uality.code>
SOLUTIONS FOR YOUR BUSINESS

Introducción a BigData

ANDRES TRIANA
Microsoft Certified Trainer (MCT)

CONTACTO
andres.triana@qualitycode.com.co
319 506 0662
www.qualitycode.com.co

Temario del Curso

Modulo 1
Fundamentos I

Modulo 2
Fundamentos II

Modulo 3
Hadoop - Taller Práctico

Modulo 4
Hive - Taller Práctico

Modulo 5
HBase - Taller Práctico

Modulo 6
Spark - Taller Práctico

BIG DATA

Quality.code>
SOLUTIONS FOR YOUR BUSINESS

¿Que es Big Data?

Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día. Pero no es la cantidad de datos lo que es importante. **Lo que importa con el Big Data es lo que las organizaciones hacen con los datos**. Big Data se puede analizar para obtener ideas que conduzcan a mejores decisiones y movimientos de negocios estratégicos.

Capturar · almacenar · buscar · compartir · analizar · visualizar · procesar · entender

Características del Big Data

- Es a menudo generada **automáticamente** por una máquina o proceso (video, sensores, web data)
- Es típicamente una **nueva** fuente de datos (como la captura de comportamiento de exploración de los clientes)
- **No** está diseñada para ser **amigable**
- Es descrita como no estructurada aunque la mayoría está al menos **semi-estructurada**
- Las fuentes estructuradas son aquellas que ya conocemos de manera tradicional

Ejemplo

```
2010-02-10 00:01:07 W3SVC1446 WEB100 216.167.204.29 GET /tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-
downloaded-error-in-ie/ - 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+
(KHTML,+like+Gecko)+Version/4.0.3+Safari/531.9 - http://www.google.com/search?hl=en&client=safari&rls=en&q=Itunes%
3A+your+current+security+settings+do+not+allow+this+program+to+be+downloaded&aq=f&aqi=&oq= blog.caneja.com 200 0 0 8530 621 982
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/contact-form-7/stylesheet.css ver=2.0.7 80 - 12.178.189.252
HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)+Version/4.0.3+Safari/531.9 -
http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com
200 0 0 811 475 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-includes/js/wp-ajax-response.js ver=2.9.1 80 - 12.178.189.252 HTTP/1.1
Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)+Version/4.0.3+Safari/531.9 -
http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com
200 0 0 1537 446 124
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/js/wp-ajax-edit-comments.js ver=2.3 80 -
12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)
+Version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-
error-in-ie/ blog.caneja.com 200 0 0 5941 478 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/lightbox-plus/css/elegant/colorbox.css - 80 - 12.178.189.252
HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)+Version/4.0.3+Safari/531.9 -
http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com
200 0 0 1365 474 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/css/themes/circular/edit-comments.css - 8
- 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)
+Version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-
error-in-ie/ blog.caneja.com 200 0 0 1414 495 109
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/css/colorbox/colorbox.css - 80 -
12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)
+Version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-
error-in-ie/ blog.caneja.com 200 0 0 1443 483 109
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/lightbox-plus/js/jquery.colorbox-min.js ver=1.3.1 80 -
12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)
+Version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-
error-in-ie/ blog.caneja.com 200 0 0 4490 470 296
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/js/jquery.colorbox-min.js ver=2.9.1 80 -
12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+U;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+like+Gecko)
```

¿El tamaño importa?

¿El tamaño importa?

15 Mb

¿El tamaño importa?

15 Mb

1.5 Gb

¿El tamaño importa?

15 Mb

1.5 Gb

15 Tb

¿La velocidad importa?

NETFLIX

amazon

4 PetaBytes
X día

¿La velocidad importa?

4 PetaBytes
X día

NETFLIX

3 PetaBytes
X día

amazon

¿La velocidad importa?

4 PetaBytes
X día

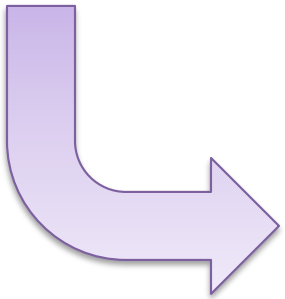
3 PetaBytes
X día

2.5 PetaBytes
X día

¿Por qué el Big Data es tan importante?

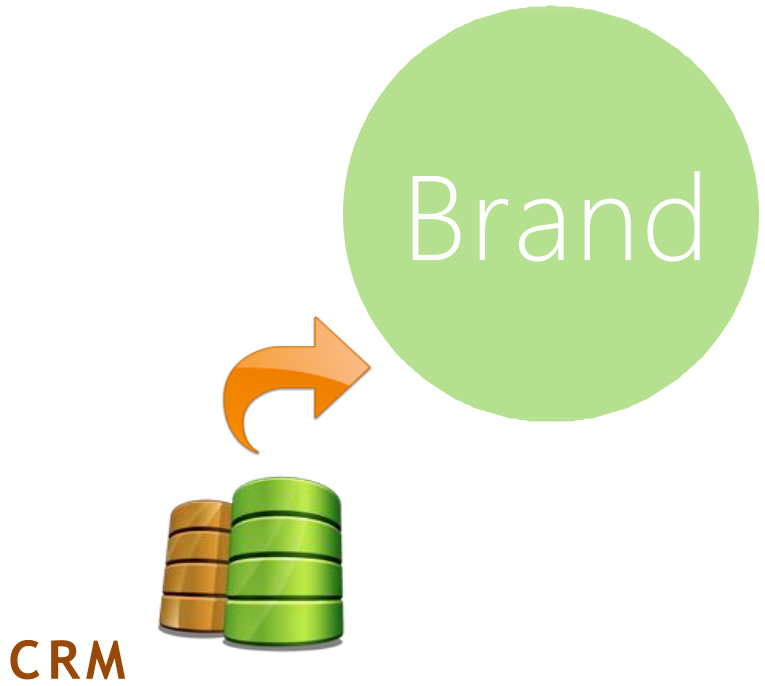
El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades.

- Reducción de coste
- Más rápido, mejor toma de decisiones.
- Nuevos productos y servicios



- Turismo
- Cuidado de la Salud
- Administración
- Retail
- Empresas Manufactureras
- Publicidad

Brand



Mobile

CRM

Web / Communities

Mobile

Brand

CRM

Web / Communities

Información Online

Mobile

Brand

CRM

Quality.code>
SOLUTIONS FOR YOUR BUSINESS

Web / Communities

Información Online

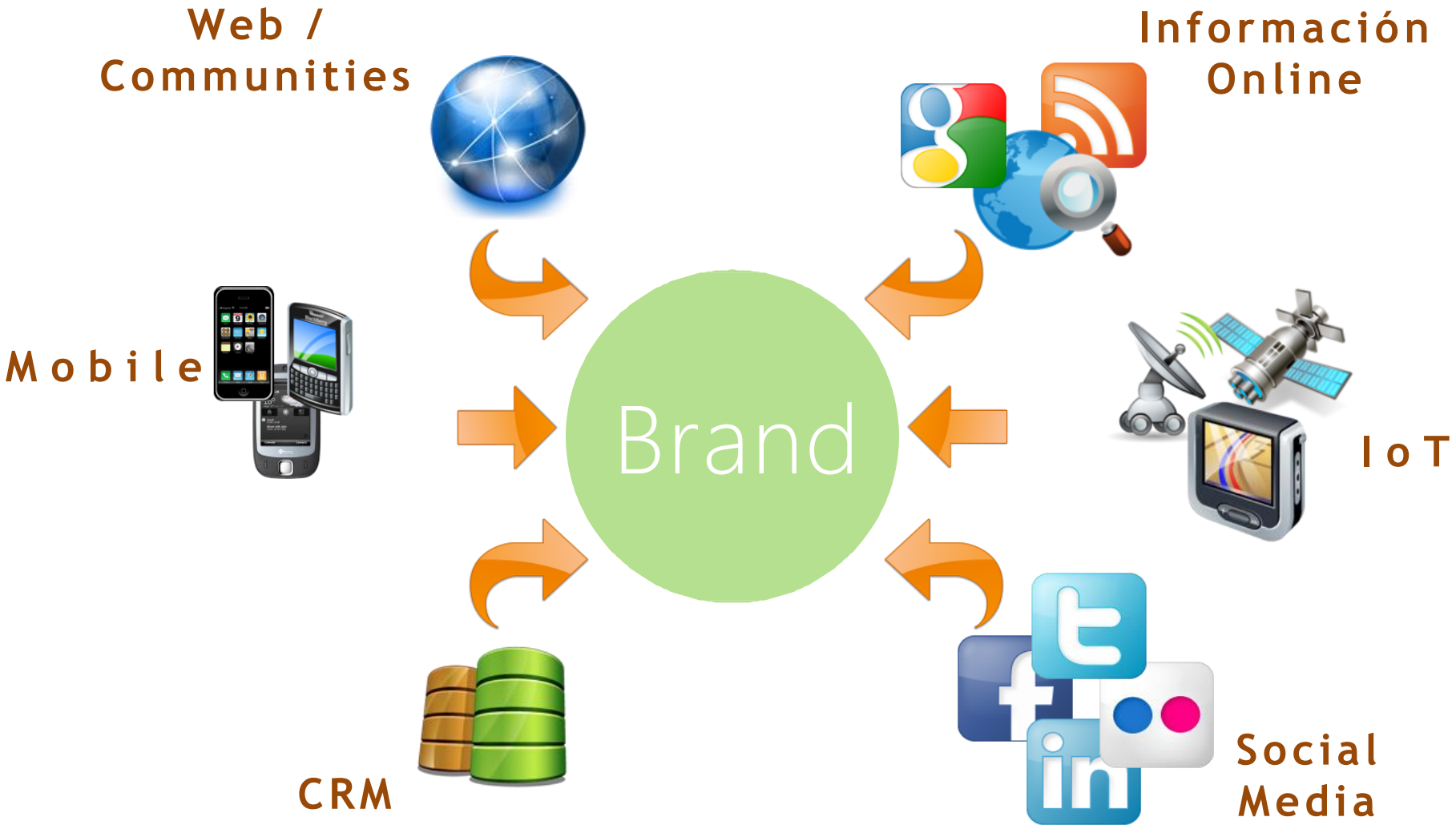
Mobile

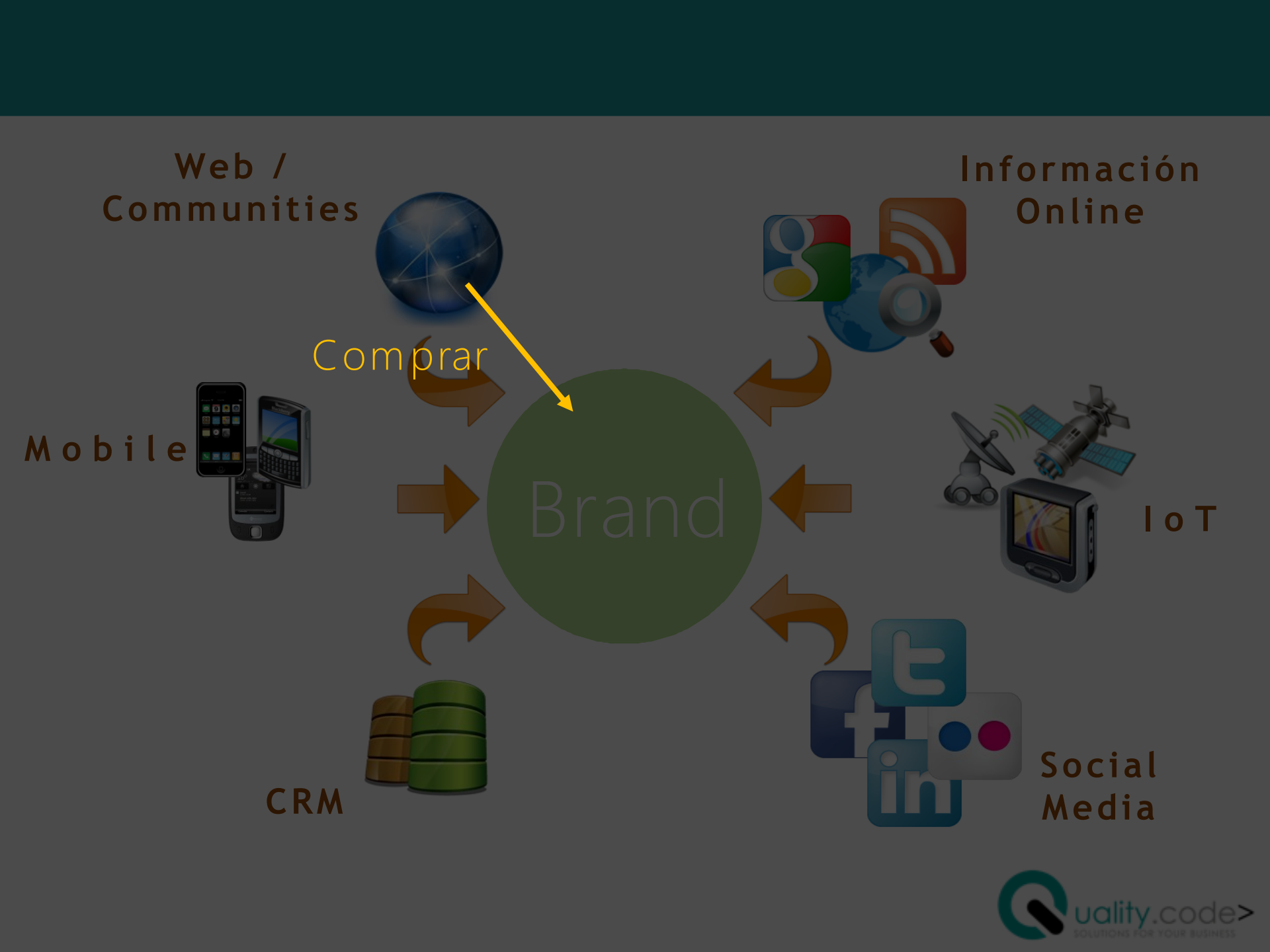
Brand

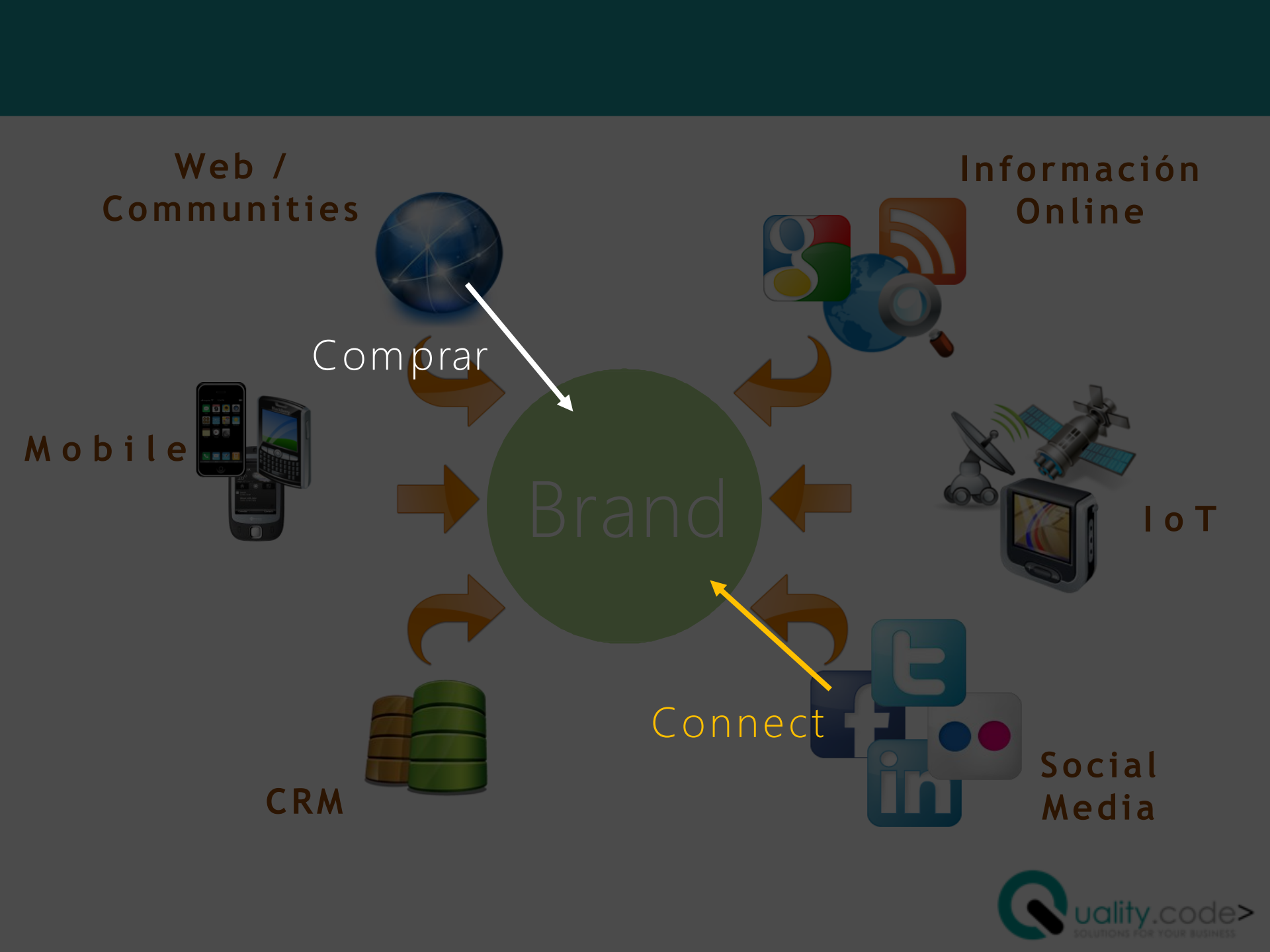
IoT

CRM

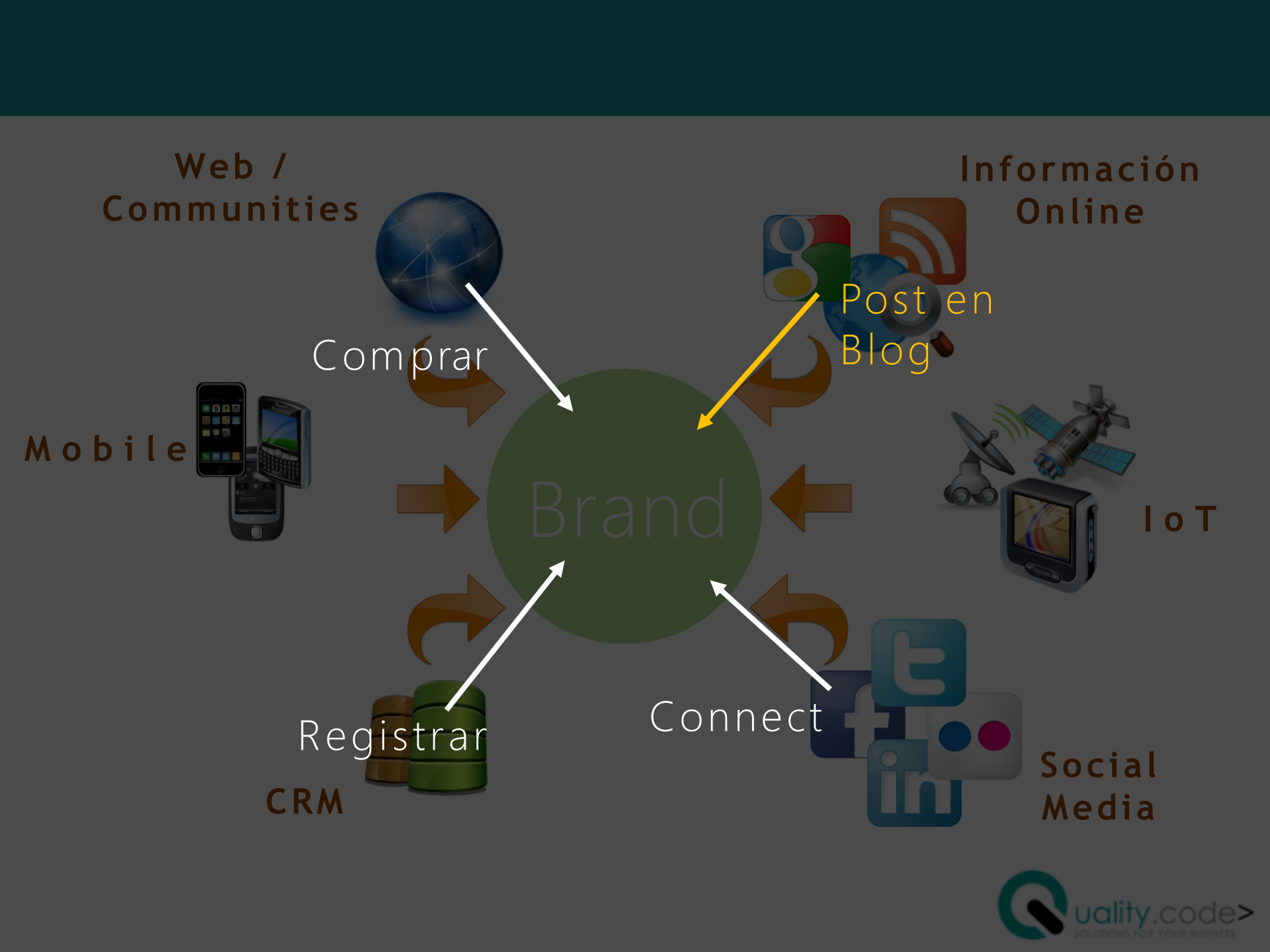
Quality.code>
SOLUTIONS FOR YOUR BUSINESS

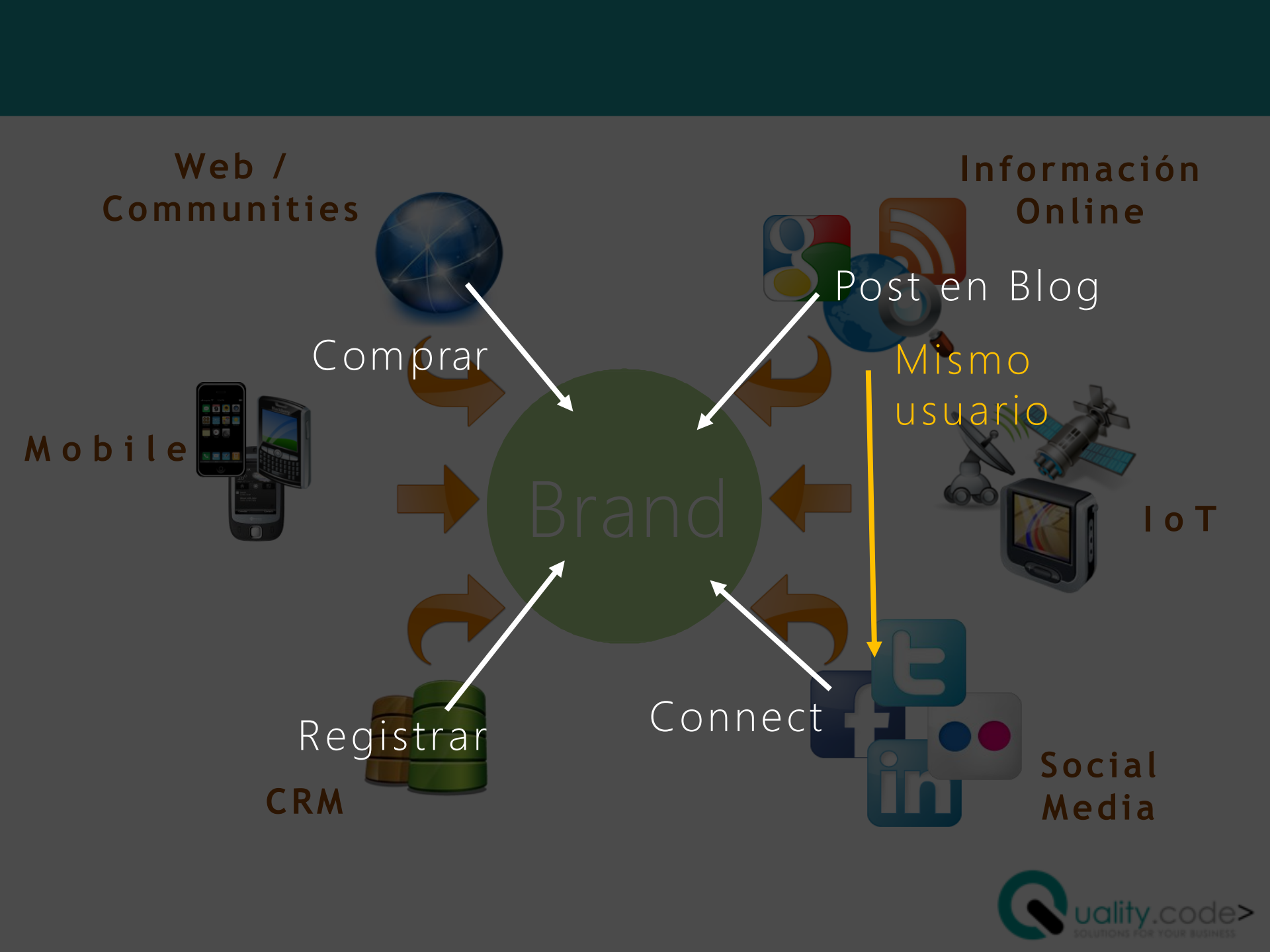

Web / Communities
Información Online
Mobile
Brand
IoT
CRM
Social Media

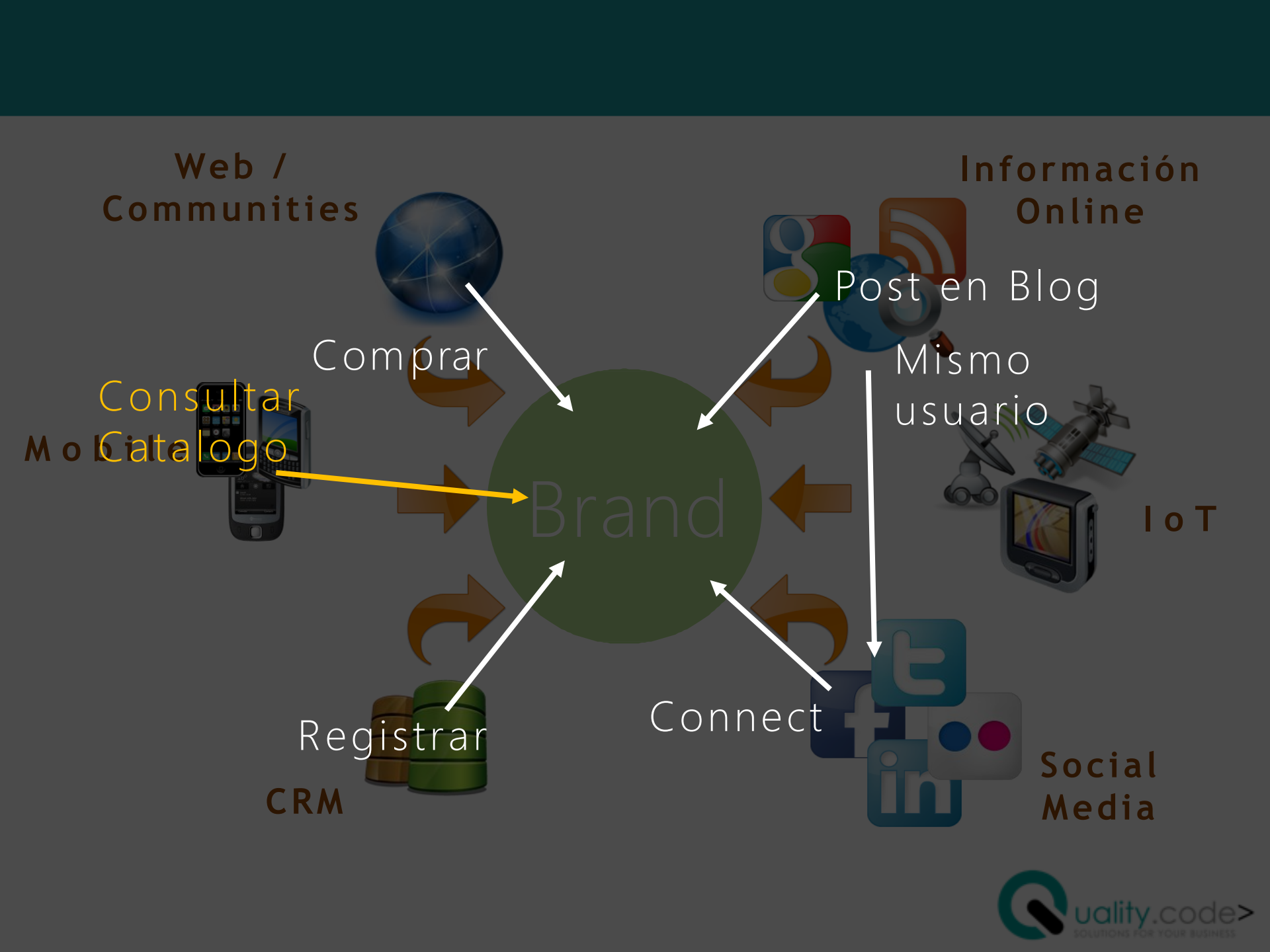

Web / Communities
Información Online
Comprar
Mobile
Brand
IoT
CRM
Social Media

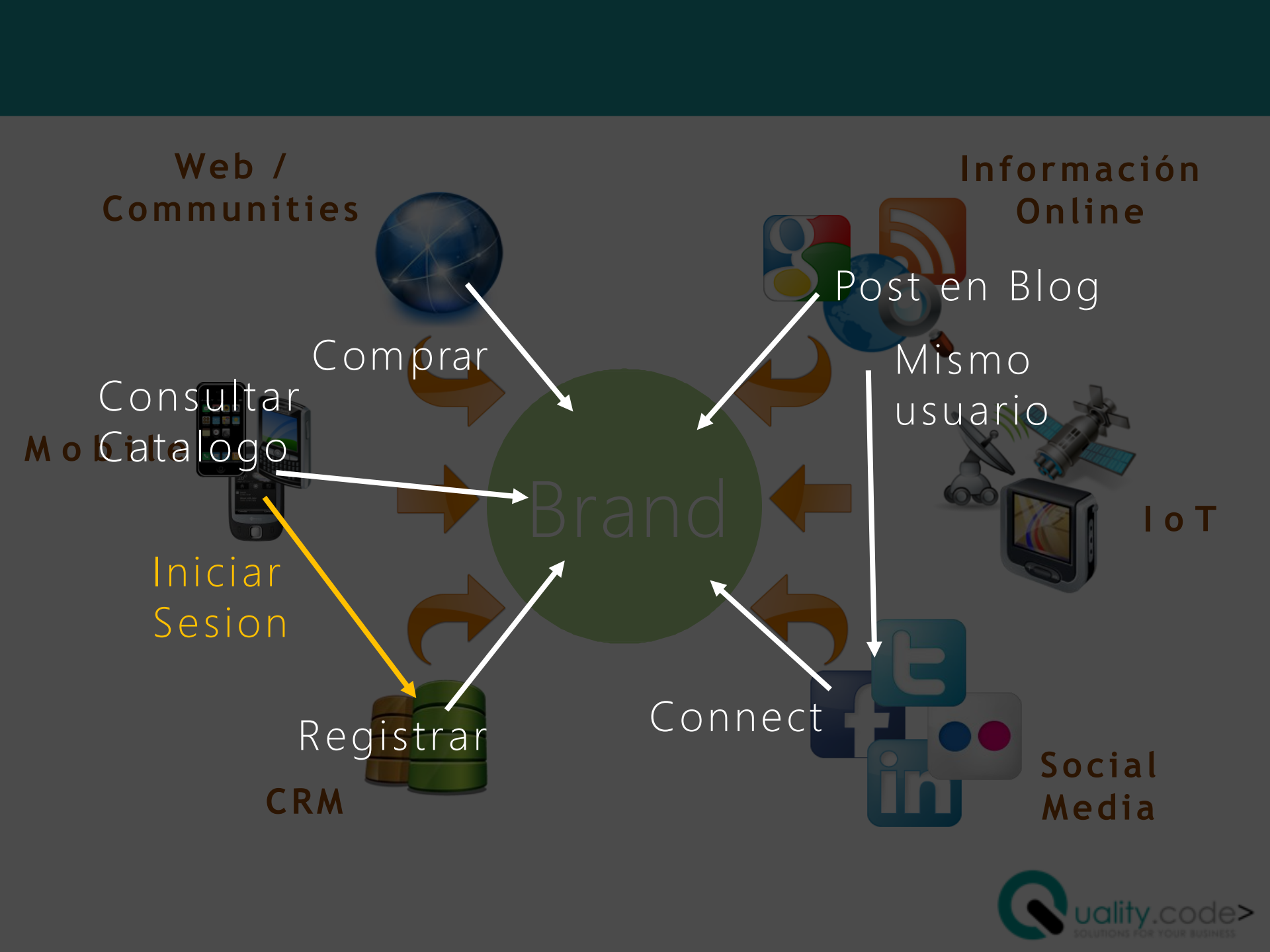

Web / Communities
Información Online
Comprar
Mobile
Brand
IoT
Connect
CRM
Social Media

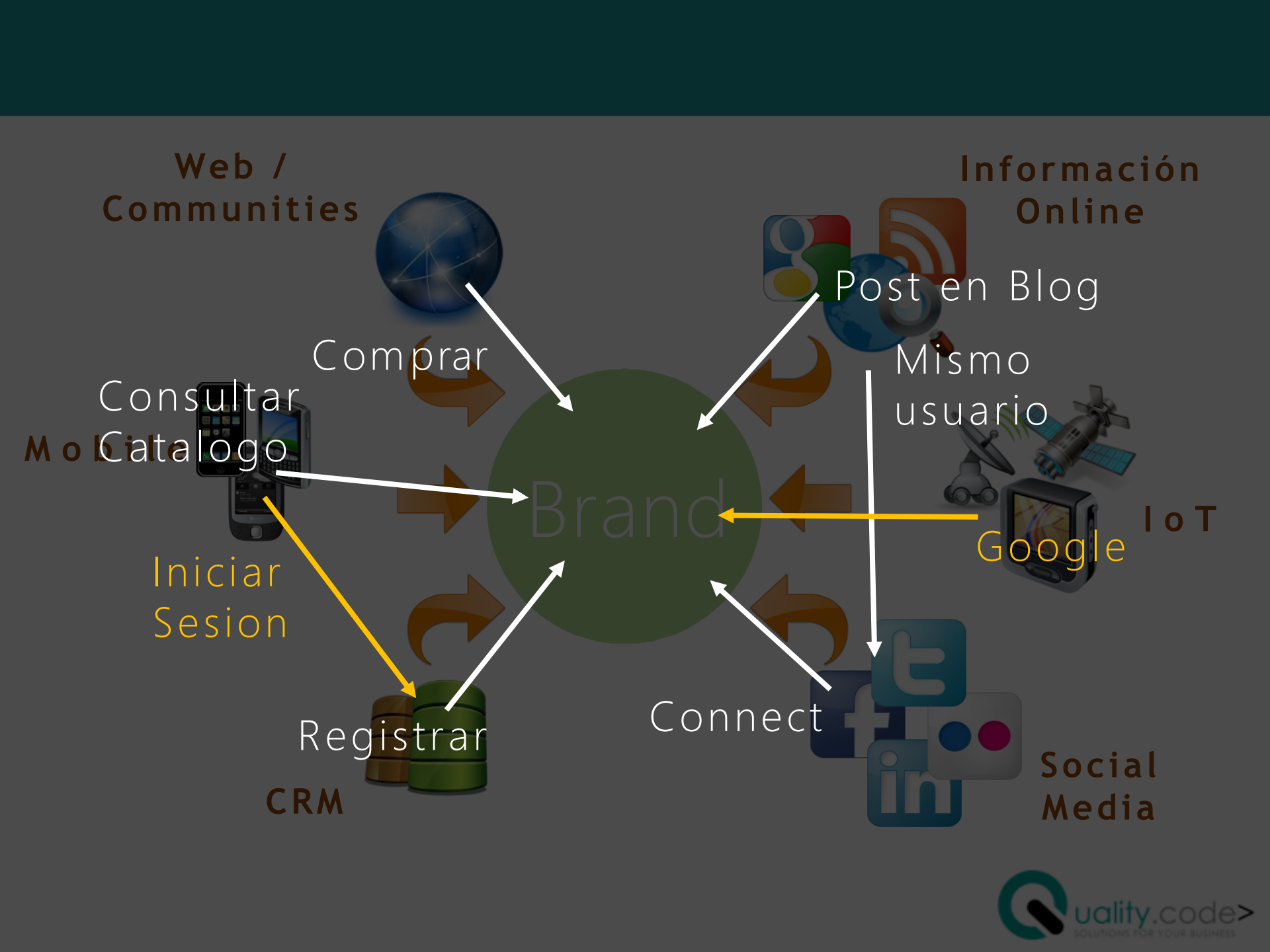

Web / Communities
Información Online
Comprar
Mobile
Brand
IoT
Registrar
Connect
CRM
Social Media

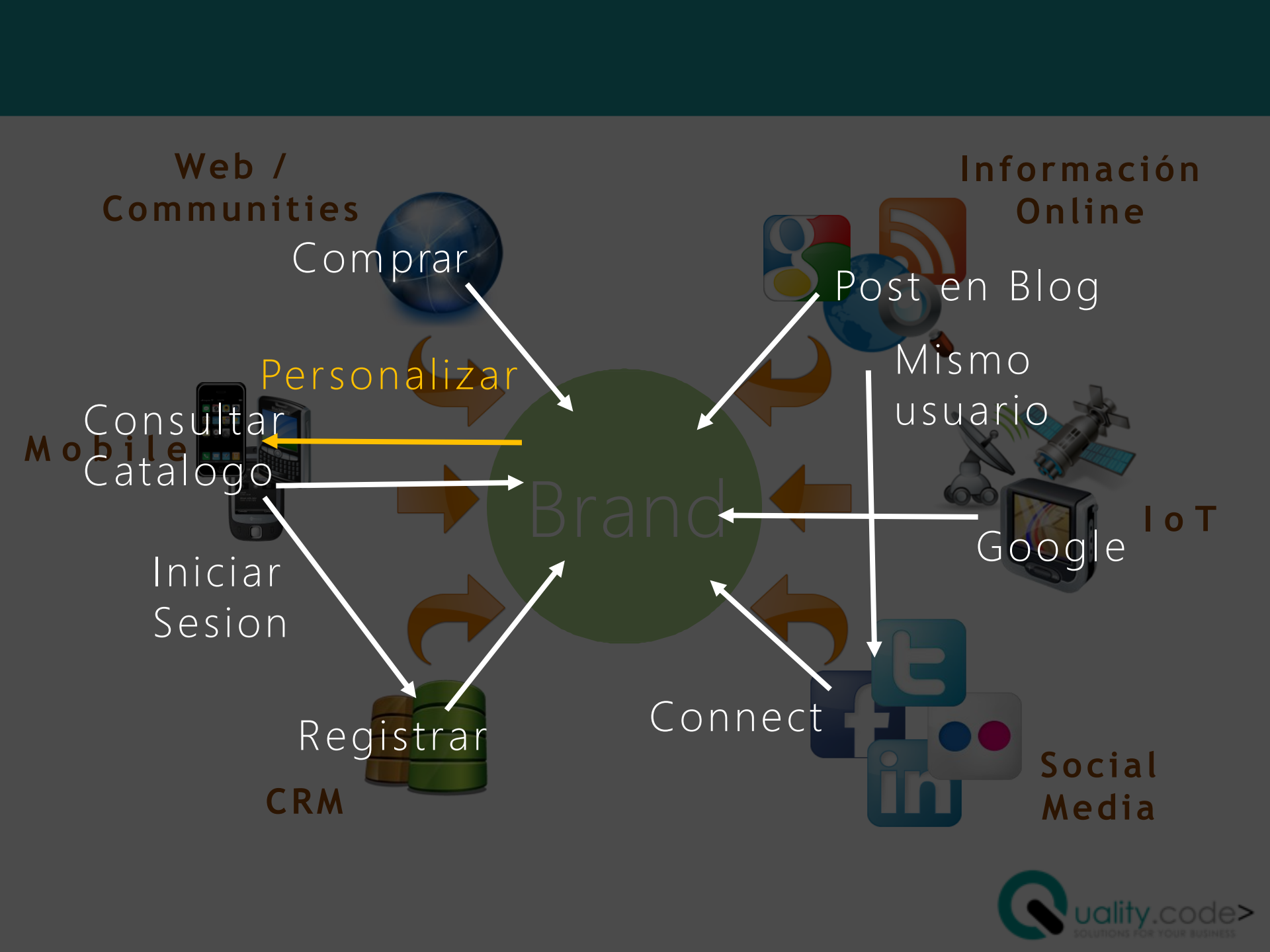

Web / Communities
Información Online
Comprar
Post en Blog
Mobile
Brand
IoT
Registrar
Connect
CRM
Social Media

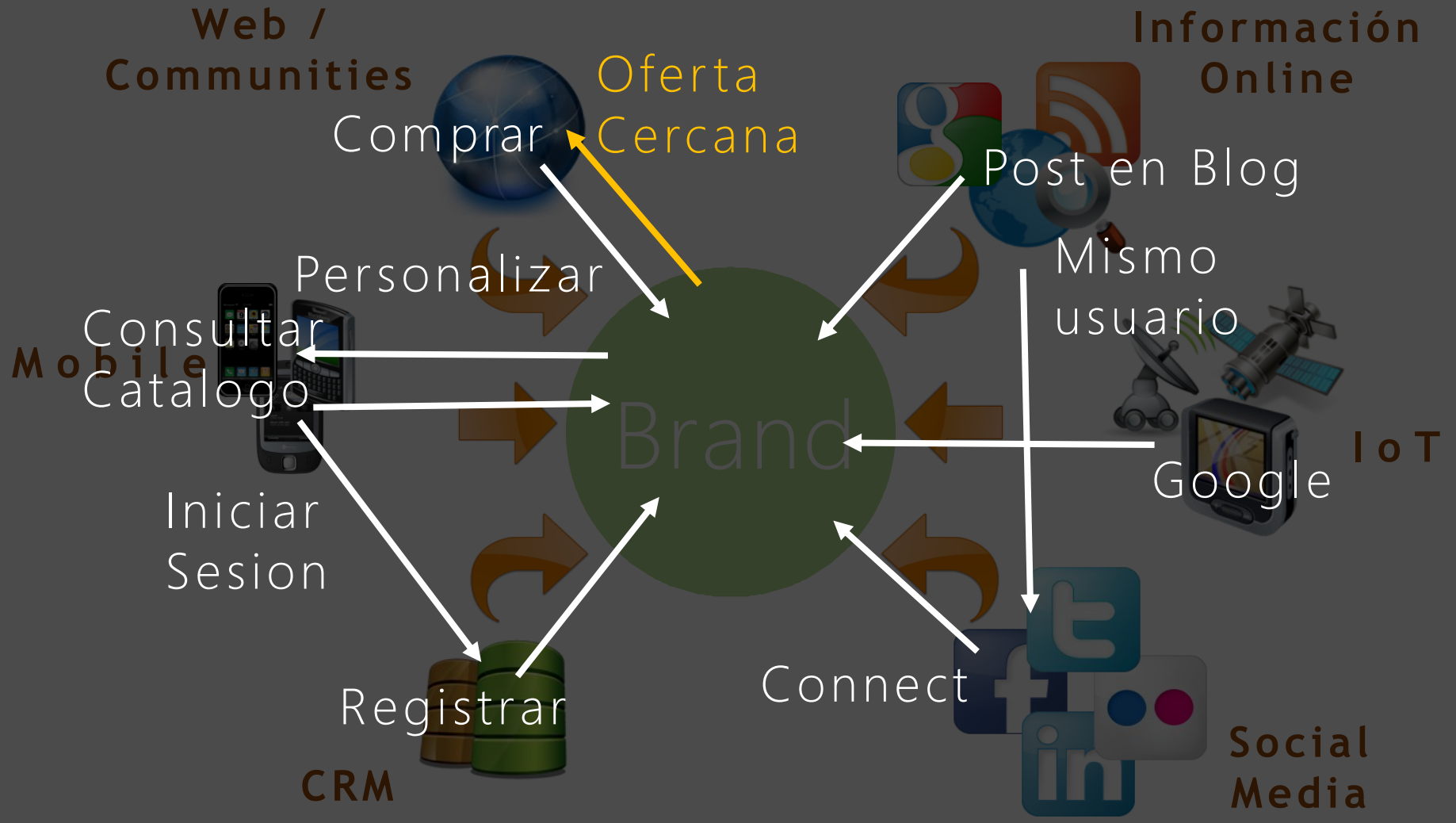

Web / Communities
Información Online
Post en Blog
Comprar
Mismo usuario
Mobile
Brand
IoT
Registrar
Connect
CRM
Social Media


Web / Communities
Información Online
Post en Blog
Comprar
Mismo usuario
Consultar Catalogo
Mobile
Brand
IoT
Registrar
Connect
CRM
Social Media


Web / Communities
Información Online
Post en Blog
Comprar
Mismo usuario
Consultar Catalogo
Mobile
Brand
IoT
Iniciar Sesion
Registrar
CRM
Connect
Social Media


Web / Communities
Información Online
Post en Blog
Comprar
Mismo usuario
Consultar Catalogo
Mobile
Brand
IoT
Google
Iniciar Sesion
Registrar
CRM
Connect
Social Media

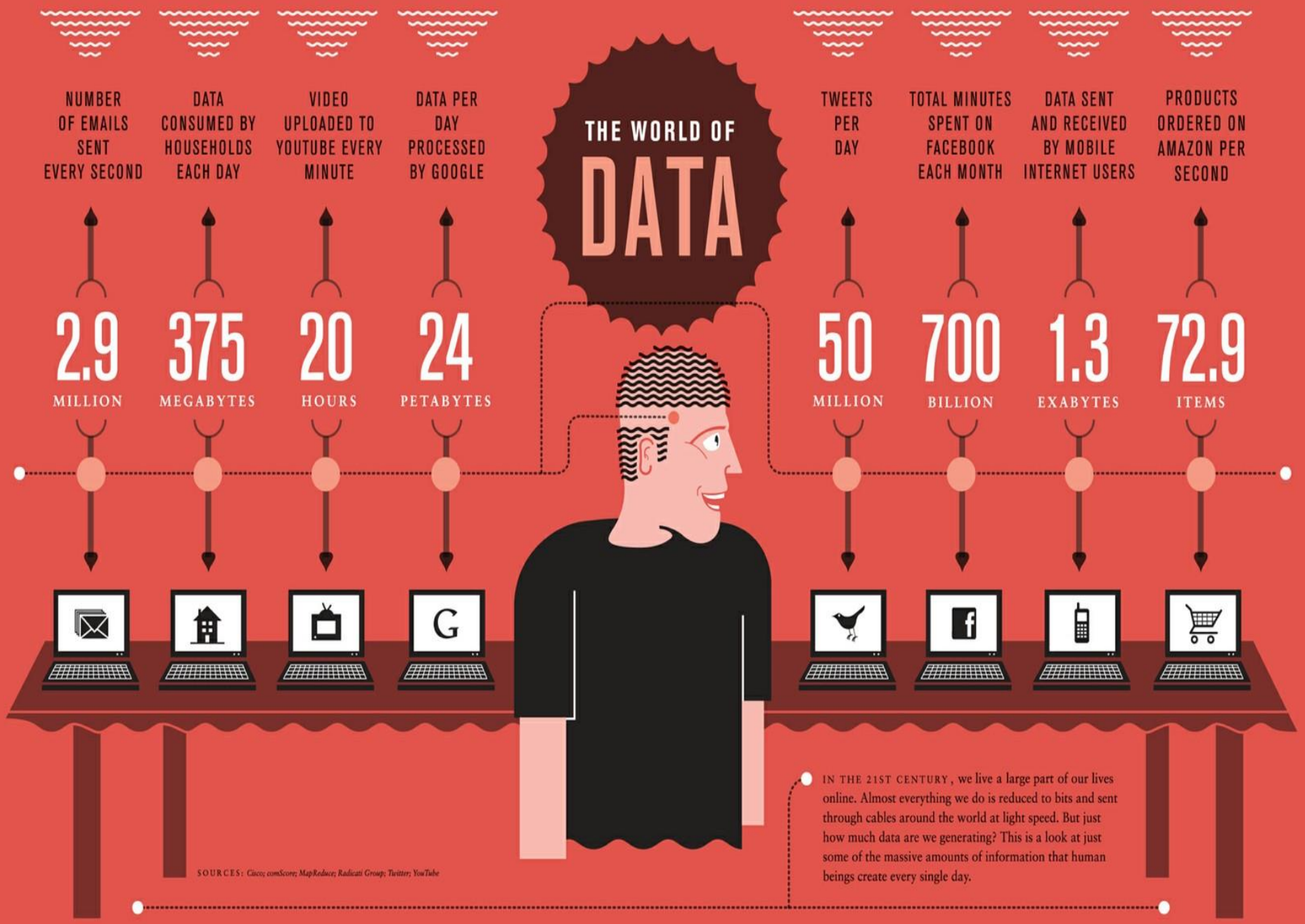

Web / Communities
Comprar
Personalizar
Consultar Catalogo
Mobile
Iniciar Sesion
Registrar
CRM
Brand
Información Online
Post en Blog
Mismo usuario
Google
IoT
Connect
Social Media



La información es Poder

Un informe del Foro Económico Mundial declaró que los datos constituyen una nueva clase de activo económico, como la moneda o el oro.

THE WORLD OF DATA



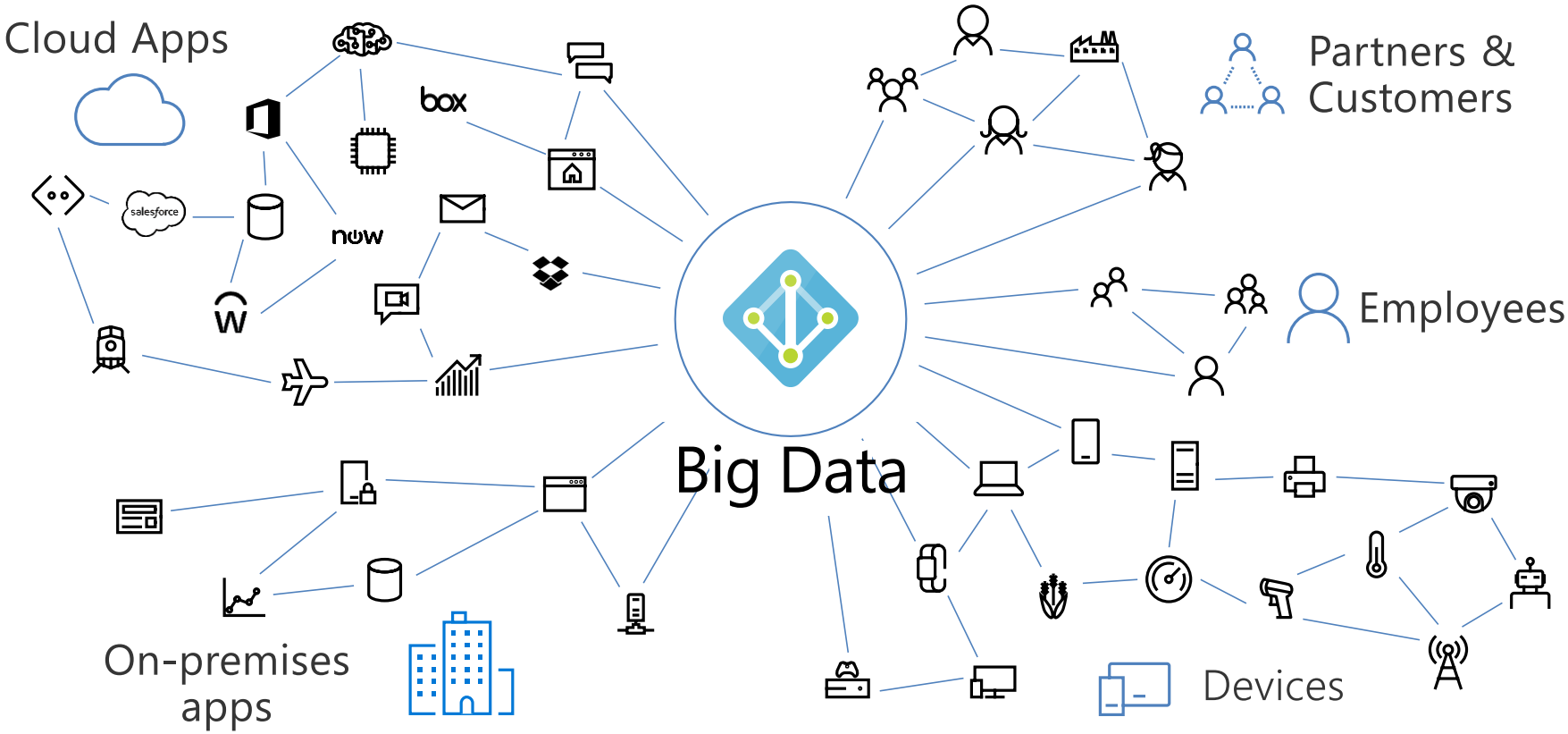
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

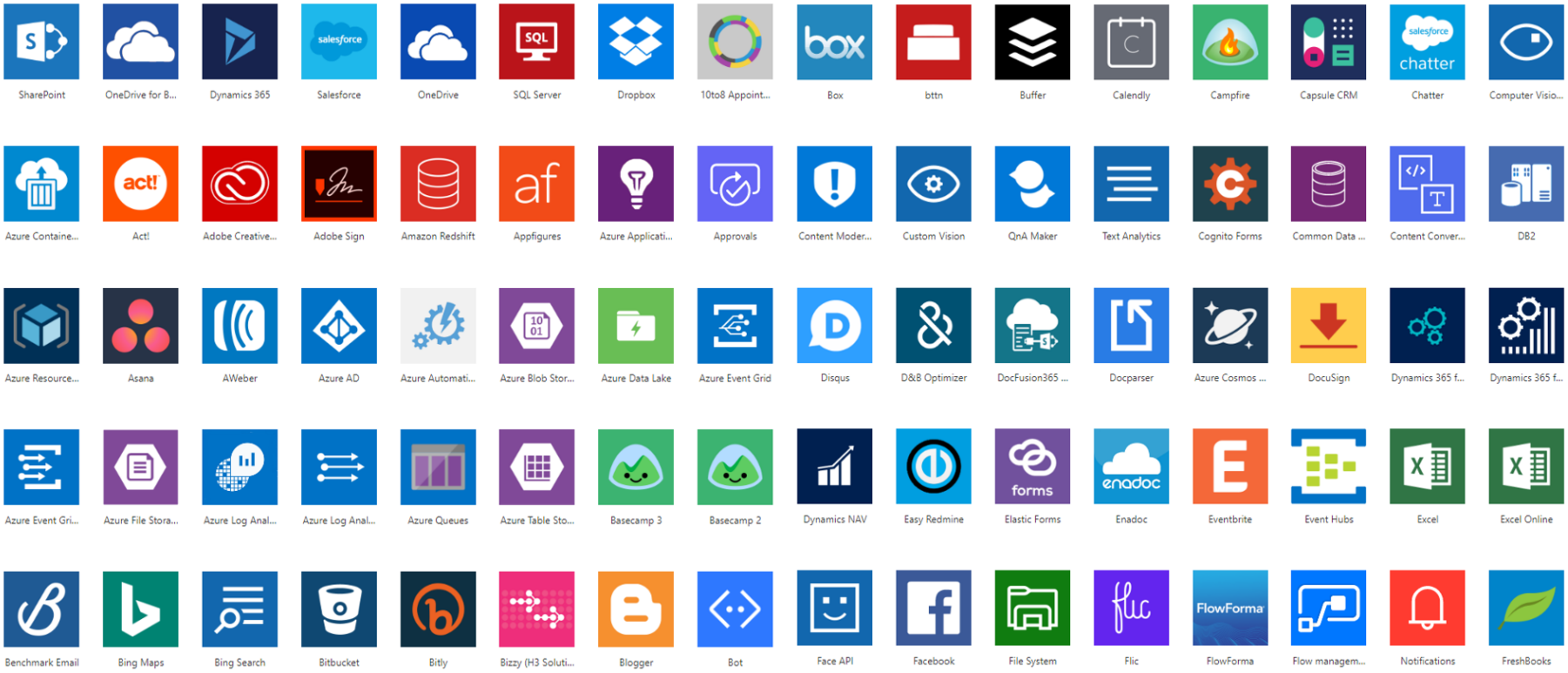
SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

Orígenes de Datos






QUIZ
uality.code>

Las 4V del Big Data

Volumen

Velocidad

Variedad

Veracidad

Las 4V del Big Data

Volumen: se refiere a la acumulación de gran cantidad de datos.

Según Gartner, en torno a 25 mil millones de dispositivos estarán conectados a la red en 2020.

Cual es el reto?

Las 4V del Big Data

Velocidad: se refiere a la rapidez a la que se generan los datos en la actualidad.

La clave está en saber recopilar, gestionar, analizar y sobre todo extraer información de negocio de valor en tiempo real.

Las 4V del Big Data

Variedad: se refiere a la gran diferencia que hay en el formato, el origen (las características generales del dato), etc.

Formato estructurado, fácil de tratar (bases de datos estructuradas, formatos predefinidos, etc.)

No estructurados extraídos de redes sociales, de videos, etc

Las 4V del Big Data

Veracidad: se refiere a la calidad, la predictibilidad y la disponibilidad del dato.

Es la variable menos uniforme y menos sencilla de controlar, debido a la dificultad de cerciorarnos de que un dato es 100% fiable.

Eventos raros en Big Data

- Suponga que arroja una moneda 10 veces: considere los eventos en los que todos estos lanzamientos de monedas son caras.
  - Si repites este experimento mil millones de veces: esperarías ver 10 cabezas aproximadamente 1 millón de veces
- Reconocimiento facial: Sistema utilizado en un aeropuerto. Suponga que tiene una cámara de reconocimiento facial, sistema que es 99.9% confiable.

Suponga que el aeropuerto usa el sistema para detectar criminales conocidos mediante la comparación de rasgos faciales de cada persona a una cara de base de datos de delincuentes conocidos.

Si una persona es un criminal conocido, entonces la probabilidad que no se detectan es 0.01% - un falso negativo.

Si una persona no es un criminal, entonces él o ella está marcado como criminal con probabilidad 0.01% - un falso positivo.

Suponga que una de cada 10 millones de personas que ingresan al sistema son conocidos criminales. Esto significaría que 1,000 personas que no son criminales son marcados como delincuentes por el sistema.

MapReduce

Framework de hardware distribuido (cluster o grids) que divide los problemas en subproblemas (Map) y luego recopilas las mini-respuestas (Reduce) para generar conclusiones. La solución más común es Hadoop. Este modelo fue creado y promovido por Google.

NoSQL Database

Amplia clase de sistemas de gestión de bases de datos que difieren del modelo clásico del sistema de gestión de bases de datos relacionales (RDBMS) en múltiples aspectos: no usan SQL como el principal lenguaje de consultas, los datos almacenados no requieren estructuras fijas como tablas, no garantizan ACID y escalan bien horizontalmente (ej: MongoDB, Cassandra, BigTable)

Algoritmos genéticos

Un algoritmo es una serie de pasos organizados que describen el proceso que se debe seguir, para dar solución a un problema específico. Con la inteligencia artificial, surgieron los algoritmos genéticos, inspirados en la evolución biológica, quienes evolucionan con sometidos a mutaciones y recombinaciones genéticas.

Valor para el Cliente

Combinar la información de Social Media con analítica para ofrecer productos a sus clientes.

Correr su código de análisis en segundos en lugar de horas y días.

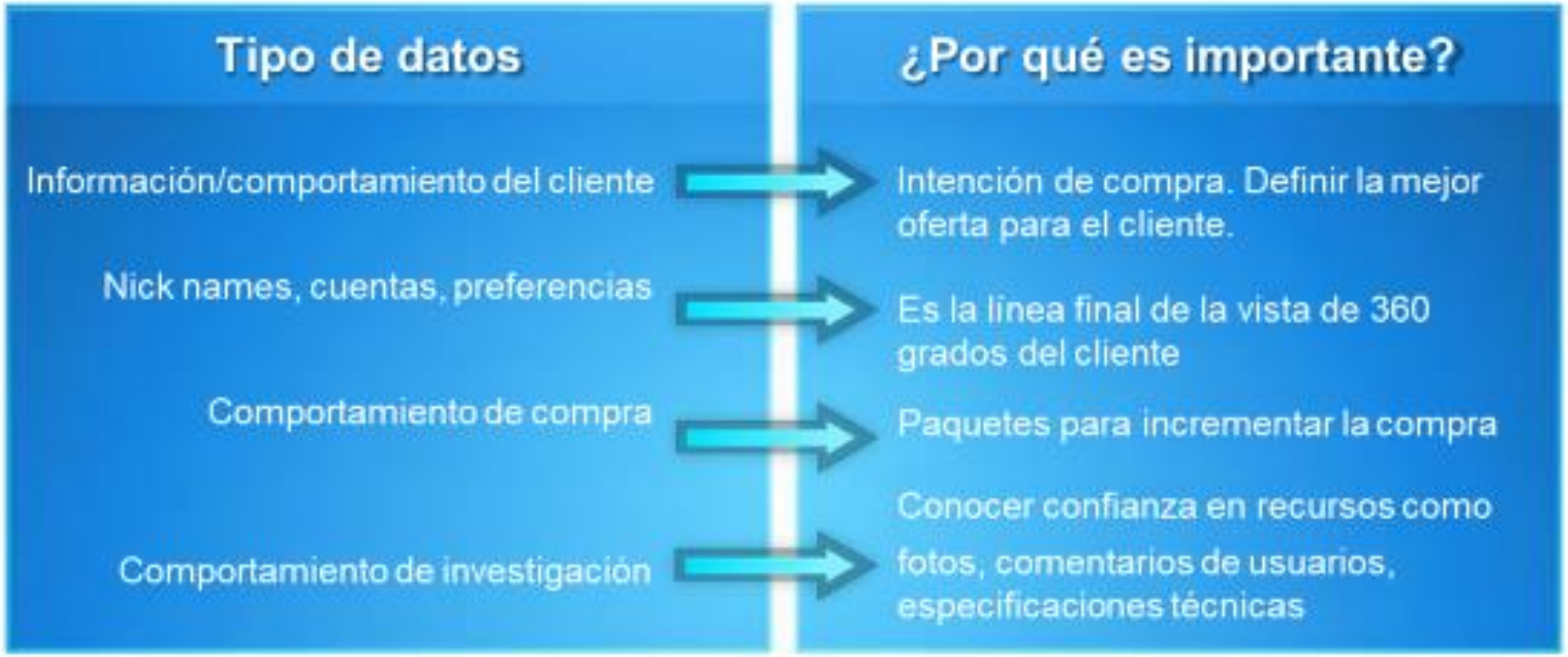
Predecir el comportamiento de compra y criterio de decisión de sus clientes varias semanas antes que la competencia.

Obtener el beneficio de ser el primero en ofrecer algo a tus clientes que no han sido identificados por sus competidores.

Responder a las necesidades del cliente con la última información generada.

Mejorar la experiencia de los clientes para incrementar su valor.

Web Data



Quality.code>
SOLUTIONS FOR YOUR BUSINESS

¡GRACIAS!