Introducción

Gonzalo Rivero

Big Data en investigación social y opinión pública

22 a 26 de julio, 2019

¿Qué es el big data?

- Hay demasiadas definiciones.
- Intuitivamente, parece incluir formas de referirse a:
  1. Una categoría de analisis: big data por oposición a datos tradicionales de menos tamaño y más costosos.
  2. Un termino de marketing: compañías que pueden extraer valor de los datos y que nos permiten ser data-driven.
  3. Un "cambio de paradigma": recolección de datos masivos y poblacionales, predicción en lugar de estimación, búsqueda de patrones en lugar de estimación.

¿Qué esperamos del big data?

- El término refleja cambios reales:
  - Multiplicación de sensores y aparatos recogiendo datos.
  - Mayor creación de datos digitales en origen.
  - Incremento de la capacidad de almacenar datos.
  - Incremento de la capacidad de procesamiento.
  - Nuevas técnicas de análisis.
- Numerosos casos de éxito populares como Google.
- Expectativa de una transformación revolucionaria en industria y academia.

Poniendo un poco de orden

- Existen dos piezas interdependientes:
  - Énfasis en *big*
    - Bases de datos masivas
    - Capacidad de capturar la población y no una muestra
    - Nuevas estructuras de datos
    - $\Rightarrow$ Componente ingenieril
  - Énfasis en *data*
    - Nuevas tecnicas y modelos computacionales
    - Énfasis en predicción y no en estimación
    - Capacidad de gestionar tipos de datos antes imposibles
    - $\Rightarrow$ Componente estadístico/analítico

La primera visión

- Garner: "Big data" como datos en gran volumen, velocidad y variedad que requiren nuevas estrategias para poder ser analizados.
- O'Reilly: "Big data" como datos que exceden la capacidad de procesamiento una base de datos típica. Los datos son demasiados grandes, entran demasiado rápido, no se ajustan a la estructura tradicional.
- McKinsey: "Big data" como datos que exceden la capacidad de una base tradicional para capturar, almacenar, gestionar y analizar.

Investigación tradicional

Es útil ver la investigación con big data en el contexto de la investigación tradicional

1. Encuestas
   - Muy costosas
     - Para el investigador
     - Para los sujetos
   - Dependen de información auto-reportada
     - Estudios de opinión pública
     - Estudios de comportamientos en salud
2. Experimentos
   - (Potencialmente) costosos
   - Riesgos y límites de aleatorización

Características de los datos digitales

- *Volumen*
  - Explosión de la cantidad de data disponible
  - Más información es registrada digitalmente
  - Ejemplos típicos de industrias tecnológicas
    - 300 horas de video subidas a YouTube por minuto
    - 1 millon de transacciones diarias en Walmart
    - 1GB de datos por segundo en vehículos autopilotados
  - También una realidad en investigación social
    - Registros administrativos para estudiar movilidad social
    - Twitter para estudiar comunicación política
    - Metadatos telefónicos para estudiar redes sociales
  - El tamaño de los datos son un fin para una causa, no un objetivo

Características de los datos digitales (y II)

- *Permanentemente conectado*
  - Tenemos acceso en tiempo real a fenómenos
  - Nos permite estudiar fenómenos inesperados
  - Posibilidad de estudiar comportamiento actitudes antes, durante y después de protestas
- *Pasividad*
  - Información es recogida sin participación del sujeto
  - Facilita estudios que de otra forma alterarían el comportamiento del sujeto
  - No está sujeto a deseabilidad social
    - Estudio de perspectivas sobre raza usando búsquedas en Internet
  - **Comportamiento** más accesible que **actitudes**
  - Sujetos y performance pública
    - Expresiones públicas en Twitter

Características de los datos digitales (y III)

- *Exhaustividad parcial*
  - Análisis tradicionales dependen de muestras debido a costes
  - Tendencia a capturar todos los datos disponibles
  - Abaratamiento de coste de unidad de almacenamiento
  - En muchos casos es inevitable:
    - Registro de transacciones bancarias
    - Metainformación en mapas
    - Datos sobre el genoma
  - Habitualmente limitados por limitaciones ajenas a la investigación
    - Tener muchas observaciones no quiere decir que tengamos todas las variables

Características de los datos digitales (y IV)

- *Limitaciones en el acceso*
  - Generalmente los datos son recopilados por agentes sin incentivos para compartir
    - Empresas tecnológicas/redes sociales
    - Datos administrativos
  - Relacionado con dificultad de anonimizar datos
    - Retos de Statistical Disclosure
- *Selección*
  - Los datos no tienen por qué ser representativos de la población de interés
    - Estudio de opinión pública en Twitter
  - Es un problema relativo a la pregunta de investigación

Características de los datos digitales (y V)

- *Deriva*
  - Los datos están sujetos a cambios en
    - La composición de los usuarios
    - El comportamiento de los usuarios
    - Las propiedades de la plataforma
  - Estudios sobre actitudes raciales usando bots en Twitter
  - Cambios en Google y predicción de gripe
    - Los datos recopilados están modificados por los incentivos de quién los recopila (comportamiento y amistad en Facebook)
- *Variedad*
  - Frecuencia de datos *desestructurados*
  - Importancia del modelo de datos en las bases tradicionales
    - Información geolocalizada en Google Maps
    - Textos, imágenes y vídeos en redes sociales
    - Registros de actividad en Apple Health
  - Incluye información no deseada/deseable
    - Ruido debido a la naturaleza del problema
    - Ruido debido a falta de control en la recolección
    - Posible información sensible

Qué nos ha traído aquí

- Incremento en la capacidad de procesamiento en computadores
  - Un smartphone de baja gama es "ten times more powerful than the Cray-1 supercomputer installed at Los Alamos in 1976" usando una fracción de la energía.
- Desarrollo de las redes informáticas
- Digitalización y sensorización
  - Sensores en electrodomésticos
  - Asistentes personales como Siri
  - Tecnologías en tratamientos médicos como fMRI
  - Cámaras públicas
- Desarrollo de identificadores únicos
  - Códigos de barras e RFID
  - Información biométrica
- Caída del coste del almacenamiento
  - \$437.000 en 1980 $\Rightarrow$ \$0.019 en 2016

La segunda visión

- Sinan Aral (2010): "Revolutions in science have often been preceded by revolutions in measurement"
- Evolución de la ciencia en Jim Gray

1	Experimental	Empiricismo	Pre-Renacimiento
2	Teórica	Modelización y generalización	Pre-informática
3	Computacional	Simulación de fenémonos complejos	Pre-big data
4	Explorativa	Intensiva en datos	Ahora

Tabla: Cuatro paradigmas de ciencia

¿Qué ha cambiado?

- Retorno del empiricismo
  - Llamadas a terminar con el desarrollo de teorías
  - Minería de datos para descubrir relaciones
  - Transversalidad de las tareas de predicción
    - Algoritmo de recomendación de Amazon no sabe nada sobre gustos o convenciones sociales
  - Objetivo de descubrir relaciones sin desarrollar preguntas
- Problemas
  - Captura de datos es orientada y limitada, no omnisciente
  - Agenda y pregunta de investigación guían desarrollo empírico
  - Los datos contienen sesgos en su captura y análisis
    - Los datos no son elementos naturales
    - Decisión de qué constituye un dato es social
    - Riesgo de confundir correlación con causación
    - Evaluación en función de objetivos y no DGP observado

El nuevo papel de los datos

- Aproximación empírica
- Generar hipótesis provenientes de los datos y no de la teoría
- Objetivo no es solo identificar relaciones sino dirigir proceso de descubrimiento
- Complemento de la teoría y basado en teoría
  - Datos no sujetos a todos los marcos ontológicos concebibles
- Lo que tenemos son: *Herramientas para explorar, extraer valor e interpretar bases de datos desestructuradas, conectadas y de gran tamaño.*

¿Dónde nos deja esto? Retos

1. Complicaciones tecnicas
   - Herramientas tradicionales de análisis no sirven
   - Explosión de nuevas tecnologías
   - Nuevo problema: difícil de navegar el panorama de herramientas
2. Los modelos estadísticos anteriores no sirven
   - Necesidad de trabajar con datos no estructurados
   - Captura de datos vs diseño de recogida
   - El problema fundamental de la inferencia causal

¿Dónde nos deja esto? Oportunidades

1. Datos generados orgánicamente por los usuarios
   - Permite acceso a nuevas dimensiones de comportamiento
   - Recolección observacional de datos poco agresiva y poco costosa
     - Para el investigador
     - Para el investigado
2. Fuentes alternativas de datos
   - Capacidad de responder a nuevas preguntas de investigación

Preocupaciones éticas

- El modo en el que recogemos los datos y el tipo de trabajo que hacemos tiene implicaciones éticas
- Captura de datos es inevitable a nivel individual sin noción de consentimiento explícito
- Incluso aunque los datos estén anonimizados es posible reconstruir individuos
  - Ciudadano medio holandés está entre 250 y 1000 bases de datos dependiendo de su actividad online
  - Cesiones parciales y desinformadas de privacidad

Preocupaciones éticas

- Esencia de los modelos predictivos (segunda mitad del curso) es adelantarse al comportamiento individual
- Modelos corren el riesgo de replicar comportamientos discriminatorios
  - Concesión de créditos mediate algoritmos y pertenencia a minorías
- Problema no son los algorimos sino el proceso de generación de datos y tener herramientas para identificar sesgos
- Analista debe ser consciente de esos riesgos

Data scientist: el nuevo estadístico

- Los cambios anteriores producen la aparición de un nuevo profesional que es *mejor estadístico que un desarrollador y mejor desarrollador que un estadístico*
- En el mercado se refiere a tres figuras:
  1. Data engineer
     - Especialización del backend developer y el database administrator
     - Crear y mantener la estructura que alberga los datos
  2. Data analyst
     - Exploración y visualización de datos
     - Más relacionado con el estadístico
  3. Data scientist
     - Implementación de algoritmos
     - Interfaz entre backend y cliente

Data scientist: el nuevo estadístico

- Capaz de:
  1. Interaccionar con bases de datos mas grandes
  2. Integrar modelos estadísticos en sistemas automáticos
  3. Implementar soluciones computacionales
- Importancia de la comunicación con:
  - Especialistas de área
  - Ingenieros y desarrolladores
  - Estadísticos y analistas

Administración

- Tutorías
  - Oficina en Sala Santa María
  - Política de puerta abierta
  - Contáctenme para horas
- Materiales
  - Código posteado al final del día en griverorz.net/big-data

Plan del curso (y I)

Primera mitad: Infraestructura

- Fundamentos de programación en R
- Captura de datos de Internet
  - REST API
  - Web Scraping
  - Streaming
- Almacenamiento y procesamiento
  - Bases de datos
  - Procesamiento paralelo

Plan del curso (y II)

Segunda mitad: Análisis

- Análisis de redes sociales
  - Descriptivos de redes
  - Inferencia
- Métodos de aprendizaje no supervisado
- Análisis de textos
  - Expresiones regulares
  - Pre-trained taggers
  - Análisis de temas
- Fusión de registros
  - Fellegi-Sunter