Intro

Ciencia de Datos y Economía Aplicada

Ignacio Sarmiento-Barbieri

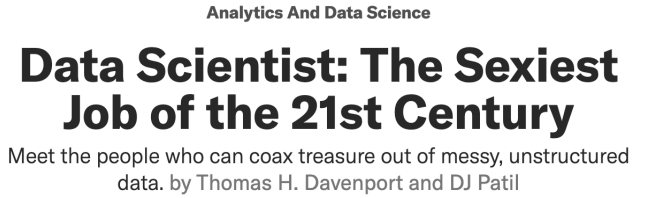
Universidad de los Andes

Agenda

1. ¿Qué es la Ciencia de Datos?
   - ¿Que es Big Data?
   - ¿Que es Aprendizaje de Máquinas?
   - Econometría vs Machine Learning
2. Predicción vs Causalidad
3. Sobre el Curso
4. Review

¿Qué es la Ciencia de Datos?

Ciencia de Datos es el proceso de usar datos para **entender el mundo**, **predecir comportamientos** y **tomar decisiones**.



Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

Source: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

¿Qué es la Ciencia de Datos?

¿Qué implica?

- Recolección y análisis de ”big data”
- Herramientas de múltiples disciplinas:
  - Estadística y **econometría** para inferencia y causalidad
  - **Aprendizaje de Máquinas** (*(Machine learning)*) para predicción
  - Programación y bases de datos para manipulación y escalabilidad
  - Visualización para comunicar hallazgos

Agenda

1. ¿Qué es la Ciencia de Datos?
   - ¿Que es Big Data?
   - ¿Que es Aprendizaje de Máquinas?
   - Econometría vs Machine Learning
2. Predicción vs Causalidad
3. Sobre el Curso
4. Review

¿Que es Big Data?

- Big n, es solo parte de la historia
- Big también es big k, muchos covariates, a veces $n << k$
- Vamos a entender Big también como datos que no surgen de fuentes tradicionales. Donde enfrentamos las 4 V:
  - **Volumen**: grandes cantidades de datos
  - **Variedad**: múltiples tipos y fuentes
  - **Velocidad**: generación continua
  - **Veracidad**: calidad y ruido

Agenda

1. ¿Qué es la Ciencia de Datos?
   - ¿Que es Big Data?
   - ¿Que es Aprendizaje de Máquinas?
   - Econometría vs Machine Learning
2. Predicción vs Causalidad
3. Sobre el Curso
4. Review

¿Qué es Aprendizaje de Máquinas?

El Aprendizaje de Máquinas trata sobre predicción

- El aprendizaje de máquinas es una rama de la informática y la estadística, dedicada a desarrollar algoritmos para predecir resultados y a partir de variables observables X.
- Esto se deja como un problema empírico que la computadora puede "aprender".
- En general, esto significa que abstraemos del modelo subyacente, el enfoque es pragmático.

¿Qué es Aprendizaje de Máquinas?

El Aprendizaje de Máquinas trata sobre predicción

- El aprendizaje de máquinas es una rama de la informática y la estadística, dedicada a desarrollar algoritmos para predecir resultados y a partir de variables observables X.
- Esto se deja como un problema empírico que la computadora puede "aprender".
- En general, esto significa que abstraemos del modelo subyacente, el enfoque es pragmático.

"Lo que funciona, funciona..."

“Lo que funciona, funciona...”

Source: https://xkcd.com/1838/

“Lo que funciona, funciona...”

La primera victoria y derrota de ML

- Contexto ¿similar? al de Covid 2020: Epidemia de la gripe A en 2009
- En EEUU la forma de monitorear es a través de reportes de la CDC
- La CDC agrega a nivel de ciudad, condado, estado, región y a nivel nacional
- Todo esto llevaba aproximadamente 10 días $\rightarrow$ demasiado tiempo para una epidemia

“Lo que funciona, funciona...”

Google se ha unido a la conversación

- Google propuso un mecanismo ingenioso: **Google Flu Trends**
- Punto de partida:
  - Proporción de visitas semanales por Gripe A en hospitales
  - 9 regiones $\times$ 5 años (2003-2007) = $2,340$ datos
  - Estos son los datos que tomaban 10 dias en elaborarse (comparemos con la Colombia de 2009)
- Google cruzó estos datos con las búsquedas sobre la gripe A
- Con estos datos, construyeron un modelo para predecir intensidad de gripe A

“Lo que funciona, funciona...”

Google se ha unido a la conversación

- ¿Un sólo modelo?

“Lo que funciona, funciona...”

Google se ha unido a la conversación

- ¿Un sólo modelo?
- Los investigadores de Google estimaron **450 millones** de modelos
- Eligieron el que mejor predice sobe la intensidad de búsqueda
- Les permite tener información diaria, semanal o mensual para cualquier punto de EEUU y el mundo
- A Google le toma 1 día lo que a la CDC 10!

“Lo que funciona, funciona...”

Google se ha unido a la conversación

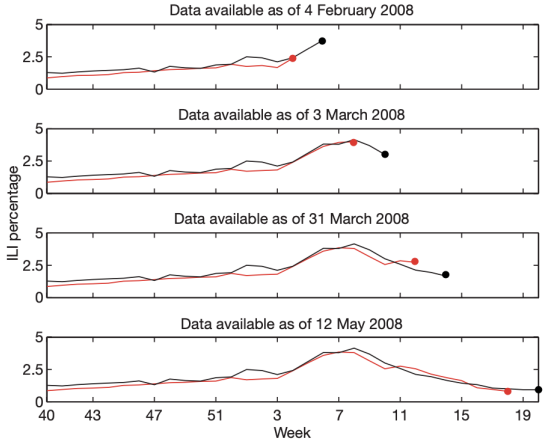


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3

“Lo que funciona, funciona...”

El rey ha muerto, larga vida al rey

- ¿Qué tienen en común Google Flu y Elvis?
  - Abanderados de la revolución
  - Definió y redefinió las reglas sistemáticas para hallar la solución a un problema
  - Éxito rotundo $\rightarrow$ Publicación en Nature!
    https://www.nature.com/articles/nature07634
  - Pero como a Elvis el éxito fue efímero
  - La predicciones comenzaron a sobre-estimar considerablemente la incidencia de la gripe A
  - Google Flu esta ahora archivado (disponible al público)
  - Continúa recolectando datos pero solo algunas instituciones científicas tienen accesso

“Lo que funciona, funciona...”

Thanks to machine-learning algorithms,
the robot apocalypse was short-lived.

Source: https://smbc-comics.com/comic/rise-of-the-machines

Agenda

1. ¿Qué es la Ciencia de Datos?
   - ¿Que es Big Data?
   - ¿Que es Aprendizaje de Máquinas?
   - Econometría vs Machine Learning
2. Predicción vs Causalidad
3. Sobre el Curso
4. Review

Econometría, Machine Learning y el Reto de los Datos

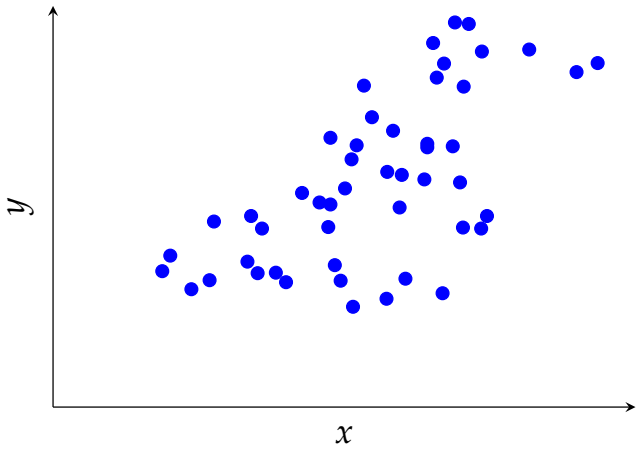
¿Por qué importa la econometría?

- Aporta **rigor** para pensar en relaciones causales
- Complementa el ML, que se enfoca en **predicción**, con una mirada sobre el **"what if"**

La ciencia de datos une el poder predictivo del ML con la interpretación causal de la econometría.

Agenda

Predicción vs Causalidad



Predicción vs Causalidad

Agenda

Evaluaciones

Table 1: Puntajes

	Puntaje Individual	Puntaje Total
Talleres grupales	10%	40%
Presentaciones	15%	30%
Proyecto Final		20%
Participación		10%
Total		100%

Agenda

Review

- Hoy: el paradigma predictivo y regresion lineal
  - Machine Learning es sobre predicción
  - Regresión lineal es un herramienta poderosa.
- Proxima clase: Actividad (traer compu con `R`) Grupal (grupos en BN)