Programa de Metodología Aplicada en Investigación Política y Social (PMet)

Política con Big Data

Prof. @Ernesto Calvo
University of Maryland
ecalvo@umd.edu

Universidad Nacional de San Martin
30 de junio: 9:30 a 12:30 y 14:00 a 17:00
3 y 5 de julio: 18:30 a 21:00
7 de julio: 10:30 a 12:30 y 14:00 a 17:00

Este seminario tiene como objetivo introducir a sus participantes al procesamiento de grandes bases de datos en R/R-Studio; el uso de ggplot, stargazer, knitr y rvest; el procesamiento de filas JSON; el uso de APIs de organizaciones como Facebook, Twitter y el NYT. El seminario está estructurado en 12 módulos, concluyendo con un hackaton de datos y la presentación de resultados. El curso será dictado integralmente en R/R-Studio.

Objetivos del Curso:

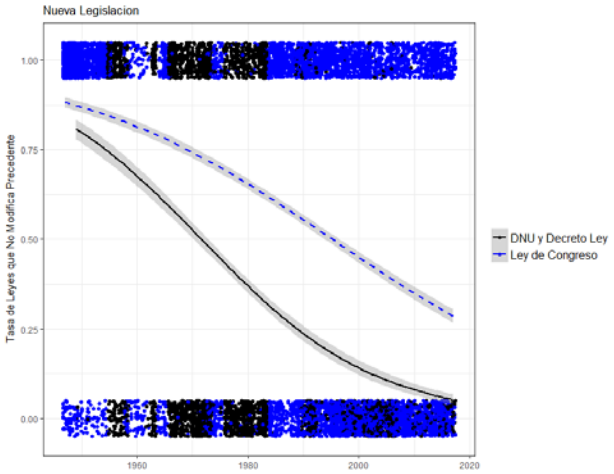
Los @participantes adquirirán las herramientas para:

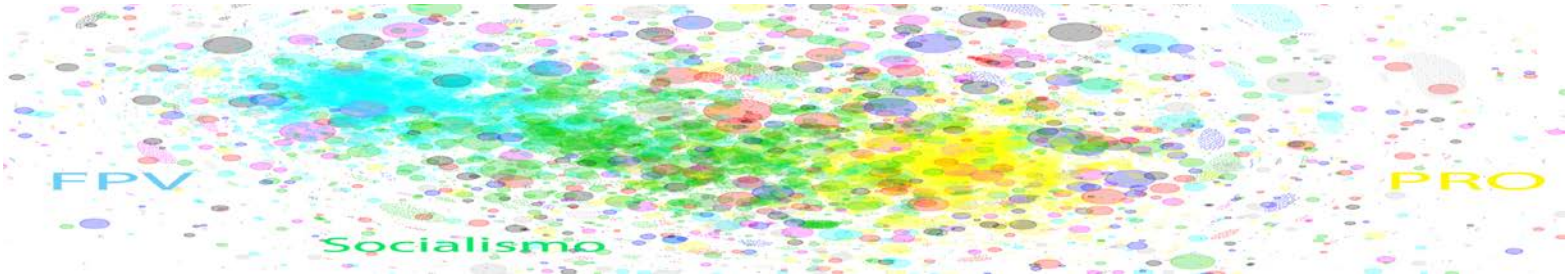
1. Utilizar R/R-Studio para crear listados, arrays, vectores, matrices, estimar modelos estadísticos lineales (lm) y modelos lineales generalizados (glm);

2. Graficar resultados en ggplot, producir tablas con stargazer, publicar resultados en HTML usando knitr;

3. Conectar con APIs para extraer datos de internet;

4. Utilizar el paquete "twitteR" para analizar timelines;





5. Finalmente, los participantes producirán un blog colaborativo con un resultado relevante utilizando big data sobre Legislación Argentina, Elecciones, Congreso o Twitter (otras opciones disponibles).

Textos de apoyo:

i) **An introduction to R:**
https://cran.r-project.org/doc/manuals/R-intro.pdf

ii) **Ggplot:**
https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

iii) **Stargazer:**
https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

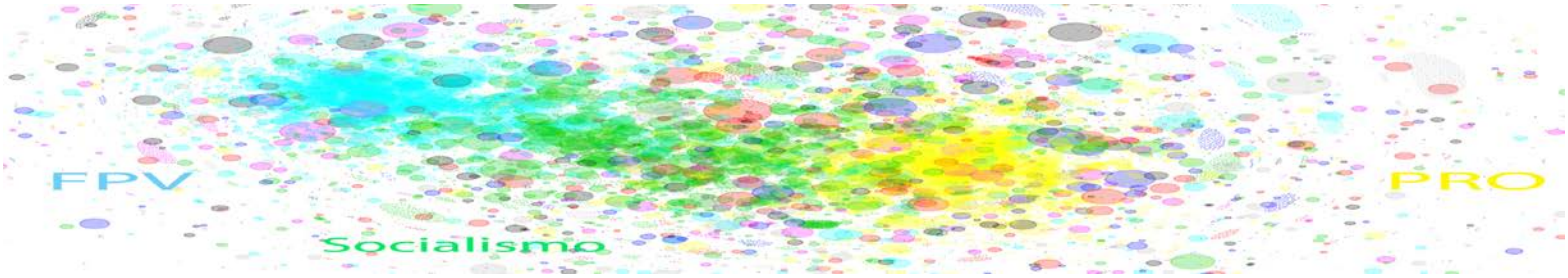
iv) **Knit:**
https://yihui.name/knitr/

Dependiendo de la base de datos que utilice cada equipo en este seminario, distribuiré materiales sobre Congreso, Elecciones o Redes Sociales.

Software:
En el seminario utilizaremos R/R-Studio (cualquiera de las dos opciones).

- **Para bajar R-Studio: https://www.rstudio.com/**
- **Para bajar R: https://cran.r-project.org/**
- **Para instalar los paquetes requeridos para la clase correr en R o R-Studio el código:**

```
sapply(c("rvest","httr","knitr","tm","igraph","foreign","twitteR","stargazer"
),install.packages)
```



Programa

Módulo 1, Sábado Julio 1, Mañana I: Presentación e Introducción a R, 1 ½ horas

- Una introducción a R/R-Studio. Tipos de datos, vectores, matrices, arrays, listas, loops, apply, tapply/sapply. Programación de funciones.

Módulo 2, Sábado Julio 1, Mañana II: Trabajando con Modelos en R, 1 ½ horas

- LM y GLM. Estimados y gráficos. Una revisión de modelos estadísticos utilizando R. Algunas reglas para trabajar con Big Datasets.

Módulo 3, Sábado Julio 1, Tarde I: Selección de Datasets, 1 ½ Hours.

- Descripción de datasets alternativos: Congreso, Legislación, Elecciones, Twitter.

Módulo 4, Sábado Julio 1, Tarde II: Usando ggplot para explorar nuestras variables dependientes y nuestros covariados.

- La semántica de gráficos de ggplot. Un modelo, un gráfico, una idea.

Módulo 5, Miércoles 4 de Julio, Tarde I: Una Introducción a web scrapping con APIs, 1 Hora.

- RSS feeds, HTTR, rvest. El API del NYT.

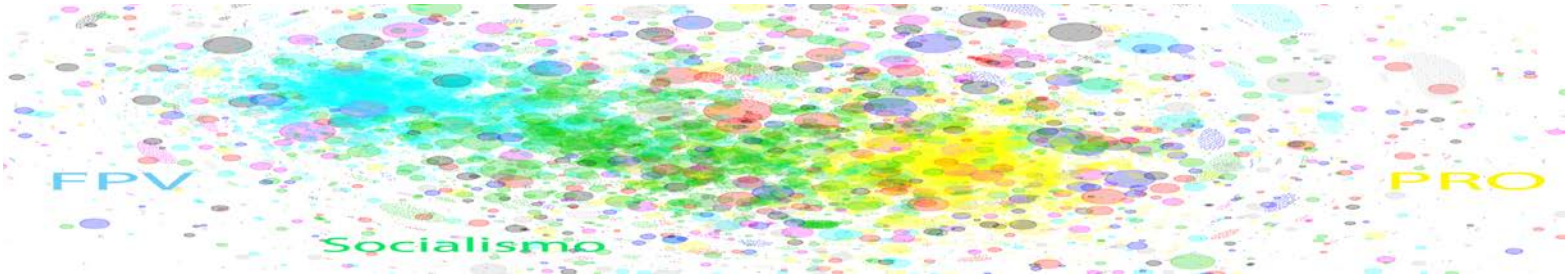
Módulo 6, Miércoles 4 de Julio, Tarde II: Elección de Modelos, 1 Hora.

- Distribuciones, efectos directos, relaciones condicionales (interacciones).

Módulo 7, Viernes 6 de Julio, Tarde I: Otro ejemplo de Web Scrapping, Twitter API, 1 hora.

- El uso de APIs para procesar timelines en Twitter. El formato JSON,

Módulo 8, Viernes 6 de Julio, Tarde II: Preparación de los materiales para el Hackaton, 1 hora.



Módulo 9, Sábado, Mañana I: Preparación del Articulo, 1 ½ horas.

- Estructura y Modelos. Una introducción a Knitr.

Módulo 10, Sábado, Mañana II: Preparación del Articulo, 1 ½ horas.

- Producción de tablas definitivas. Una introducción a Stargazer.

Módulo 11, Sábado, Tarde I: Preparación de presentación en HTML, 1 ½ horas.

- Trabajo de laboratorio.

Módulo 12, Sábado, Tarde II: Presentaciones y Cierre, 1 ½ horas.

- Cierre y Presentación de resultados.