MEU (Maximum Expected Utility):

$$RE(A, \geq) = \{a' \in A \mid \hat{u}(a') \geq \hat{u}(a), \forall a \in A\} \Leftrightarrow a' = \arg \max_{a \in A} \hat{u}(c, a)$$

Con c indicando el usuario, a una alternativa, y \hat{u} la utilidad esperada (expected Utility).

La utilidad esperada \hat{u} se calcula en base a los posibles resultados de una Decision. Cada resultado r tiene una probabilidad de ocurrencia y una utilidad asociada. Por tanto, la utilidad esperada \hat{u} asociada a la elección de la alternativa a se calcula:

$$\hat{u}(c, a) = \sum_r P(Result = r) \hat{u}(c, r)$$

Teoría de la Utilidad Esperada



Ejercicio: ¿Qué alternativa prefieres?

Opción A:

70% probabilidad de ganar 1.000 euros
30% probabilidad de no ganar nada

Opción B:

50% probabilidad de ganar 500 euros
50% probabilidad de ganar 200 euros.

Opción A:

33% probabilidad de ganar 2.500 euros
67% probabilidad de no ganar nada

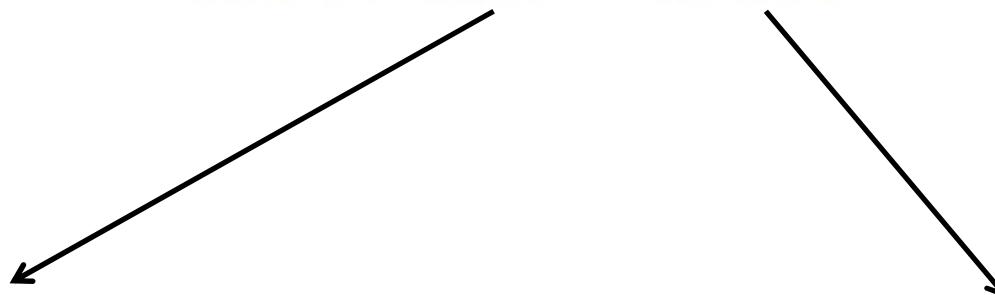
Opción B:

34% probabilidad de ganar 2.400 euros
66% probabilidad de no ganar nada.

Predicción: el problema del científico

Problema: En muchas situaciones, el científico no conoce directamente los valores de $P(\text{Result}=r)$ y/o de utilidad esperada \hat{u} :

$$\hat{u}(c, a) = \sum_r P(\text{Result} = r) \hat{u}(c, r)$$



¿Y si esta probabilidad no es constante, sino que puede variar si tenemos en cuenta evidencias disponibles del problema?

¿Y si esta utilidad depende de factores del usuario que el científico desconoce?

Estimando probabilidades con datos



Problema: ¿Y qué pasa si la probabilidad $P(\text{Result}=r)$ está condicionada por evidencias o factores externos? Es decir, tenemos que estimar $P(\text{Result}=r | \text{Evidencias})$.

Ejercicio:

Un taxi es el responsable de atropellar a un peatón de noche, dándose a la fuga tras el incidente. Dos compañías de taxi, la Verde y la Azul, son las que operan en la ciudad.

Te dan los siguientes datos:

1. El 85% de los taxis en la ciudad son Verdes y el 15% de ellos son Azules.
2. Un testigo identificó el taxi como un taxi Azul. En la investigación del suceso se ha comprobado la precisión del testigo en circunstancias similares a la de aquella noche, y se concluyó que el testigo identifica correctamente cada uno de los dos colores el 80% de los casos, y falla el 20% restante.

Cual es la probabilidad de que el taxi responsable del accidente sea Verde o Azul?

Problema: ¿Y qué pasa si la probabilidad $P(\text{Result}=r)$ está condicionada por el contexto o por otros factores externos? Es decir, tenemos que estimar $P(\text{Result}=r | \text{Evidencias})$.

Solución: *Teorema de Bayes e inferencia bayesiana*

Teorema de Bayes:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Teorema de Bayes

¿De dónde viene el Teorema de Bayes?:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A|B) P(B) = P(A \cap B) = P(B|A) P(A)$$

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Interpretación del Teorema de Bayes con Hipótesis y Evidencias:

$$P(H/E) = \frac{P(E/H) \cdot P(H)}{P(E)}$$

where $P(H)$ = the previous or *a priori* probability that the hypothesis is true

$P(E)$ = the probability that an event will occur

$P(E/H)$ = the probability that the event will occur given that the hypothesis is true

Y el denominador es un factor de normalización que se calcula:

$$P(E) = \sum_i P(E|H_i) \cdot P(H_i)$$

Metodología: (1) Identificar las hipótesis, (2) Identificar las evidencias, (3) Identificar las probabilidades a priori, (4) Identificar las verosimilitudes, (5) Normalizar la ecuación

Ejercicio:

Un taxi es el responsable de atropellar a un peatón de noche, dándose a la fuga tras el incidente. Dos compañías de taxi, la Verde y la Azul, son las que operan en la ciudad.

Te dan los siguientes datos:

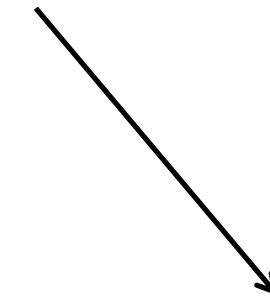
1. El 85% de los taxis en la ciudad son Verdes y el 15% de ellos son Azules.
2. Un testigo identificó el taxi como un taxi Azul. En la investigación del suceso se ha comprobado la precisión del testigo en circunstancias similares a la de aquella noche, y se concluyó que el testigo identifica correctamente cada uno de los dos colores el 80% de los casos, y falla el 20% restante.

Cual es la probabilidad de que el taxi responsable del accidente sea Verde o Azul?

Utilidades con variables no observadas

Problema: ¿Y qué pasa si la utilidad esperada depende de factores que el científico desconoce?

$$\hat{u}(c, a) = \sum_r P(\text{Result} = r) \hat{u}(c, r)$$



¿Y si esta utilidad depende de factores que el científico desconoce?

Utilidades con variables no observadas

Problema: ¿Y qué pasa si la utilidad esperada depende de factores que el científico desconoce?

Solución: *Modelos de utilidad aleatorios (MUA)*

$$u(c, a) = u_r(c, a) + \varepsilon_{c,a}$$

- $u_r(c, a)$ = utilidad determinista. Variable que puede calcular el científico utilizando los datos disponibles.
 - ε = utilidad no determinista. Variable aleatoria que depende de factores no observables.

¡¡¡La variable $u(c, a)$ se convierte entonces en una variable aleatoria!!!

El Principio de Decisión o Regla de Elección, queda:

$$RE(A, \geq) = \{a' \in A \mid P(\text{eleccion}(c) = a') \geq P(\text{eleccion}(c) = a), \forall a \in A\}$$

Y las probabilidades se calculan:

$$\begin{aligned} P(\text{eleccion}(c) = a') &= P(u(c, a') > u(c, a), \forall a \in A \text{ y } a \neq a') \\ &= P(u_r(c, a') + \varepsilon_{c,a'} > u_r(c, a) + \varepsilon_{c,a}, \forall a \in A \text{ y } a \neq a') \\ &= P(\varepsilon_{c,a} - \varepsilon_{c,a'} < u_r(c, a') - u_r(c, a), \forall a \in A \text{ y } a \neq a') \end{aligned}$$

Modelos de utilidad aleatorios

Y la probabilidad se calcula integrando sobre ε :

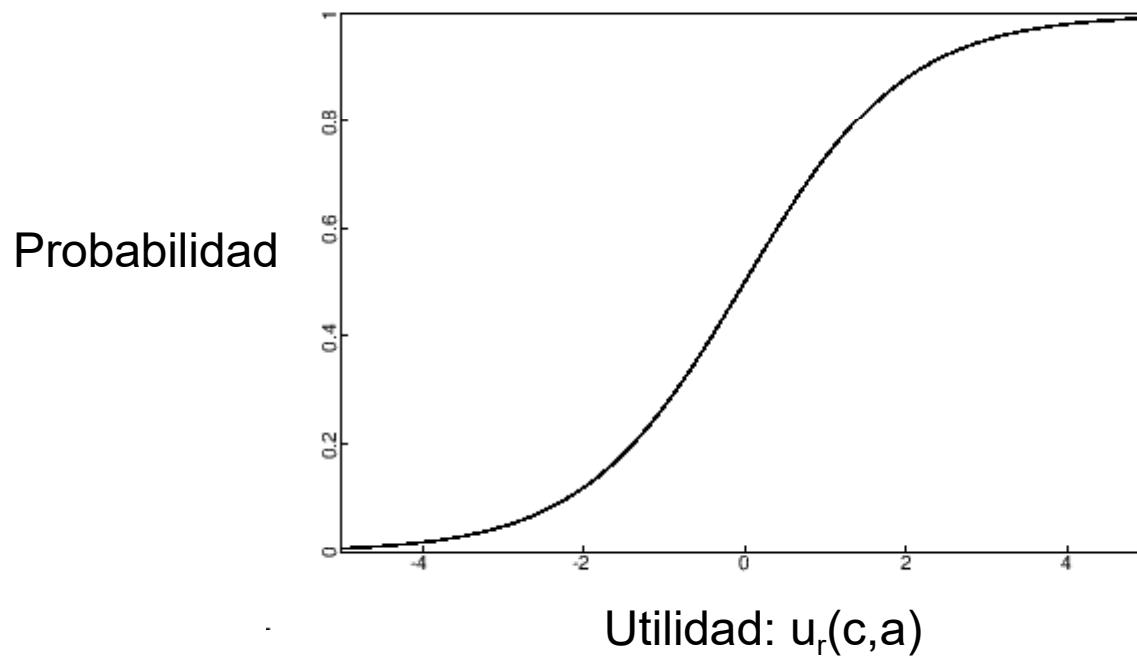
$$P(\text{elección}(c) = a') = \int_{\varepsilon_c = -\infty}^{\infty} I(\varepsilon_{c,a} - \varepsilon_{c,a'} < u_r(c, a') - u_r(c, a), \forall a \neq a') f(\varepsilon_c) d\varepsilon_c$$

Resolvemos la integral asumiendo una cierta distribución de probabilidad para ε , y obtenemos una función logística:

$$P(\text{elección}(c) = a') = \frac{e^{u_r(c, a')}}{\sum_j e^{u_r(c, a_j)}}$$

Modelos de utilidad aleatorios

Interpretación de la función logística:



Resultados



Rational choice-based models:

- Easy to explain
- Superior performance compared with basic baseline algorithms

Choice-based recommender systems

Paula Saavedra
CITIUS
University of Santiago de Compostela
Santiago de Compostela,
Spain
paula.saavedra@usc.es

Rosa Crujeiras
School of Mathematics
University of Santiago de Compostela
Santiago de Compostela,
Spain
rosa.crujeiras@usc.es

Pablo Barreiro
CITIUS
University of Santiago de Compostela
Santiago de Compostela,
Spain
pablobv70@gmail.com

Maria Loureiro
School of Business
University of Santiago de Compostela
Santiago de Compostela,
Spain
maria.loureiro@usc.es

Roi Durán
CITIUS
University of Santiago de Compostela
Santiago de Compostela,
Spain
roiduram@gmail.com

Eduardo Sánchez Vila
CITIUS
University of Santiago de Compostela
Santiago de Compostela,
Spain
eduardo.sanchez.vila@usc.es

ABSTRACT

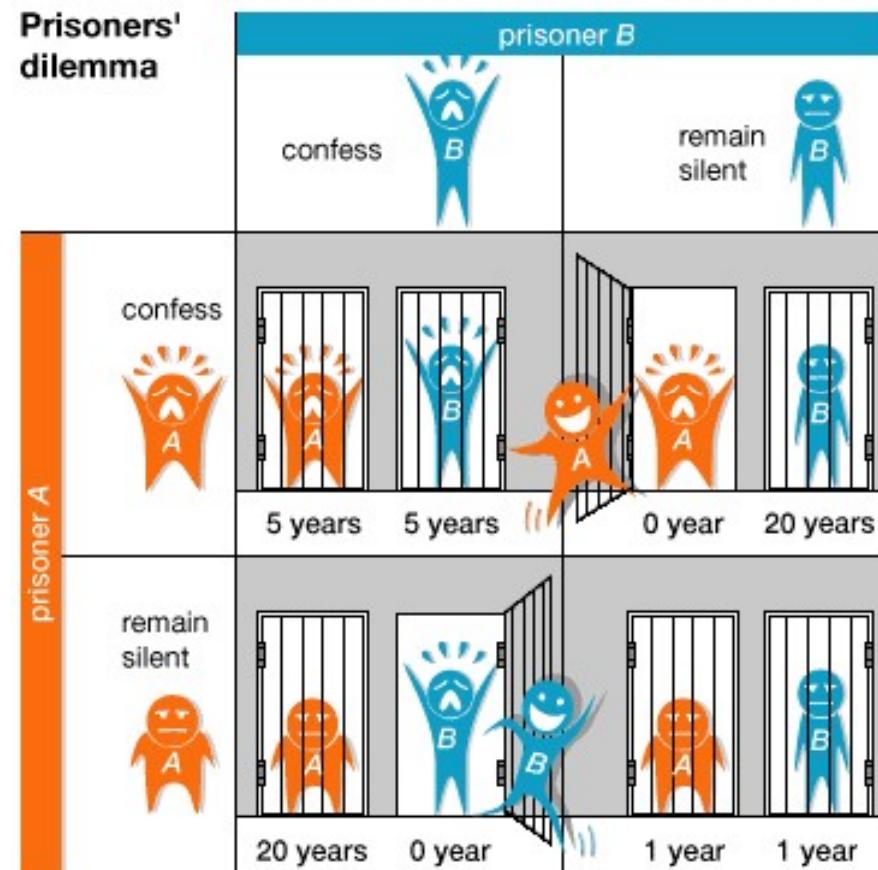
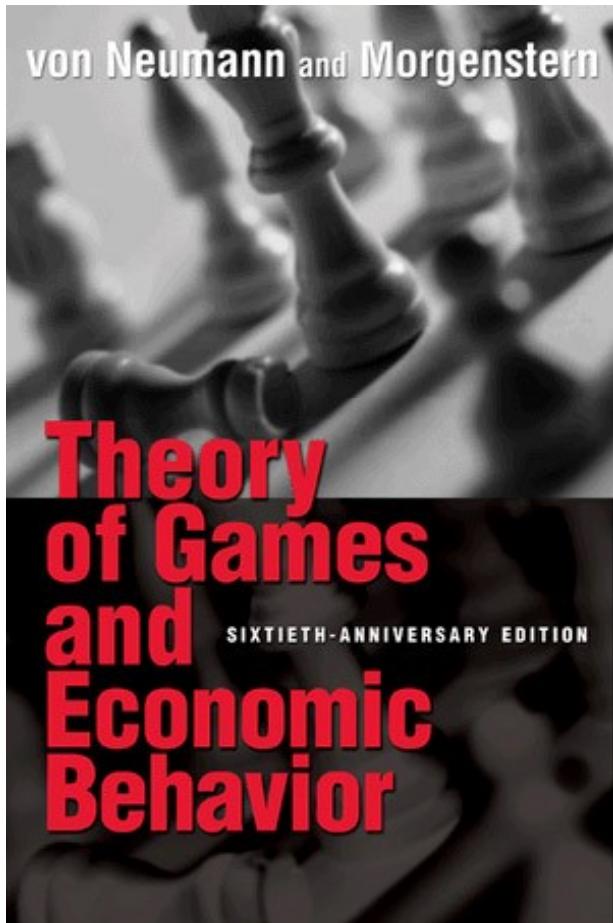
Choice-based models are proposed to overcome some of the limitations found in traditional rating-based strategies. The new approach is grounded on decision-making paradigms, such as choice and utility theories. Specifically, random utility models were applied in a recommendation problem. Prediction accuracy was compared with state-of-art rating-based algorithms in a gastronomy dataset. The results show the superior performance of choice-based models, which may suggest that real choices could bring more predictive power than ratings.

item's attributes [2]. These preferences can be used to predict the utility of any given item by comparing them with the values of item's attributes. Collaborative recommenders, on the other hand, take advantage of previous ratings provided by the available decision-makers to predict the utility of any given user-item pair [6]. This approach has been widely adopted as it removes the burden of knowing and managing item attributes as well as their corresponding values.

Many algorithms and models have been proposed under the collaborative paradigm. Among them, two families have gained major attraction: neighborhood algorithms and latent factor models. The neighborhood approach was the

(RECSYS, 2016)

Decisiones que dependen de otras decisiones



1944

Juego de Emociones

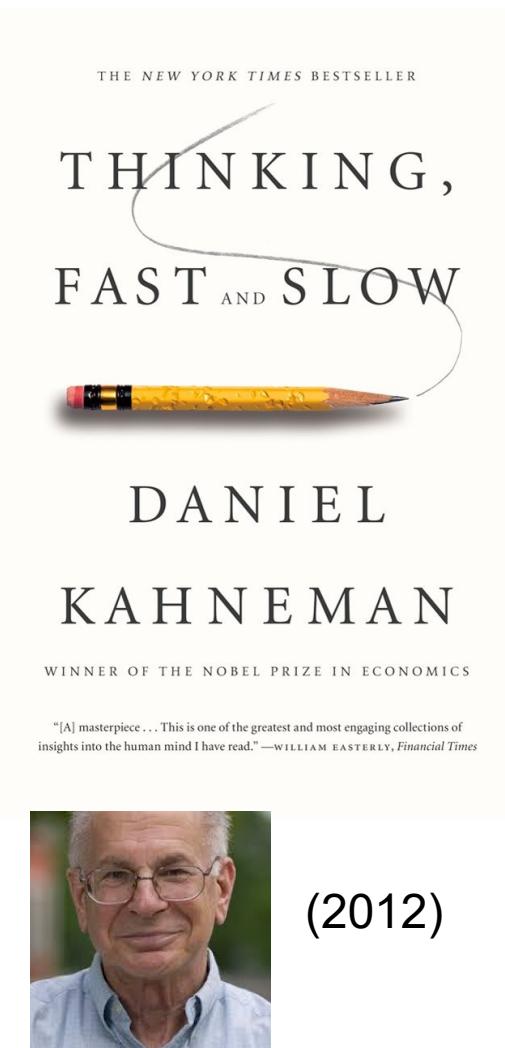
Estudiando el Sistema I

Nuestra mente tiene dos sistemas:

- Sistema I: Rápido, automático, poco esfuerzo, sin control consciente.
- Sistema II: Lento, racional, costoso, con control consciente.

Psicólogos ante la teoría económica:

- Econos Vs. Humanos.
- El poder de la multidisciplinariedad



Sistema I: Hipótesis del marcador somático

Emotions and decision-making:

- Emotions play a key role in making decisions
- Impairments/defects in emotional system reduces performance on decision-making tasks

The somatic marker hypothesis and the possible functions of the prefrontal cortex

ANTONIO R. DAMASIO*

Department of Neurology, Division of Behavioral Neurology and Cognitive Neuroscience, University of Iowa College of Medicine, Iowa City, Iowa, U.S.A.

SUMMARY

In this article I discuss a hypothesis, known as the somatic marker hypothesis, which I believe is relevant to the understanding of processes of human reasoning and decision making. The ventromedial sector of the prefrontal cortices is critical to the operations postulated here, but the hypothesis does not necessarily apply to prefrontal cortex as a whole and should not be seen as an attempt to unify frontal lobe functions under a single mechanism.

The key idea in the hypothesis is that 'marker' signals influence the processes of response to stimuli, at multiple levels of operation, some of which occur overtly (consciously, 'in mind') and some of which occur covertly (non-consciously, in a non-minded manner). The marker signals arise in bioregulatory processes, including those which express themselves in emotions and feelings, but are not necessarily confined to those alone. This is the reason why the markers are termed somatic: they relate to body-state structure and regulation even when they do not arise in the body proper but rather in the brain's representation of the body.

Examples of the covert action of 'marker' signals are the undeliberated inhibition of a response learned previously; the introduction of a bias in the selection of an aversive or appetitive mode of behaviour, or in the otherwise deliberate evaluation of varied option-outcome scenarios. Examples of overt action include the conscious 'qualifying' of certain option-outcome scenarios as dangerous or advantageous.

The hypothesis rejects attempts to limit human reasoning and decision making to mechanisms relying, in an exclusive and unrelated manner, on either conditioning alone or cognition alone.



(1996)

- Hypothesis 1: System I models (Emotional, Attentional) will show better performance than System II models (Rational).

- Hypothesis 2: Emotional models will ouperform Attentional models.

Métodos



HOME JOBS FINANCE TABLES AND CHARTS EVENTS PAIDCONTENT 50

Sep 30, 2010 - 12:12AM

ADVERTISEMENT

Facebook's Sandberg: In The Future, All Media Will Be Personalized

BY David Kaplan

5 Comments +1

There will always be a place for mass marketing, but in the next three- to five years, a website that isn't tailored to a specific user's in...



There will always be a place for mass marketing, but in the next three- to five years, a website that isn't tailored to a specific user's interest will be an anachronism, Facebook COO Sheryl Sandberg told Arianna Huffington at the latter's Advertising Week event. "People don't want something targeted to the whole world — they want something that reflects what they want to see and know," she said. Sandberg and Huffington also discussed Facebook's privacy issues and a certain movie being released this weekend.

RELATED

France calms fears over Facebook Timeline scare

A French tabloid set off a temporary worldwide panic that Facebook had published the private messages of...

Newsprint joins the internet of things
Print and digital media are so disconnected, they often appear a lifetime apart. But new technology promises...

No... an internet you won't leave journalism



The Premium Experience:
Neurological Engagement
on Premium Websites

Experimental Setup

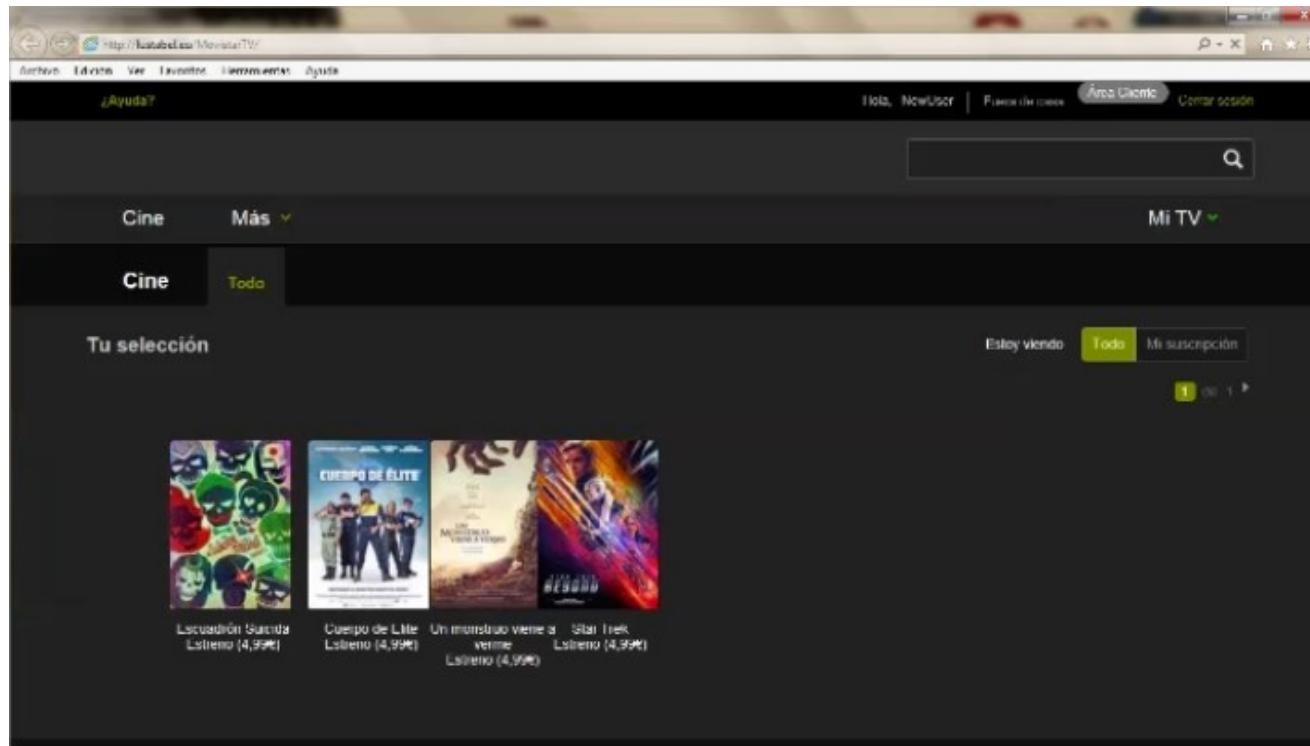


Experimental Setup

Table 1. Devices and recorded variables.

| Devices | Models | Variables | Variable Type |
|------------------------|-----------------------------|--|---------------|
| EEG | Emotiv EPOC Headset | Frustration, Excitement and Engagement | Emotional |
| Facial Coding (FACEIT) | Logitech HD Pro webcam C920 | Joy, Anger, Fear, Surprise, Contempt, Disgust, Sadness, Neutral, Positive and Negative | Emotional |
| Eye-Tracking | TOBii X2 @30Hz | TimeSpend and Fixations | Attentional |
| iMotions (Software) | Version 6.2 | | |

Experimental Design



1 Trial per user:

- 4 Movies per situation
- 20 choice situations
- 80 Movies per Trial

Experimental Design

Table 2. Characterization of movies: attributes and values.

| Feature | Values |
|---------|---|
| Genre | Action, Comedy, Science Fiction, Drama |
| Novelty | Release, Catalog |
| Price | 4,99 euros (Release), 0 euros (Catalog) |

Factorial Design:

- 8 Profiles
- 10 Movies per Profile
- 80 Movies per Trial

Experiments



Choice Rule in probabilistic form:

$$CR(A, \geq) = \{a_i \in A \mid \mathbb{P}_i \geq \mathbb{P}_j, \forall a_j \in A\}$$

Form of probabilities: Standard Logit Function

$$\mathbb{P}_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}. \quad \text{with } V_{nj} = \beta_{nj} \times x_j$$

V_{nj} , the utility of alternative a_j for decision maker c_n

β_{nj} , the vector of preferences of decision maker c_n

x_j , the observed features of alternative a_j

Observed features

Table 3. Observed features.

| Models | Features |
|-------------|---|
| Rational | Action, Comedy, Science Fiction, Drama, Release and Catalog |
| Emotional | Frustration, Excitement, Engagement, Joy, Anger, Fear, Surprise, Contempt, Disgust, Sadness, Neutral, Positive and Negative |
| Attentional | TimeSpend and Fixations |

Specific Models

Table 4. Choice models used in this work: Rational (R), Attentional (A), and Emotional (E).

| Model | Features |
|-------|---|
| R | Action, Comedy, Fiction, Premiere |
| A1 | TimeSpend |
| A2 | FixationTime |
| E1 | Action, Comedy, Fiction, Premiere, Anger, Frustration, Negative, Fear |
| E2 | Action, Comedy, Fiction, Premiere, Excitement, Joy, Engagement |
| E3 | Action, Comedy, Fiction, Premiere, Frustration, Excitement |
| E4 | Action, Comedy, Fiction, Premiere, Frustration, Excitement, Engagement |
| E5 | Action, Comedy, Fiction, Premiere, Anger, Fear |
| E6 | Action, Comedy, Fiction, Premiere, Fear, Contempt, Disgust, Sadness |
| E7 | Action, Comedy, Fiction, Premiere, Joy, Anger, Fear, Surprise |
| E8 | Action, Comedy, Fiction, Premiere, Frustration, Excitement, Engagement, Joy, Anger |
| E9 | Action, Comedy, Fiction, Premiere, Joy, Anger, Fear, Surprise, Contempt, Disgust, Sadness |
| E10 | Action, Comedy, Fiction, Premiere, Sadness, Neutral, Positive, Negative |
| E11 | Action, Comedy, Fiction, Premiere, Frustration, Excitement, Engagement, Neutral, Positive, Negative |
| E12 | Action, Comedy, Fiction, Premiere, Anger, Fear, Surprise, Contempt, Disgust, Sadness, Neutral, Joy, Negative |
| E13 | Action, Comedy, Fiction, Premiere, Frustration, Excitement, Engagement, Joy, Anger, Fear, Surprise, Contempt, Disgust, Sadness, Neutral, Negative |

Evaluation

$$Accuracy = \frac{T_{CorrectPrediction}}{T_{AllPrediction}}$$

Resultados

Best Model for Single Subject

Table 5. Performance of choice-based models for each user. Accuracy results for top five models.

| Users | Number of Choices | Top Five Models | Accuracy |
|-------|-------------------|-----------------|----------|
| J02 | 12 | E8 | 0.787879 |
| | | E11 | 0.787879 |
| | | E4 | 0.757576 |
| | | E2 | 0.666667 |
| | | E6 | 0.636364 |
| J06 | 14 | A2 | 0.681818 |
| | | A1 | 0.621212 |
| | | E4 | 0.560606 |
| | | E9 | 0.560606 |
| | | R | 0.545455 |
| V09 | 20 | R | 0.522727 |
| | | E1 | 0.477273 |
| | | E0 | 0.454545 |
| | | E5 | 0.431818 |
| | | E6 | 0.431818 |

Best Model for Single Subject

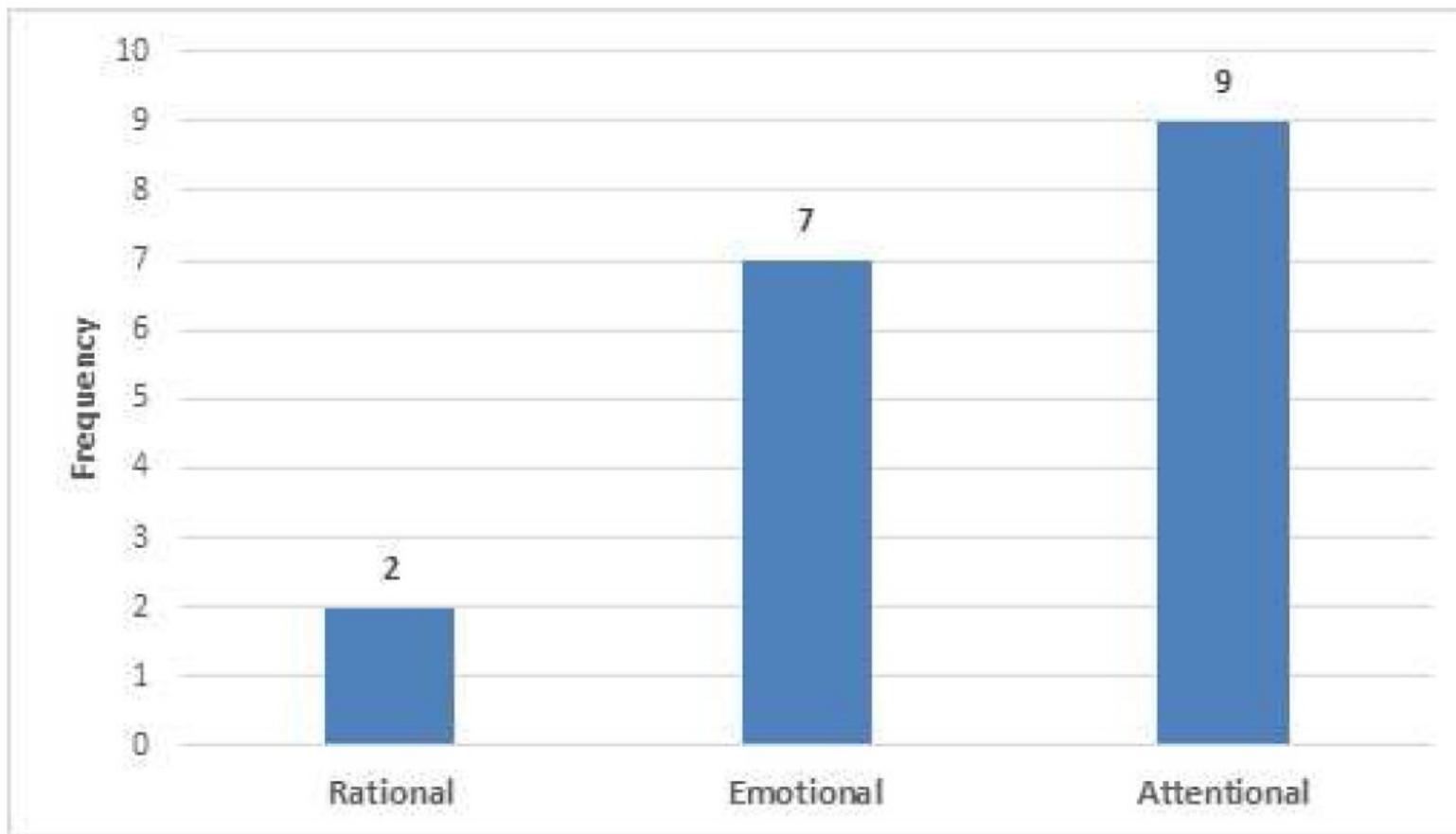


Fig. 3. Frequency of the Rational, Emotional, and Attentional models as best model.

Best Model for All Subjects

Table 6. Performance of choice-based models for average model for all users. Accuracy results for top five models.

| Models | Average of Accuracy |
|------------|---------------------|
| Fixation | 0.528282828 |
| TimeSpend | 0.522306397 |
| Rational | 0.520911496 |
| Emotional6 | 0.464850890 |
| Emotional4 | 0.448328523 |

Best Model for All Subjects

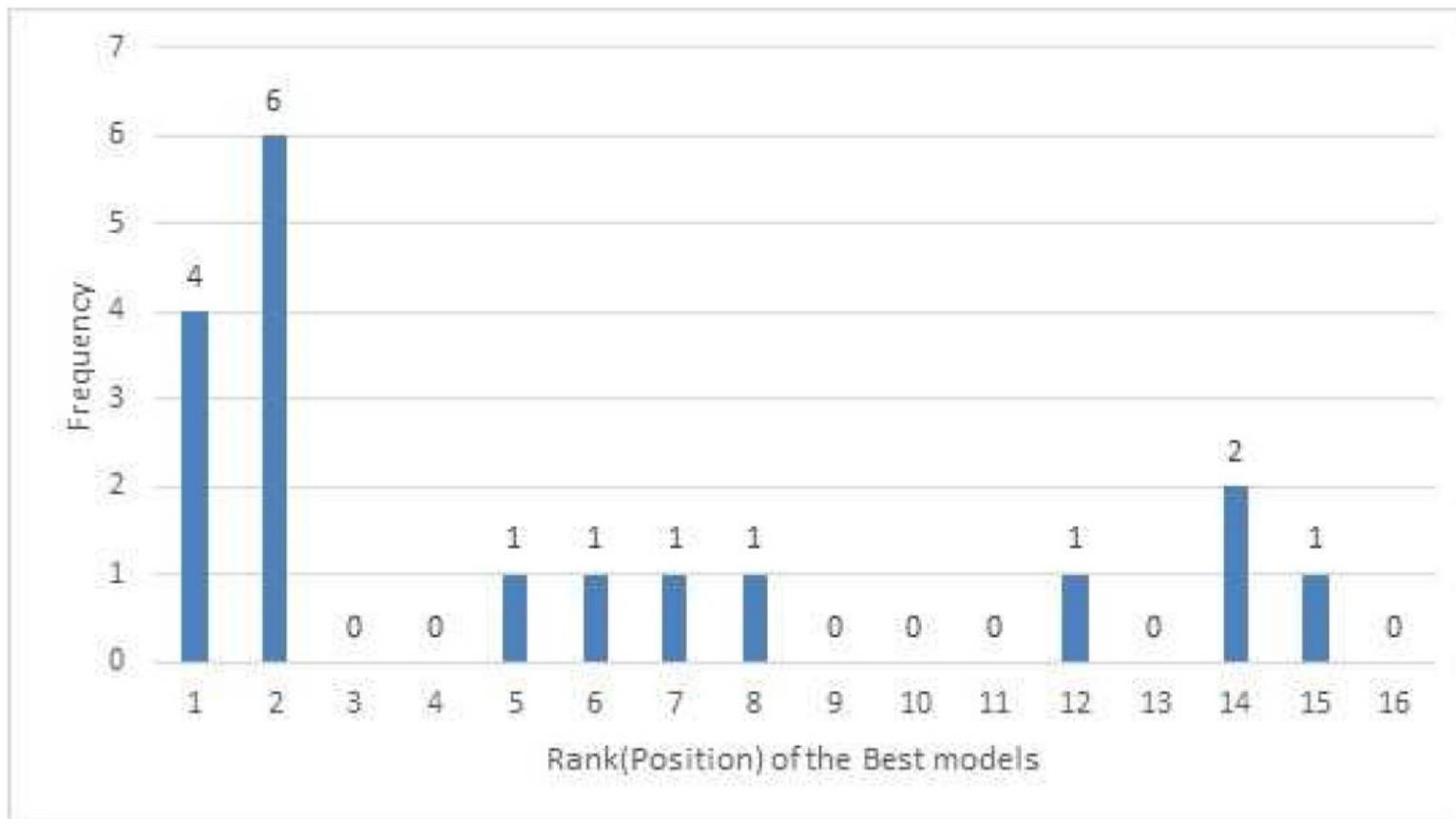


Fig. 4. Frequency of the rank of the best average model(Fixation model).

Discusión

Conclusiones principales

- Hypothesis I: System I models (Emotional, Attentional) will show better performance than System II models (Rational).
 - **RESULT: YES!**
- Hypothesis 2: Emotional models will ouperform Attentional models.
 - **RESULT: NO!**
- FINDING: No model is the best for every single subject!
- FUTURE WORK: Factors underlying the attention process?

Conexión con el Thick Data

Modelos Toma Decisiones => Thick Data!



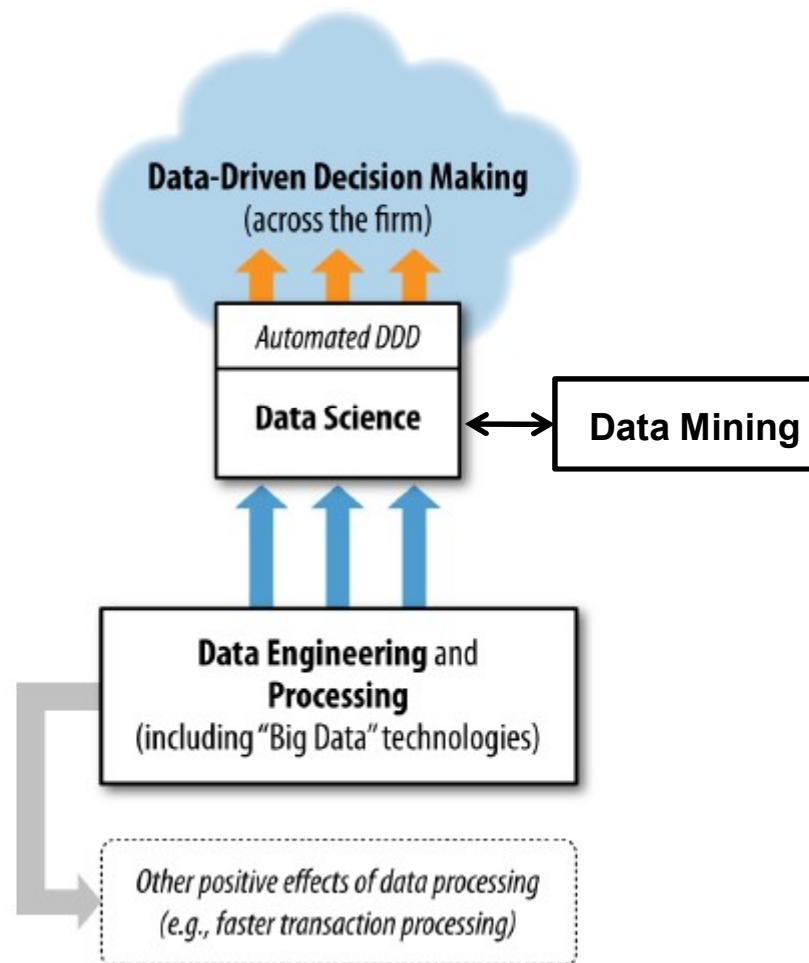
Modelos Toma Decisiones + Big Data => pensar analíticamente!

1. HIPÓTESIS

2. MODELOS CAUSALES / PREDICTIVOS

3. DATOS

Modelos Toma Decisiones + Big Data



Modelos de Toma de decisiones



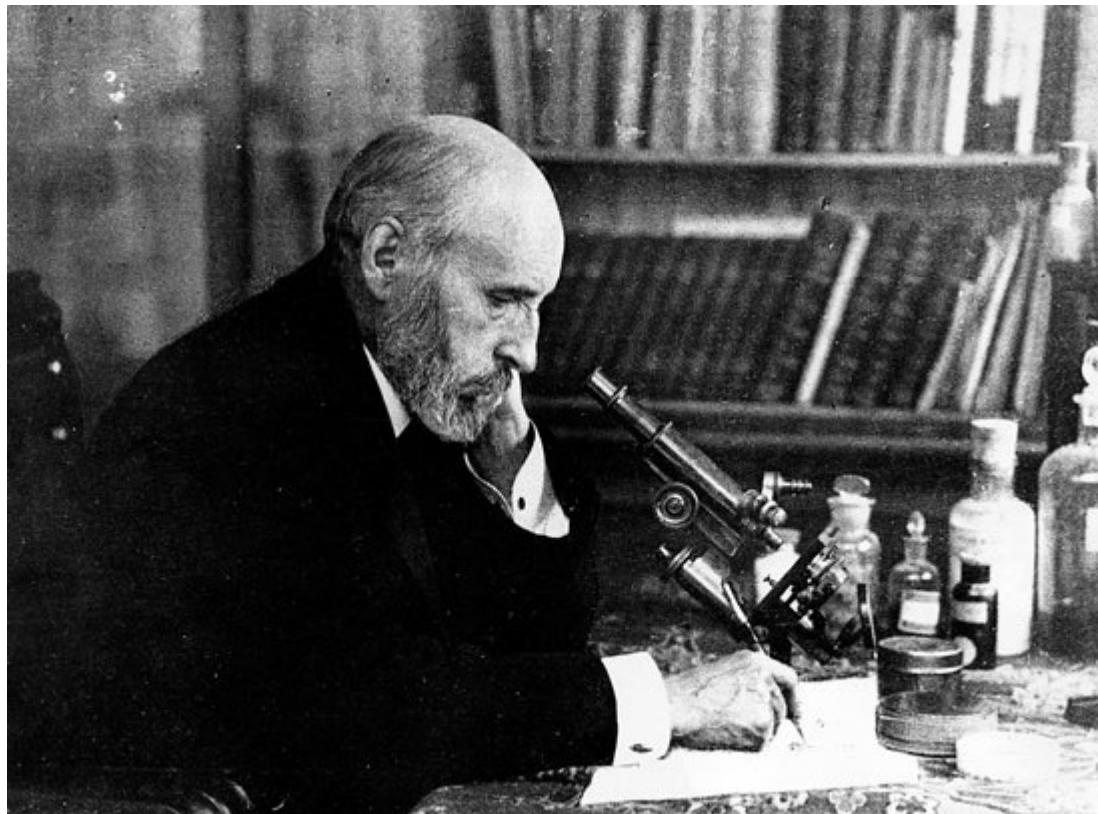
Grupo de Sistemas *inteligentes*



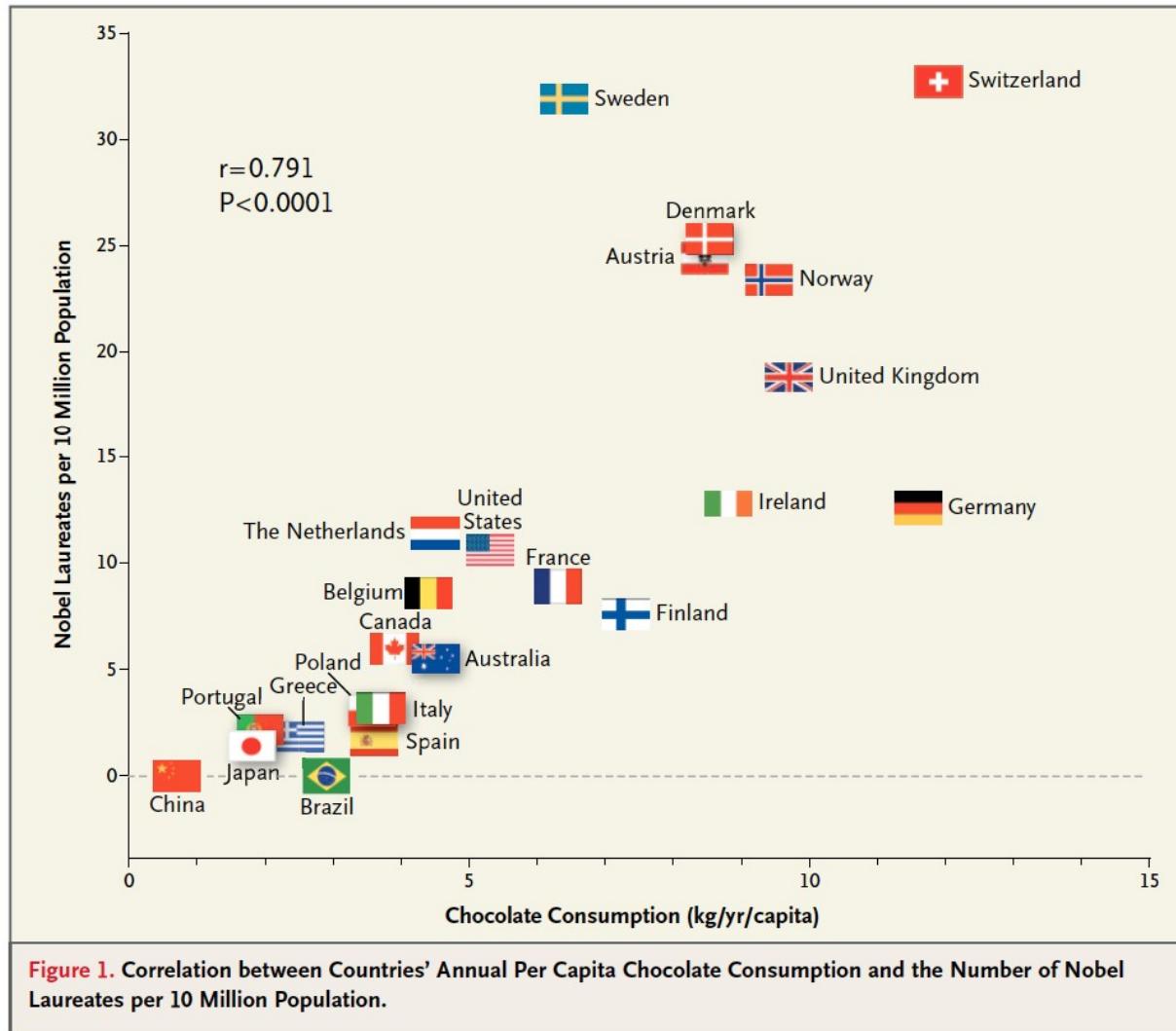
Eduardo M. Sánchez Vila
eduardo.sanchez.vila@usc.es

Grupo de Sistemas Inteligentes
Universidad de Santiago de Compostela

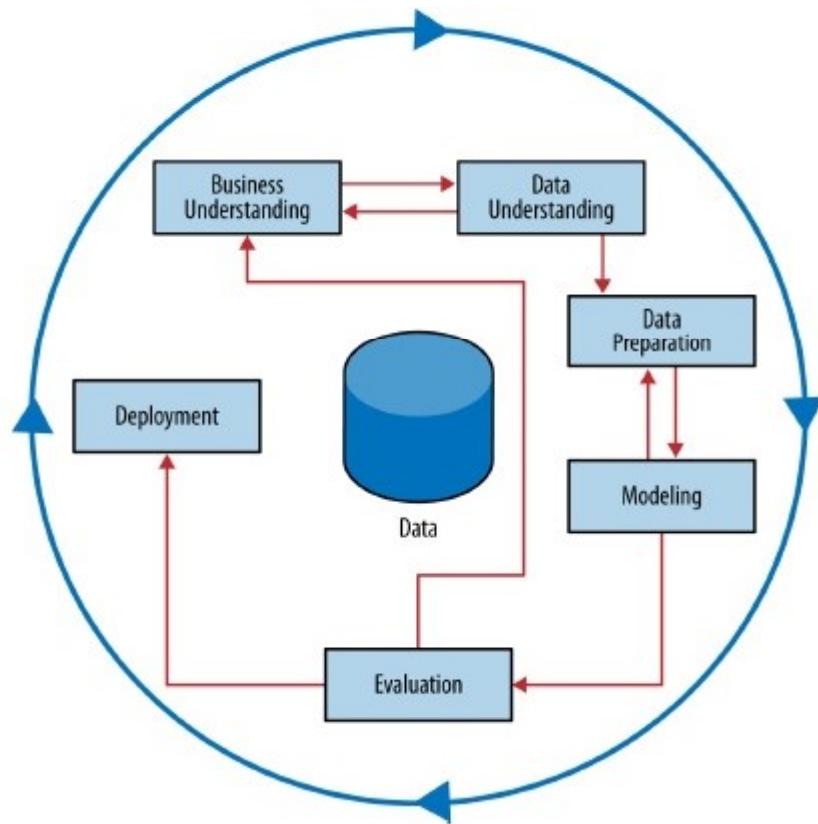
¿Factores de éxito para ganar un Nobel?



¿Factores de éxito para ganar un Nobel?



Del Business Understanding al Data Mining



Problema: Campaña de Fundraising

For this case study, let's consider a real example of targeted marketing: targeting the best prospects for a charity mailing. Fundraising organizations (including those in universities) need to manage their budgets and the patience of their potential donors. In any given campaign segment, they would like to solicit from a “good” subset of the donors. This could be a very large subset for an inexpensive, infrequent campaign, or a smaller subset for a focused campaign that includes a not-so-inexpensive incentive package.

So let's get specific. A data miner might immediately think: we want to model the probability that each prospective customer, a prospective donor in this case, will respond to the offer. However, thinking carefully about the business problem we realize that in this case, the response can vary—some people might donate \$100 while others might donate \$1. We need to take this into account.

Objetivo: Diseñar una campaña para maximizar la captación de fondos, donde hay que descontar a cada donación el coste de la campaña.

Propuesta de modelo

Se propone un modelo de toma de decisiones basado en la teoría de utilidad esperada:

$$\text{Expected benefit of targeting} = p(R | x) \cdot v_R(x) + [1 - p(R | x)] \cdot v_{NR}(x)$$

donde:

$P(R | x)$ es la probabilidad de que la persona x responda positivamente.

$v_R(x)$ es el valor obtenido si la respuesta de x es positiva

$v_{NR}(x)$ es el valor obtenido si la respuesta de x es negativa

$v(x) = \text{donacion}(x) - \text{coste}(x)$

El objetivo consiste en diseñar una campaña tal que:

Expected benefit of targeting > 0

Lo que implica que el beneficio tiene que ser superior al coste:

$$\begin{aligned} p(R | \mathbf{x}) \cdot (d_R(\mathbf{x}) - c) + [1 - p(R | \mathbf{x})] \cdot (-c) &> 0 \\ p(R | \mathbf{x}) \cdot d_R(\mathbf{x}) - p(R | \mathbf{x}) \cdot c - c + p(R | \mathbf{x}) \cdot c &> 0 \\ p(R | \mathbf{x}) \cdot d_R(\mathbf{x}) &> c \end{aligned}$$

Responde a estas cuestiones:

- 1. ¿El modelo representa correctamente el problema? ¿Tiene el nivel de detalle suficiente?**
- 2. ¿Qué supuestos asume el modelo? ¿Son correctos?**
- 3. ¿Qué datos necesitamos para conocer las variables $p(R|x)$, d_R y c ? ¿Cómo obtenemos esos datos?**
- 4. ¿Cómo se puede evaluar el modelo antes de lanzar la campaña?**

Respuestas:

1. ¿El modelo representa correctamente el problema? ¿Tiene el nivel de detalle suficiente?

No, el modelo no representa correctamente el problema. Habría que considerar diferentes tipos de potenciales donantes.

2. ¿Qué supuestos asume el modelo? ¿Son correctos?

Se asume un perfil de donante medio, con una donación media y coste de la campaña similar. Este supuesto simplifica el modelo pero no tiene en cuenta las diferencias entre potenciales donantes.

3. ¿Qué datos necesitamos para conocer las variables $p(R|x)$, d_R y c ? ¿Cómo obtenemos esos datos?

Se necesitan datos de campañas anteriores para conocer $p(R|x)$ y d_R . Las campañas deberían ser similares a la actual, dado que la afinidad/interés por la temática de la campaña pueden cambiar mucho los valores de $p(R|x)$ y d_R .

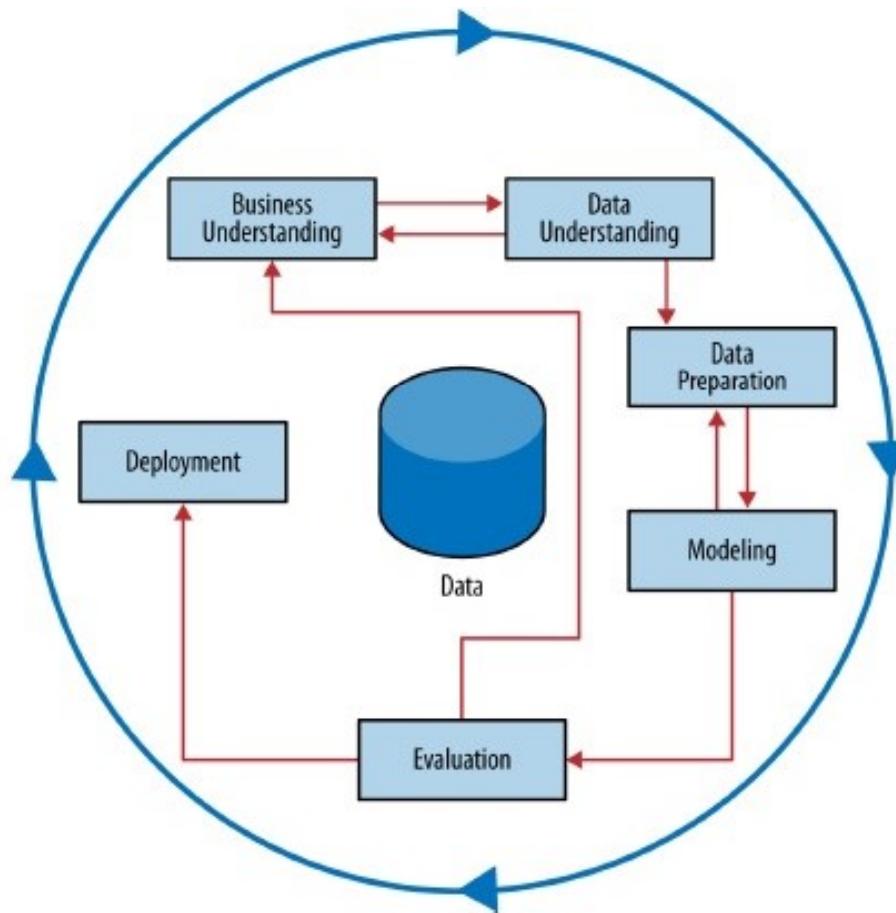
Respuestas:

4. ¿Cómo se puede evaluar el modelo antes de lanzar la campaña?

Si se dispone de una campaña similar realizada anteriormente, se puede evaluar el modelo con los datos de esa campaña.

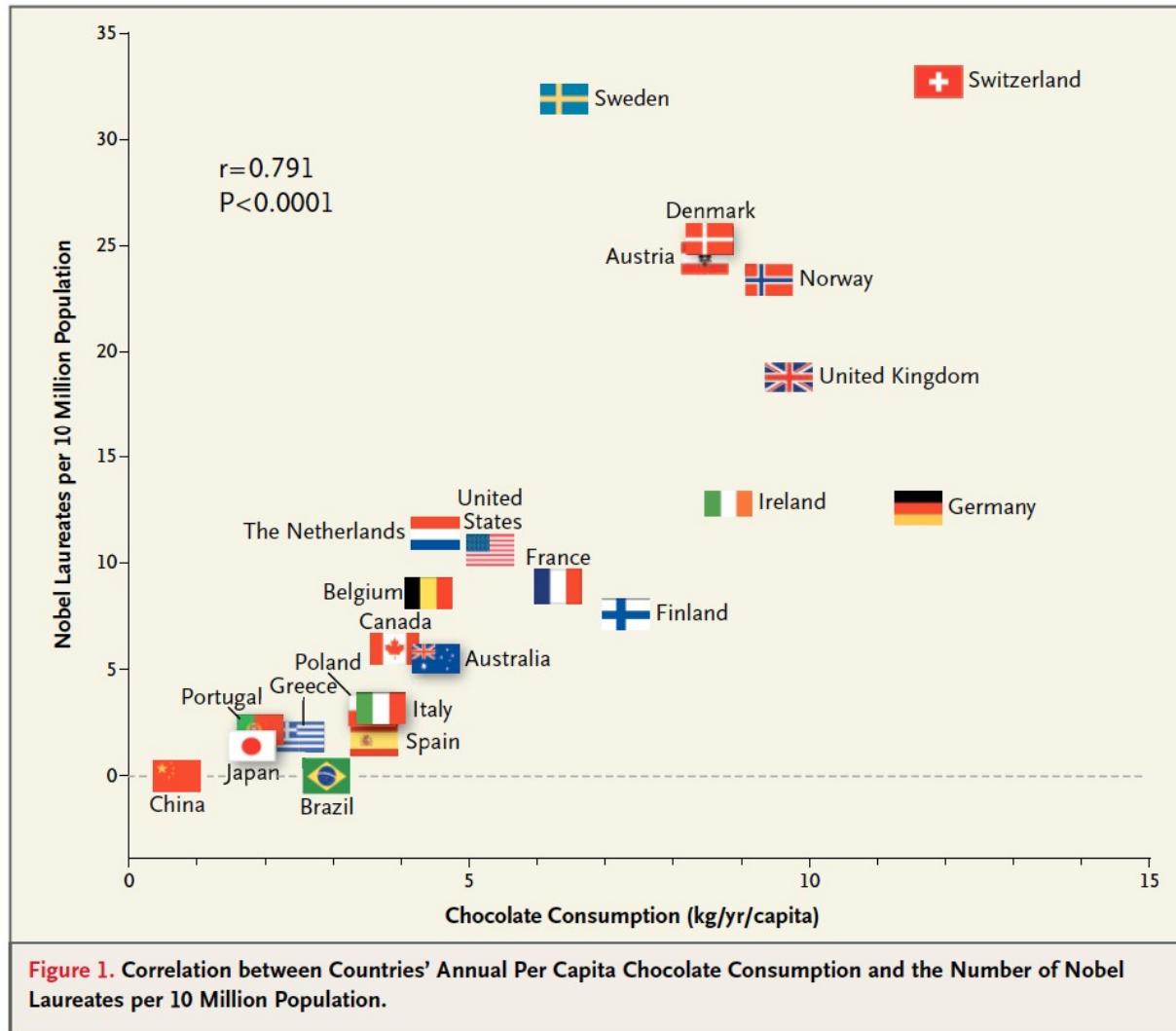
En caso contrario, habría que hacer una prueba piloto con una campaña reducida para obtener los datos y evaluar el modelo.

Guía para la creación/evaluación de proyectos



Piensa, luego haz todo lo demás

¿Factores de éxito para ganar un Nobel?



1. Entender el problema. Preguntas:

- ¿En qué consiste el problema de forma precisa?
- ¿Qué entidades incluye el problema?
- ¿Cuáles son las variables/atributos relevantes del problema?

2. Entender los datos que necesitamos. Preguntas:

- ¿Tengo los datos?
- ¿Puedo acceder a todas las variables?
- ¿Cómo puedo obtener los datos?

3. Preparar los datos. Preguntas

- ¿En qué formato tengo los datos?
- ¿Tengo muchas variables? ¿Son todas necesarias?
- ¿Tengo muchas/pocas observaciones? ¿Cuántas necesito?

4. Crear un modelo. Preguntas:

- ¿Qué modelo(s) es(son) el(los) más apropiado(s)?
- Si hay varios posibles, ¿cuál escoger?
- ¿Tengo datos para estimar las variables del modelo?
- ¿Cumple el modelo aspectos no funcionales (generalización, escalabilidad, etc...)?

5. Evaluar el modelo. Preguntas:

- ¿Plan para evaluar el modelo?
- ¿Hay datos para evaluar el modelo?
- ¿Son apropiados los criterios (métricas) para evaluar el modelo?

Datos: ¿Calidad o cantidad?

Small Data y Thick Data



Datos: ¿implícitos o explícitos?

Elecciones vs Valoraciones

| Producto | Autor | Fecha | Precio | Género | Acción usuario |
|----------|-------|-------|--------|-----------|----------------|
| Libro1 | --- | --- | 15 | Ficción | Comprado |
| Libro2 | --- | --- | 50 | Histórico | Comprado |
| | | | | | |
| LibroN | --- | --- | 10 | Ficción | Comprado |

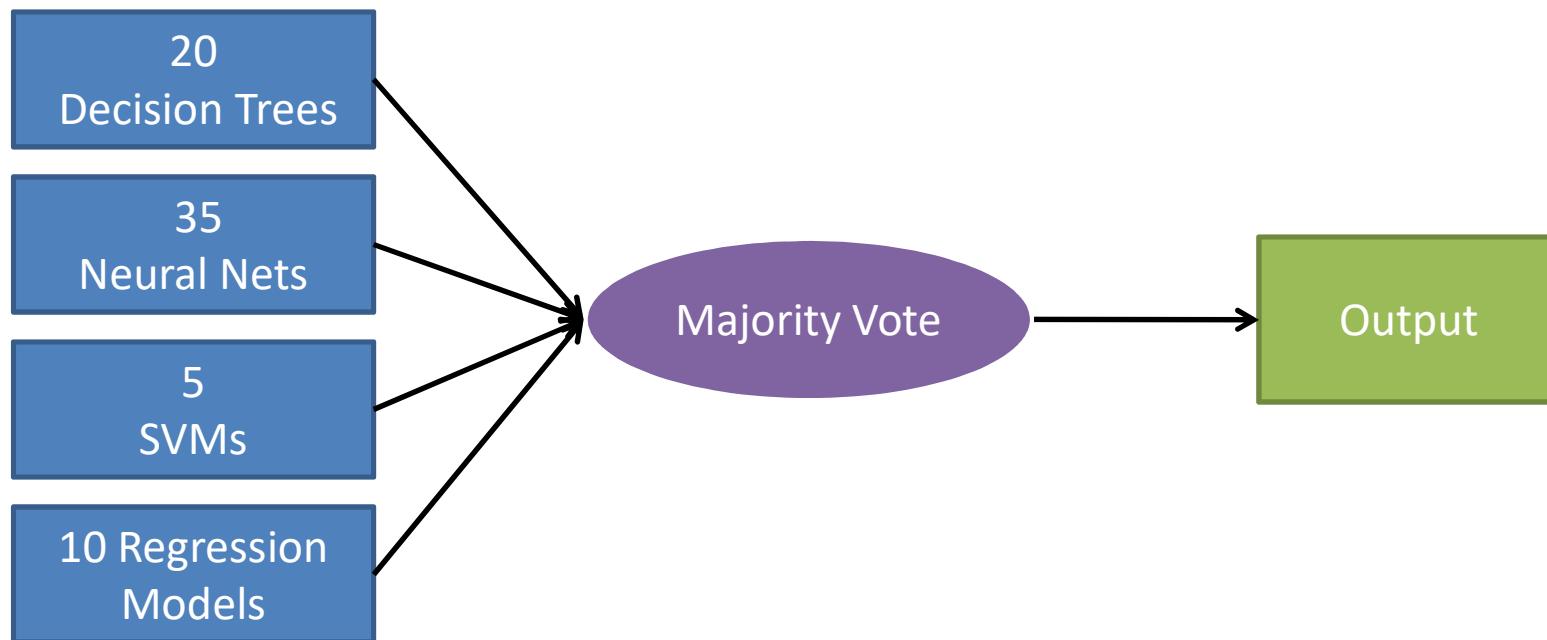
Elecciones

| | Item1 | Item2 | Item3 | Item4 | Item5 |
|-------|-------|-------|-------|-------|-------|
| Alice | 5 | 3 | 4 | 4 | ? |
| User1 | 3 | 1 | 2 | 3 | 3 |
| User2 | 4 | 3 | 4 | 3 | 5 |
| User3 | 3 | 3 | 1 | 5 | 4 |
| User4 | 1 | 5 | 5 | 2 | 1 |

Valoraciones

Modela “con fundamento”

Problemas con las Recomendaciones



¡¡Cómo lo explicamos!!

Modelos Toma Decisiones + Big Data => pensar analíticamente!

1. HIPÓTESIS

2. MODELOS CAUSALES / PREDICTIVOS

3. DATOS

El modelo único no existe

Best Model for Single Subject

Table 5. Performance of choice-based models for each user. Accuracy results for top five models.

| Users | Number of Choices | Top Five Models | Accuracy |
|-------|-------------------|-----------------|----------|
| J02 | 12 | E8 | 0.787879 |
| | | E11 | 0.787879 |
| | | E4 | 0.757576 |
| | | E2 | 0.666667 |
| | | E6 | 0.636364 |
| J06 | 14 | A2 | 0.681818 |
| | | A1 | 0.621212 |
| | | E4 | 0.560606 |
| | | E9 | 0.560606 |
| | | R | 0.545455 |
| V09 | 20 | R | 0.522727 |
| | | E1 | 0.477273 |
| | | E0 | 0.454545 |
| | | E5 | 0.431818 |
| | | E6 | 0.431818 |

Best Model for Single Subject

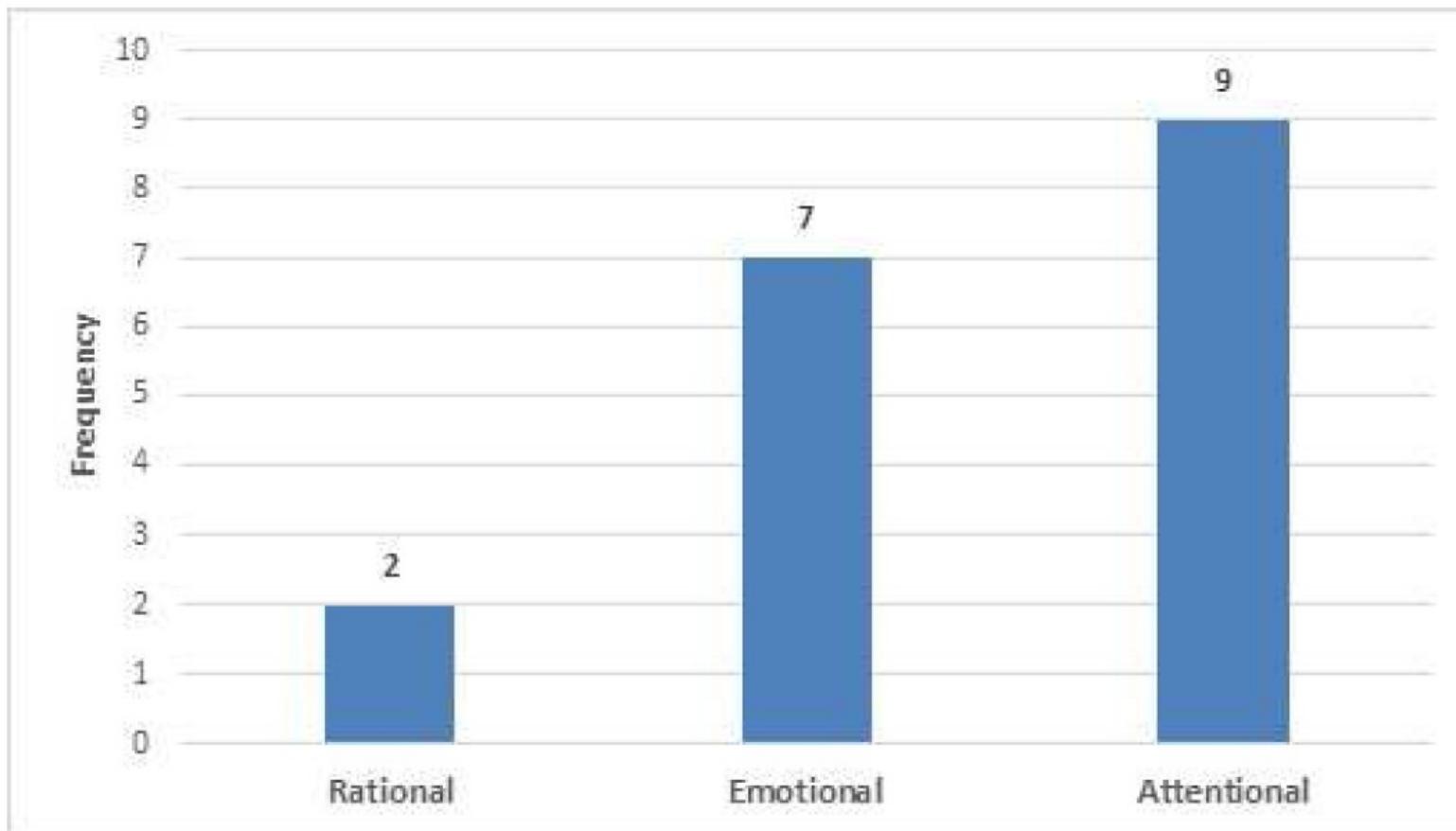


Fig. 3. Frequency of the Rational, Emotional, and Attentional models as best model.

Visualiza que no es poco

NETFLIX vs MOVISTAR

The image shows two screenshots side-by-side for comparison.

Left Side (Netflix):

- Recently Added:** A section showing recently added content like "ARQ", "Katy Perry: Part of Me 50", "THE BIG SHORT", and "NETFLIX CHEF'S TABLE FRANC". A green arrow points to the "Recently Added" heading.
- Recommendations:** A section titled "Because you added To Kill a Mockingbird to your list" showing "A CONVERSATION WITH GREGORY PECK", "GOOD WILL HUNTING", and "GONE WITH THE WIND". A green arrow points to this section.
- Watch History:** A section titled "Because you watched Helmut Schmidt – Lebensfragen" showing a thumbnail for the show.

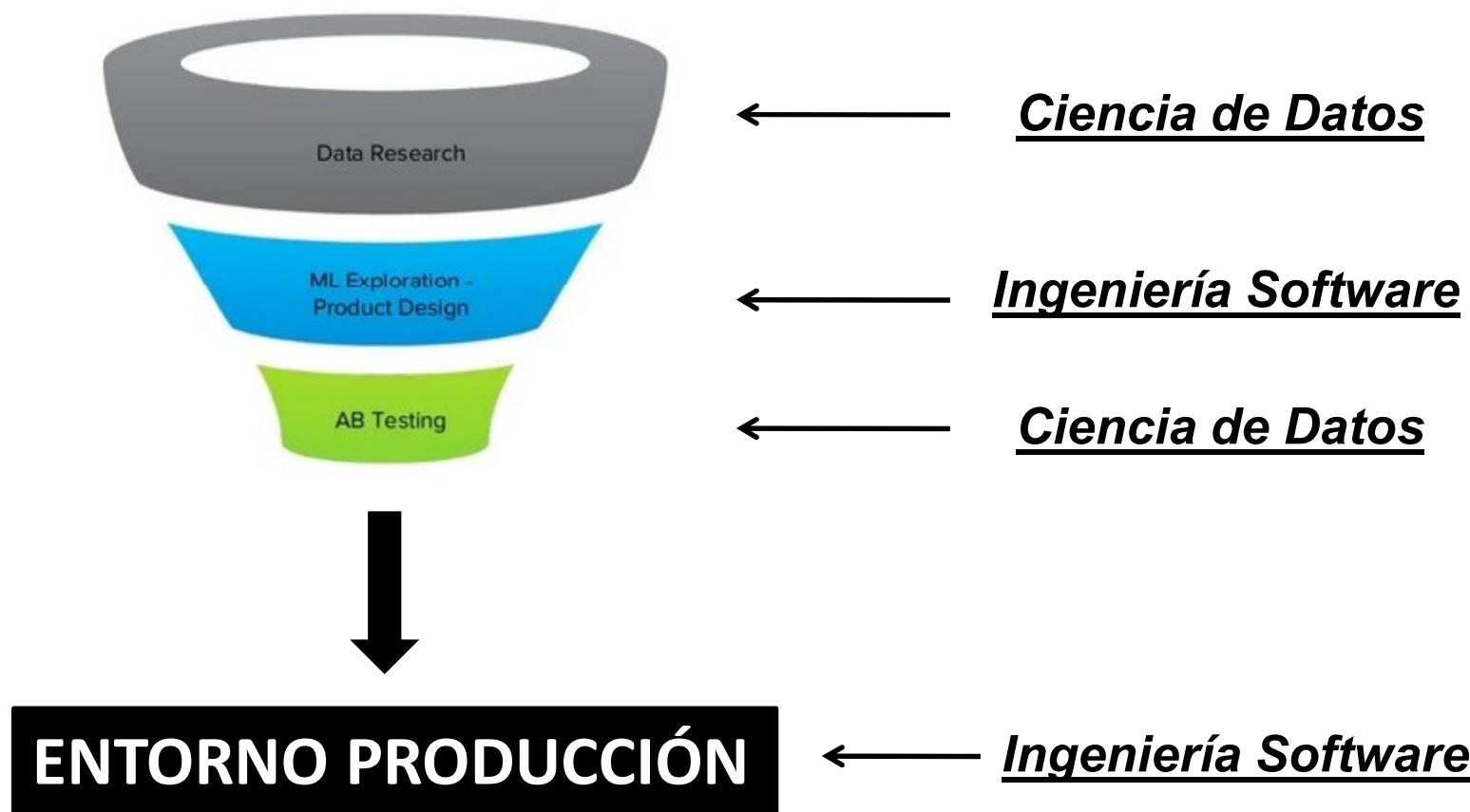
Right Side (Movistar+):

- Shows:** A large image for the TV show "DIME QUIÉN SOY" with the subtitle "Uma serie original Movistar+ BASADA EN LA NOVELA DE JULIA NAVARRO".
- Navigation:** A navigation bar at the bottom with links for "Apps", a search icon, "Originals", "Cine", "Series", "Deportes", "Netflix", "Infantil", and "Documentales".
- Partners:** A row of logos for "Ahora en TV" partners: "#0 de movistar+", "#VAMOS de movistar+", "M Series", "Seriesmania", "FOX", "AXN", "TNT", and "CCTV".
- Popular Content:** A section titled "Popular en Movistar+" showing thumbnails for "M+", "M+", and "M+".

Una gran pareja:
Ciencia de Datos + Ingeniería Software

The data-driven ML innovation funnel

Quora



La solución más simple
puede ser la más eficiente

1. Lo que vende:

“Solución basada en Hadoop/Spark,
cluster big Data con N servidores,
arquitectura lambda para procesamiento real-time,
frameworks Kafda y Databrics,
y bases de datos noSQL”

2. Proyecto real en Quora (Ref: Xavier Amatrain)

- **Implementación Spark:** 6 horas, 15 máquinas.
- **Implementación C++:** 10 minutos, 1 máquina

Ejercicio I: Churn reduction!

Scenario and Proposal

You've landed a great job with Green Giant Consulting (GGC), managing an analytical team that is just building up its data science skill set. GGC is proposing a data science project with TelCo, the nation's second-largest provider of wireless communication services, to help address their problem of customer churn. Your team of analysts has produced the following proposal, and you are reviewing it prior to presenting the proposed plan to TelCo. Do you find any flaws with the plan? Do you have any suggestions for how to improve it?

Churn Reduction via Targeted Incentives — A GGC Proposal

We propose that TelCo test its ability to control its customer churn via an analysis of churn prediction. The key idea is that TelCo can use data on customer behavior to predict when customers will leave, and then can target these customers with special incentives to remain with TelCo. We propose the following modeling problem, which can be carried out using data already in TelCo's possession.

We will model the probability that a customer will (or will not) leave within 90 days of contract expiration, with the understanding that there is a separate problem of retaining customers who are continuing their service month-to-month, long after contract expiration. We believe that predicting churn in this 90-day window is an appropriate starting point, and the lessons learned may apply to other churn-prediction cases as well. The

Ejercicio I: Churn reduction!

model will be built on a database of historical cases of customers who have left the company. Churn probability will be predicted based on data 45 days prior to contract expiration, in order for TelCo to have sufficient lead time to affect customer behavior with an incentive offer. We will model churn probability by building an ensemble of trees (random forest) model, which is known to have high accuracy for a wide variety of estimation problems.

We estimate that we will be able to identify 70% of the customers who will leave within the 90-day time window. We will verify this by running the model on the database to verify that indeed the model can reach this level of accuracy. Through interactions with TelCo stakeholders, we understand that it is very important that the V.P. of Customer Retention sign off on any new customer retention procedures, and she has indicated that she will base her decision on her own assessment that the procedure used for identifying customers makes sense and on the opinions about the procedure from selected firm experts in customer retention. Therefore, we will give the V.P. and the experts access to the model, so that they can verify that it will operate effectively and appropriately. We propose that every week, the model be run to estimate the probabilities of churn of the customers whose contracts expire in 45 days (give or take a week). The customers will be ranked based on these probabilities, and the top N will be selected to receive the current incentive, with N based on the cost of the incentive and the weekly retention budget.

Ejercicio II: Whiz-Bang Widget

Your company has an installed user base of 900,000 current users of your Whiz-bang® widget. You now have developed Whiz-bang® 2.0, which has substantially lower operating costs than the original. Ideally, you would like to convert (“migrate”) your entire user base over to version 2.0; however, using 2.0 requires that users master the new interface, and there is a serious risk that in attempting to do so, the customers will become frustrated and not convert, become less satisfied with the company, or in the worst case, switch to your competitor’s popular Boppo® widget. Marketing has designed a brand-new migration incentive plan, which will cost \$250 per selected customer. There is no guarantee that a customer will choose to migrate even if she takes this incentive.

An external firm, Big Red Consulting, is proposing a plan to target customers carefully for Whiz-bang® 2.0, and given your demonstrated fluency with the fundamentals of data science, you are called in to help assess Big Red’s proposal. Do Big Red’s choices seem correct?

Ejercicio II: Whiz-Bang Widget

Targeted Whiz-bang Customer Migration—prepared by Big Red Consulting, Inc.

We will develop a predictive model using modern data-mining technology. As discussed in our last meeting, we assume a budget of \$5,000,000 for this phase of customer migration; adjusting the plan for other budgets is straightforward. Thus we can target 20,000 customers under this budget. Here is how we will select those customers:

We will use data to build a model of whether or not a customer will migrate given the incentive. The dataset will comprise a set of attributes of customers, such as the number and type of prior customer service interactions, level of usage of the widget, location of the customer, estimated technical sophistication, tenure with the firm, and other loyalty indicators, such as number of other firm products and services in use. The target will be whether or not the customer will migrate to the new widget if he/she is given the incentive. Using these data, we will build a linear regression to estimate the target variable. The model will be evaluated based on its accuracy on these data; in particular, we want to ensure that the accuracy is substantially greater than if we targeted randomly.

To use the model: for each customer we will apply the regression model to estimate the target variable. If the estimate is greater than 0.5, we will predict that the customer will migrate; otherwise, we will say the customer will not migrate. We then will select at random 20,000 customers from those predicted to migrate, and these 20,000 will be the recommended targets.

Big Data: Lecciones aprendidas



Grupo de Sistemas *inteligentes*



Eduardo M. Sánchez Vila
eduardo.sanchez.vila@usc.es

Grupo de Sistemas Inteligentes
Universidad de Santiago de Compostela