



GRANDATA[®]

Personas y Datos

Jornada Big Data 27/8/2013

Martin Minnoni

Qué hacemos en Grandata?

Entender comportamientos humanos con datos

Qué preguntas nos interesan?

Encontrar regularidades de cómo las personas:

- Se mueven en el espacio (**movilidad**)
- Se relacionan entre sí (**social**)
- Adquieren productos (**compra**)

Qué datos usamos?

- Información pública de Internet
- Información propietaria de las Telcos
 - **Metadata de llamados, texto**
 - Información de los clientes (CRM)
- Información propietaria de los Bancos
 - Transacciones de tarjetas de crédito, débito
 - Riesgo crediticio

Qué tan predecible es la ubicación futura de las personas?

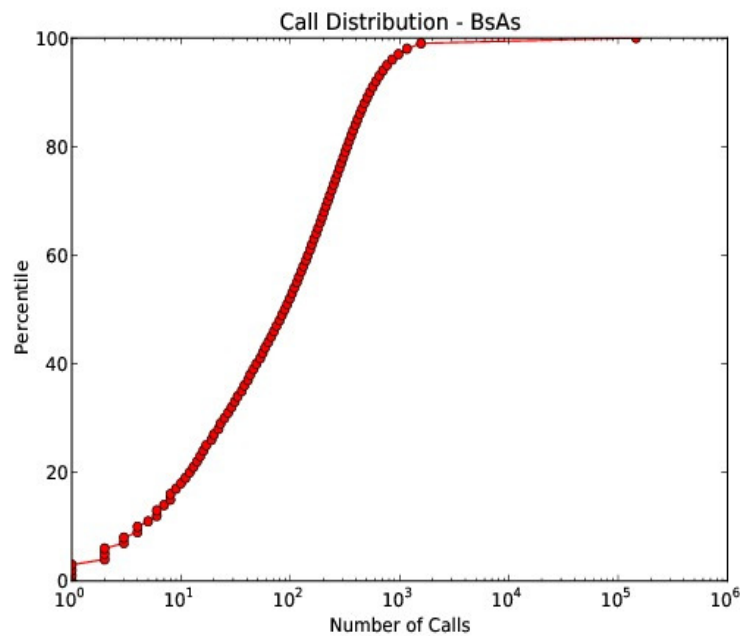
Con qué datos contamos?

- Información de tráfico (llamadas) anonimizada (CDR).
- Nos enfocamos en la región metropolitana de Buenos Aires, en un período de 5 meses.
- Data cruda: ~50 millones de llamados por día

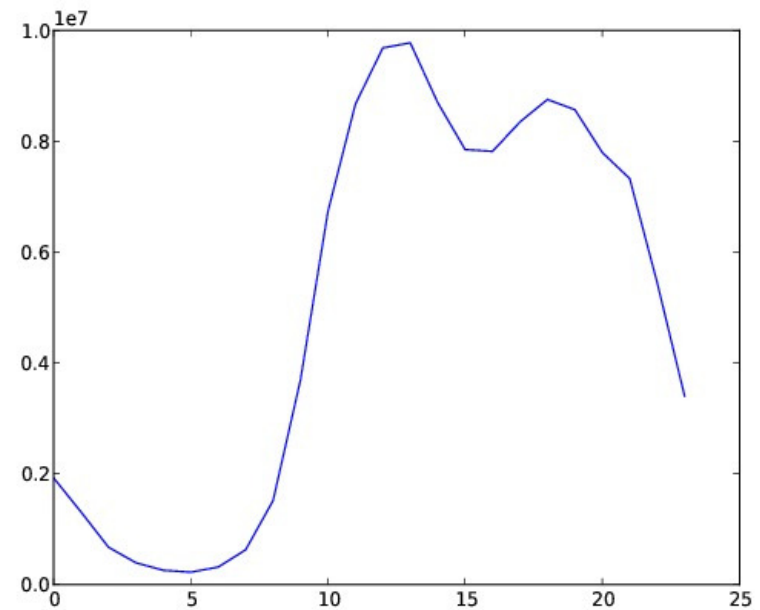
De los CDR, usaremos:

- Usuario emisor, Usuario receptor, Dirección de la llamada (entrante/saliente), Fecha y hora, Ubicación de la antena utilizada.

Algunas Métricas y Distros

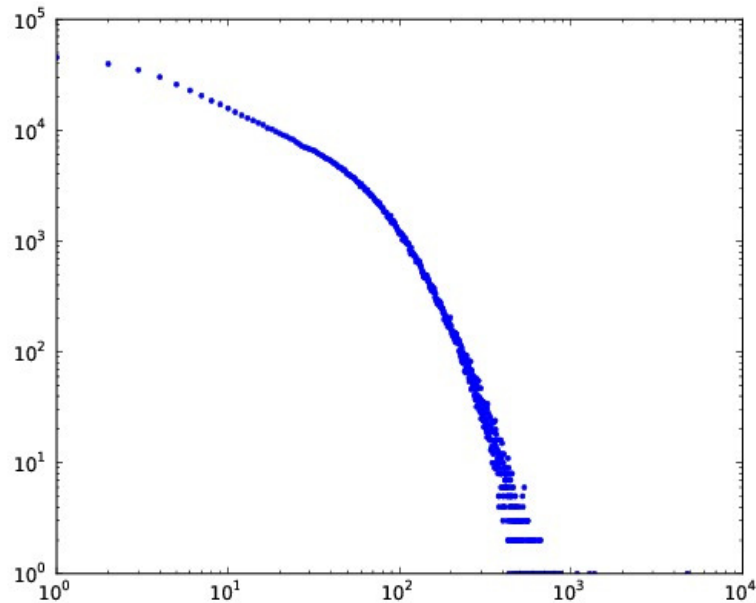


Número de Llamadas por Percentil



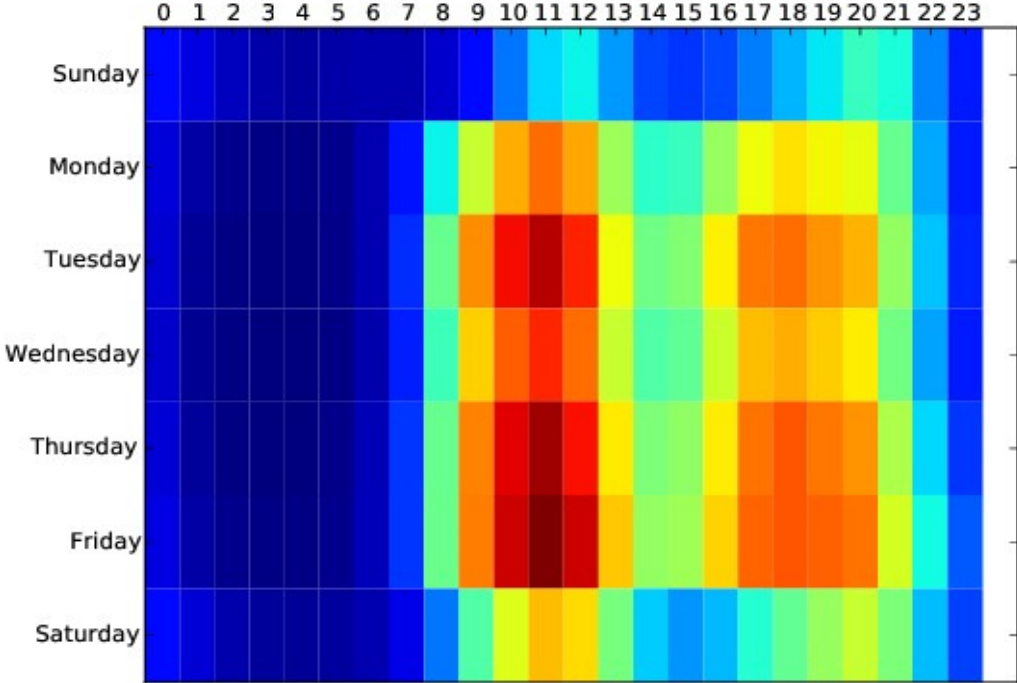
Llamados por hora - Día de Semana

Algunas Métricas y Distros



Número de usuarios vs cantidad de lugares distintos visitados.

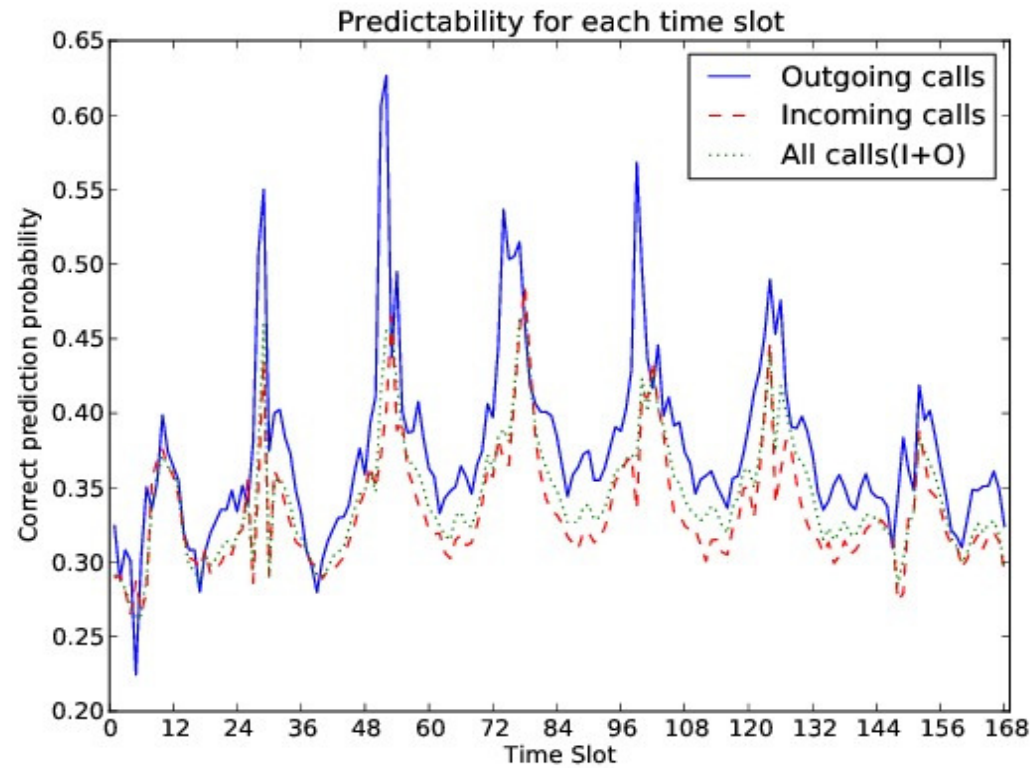
Distribución de Llamadas



Modelo Predictivo

- Modelo simple basado en la antena más frecuente usada previamente, para cada hora.
- Solemos tener patrones de movilidad muy predecibles.
- Entrenamos el modelo con 15 semanas de datos.
- Lo testeamos con 2 semanas de datos.
- En promedio, ~35% de predicciones correctas.
- Picos por encima del 50%

Modelo Predictivo



Modelo Predictivo - Conclusiones

- La predictabilidad es al menos 25% mayor durante los días de semana.
- Picos de predictabilidad durante la noche.
- Mayor predictabilidad para llamadas salientes.
- Filtrando los usuarios, se puede obtener una mayor predictabilidad, y una curva más “suave”.
- Podemos clusterizar las antenas para mejorar la predictabilidad.

Podemos visualizar el commute a la ciudad de Buenos Aires a través de nuestros datos?

Que métricas podemos obtener?

Flujo E/S en Buenos Aires

- El fenómeno de commute se da en las grandes áreas metropolitanas.
- 3 millones de personas hacen commute cada día a Bs As, según encuestas y estimaciones.
- Con los datos de celulares, podemos cuantificar el fenómeno de forma más precisa y simple.
- Esperamos ver un flujo entrante durante las mañanas y saliente durante las tardes.

(videíto...)

Flujo E/S en Buenos Aires

- Identificamos la casa y el trabajo para cada usuario.
- Definimos un usuario como 'commuter' si su casa se encuentra fuera de la CABA y su trabajo dentro.

Métricas del Radio del Commute (ROC):

- ROC promedio: 7.8 km
- Diámetro de la CABA: 14 km
- Diámetro del área metropolitana: 90 km
- ROC randomizado: 32.9 km

Qué podemos decir de un evento social masivo?

Juega Boca en La Bombonera...

(videíto...)

Fútbol a través de celulares

- Asistencia promedio a partidos de local: 42,000 personas
- Postseleccionamos usuarios que asisten al evento. Esto produce el efecto de no tener llamados fuera del área del estadio durante el partido.
- Se marcan claramente patrones de convergencia hacia el estadio antes del partido, y divergencia después.
- Marcadamente diferente a los patrones para la misma zona y horario en un día en que no hay partido.

Podemos usar información social externa para hacer mejores predicciones?

Enriqueciendo la Data

- Intentaremos usar nuestro conocimiento de eventos sociales para mejorar la precisión de nuestro modelo predictivo.
- Proof of Concept: Hinchas de Boca.
- Taggeamos los usuarios si y sólo si tienen llamados alrededor del estadio donde juega Boca durante 3 partidos consecutivos.
- Luego usaremos el fixture y el tag generado para mejorar el modelo.

Enriqueciendo la Data



Enriqueciendo la Data

- Sólo para usuarios taggeados.
- Predicciones hechas:
Modelo Básico: 63 %
Modelo Enriquecido: 100 %
- Precisión en las predicciones:
Modelo Básico: 19 %
Modelo Enriquecido: 38%
- El modelo enriquecido TRIPLICA la eficiencia total, sobre TODOS los eventos.

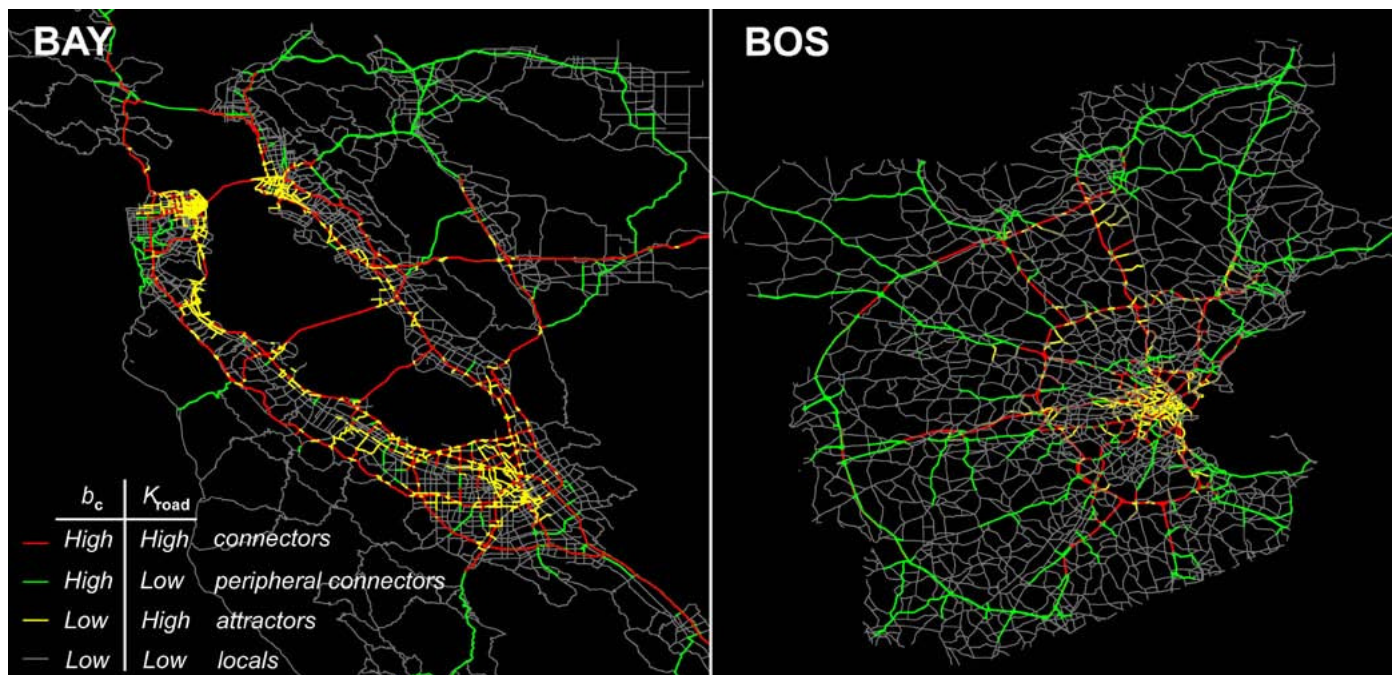
Smart Cities – Smart Drivers

Colaboración de investigación con el grupo de HumNet (Human Mobility and Networks Lab) del MIT, a cargo de Marta González

Smart Cities – Smart Drivers

- Hacer un estudio del uso de calles en distintas ciudades latinoamericanas a través de la data de las Telcos.
- Mejorar la situación de dichas ciudades respecto al tráfico y los embotellamientos en hora pico.

“... our ability to pinpoint the few driver sources contributing to the major traffic flow allows us to create a strategy that achieves a significant reduction of the travel time across the entire road system, compared to a benchmark approach.”
Understanding Road Usage Patterns in Urban Areas, Wang et. al.



Gracias !

martin@grandata.com

grandata.com
[@grandatacorp](https://twitter.com/grandatacorp)
[facebook/grandata](https://facebook.com/grandata)