

# Lecture 0: Introducción

## Big Data and Machine Learning for Applied Economics

### Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 11, 2020

# Agenda

- 1 Motivación
- 2 Presentación: un poco sobre nosotros
- 3 Housekeeping
- 4 Recap
- 5 Next
- 6 Para seguir leyendo

# Motivación



# Motivación

- ▶ Hay datos por todos lados: redes sociales, tarjetas de crédito, teléfonos, relojes, GPS, ....
- ▶ Las maquinas estan haciendo cosas "inteligentes":
  - ▶ Facebook te reconoce las caras en las fotos,
  - ▶ Siri te entiende la voz
  - ▶ Google es capaz de traducir sitios web
- ▶ Todos estos "breakthroughs" son tanto computacionales como estadísticos

# Motivación

## La primera victoria y derrota del Big Data

- ▶ Contexto ¿similar? al de hoy: 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 → demasiado tiempo para una epidemia

# Motivación

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 días 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

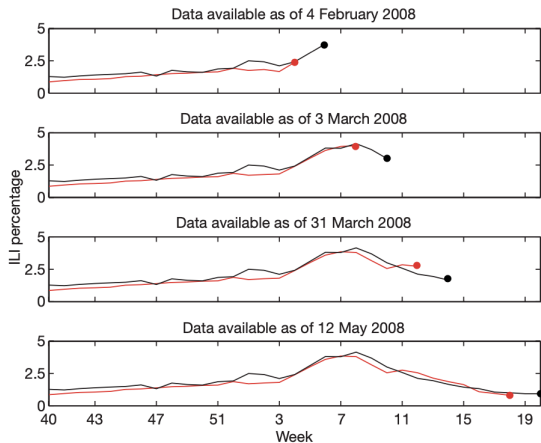
# Motivación

Google se ha unido a la conversación

- ▶ Un solo modelo?
- ▶ Los investigadores de Google estimaron **450 millones** de modelos
- ▶ Eligieron el que mejor predice sobre 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!

# Motivación

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 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by the CDC.



# Motivación

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 → Publicacion 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 publico)
  - ▶ Continúa recolectando datos pero solo algunas instituciones científicas tienen acceso

# Motivación

- ▶ Otro ejemplo, los algoritmos de reconocimiento de cara:
  - ▶ no son reglas fijas basadas en que los humanos entendemos por rostros y a partir de ello buscar combinaciones de píxeles.
  - ▶ son algoritmos que usan datos de fotos etiquetadas con un rostro y estiman una función  $f(x)$  que predice si es un rostro o no a partir de píxeles  $x$ .
- ▶ El aprendizaje de máquinas se hizo una realidad cuando los investigadores dejaron de afrontarlo de manera teórica y lo hicieron empíricamente.
- ▶ Las similitudes con la econometría plantea interrogantes:
  - ▶ ¿Estos algoritmos están simplemente aplicando técnicas estándar a nuevos y grandes conjuntos de datos?
  - ▶ Si hay herramientas empíricas fundamentalmente nuevas, ¿cómo encajan con lo que conocemos?
  - ▶ Como economistas empíricos, ¿cómo podemos utilizarlas?

# Presentación: Sobre mi

- ▶ Ignacio Sarmiento Barbieri
- ▶ <https://ignaciomsarmiento.github.io/>
- ▶ [i.sarmiento@uniandes.edu.co](mailto:i.sarmiento@uniandes.edu.co)
- ▶ Intereses: Economía Publica y Urbana. Economía del Crime. Econometría Aplicada, Big Data y Machine Learning.
- ▶ Originario de Salta, Argentina

# Presentación: Sobre mi

## ► Originario de Salta, Argentina



# Presentación: Sobre el profesor asistente del curso

- ▶ Jacob Muñoz Castro
- ▶ [j.munoz14@uniandes.edu.co](mailto:j.munoz14@uniandes.edu.co)
- ▶ Intereses: financial economics, misallocation, regulation, econometrics, big data and machine learning
- ▶ Originario de Armenia, Colombia.

# Presentación: Sobre vos

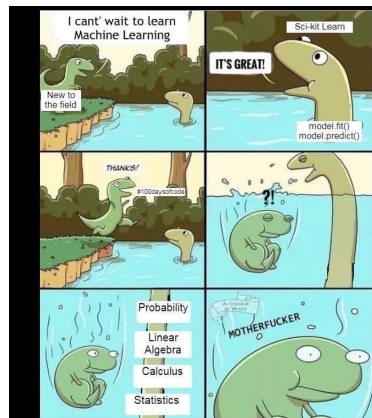
- ▶ Nombre
- ▶ Programa en el que estan inscriptos
- ▶ Un poco de su "background"
- ▶ Perfil de Slack y chat

# Clases

- ▶ Horario de Clases: Las clases serán completamente virtuales los martes y jueves de 11:30 a 12:45
- ▶ Zoom link: <https://uniandes-edu-co.zoom.us/j/92913017634>
- ▶ Github classrooms <https://github.com/ECON-4676-UNIANDES>
  - ▶ Syllabus: <https://github.com/ECON-4676-UNIANDES/Syllabus>
  - ▶ Clases
  - ▶ Talleres
  - ▶ Tutoriales (eTAs)
- ▶ Comunicacion via [Slack](#). Si no recibieron el link de la invitación por favor enviarme correo

# Lenguajes

- ▶ Matemáticas
- ▶ Inglés
- ▶ Uno de los objetivos implícitos es que mejoren sus habilidades para escribir código y usar herramientas de la industria
  - ▶ Elijan el que quieran:
    - ▶ R, Python, o cualquier otro
    - ▶ no hay restricción
    - ▶ yo me basare en R
  - ▶ Github
  - ▶ AWS
- ▶ Aprender haciendo y mucha prueba y error!





# Bibliografía

## 1 Statistical Learning (FREE!!! speech and beer)

- ▶ James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (ISLR)
- ▶ Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction
- ▶ Efron, B., & Hastie, T. (2016). Computer age statistical inference.

## 2 Libros avanzados de econometría

- ▶ Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods
- ▶ Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data.
- ▶ Hayashi, F. (2000). Econometrics

## 3 Otros tópicos: bayesiano, econometria espacial, voy a indicar y a actualizar el syllabus

# Criterios de evaluación

Clase magistral	Puntaje Individual	Puntaje Total
Participación		10%
Talleres	10%	40%
Propuesta de Trabajo		25%
Primer Entrega	5%	
Segunda Entrega	5%	
Tercer Entrega	15%	
Examen Final		25%
Total		100%

# Criterios de evaluación

- ▶ **Participación (10%)** → “tiebreaker” de la nota final.
  - ▶ Asistencia a clases: aleatoriamente voy a asignar al menos 4 personas que tengan la cámara prendida, voy a mandar mensaje por Slack
  - ▶ Slack: compartir artículos, tutoriales, material relevante para el curso en el canal #cosas interesantes
  - ▶ Github: Spell check al instructor, sobre todo acentos =), y corregir falacias/errores conceptuales
  - ▶ AWS: crédito para que todos tengan su “nube”

# Criterios de evaluación

## ▶ Talleres (40%)

- ▶ Pueden ser en grupos de no mas de 3 personas.
  - ▶ Enviar slack al profe con los miembros del grupo
  - ▶ Bonus en participación de la historia del repositorio (evaluar la contribución de cada estudiante).
- 
- ▶ Primer taller → fecha de entrega 8 de septiembre antes de la clase (pronto disponible en github)
- 
- ▶ En clase los estudiantes a designar aleatoriamente los dias previos, presentaran su solución en no mas de 3 slides.

# Criterios de evaluación

## ▶ Propuesta de trabajo (25%)

- ▶ Pueden ser en grupos de no mas de 3 personas. El mismo o no de los talleres.

### 1 Primera entrega los grupos se reunirán conmigo y presentarán brevemente (máximo 5 slides) la idea y como planean llevarla a cabo.

- ▶ Concretar cita con el profesor,
- ▶ fecha de entrega: 25 de septiembre

### 2 Segunda entrega donde se expondrán la idea con los datos propuestos.

- ▶ Concretar cita con el profesor,
- ▶ fecha de entrega: 23 de octubre

### 3 Entrega final que consolida todo el trabajo.

- ▶ fecha de entrega: 17 de diciembre
- ▶ Bonus a los estudiantes que además de presentar el plan de trabajo o propuesta, entreguen resultados concretos.

# Cláusula de ajustes razonables

- ▶ Si lo considera pertinente, siéntase en libertad de informar al profesor lo antes posible si usted tiene alguna condición, visible o invisible, por la cual requiera algún ajuste para estar en igualdad de condiciones con los y las demás estudiantes. Debido a las actuales circunstancias, barreras de conectividad o acceso a los recursos tecnológicos indispensables para la clase son parte de las condiciones que pueden requerir ajustes. Por la misma razón, no necesitará presentar documentación para solicitar esos ajustes.
- ▶ También lo invitamos a buscar asesoría y apoyo en la Dirección de su programa, en la Decanatura de Estudiantes (<http://centrodeconsejeria.uniandes.edu.co>, Bloque Ñf, ext. 2207, 2230 y 4967, horario de atención L-V 8:00 a.m. a 5:00 p.m.) o en el Programa de Acción por la Igualdad y la Inclusión Social (PAIIS) de la Facultad de Derecho ([paiis@uniandes.edu.co](mailto:paiis@uniandes.edu.co)). Si su solicitud se basa en dificultades de acceso a conectividad o tecnología, es particularmente importante que haga este contacto adicional para que pueda acceder a los recursos de apoyo que brinda la Universidad.
- ▶ Se entiende por ajustes razonables todas "las modificaciones y adaptaciones necesarias y adecuadas que no impongan una carga desproporcionada o indebida, cuando se requieran en un caso particular, para garantizar a las personas con discapacidad el goce o ejercicio, en igualdad de condiciones con las demás, de todos los derechos humanos y libertades fundamentales" Convención sobre los Derechos de las personas con discapacidad, art.2.

# Cláusula de respeto por la diversidad

- ▶ Todos debemos respetar los derechos de quienes hacemos parte de esta comunidad académica. En esta comunidad consideramos inaceptable cualquier situación de acoso, acoso sexual, discriminación, matoneo, y/o amenaza. La persona que se sienta en alguna de estas situaciones puede denunciar su ocurrencia y buscar orientación y apoyo ante alguna de las siguientes instancias: el equipo pedagógico del curso, la Coordinación o la Dirección del programa, la Decanatura de Estudiantes (DECA, Ed. Ñf-Casita amarilla), la Ombudsperson (ombudsperson@uniandes.edu.co, Edificio RGA–Pedro Navas, Of. 201, ext. 5300 y 3933) o el Comité MAAD (lineamaad@uniandes.edu.co, <https://uniandes.edu.co/MAAD> o a la ext. 2707 o 2230). Si quieren mayor información, guía o necesitan activar el protocolo MAAD pueden acudir a Nancy García (n.garcia@uniandes.edu.co) en la Facultad. También puede acudir a los grupos estudiantiles que pueden ofrecerle apoyo y acompañamiento: No Es Normal (derechoygenero@uniandes.edu.co o <https://www.facebook.com/noesnormaluniandes/?fref=ts>); Pares de Acompañamiento Contra el Acoso-PACA (paca@uniandes.edu.co o <https://www.facebook.com/PACA-1475960596003814/?fref=ts>). Para mayor información sobre el protocolo MAAD, puede visitar esta página: <https://decanaturadeestudiantes.uniandes.edu.co/index.php/es/sobre-la-decanatura/827>

# Política de momentos difíciles

- ▶ Todas las personas pueden pasar por un momento difícil que de alguna manera pueda afectar nuestra vida en la Universidad. Pueden ser problemas en casa, con la pareja, incluso estrés por esta u otra materia.
- ▶ Si usted siente que está pasando por un momento complicado, sin importar el motivo, siéntase con la tranquilidad de hablar con la profesora para pedir tiempo o apoyo. Ningún trabajo o entrega puede sobrepasar su salud mental y física.
- ▶ **Su bienestar es lo más importante.**



# Recap

- ▶ 3 Plataformas
  - ▶ Curso: Github <https://github.com/ECON-4676-UNIANDES>
  - ▶ Comunicación: Slack
  - ▶ Play: AWS
- ▶ Inglés, Estadística, Econometría y mucho coding
- ▶ Participación, prueba y error serán las banderas del curso, armarse de paciencia!
- ▶ No duden en comunicarse conmigo por cualquier tema
  - ▶ Ajustes razonables, incluyendo problemas de conectividad
  - ▶ Consultas del curso o de cualquier otra cosa
  - ▶ Momentos difíciles
- ▶ **Recordar: Tu bienestar es lo más importante!!!**

# Next

- ▶ Tutorial de GitHub
- ▶ Verificar que tiene instalado el software que usted va a usar en el curso
  - ▶ Voy a usar R con RStudio
  - ▶ Usted puede usar el que desee, altamente recomendados son R y Python
  - ▶ Tutorial de R disponible en breve en los e-TAs
  - ▶ Sera muy bienvenido y debidamente atribuido si quiere contribuir con un tutorial de Python o “traducir” el eTA a Python

# Para seguir leyendo

- ▶ Einav, Liran, and Jonathan D. Levin. The data revolution and economic analysis. No. w19035. National Bureau of Economic Research, 2013.
- ▶ Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), pp.87-106.
- ▶ Sosa Escudero, W. (2019). Big Data. Siglo Veintiuno Editores
- ▶ Varian, Hal R. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28, no. 2 (2014): 3-28.