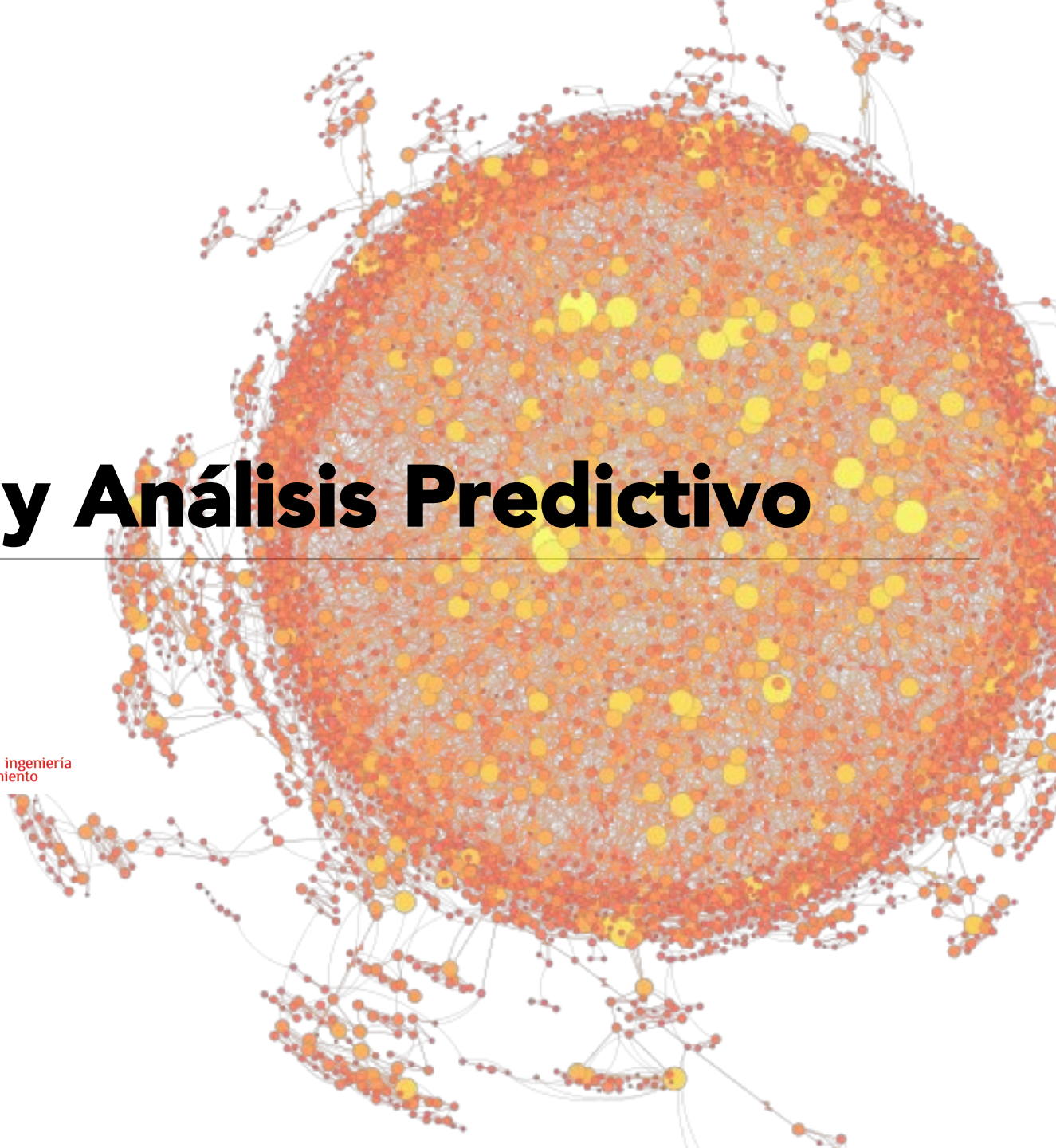
Big Data y Análisis Predictivo

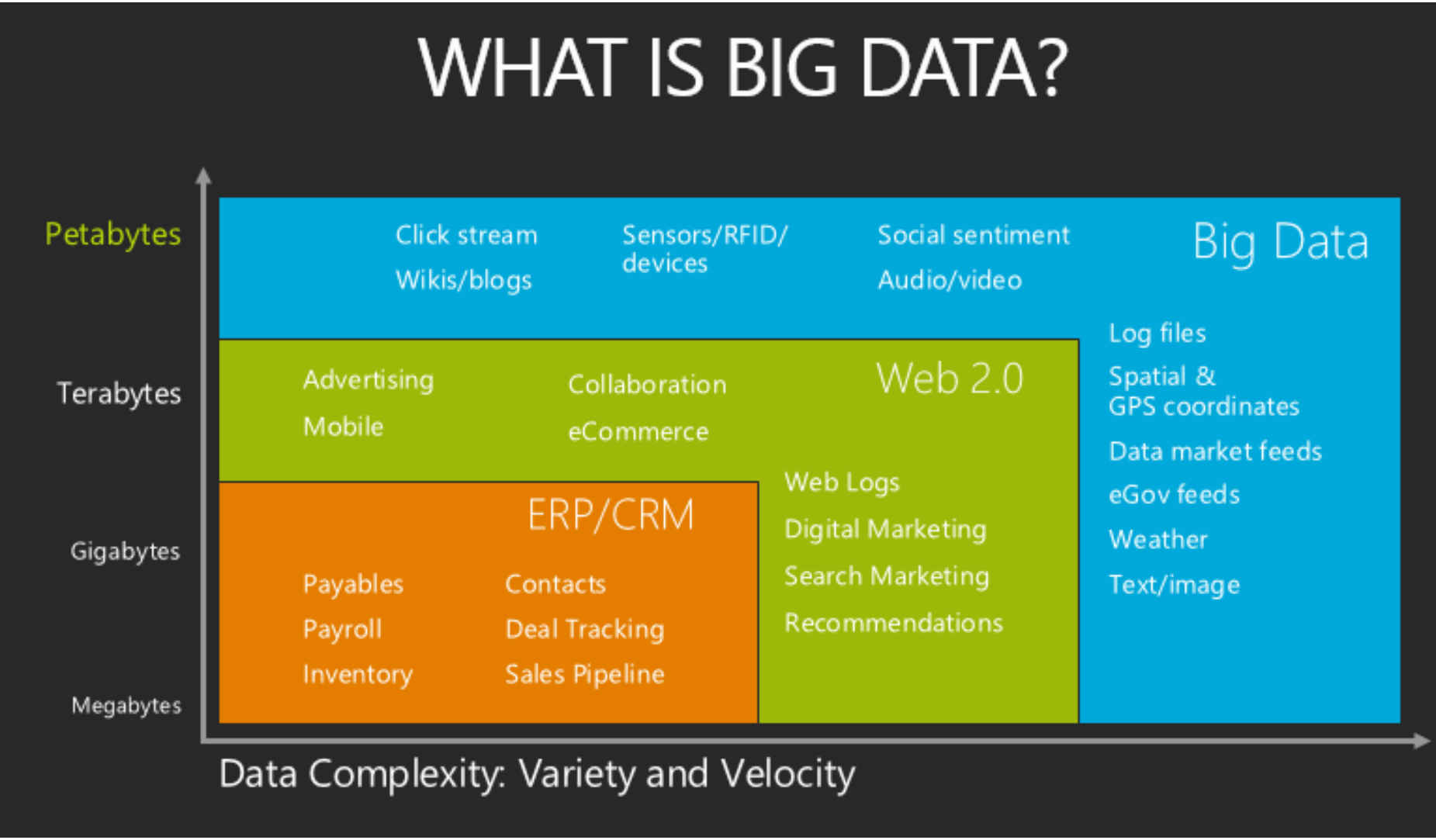
Esteban Moro
(UC3M+IIC)

Universidad Carlos III de Madrid

instituto de ingeniería del conocimiento

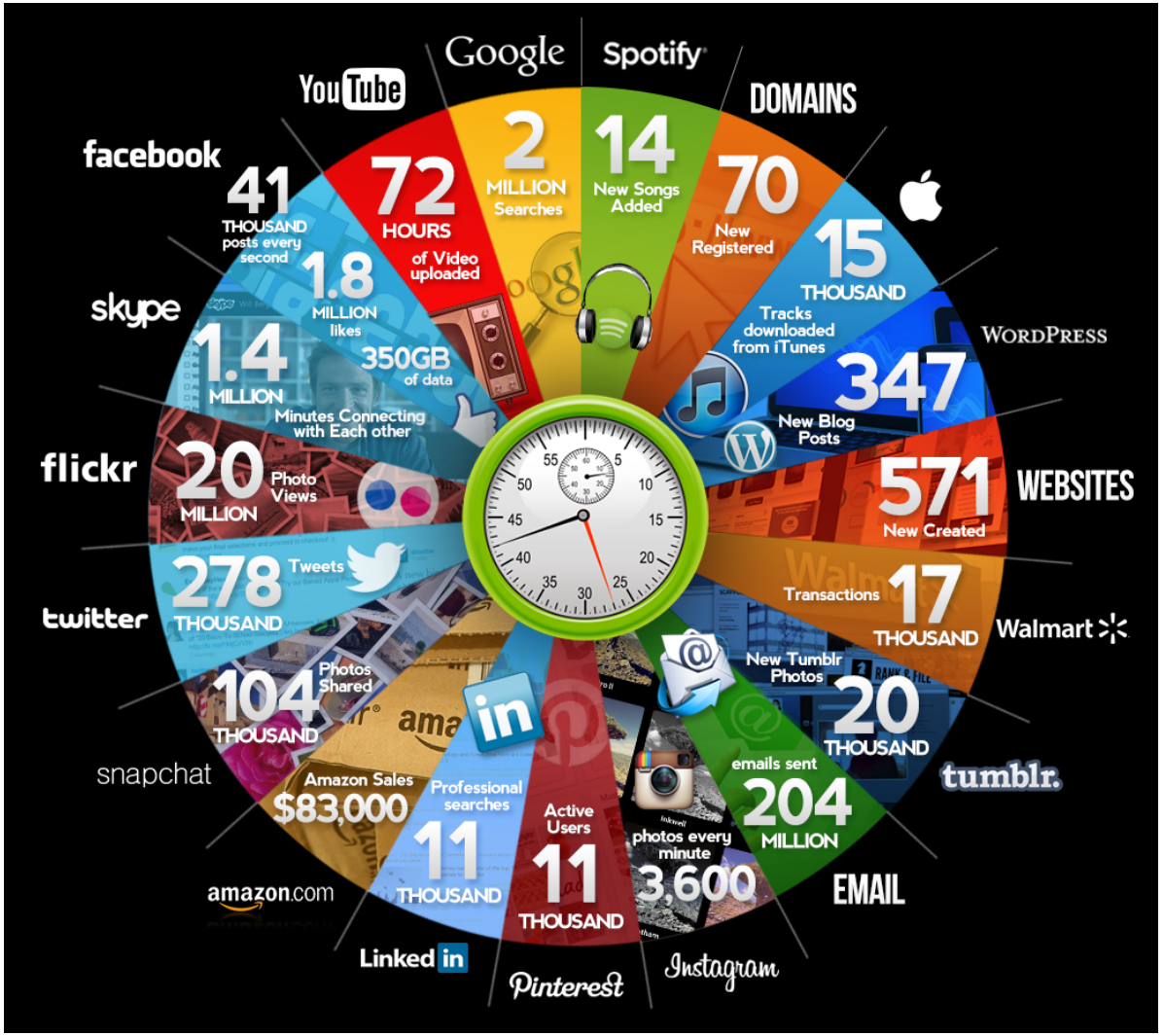


Qué Big Data?



http://blogs.msdn.com/b/data__knowledge__intelligence/archive/2013/02/18/big-data-big-deal.aspx

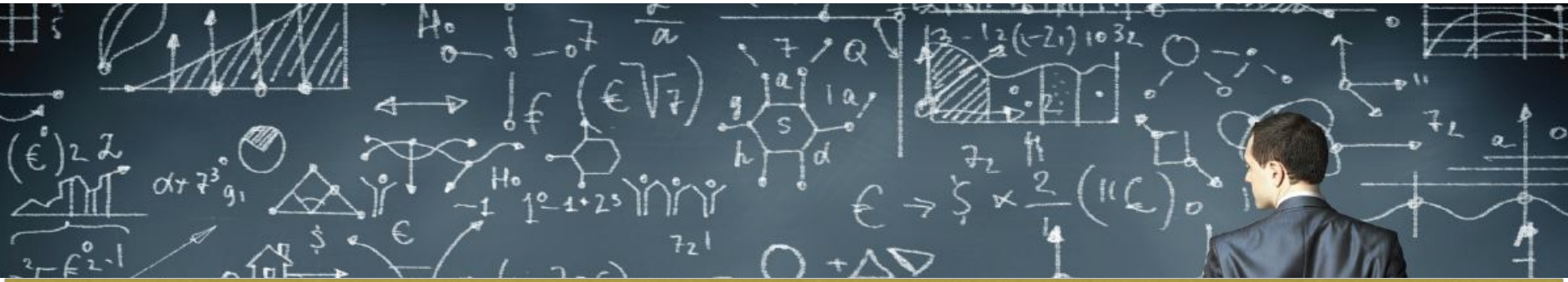
Every 60 seconds



http://blog.qmee.com/wp-content/uploads/2013/07/Qmee-Online-In-60-Seconds21.png

Por qué ahora?

Tecnología / Recursos



mahout

REVOLUTION ANALYTICS

sas

S4 distributed stream computing platform

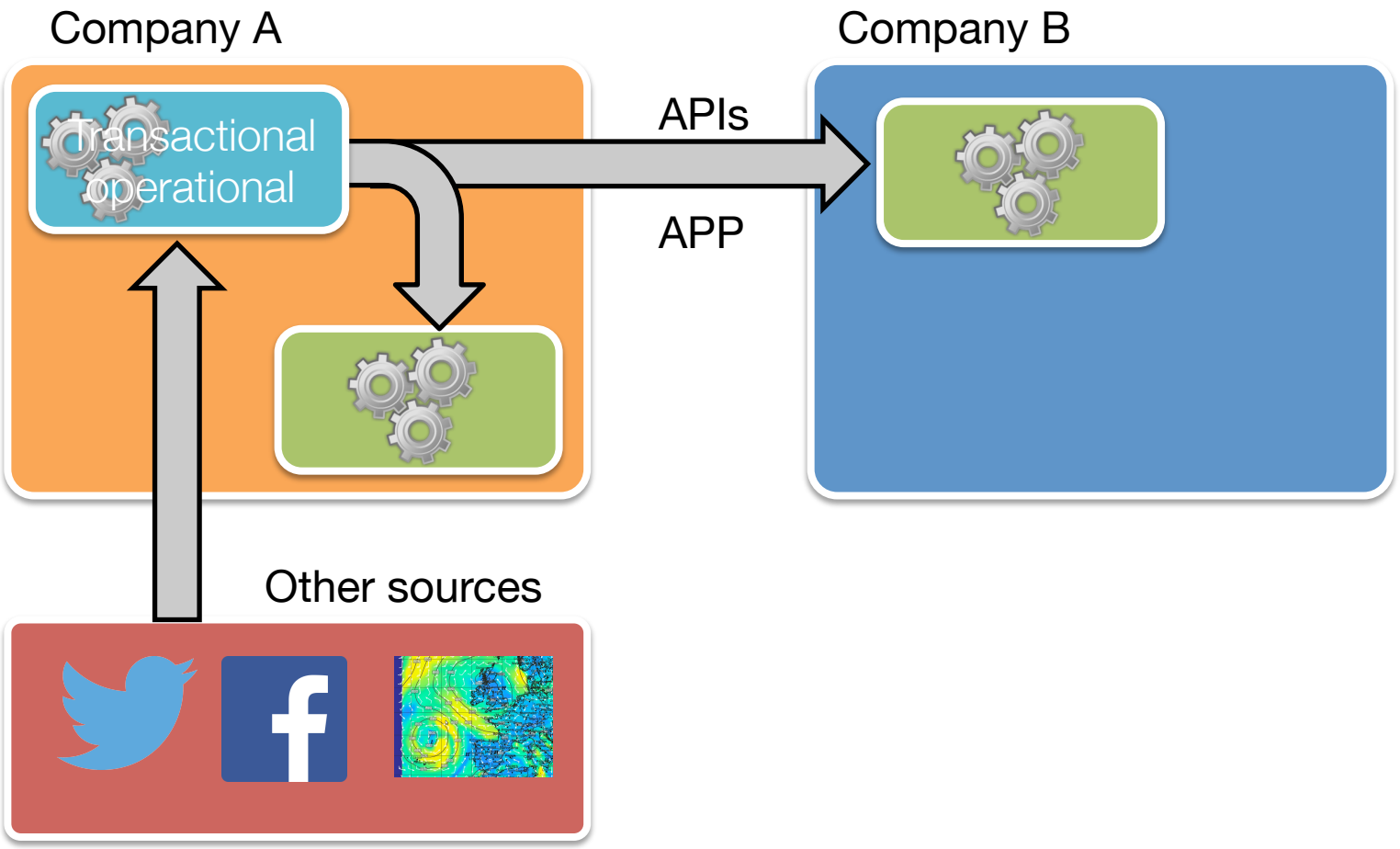
Spark

Storm

Ambari
Zookeeper
HBase
Pig
Hive
MapReduce
HCatalog
HDFS
hadoop

Google Bigtable
riak
CouchDB
redis
mongoDB
Neo4j the graph database
Cassandra
HBASE
CouchBase
amazon web services
Kyoto Cabinet

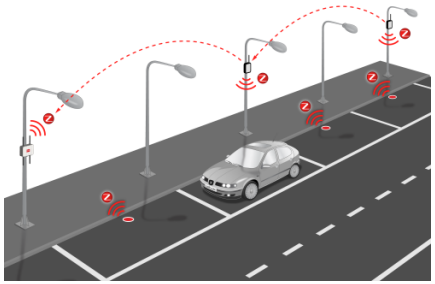
Flujo de BigData y valor



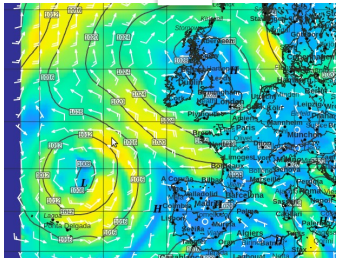
Fuentes de datos

Operacional
Transaccional

Comunicaciones



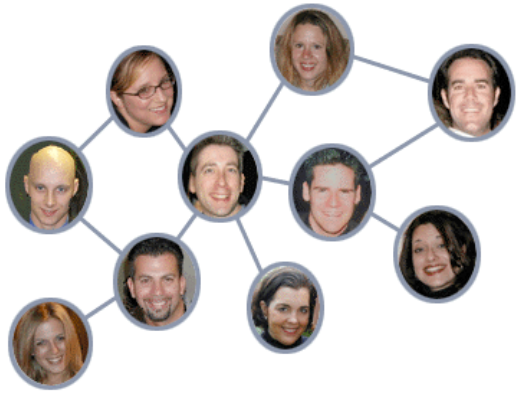
Sensores



Tiempo / Mercados

Tipos de datos

Disponemos de datos sociales, de movilidad y otros patrones de comportamiento



Interacción social

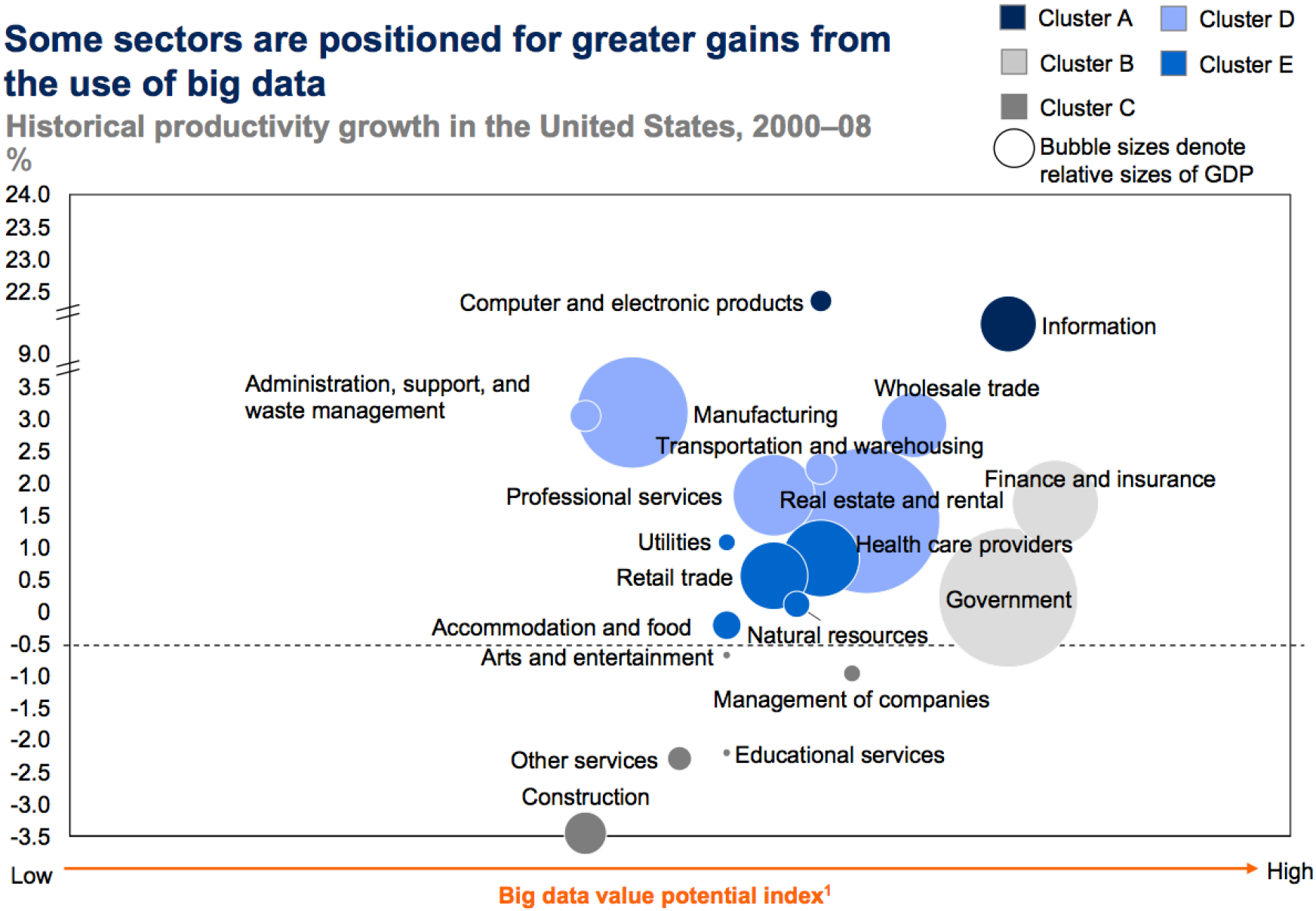
Patrones de comportamiento

Movilidad geográfica

Estas tres marcan el día a día de las personas

Some sectors are positioned for greater gains from the use of big data

Historical productivity growth in the United States, 2000–08



McKinsey Global Institute Big Data Report 2011

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Universidad Carlos III de Madrid

instituto de ingeniería del conocimiento

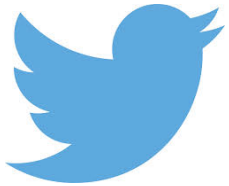
@estebanmoro

Valor

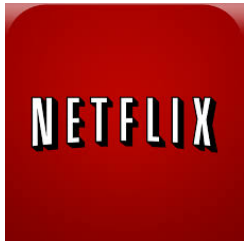
- Creando transparencia
- Reducir ineficiencias
- Permitiendo la experimentación para descubrir necesidades, exponer variabilidad y mejorar el rendimiento de procesos
- Reemplazar o ayudar decisiones humanas por algoritmos automatizados
- Crear nuevos modelos de negocio, productos o servicios

- McKinsey Global Institute Big Data Report 2011
- http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
- Demystifying Big Data: A Practical Guide To Transforming The Business of Government
- TechAmerica Foundation: Federal Big Data Commission

Casos de uso



- Social networks
- Movie recommendation
- Retail habits



  - http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=2&hp&pagewanted=all&

- Political campaigns
  - http://www.technologyreview.com/featuredstory/508836/how-obama-used-big-data-to-rally-voters-part-1/

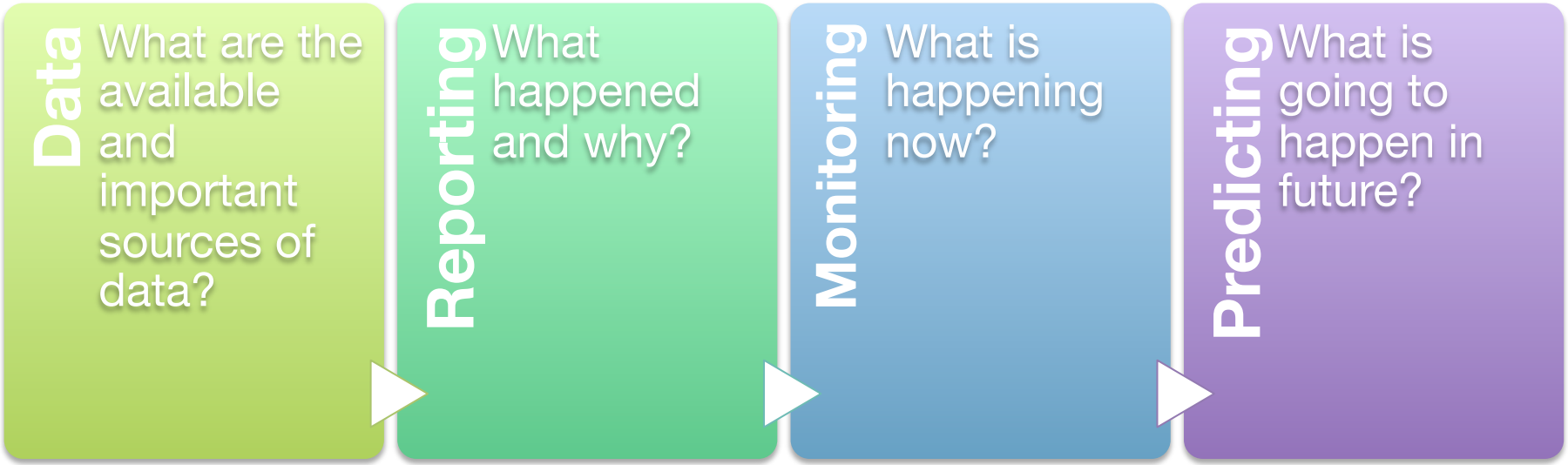
- Location-based new products
  - Telefónica “Smart Steps”
  - BBVA “Commerce360”

Big data y predicción

Universidad Carlos III de Madrid
instituto de ingeniería del conocimiento
@estebanmoro

Los pasos

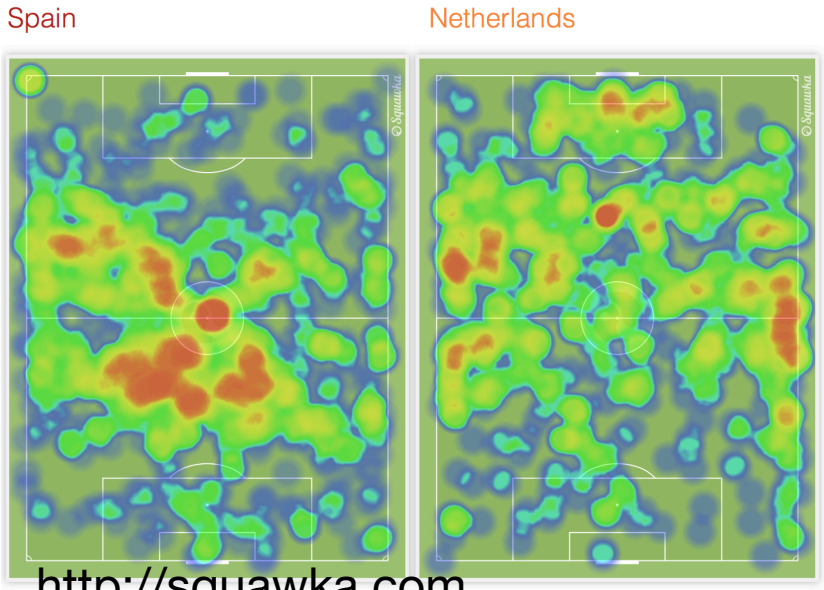
Hacia un modelo predictivo

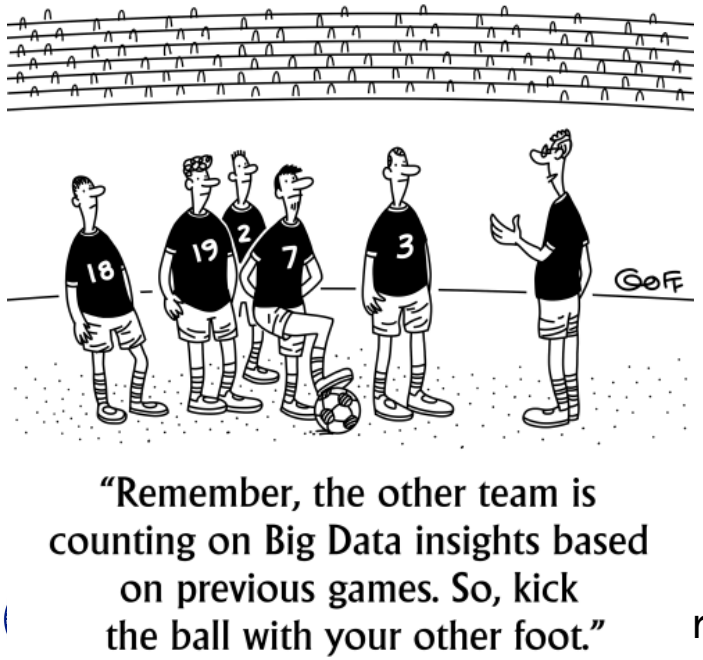


Dónde se aplican los modelos predictivos?

- **Finanzas**
  - Detección de fraude
  - Gestión riesgo
  - Seguros
- **Marketing**
  - Adopción productos/servicios
  - Mejora campañas
  - Sistemas de recomendación
- **Salud**
- **Deportes**
- **Administración**

- **Ciencia**
  - LHC
  - Biología

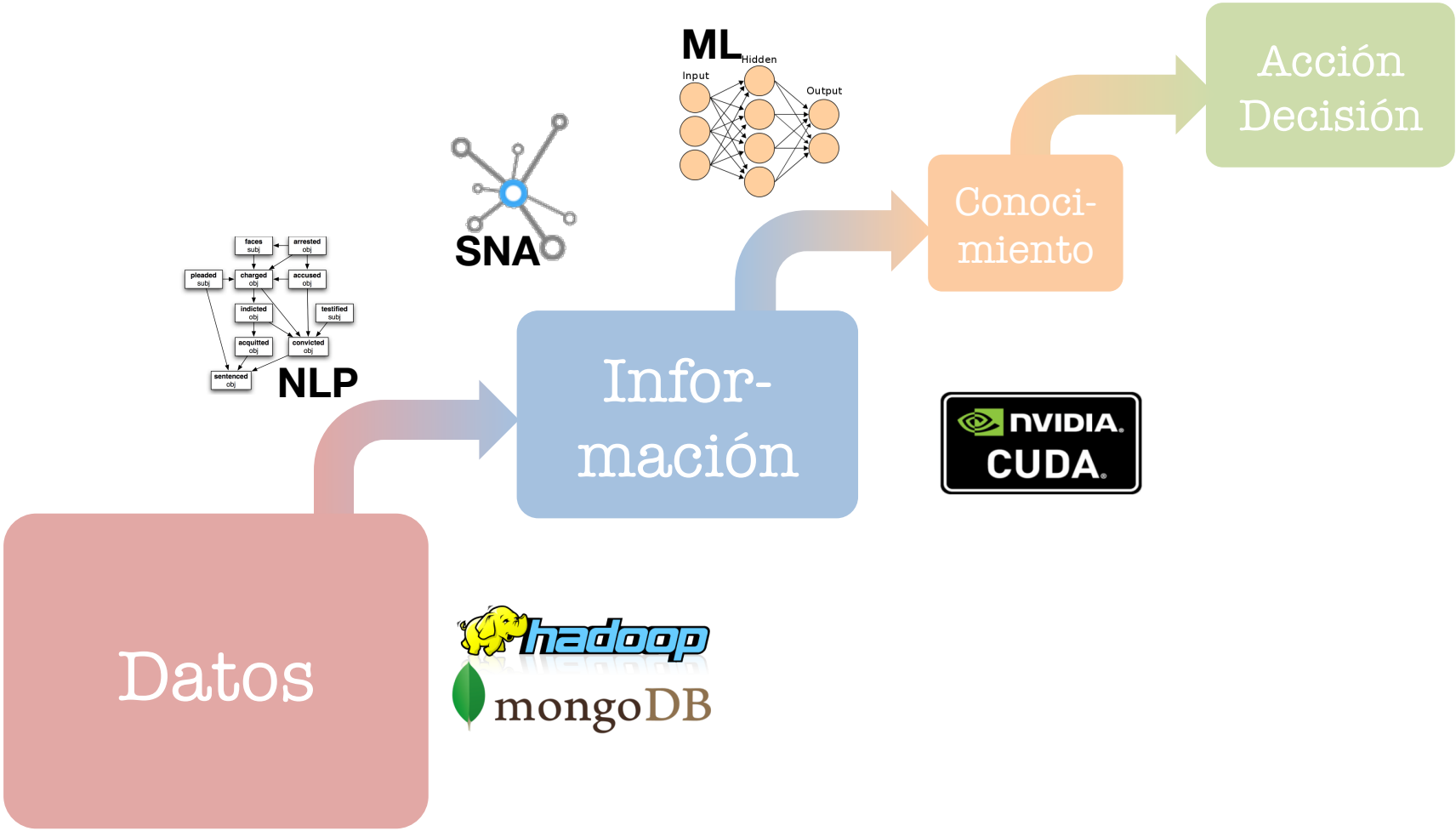




ro

Lo importante es el valor

- Datos ≠ Información ≠ **V**alor

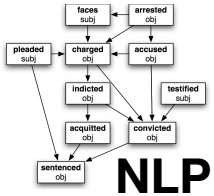


Lo importante es el valor

- Datos ≠ Información ≠ **V**alor

Interacción con usuarios

Clasificación del contenido. Generación de alertas



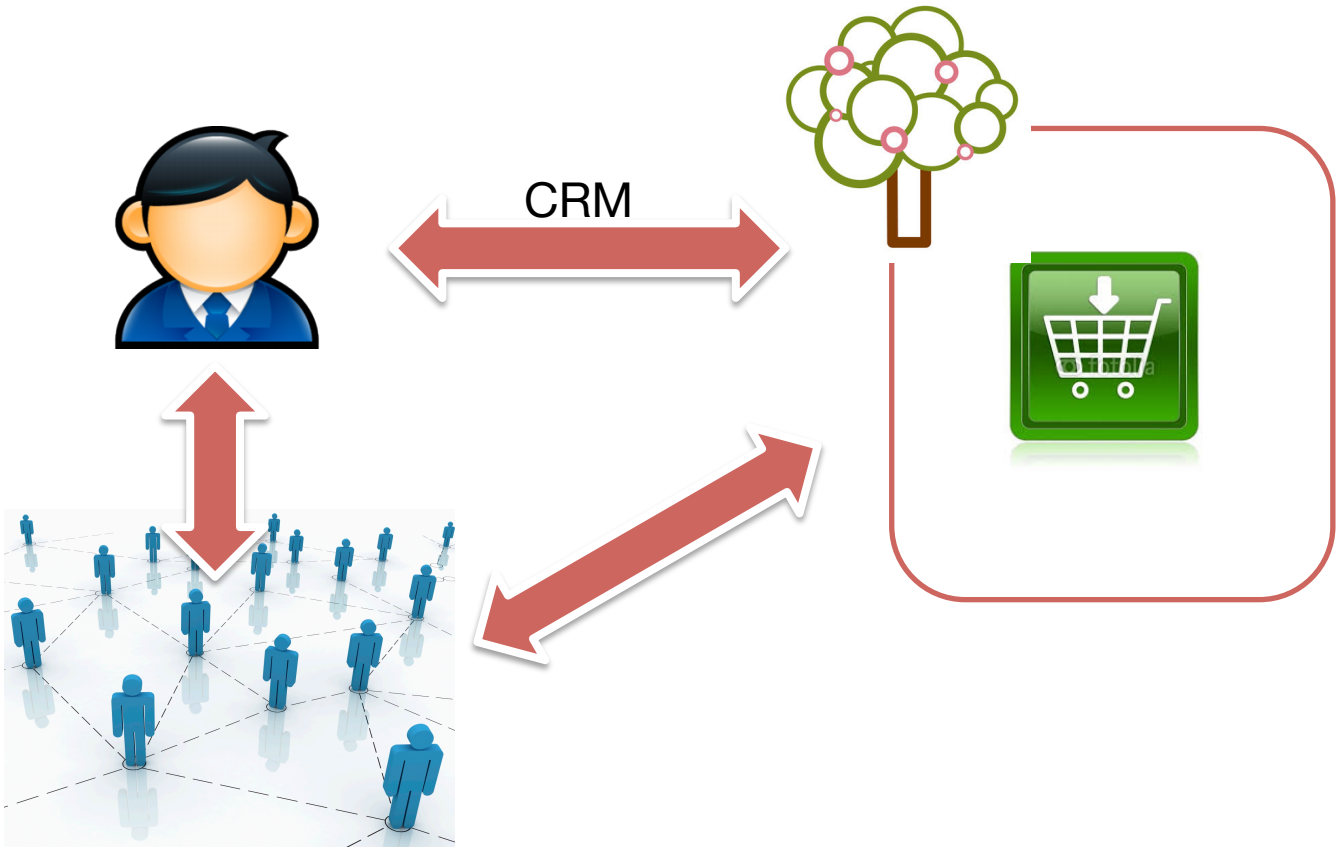
Análisis lingüístico del contenido

Tweets sobre marca/evento/tema

Ejemplos

Example 1: Telco

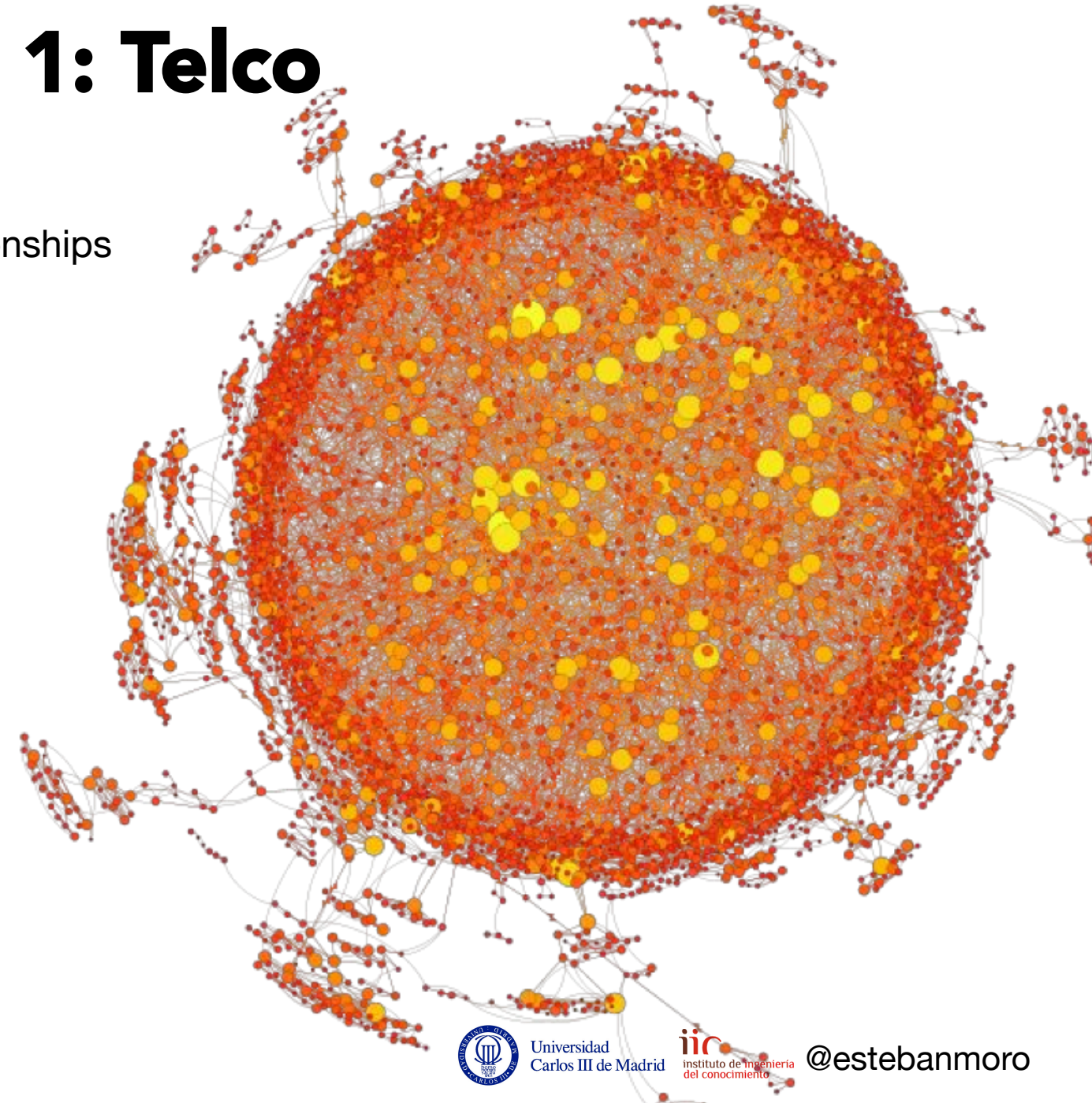
Predict adoption of product/services including social influence



Example 1: Telco

17 million of relationships
6 million users
6 months of data

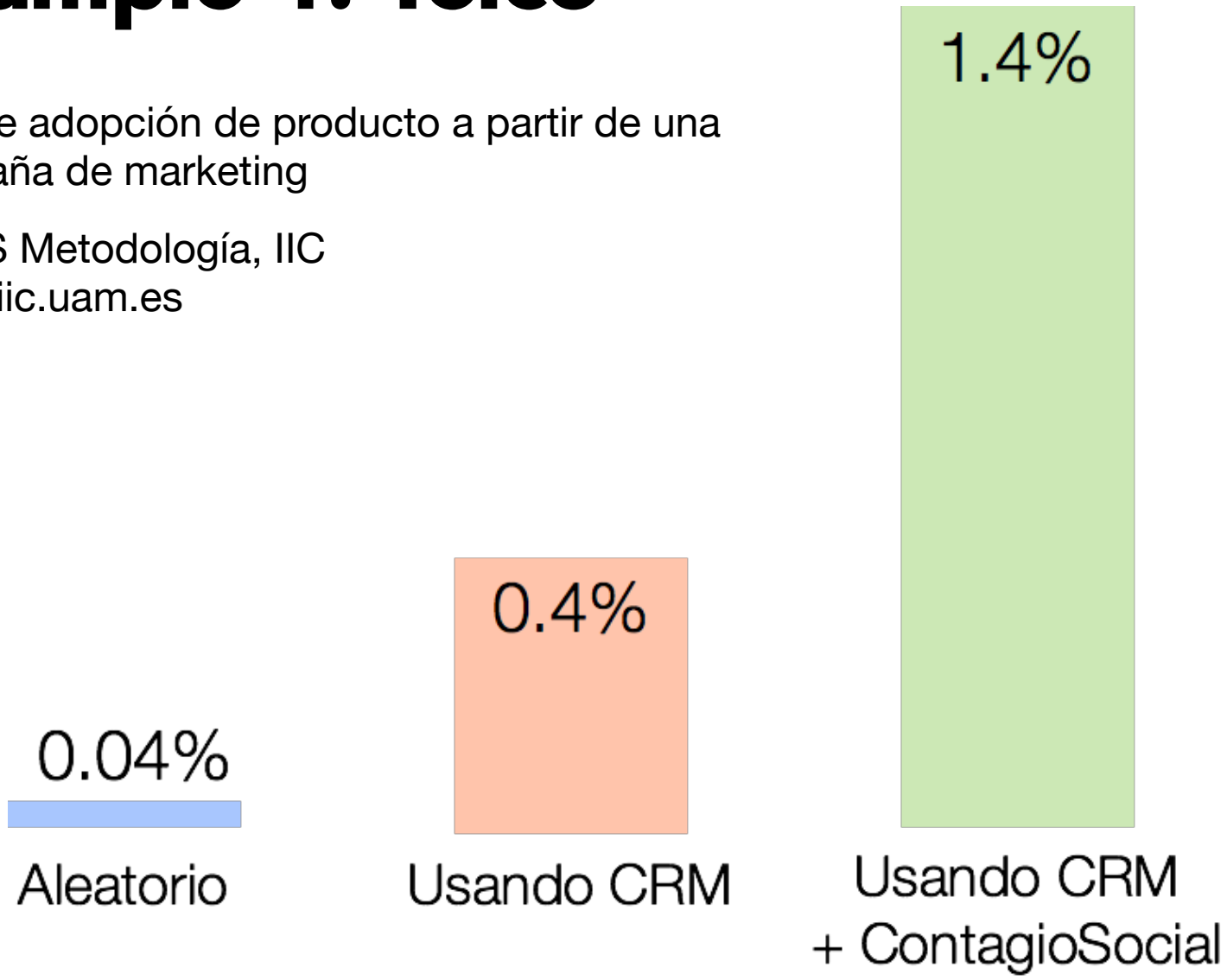
+ CRM



Example 1: Telco

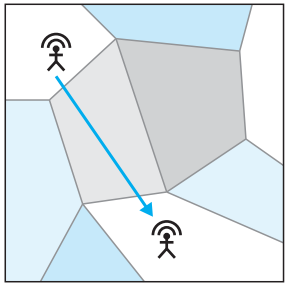
Tasa de adopción de producto a partir de una Campaña de marketing

MAAIS Metodología, IIC
www.iic.uam.es

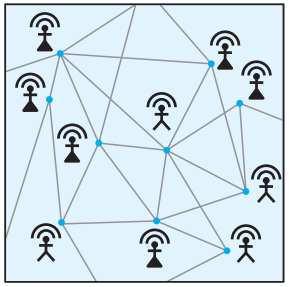


Example 1: Telco

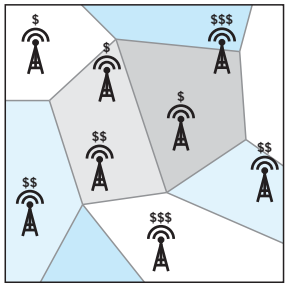
Usar las llamadas de teléfono para proyectos de desarrollo



Mobility

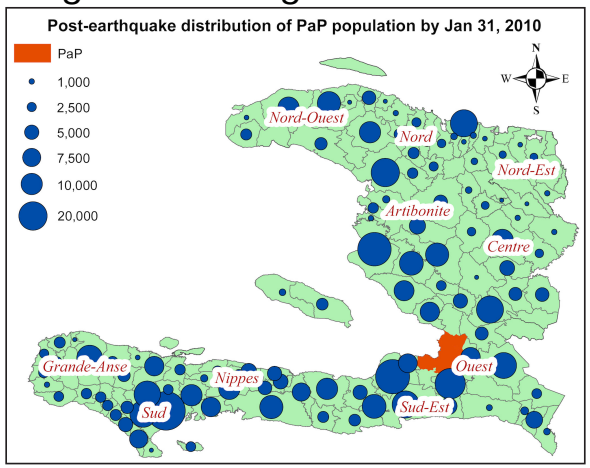


Social Interaction

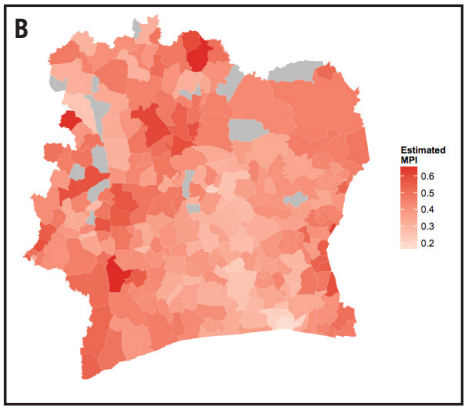


Economical activity

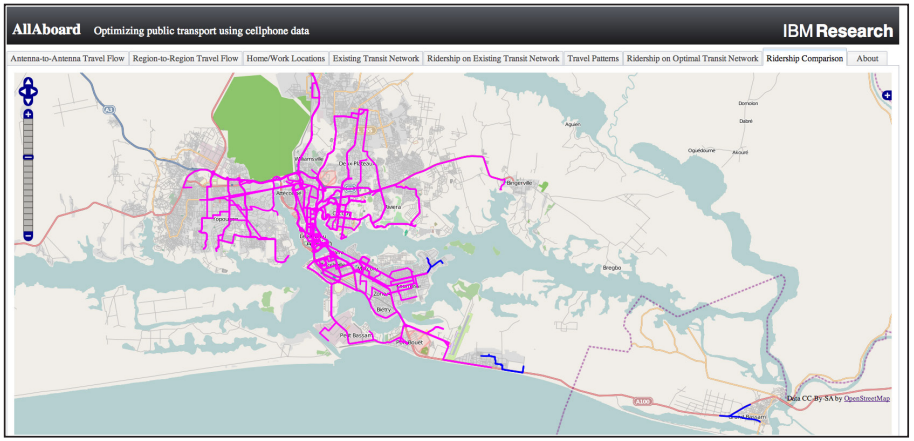
Migración emergencia en Haiti



Niveles de pobreza en Costa de Marfil



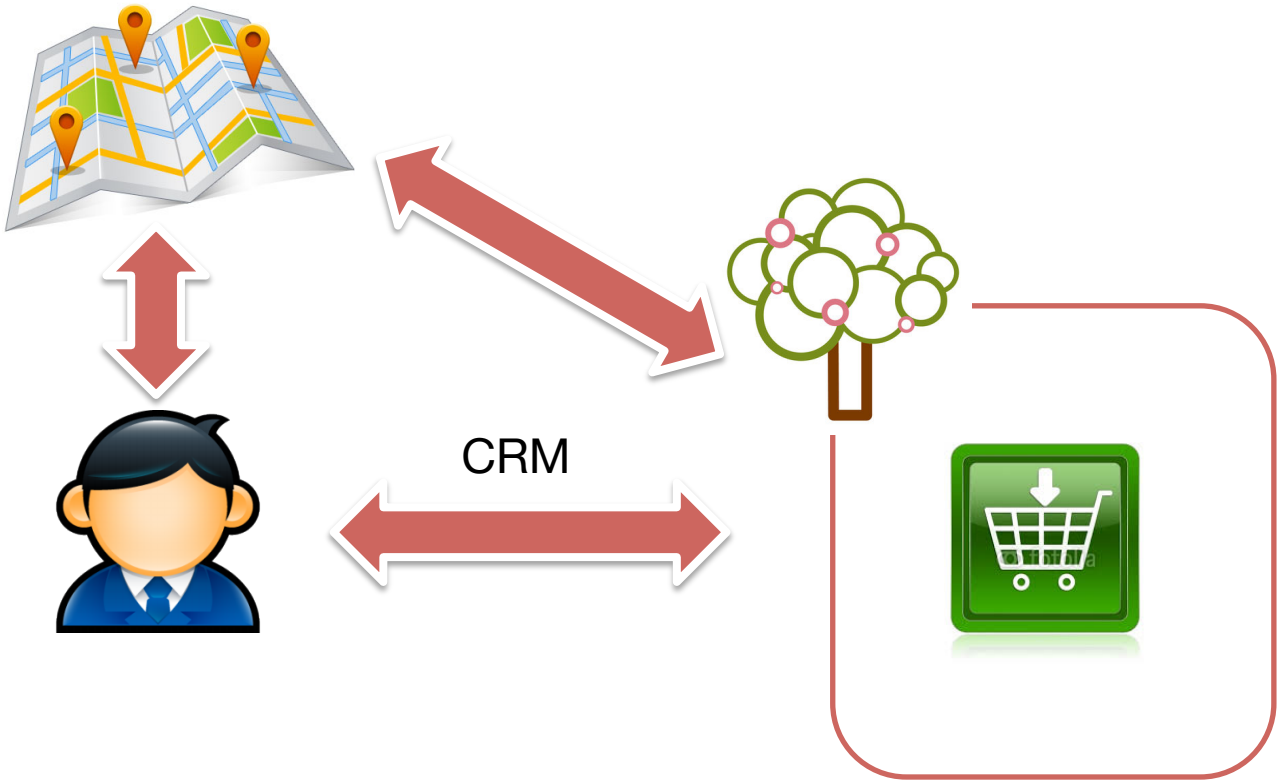
Optimización rutas transporte en Abidjan



UN GlobalPulse, 2013. MOBILE PHONE NETWORK DATA FOR DEVELOPMENT

Example 2: Bank

Predicting future purchases volume/money in city areas

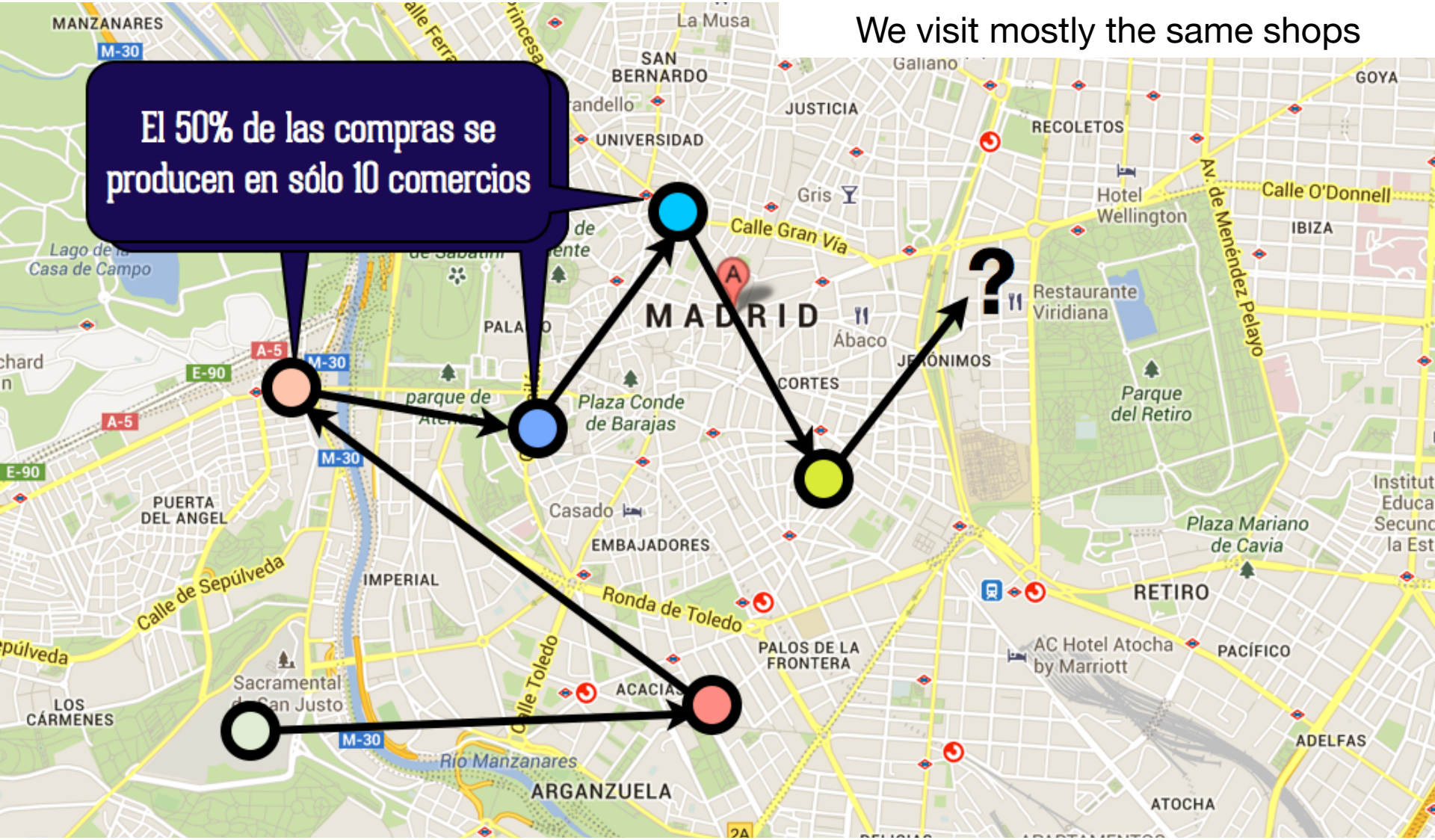


Example 2: Geomarketing/Bank

Predicting place of next purchase

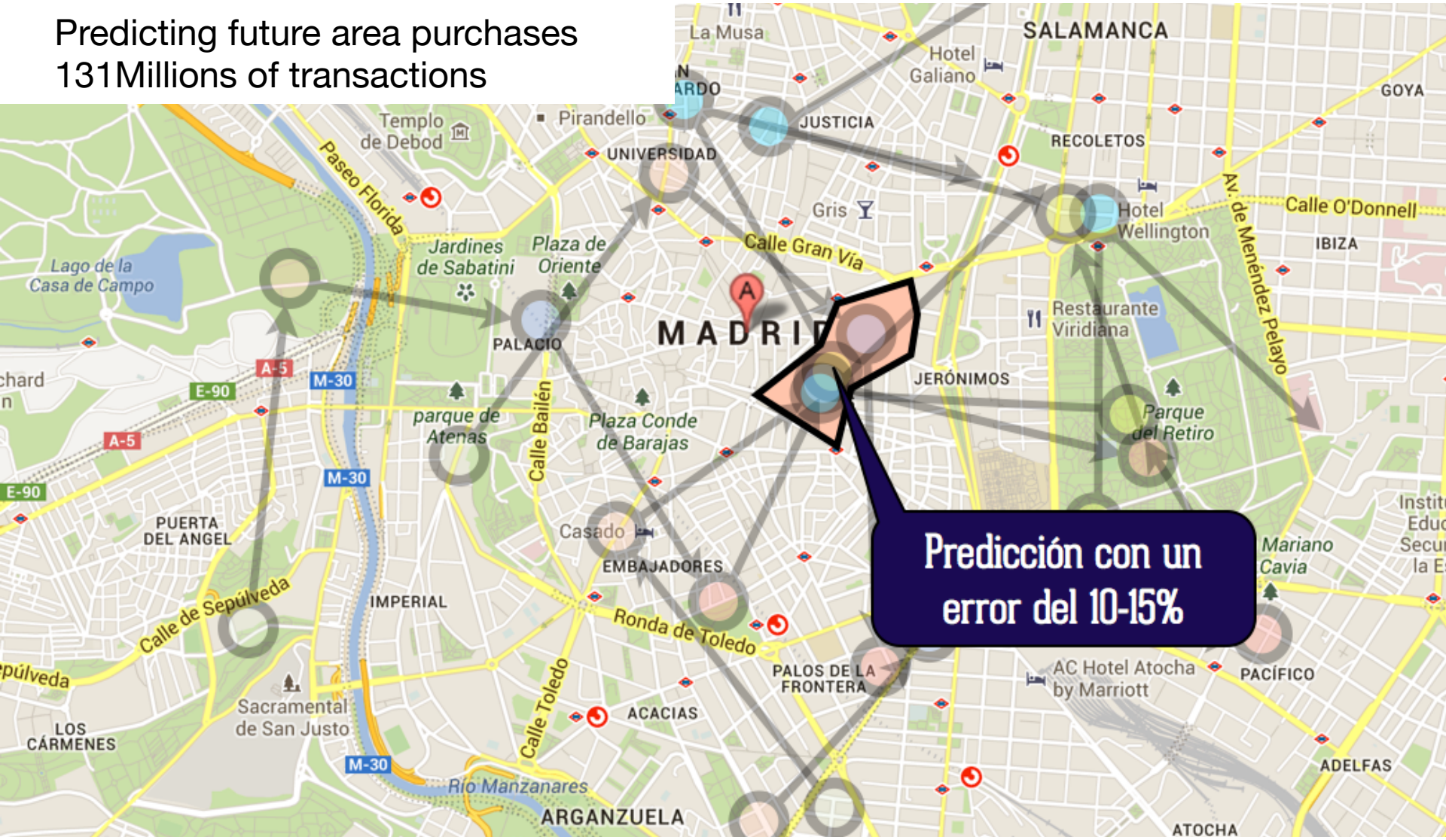
Example 2: Geomarketing/Bank

We visit mostly the same shops



Example 2: Geomarketing/Bank

Predicting future area purchases
131Millions of transactions

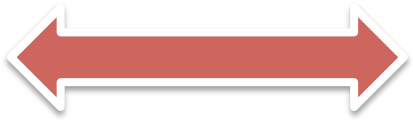


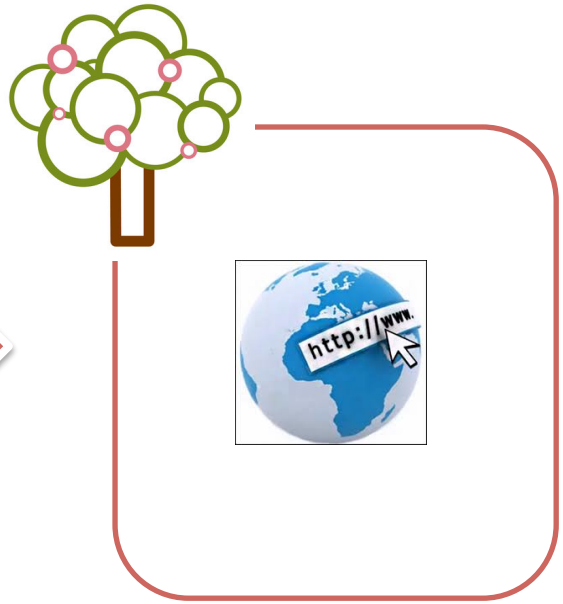
Example 2: Media company

Segmentación automática de clientes de acuerdo a sus patrones de navegación online de un medio de comunicación online

~ 0.5 Millones de usuarios
~ 17 Millones de accesos al mes

CRM





140 variables nuevas

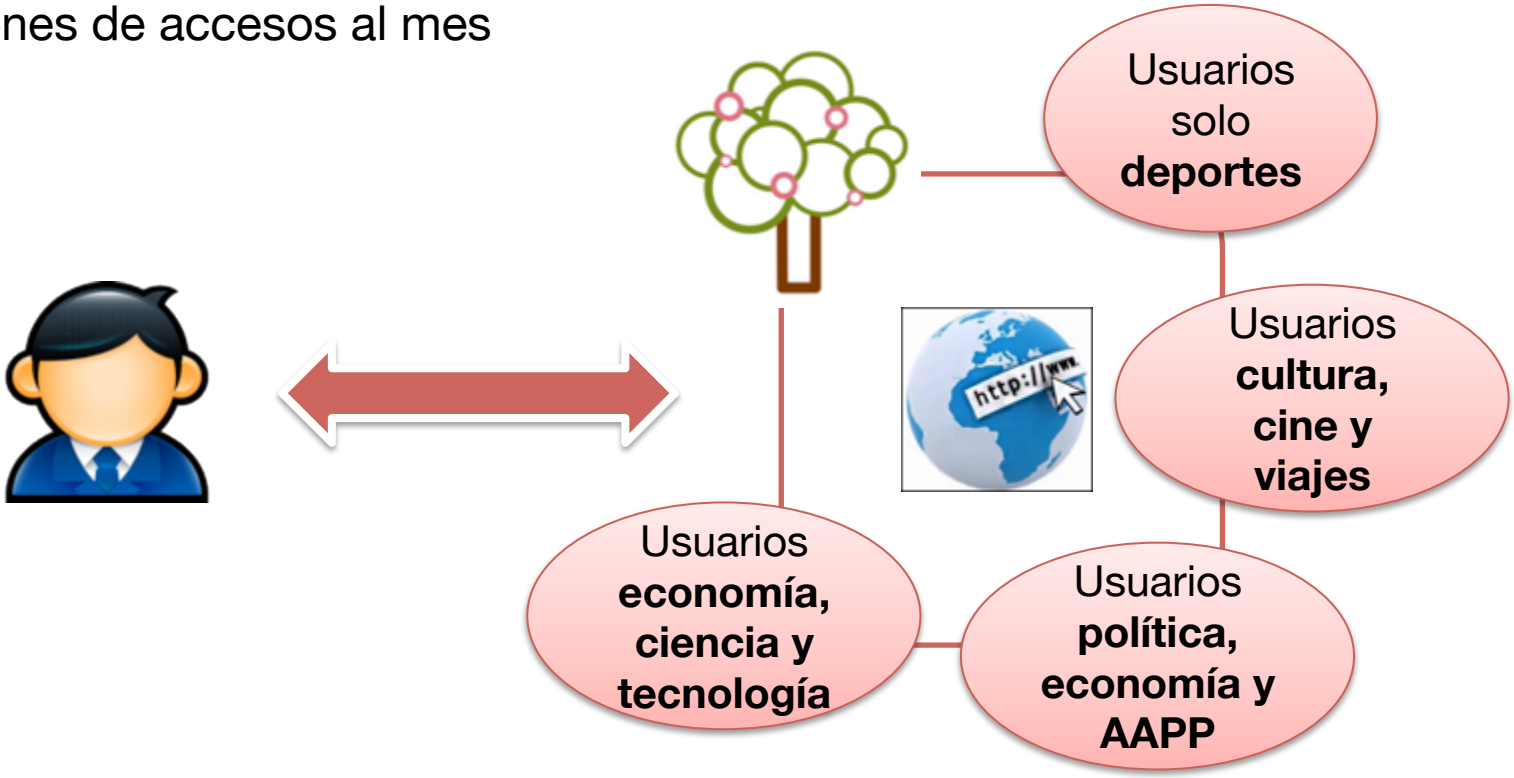
Horas a las que el usuario se conecta a la web, número de visitas, accesos únicos, categorías (política, economía, deportes, cultura…), dispositivo, localización…

- Detección de heavy users
- Usuarios multicanal

- Publicación de contenidos por franjas horarias
- Segmentación por intereses

Example 2: Media company

Segmentación automática de clientes de acuerdo a sus patrones de navegación online de un medio de comunicación online
~ 0.5 Millones de usuarios
~ 17 Millones de accesos al mes

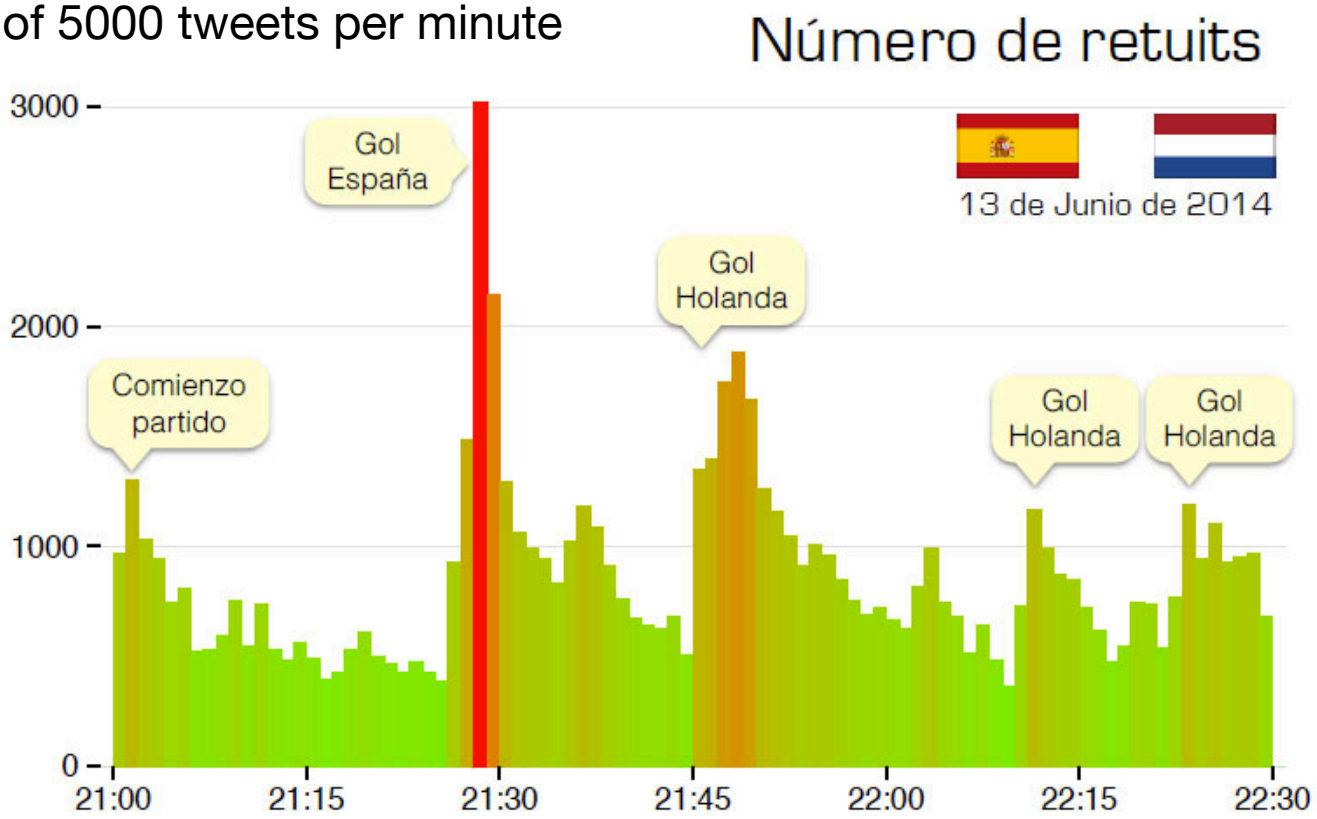


Identificación y clasificación de usuarios en grupos ocultos difícilmente alcanzables por métodos tradicionales.

Example 3: Opinion analysis in SM

Automatic detection of opinion, sentiment, brands, etc. in real time during the WorldCup 2014

~ 10Million tweets per game
~ Peaks of 5000 tweets per minute

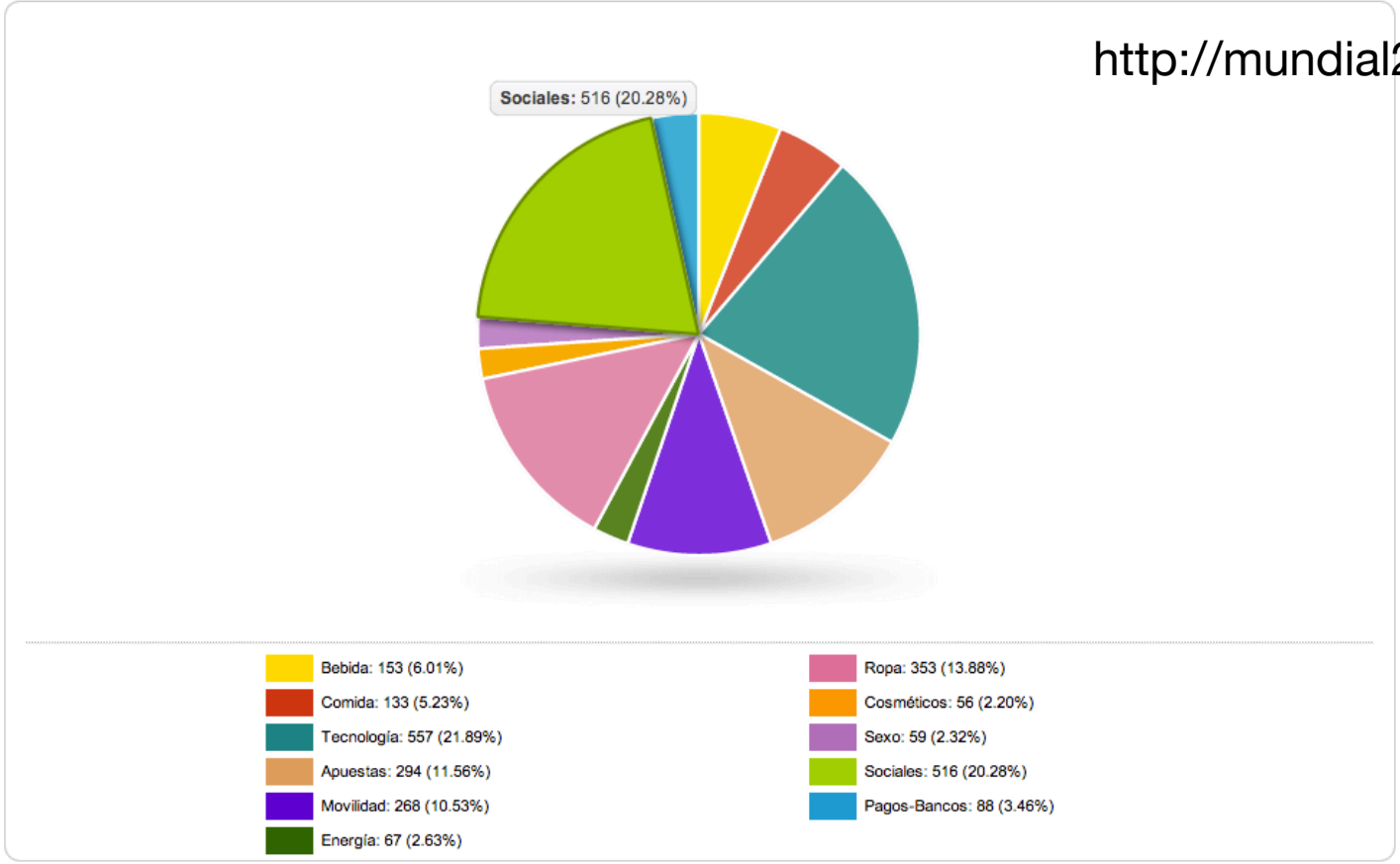


Example 3: Opinion analysis in SM

Automatic detection of opinion, sentiment, brands, etc. in real time during the WorldCup 2014

We can detect the different categories of the conversation

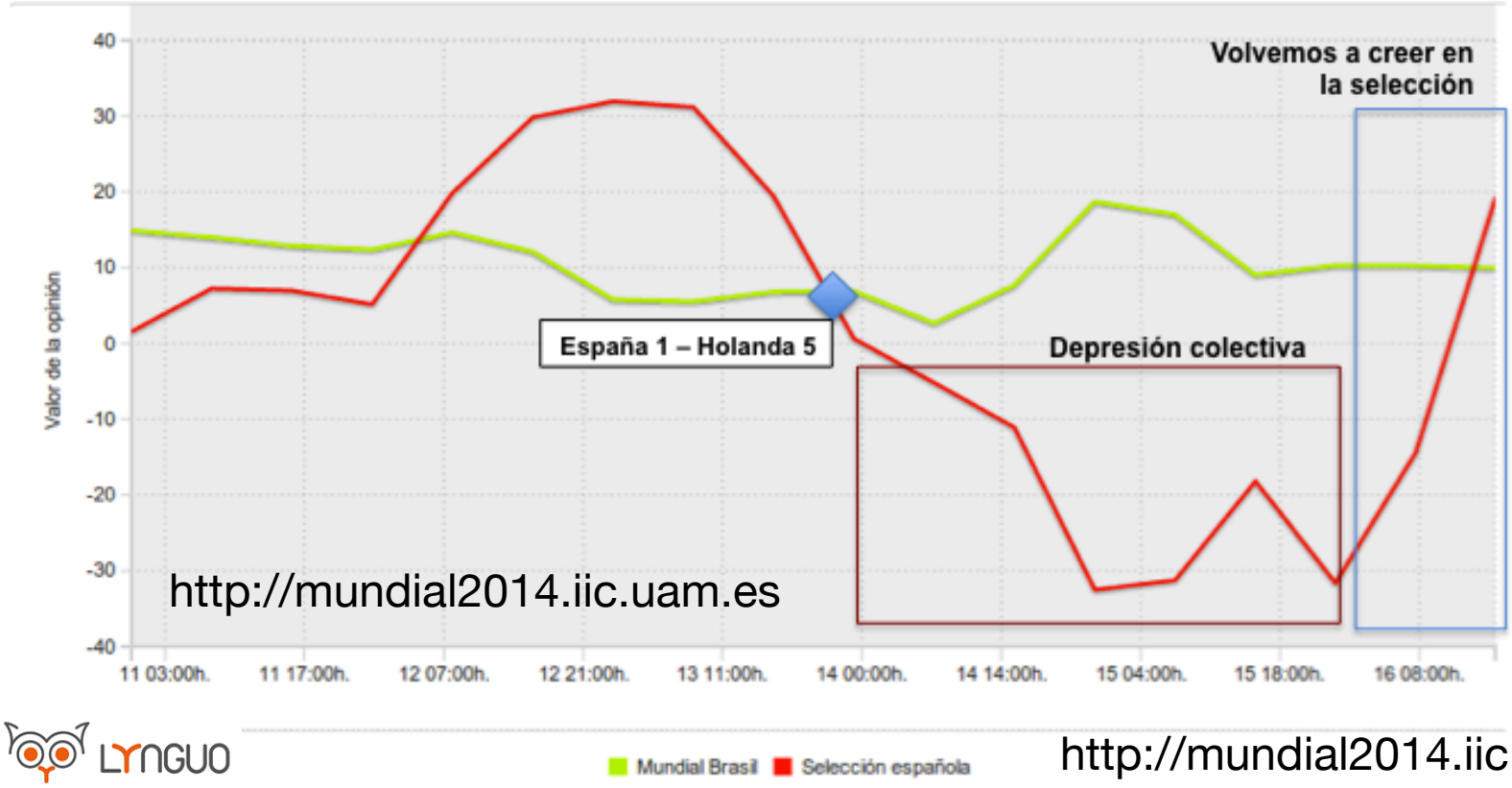
http://mundial2014.iic.uam.es

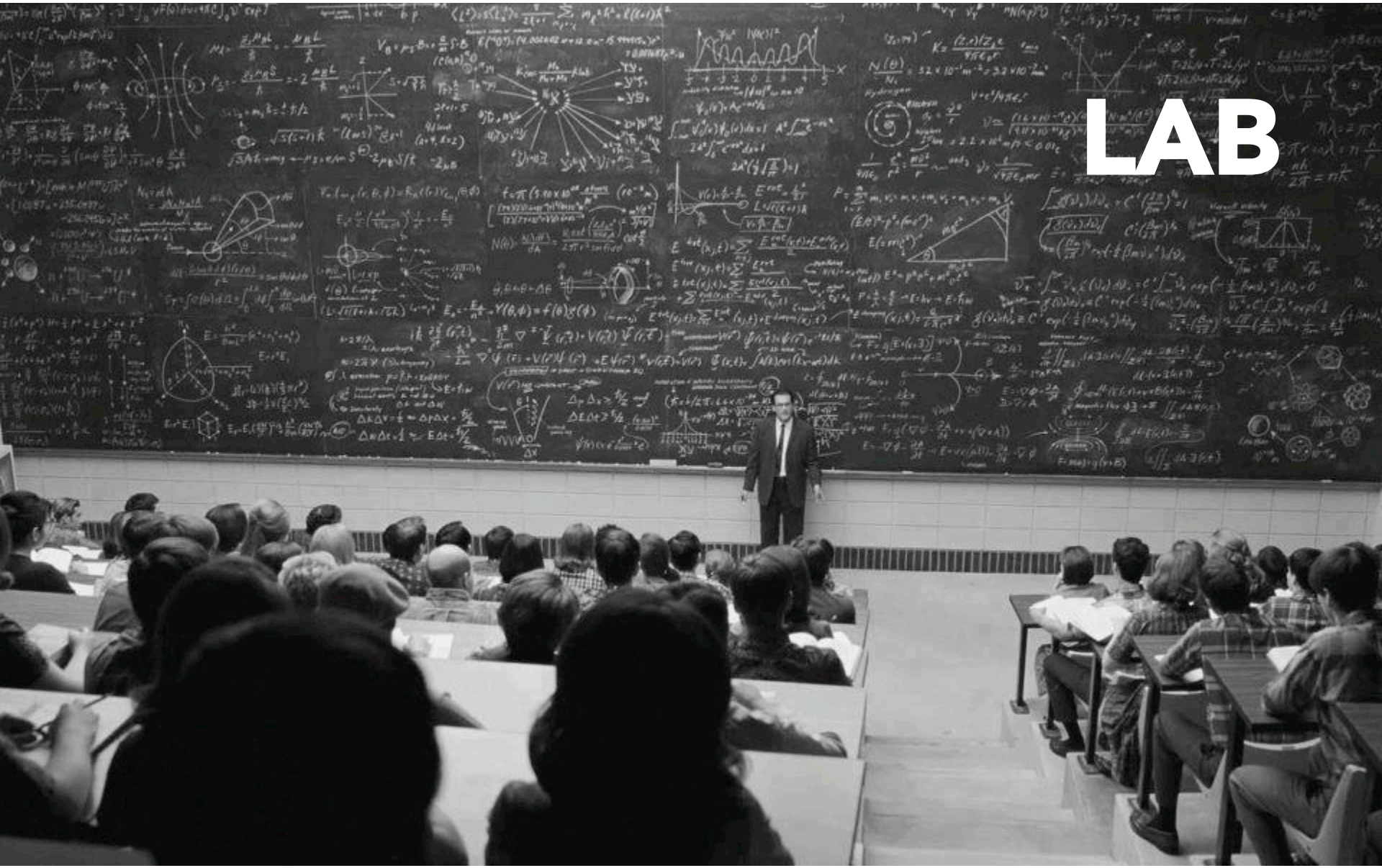


Example 3: Opinion analysis in SM

Automatic detection of opinion, sentiment, brands, etc. in real time during the WorldCup 2014

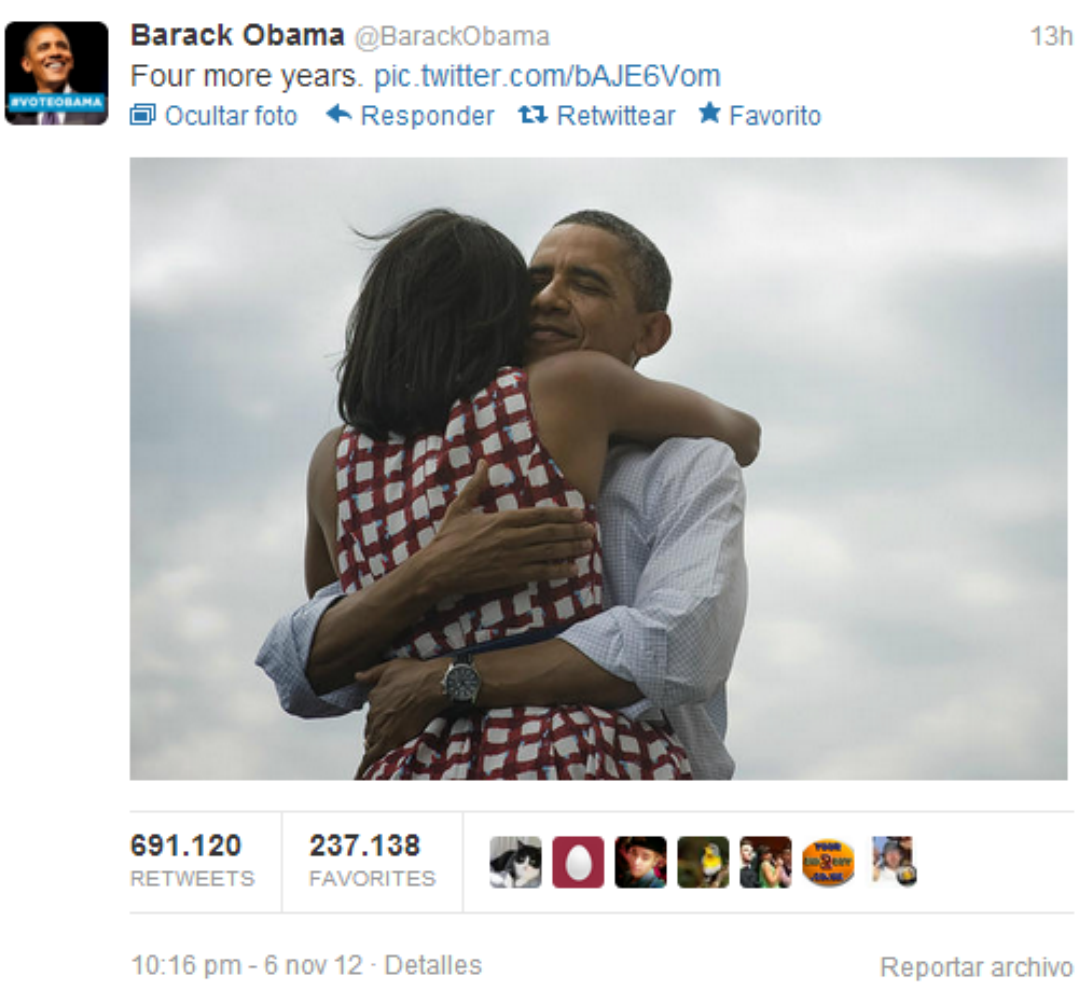
Also the sentiment about teams, brands, events, etc.



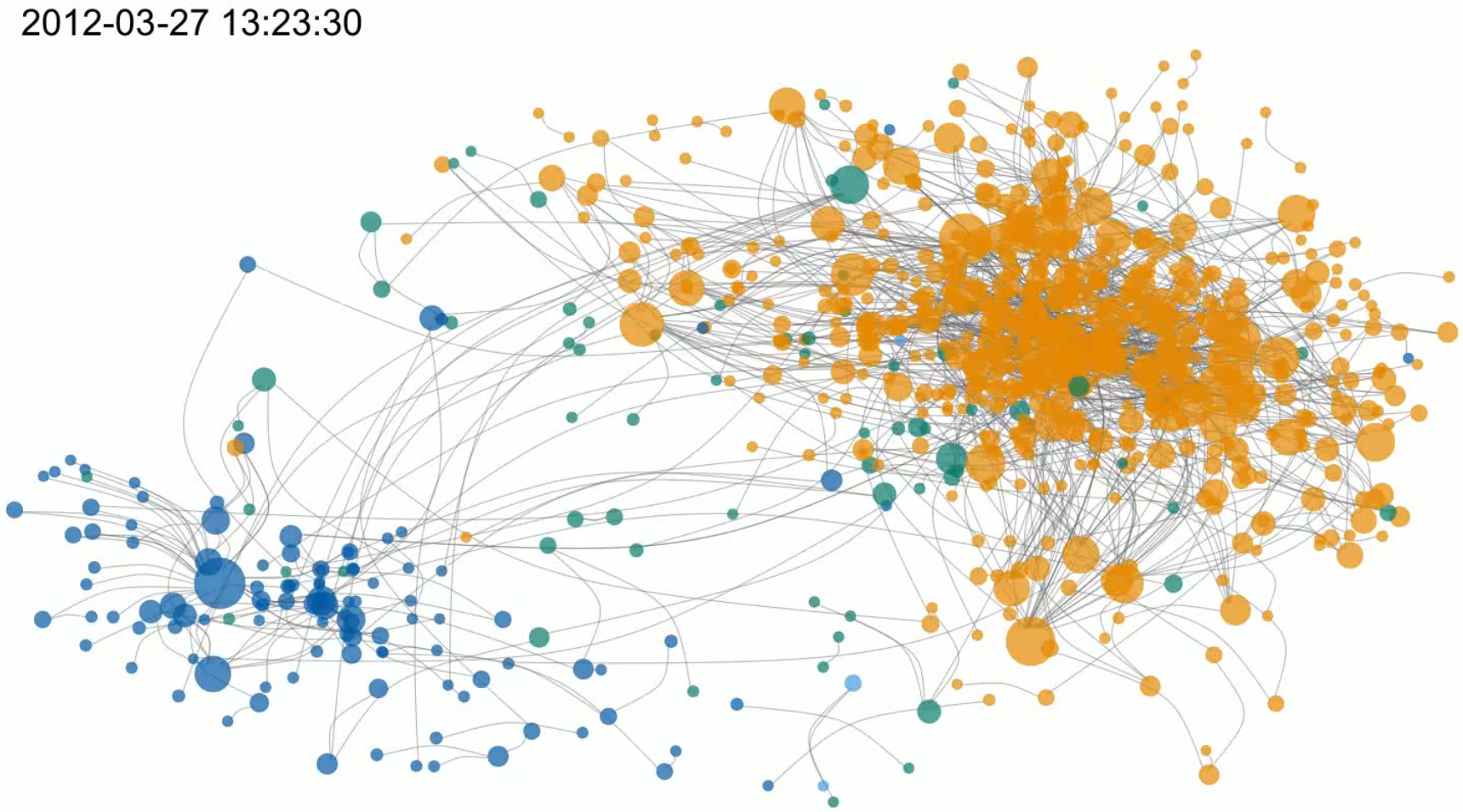

LAB

Predicción éxito contenidos en SM

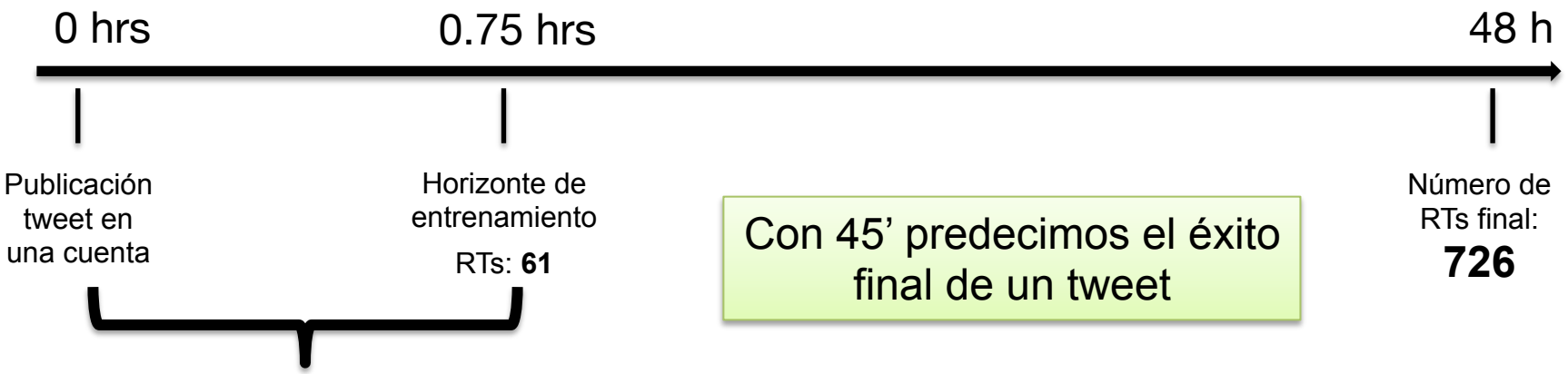
¿Podemos medir el éxito/alcance/engagement de contenidos en las redes sociales?



Predicción éxito contenidos en SM



Predicción éxito contenidos en SM



- **Variables de evolución.**
  - ✓ RTs en los primeros minutos.
- **Variables sociales.**
  - ✓ Followers y followees.
  - ✓ Número de tweets.
  - ✓ Listas en las que aparecen.
  - ✓ Klout.

Predicción éxito contenidos en SM

Delfos

Trends Analytics Sign out

Success of the last 24h tweets

Search: Show 10

Date	Age	Tweet	Prediction 24H	Current	Reach %
2014-06-06 09:30:07	00:17	¿Qué prefieres, Monarquía o República? La viñeta de hoy de El Roto http://t.co/M8bJXr7S44 http://t.co/Ve5JXaIeIB	-	179	0
2014-06-06 09:10:07	00:37	Sigue EN DIRECTO los actos conmemorativos del Desembarco de Normandía http://t.co/u3Pw54VRYs II Guerra Mundial http://t.co/IDkdvnzf8q	-	56	0
2014-06-06 08:55:08	00:52	Un veterano del desembarco de Normandía nos cuenta cómo fue aquel día. El Día D http://t.co/nNN2lqc1Py @marcbassets http://t.co/ysDYHMp9Yw	122	90	73.8
2014-06-06 08:37:50	01:09	Malula, cuna del cristianismo en Siria, está arrasada por la guerra→ http://t.co/OJb5RUME2c Todavía hablan en arameo http://t.co/S3qa2Sz1oC	79	54	68.4
2014-06-06 08:20:38	01:26	Los partidos que defienden la independencia de #Cataluña quieren mantener su hoja de ruta con Felipe VI http://t.co/QoDRxzT7Iy @maiolroger	21	11	52.4
2014-06-06 08:05:29	01:41	Juan Manuel Santos y Óscar Zuluaga se enfrentan por el fin de las FARC http://t.co/MLrOBN4RT4 Así fue el debate televisado en #Colombia	27	14	51.9
2014-06-06 05:52:37	03:54	Un monstruo con sotana http://t.co/kAYyD9nKAk El Vaticano expulsa a un sacerdote mexicano por abusar de al menos 20 menores, por @luispablob	132	182	100
2014-06-06 05:09:32	04:37	La 'burbuja inmobiliaria' brasileña comienza a desinflarse http://t.co/4GMyMIWjjH	49	59	100
2014-06-06 04:34:55	05:12	Un recorrido por Buenos Aires, su boom gastronómico y sus noches de tango http://t.co/CpRg0WNiLT En @ElViajero_Pais	47	50	100
2014-06-06 03:53:42	05:53	Nuevo papel para George Clooney: político http://t.co/DvXK8ctdSp El actor está pensando sumarse a la campaña electoral demócrata en 2016	29	41	100

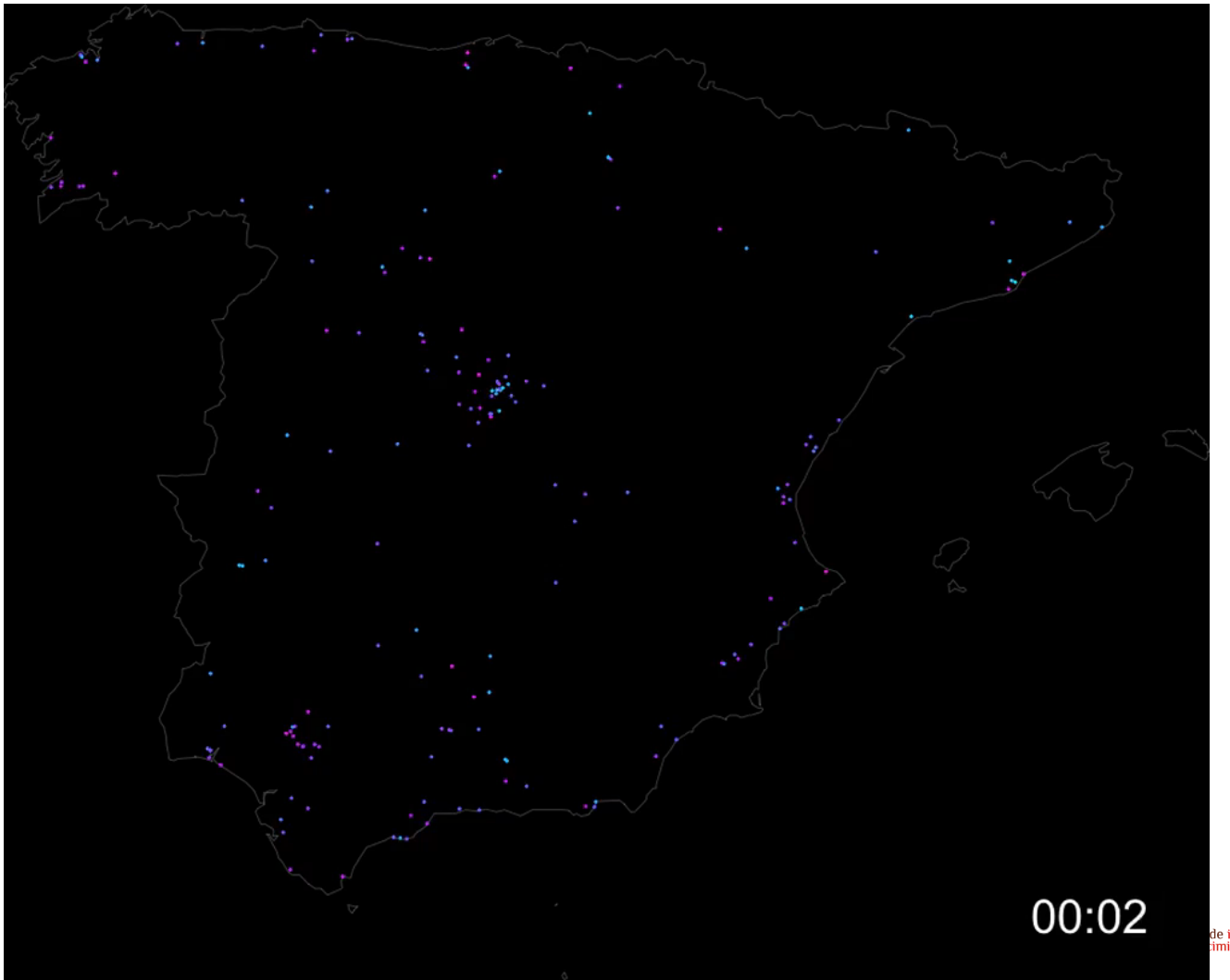
Next 5-hour trends

Key words | Hashtags | Users

Monitorización de indicadores económicos

Utilizar bigdata para crear indicadores económicos en tiempo real

Movilidad + social + contenido tweets

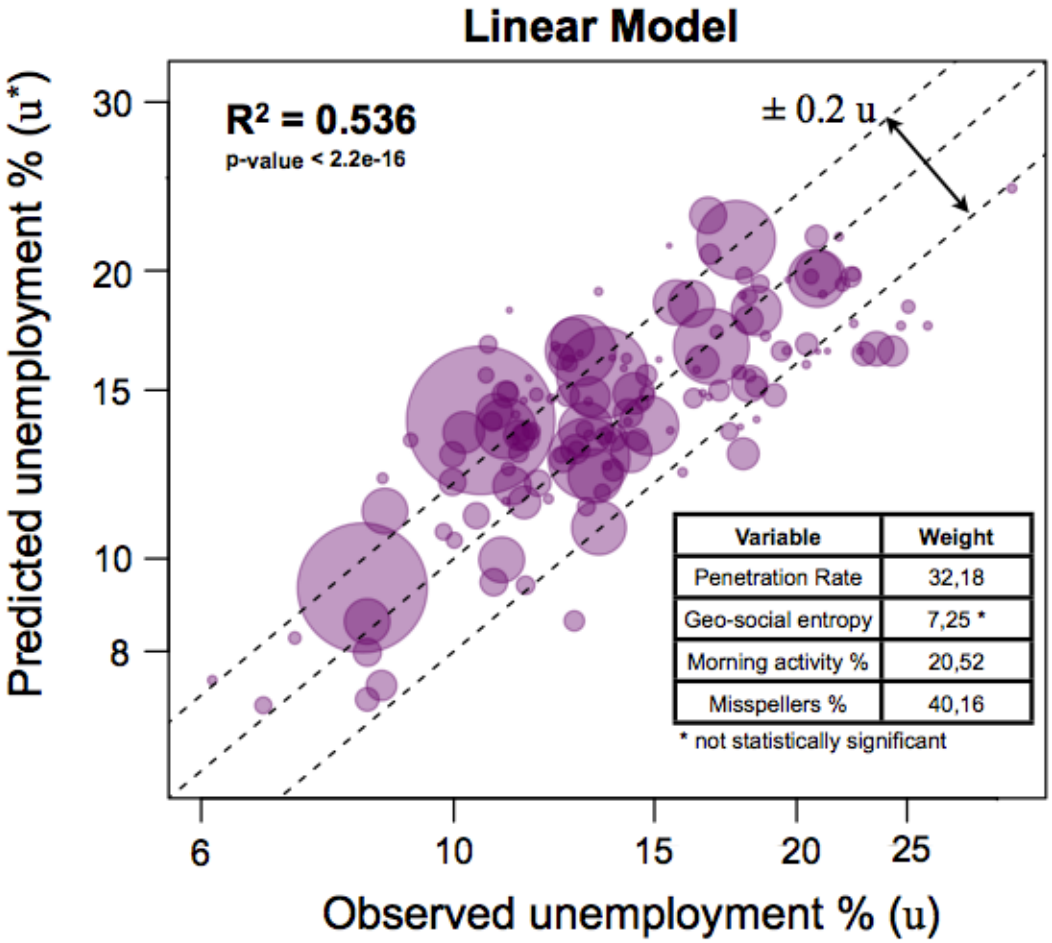


20Millones tweets
200k usuarios

Monitorización de indicadores económicos

Utilizar bigdata para crear indicadores económicos en tiempo real

Mobilidad + interacción social + contenido tweets



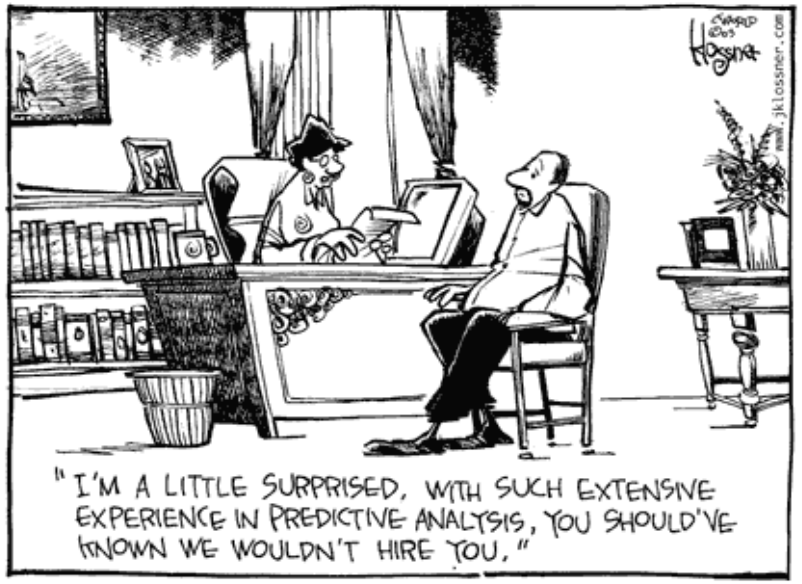
BigData o el pulpo Paul?

"Es difícil hacer predicciones,
sobre todo sobre el futuro"

BigData o el pulpo Paul?

Los peligros de usar BigData en predicción

- **Predecir supone decir lo que va a pasar y con qué probabilidad**
  - No ignoremos los falsos positivos de nuestros algoritmos
- **Los modelos de hoy no valdrán mañana**
- **Correlación ≠ Causalidad**
  - Aunque ciertas variables muestren poder predictivo, eso no significa que hayamos encontrado un mecanismo que explica lo sucedido
- **Big no es All (sesgos poblacionales)**
  - Incluso aunque tengamos millones de usuarios o de eventos puede ser que no tengamos todos



http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html

BigData o el pulpo Paul?

Tenemos que:

Comprobar las hipótesis

Utilizar modelos nulos para descartar efectos espúreos, correlaciones no deseadas, etc.

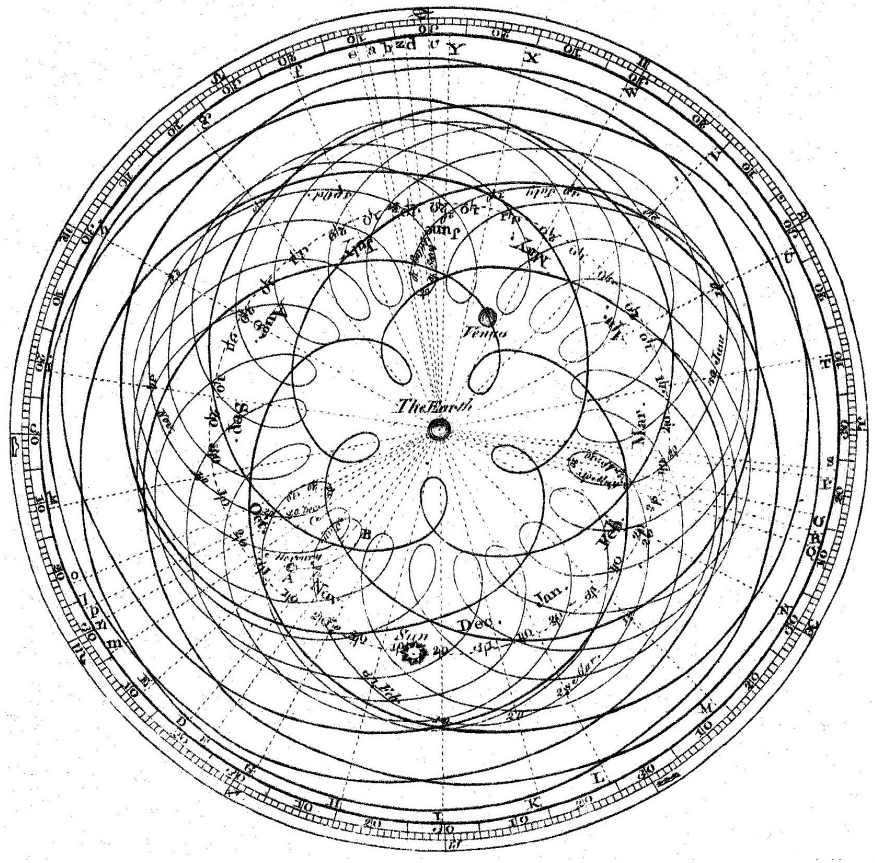
A/B testing para aislar las causas de las correlaciones

Experimentar es la única manera de encontrar las causas

Demografía

Preguntémosnos sobre el origen de los datos y su representatividad.

Método Científico, por favor!



http://en.wikipedia.org/wiki/Deferent_and_epicycle

http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html