

Laboratorio de Datos

2do. cuatrimestre 2021



Ariel Salgado arielolafsalgado@gmail.com
Martín Elias Costa mec@df.uba.ar
Matías Iglesias matuteiglesias@gmail.com
Andrés Farall afarall@gmail.com
Guillermo Solovey gsolovey@gmail.com

Ariel Salgado

Licenciado en física, doctorando del Instituto de Cálculo. Me alejé un poco de la física más tradicional para acercarme a problemas con temáticas forenses.

Mi tesis es en grafos y análisis de datos.

Además soy powerlifter aficionado!



Martín Elias Costa

Doctor en física de la UBA, trabajé en áreas de datos tanto del sector público como del sector privado. Mis principales temas de interés son: visión computacional y procesamiento de lenguaje natural



Matías Iglesias

Me interesa mucho el uso de datos para problemas públicos. Soy principalmente analista de datos.

- Lic. Física (2009-2015)
- Dr. Economía (2015-2020)

Actualmente:

- Dir. Análisis de Datos en MinSeg PBA
- Consultor (STC), en DIME World Bank Group.



Andrés Farall

Actuario con un Master en Estadística Matemática y Doctor de la UBA, cuento con más de 25 años de experiencia en Ciencia de Datos, asesorando empresas, organismos y brindando cursos de posgrado.



Guillermo Solovey

Investigador de CONICET. Físico volcado a la ciencias cognitivas. La ciencia de datos es fundamental en todas las etapas de mi trabajo sobre toma de decisiones y formación de creencias.



Objetivos de la Materia

Brindar una introducción al **Análisis Exploratorio** de Datos (EDA) y al **Modelado** de Datos, utilizando elementos básicos de matemáticas y de programación, sin el uso de nociones de Probabilidad y Estadística.

Generar una serie de **preguntas** que pueden hacerse sobre un conjunto de datos, que finalmente serán respondidas mediante modelos estadísticos o algoritmos de *machine learning*.

Introducir algunos **conceptos fundamentales** de la Ciencia de Datos, como ser: Descripción-Predicción-Explicación, significatividad estadística, sobreajuste, bondad de ajuste, funciones de pérdida, asociación entre variables, análisis supervisado vs. no supervisado, modelos paramétricos vs. no paramétricos, etc.

Programa

1. Obtención y organización de datos. Datos estructurados y no estructurados.
2. Visualización de datos como herramienta exploratoria antes del desarrollo de modelos y aprendizaje estadísticos. Análisis exploratorio de datos.
3. Introducción al modelado. Regresión Lineal Múltiple y Vecinos más Cercanos. Modelos predictivos vs modelos explicativos. Distinción entre modelos univariados y multivariados, y modelos paramétricos y no-paramétricos.
4. Herramientas de validación de un modelo. Muestras de testeo y entrenamiento. Métricas y métodos para la evaluación de algoritmos y modelos estadísticos.
5. Análisis Supervisado: Regresión y Clasificación
6. Análisis No Supervisado: Clustering y Reducción de Dimensión

Dinámica de las clases

Las clases van a ser virtuales.

Todo contenido estará grabado y subido a *YouTube* y *GitHub*.

Vamos a pasarlo oportunamente. Una parte de la clase (*en términos de tiempo*) está reservada para que vean ese material.

El resto de la clase, vamos a dedicarlo a la resolución de trabajos prácticos.

Modalidad de Evaluación

Un **Trabajo Práctico** en la mitad de la materia, en el que deberán realizar un análisis sobre un conjunto de datos, devolviendo un *notebook* como entregable.

Un **Trabajo Práctico** sobre el final de la materia, en el que deberán generar una serie de videos explicativos sobre un tema específico.

Un **Examen Final** tipo *múltiple choice* que repasa los contenidos conceptuales de toda la materia.

Referencias

- Wickham H & Grolemund R. (2016). R for data science: import, tidy, transform, visualize, and model data. RStudio, Inc.
- Antonio Vazquez Brust, (2020) Ciencia de Datos para Gente Sociable - [Libro Completo](#)
- Holger K. (2018) Data visualization: a practical introduction. Princeton University Press. El libro
- Canal de Ariel <https://www.youtube.com/user/olafsalgado>
- Canal de Andrés <https://www.youtube.com/channel/UC5Rup8Tq90zOekekNISdRQ>

La Herramienta más difundida

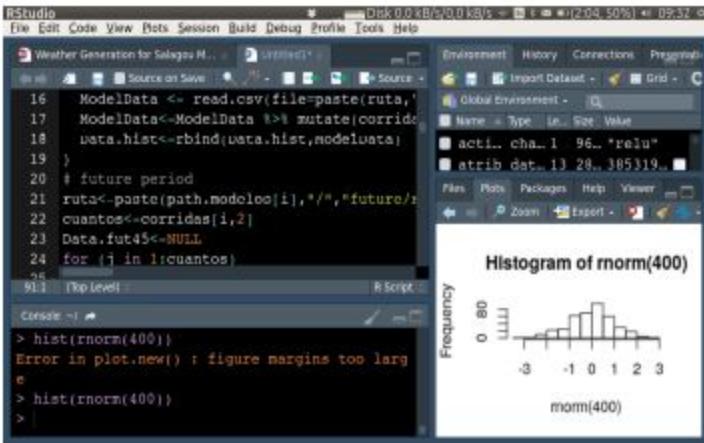
#11 conceptos

- Código Abierto (GNU-GPL v 3)
- Gratuito (GNU-GPL V 3)
- Multiplataforma (Windows, Linux, MAC/os)
- Comunitario (>2.000.000 usuarios al 2019)
- Orientado a objetos
- Especializado en el análisis de datos
- Potentes gráficos
- Flexible (interprete)
- Alto nivel de expresión
- Fuerte aceptación/intervención académica
- Facil integración vertical



Cómo se trabaja en R

Datos →
Análisis →
Preguntas →



Informes →
APIs →
Prototipos →
Resultado →
Datos →
Métodos →
Algoritmo →

UNIX



El Origen de R...

1976: John Chambers crea el Lenguaje S en **Bell Labs**.

1984: AT & T (Bell Labs) vende la Licencia de S.

1991: Ross Ihaka y Robert Gentleman lanzan el **Proyecto R** basado en el lenguaje S.

2001: La versión 1.0.0 de R es lanzada.

2019: R cuenta con,

-> 2.000.000 de usuarios

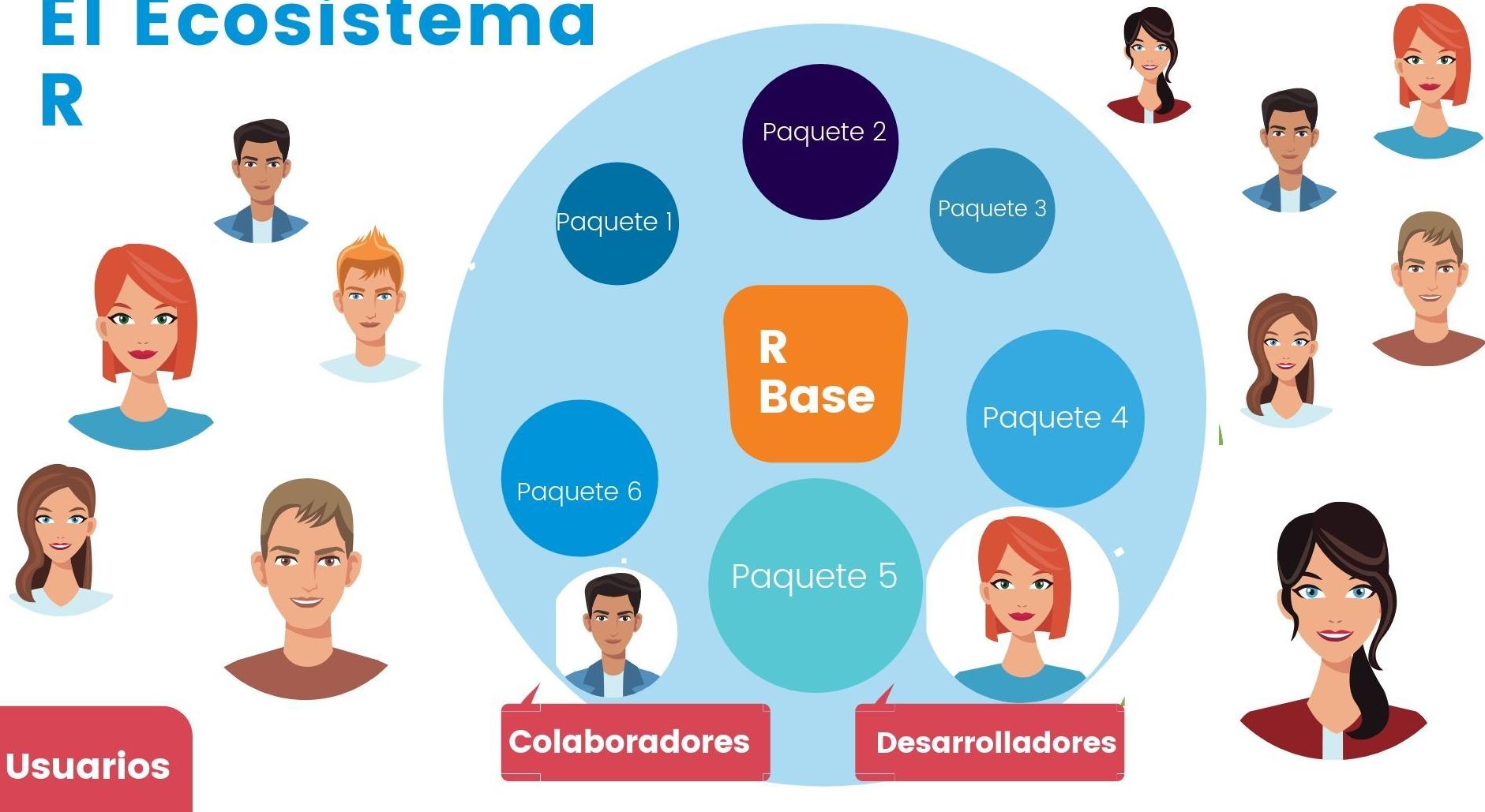
-> 10.000 colaboradores

-> 100 grandes empresas lo utilizan



El Ecosistema

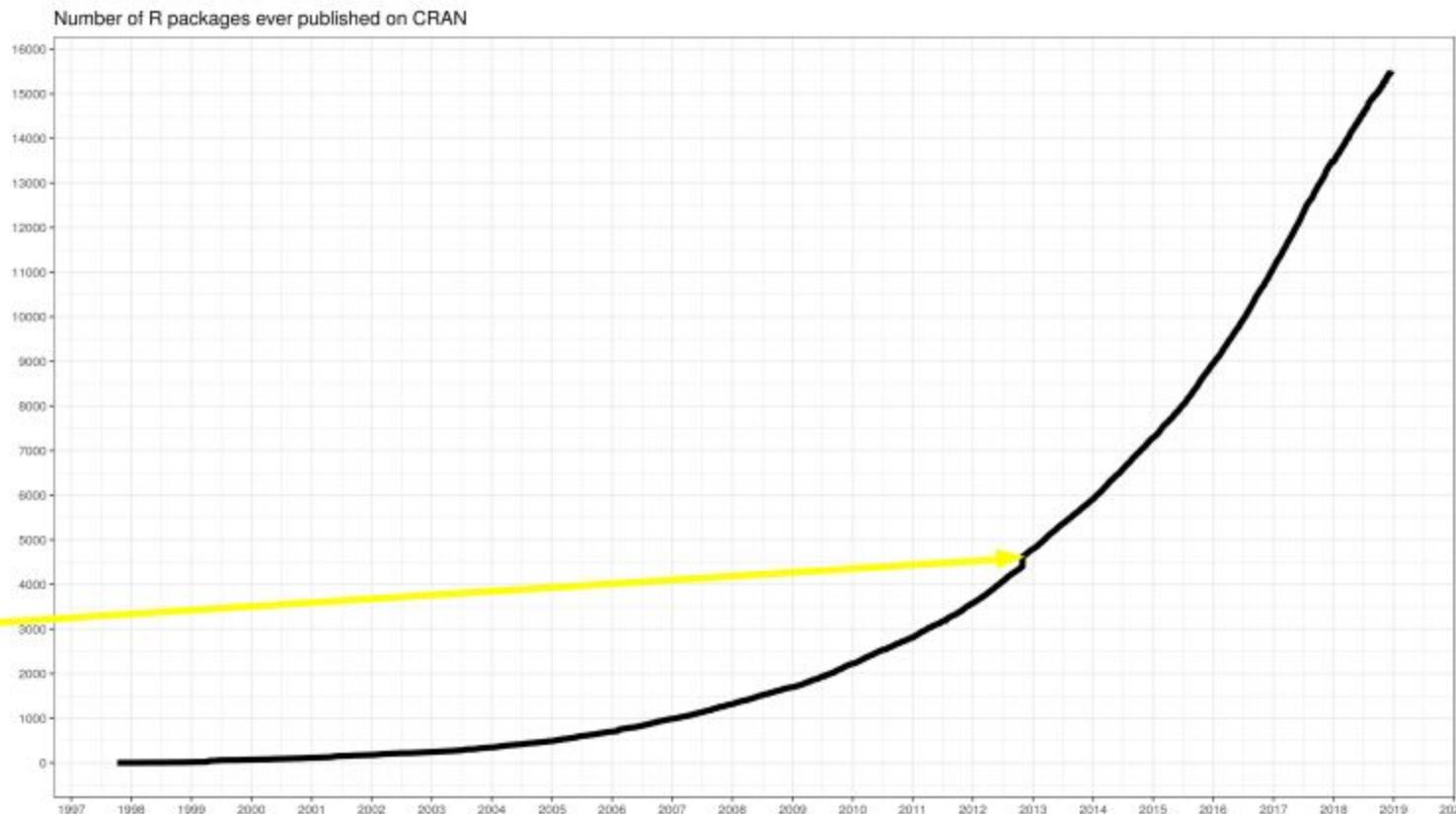
R



Crecimiento Exponencial de R

El Ecosistema crece

- Usuarios
- Paquetes
- Colaboradores
- Conexiones



¿Qué Explica el Surgimiento del Paradigma FOS?

- Compartir un desarrollo digital tiene un **costo de oportunidad negativo**. Una vez resuelto un problema propio conviene compartirlo !
- **Subsidios cruzados**: profesionales y científicos de países desarrollados generan y mantienen los proyectos financiados por organizaciones altamente lucrativas
- La existencia de un entorno tecnológico global interconectado (Internet)
- Bajo **costo** de generación de proyectos, pero **alto impacto**

¿Qué Explica el Crecimiento del Paradigma FOS?

- Una vez creada la herramienta FOS, las ventajas comparativas son INMENSAS
- Precios inferiores a la competencia (gratis)
- Mayor adaptabilidad a las necesidades de la demanda
- Facilidad de difusión, ya que el costo de adquisición es 0
- Un Desarrollador debiera preferir SIEMPRE una herramienta FOS



Las Claves del Exito Libre y Gratuito FREE

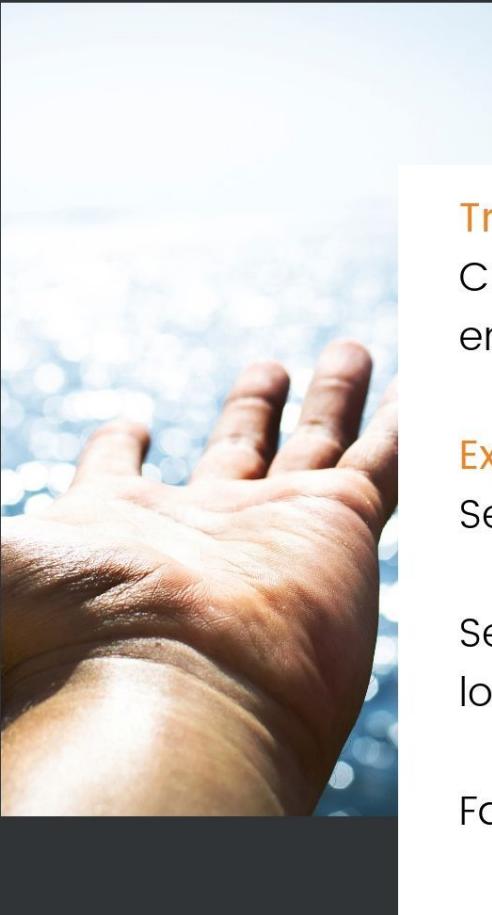
El precio es 0 (cero)!

En un mercado de competencia perfecta, el precio de los productos debe tender a su costo marginal. Los bienes digitales tiene un costo marginal de 0 (cero)!

Cualquier herramienta paga debiera ser desplazada por una gratuita, si las prestaciones son similares.

Fácil difusión de la herramienta, no requiere:

- Costosos **acuerdos corporativos**
- Distribución ilegal (**piratería**)
- Complejos esquemas de **promoción** gratuita a universidades, escuelas, fundaciones, ministerios, etc.



Las Claves del Exito Código Abierto OPEN SOURCE

Transparencia.

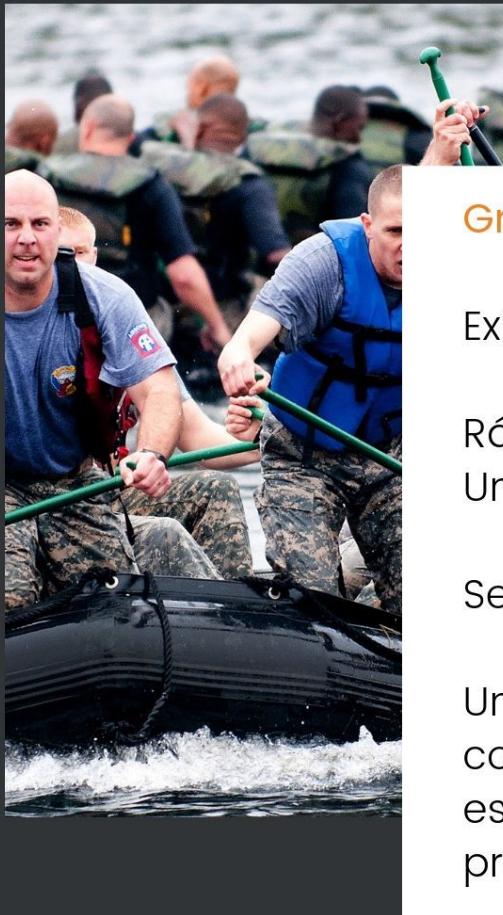
Cualquiera puede saber que está haciendo la herramienta, en cualquier situación.

Expandible y Mejorable.

Se la puede adaptar a requerimientos específicos.

Se la puede integrar en desarrollos comerciales, reduciendo los costos de operación y aumentando el beneficio.

Facilita la integración en otros sistemas informáticos.



Las Claves del Exito Comunitario

Gran red de contribuidores al proyecto.

Existencia de **foros** de ayuda y discusión.

Rápidamente asimilable en **ámbitos públicos** (Gobiernos y Universidades).

Sentimiento de **pertenencia** de los colaboradores.

Una comunidad diversa y extendida asegura la contribución de herramientas útiles en **nichos** pequeños y específicos que NO pueden ser atendidos por software propietario.

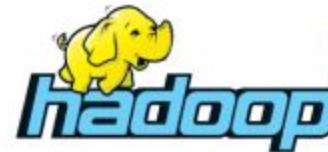
Algunos Ejemplos



Linux



ANDROID



TensorFlow



mongoDB



git



BitTorrent™





Y, Entonces... ¿Dónde está el negocio?

Modelo Freemium (No es el caso de R ni de Linux)

Potencial de Monetización:

- Servicios de Implementación e Integración
- Soporte a Empresas
- Reducción del Riesgo
- Capacitación y Difusión
- Desarrollos a medida
- Creación de Fundaciones sin fines de lucro

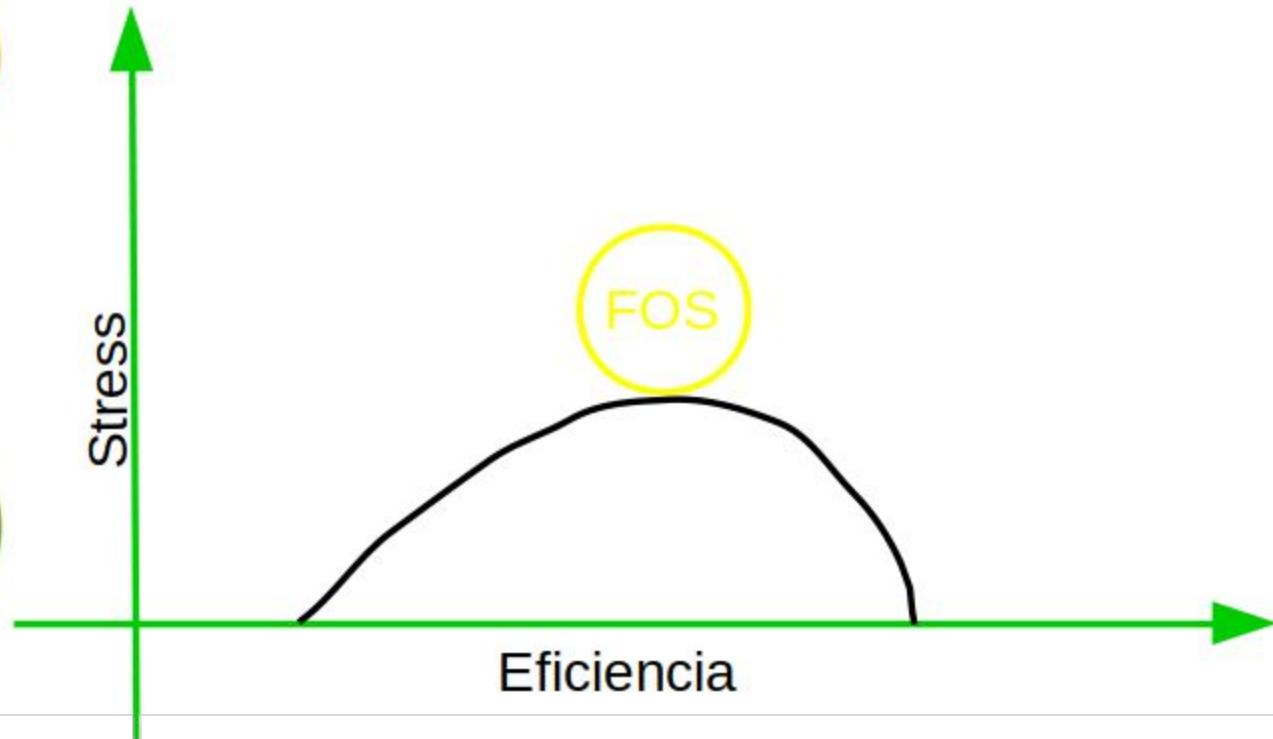
Apoyo de corporaciones que utilizan
y aprovechan la herramienta.

¿Es suficiente?

¿Que podemos esperar en el futuro?

Expertos
trabajando gratis
Empresas perdiendo
mercados

Expertos
remunerados
Empresas conservando
mercados



¿A dónde podemos llegar con esta materia?

Análisis de las Películas de los Últimos 50 Años

Internet Movie Database	
IMDb	
Información general	
Dominio IMDb (en inglés)	
Tipo	Base de datos cinematográfica Compilador de reseñas Base de datos de videojuegos Catalogación social Television series database Directorio de podcasts
Comercial	Sí
Registro	El registro es opcional para miembros para participar en discusiones, comentarios, calificaciones y votaciones, incluyendo acceso a listados de películas, catálogos y horarios ¹
Idiomas disponibles Inglés	
En español	No
Estado actual	Activo
Gestión	
Desarrollador	Col Needham
Propietario	IMDb.com, Inc.
Lanzamiento	17 de octubre de 1990
Estadísticas	
Ranking Alexa	▲ 63º (8 de marzo de 2021) ³



2

Datos + geo info geográfica

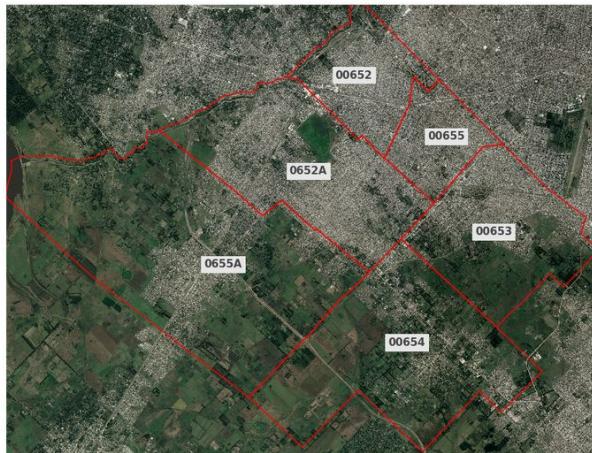
Merlo 

Percentil ingresos por circuito:

Ingresos en pesos corrientes de Sep-2018. Quantiles 25%, 75% y mediana.

distrito	ingreso 25% (\$)	ingreso medio (\$)	ingreso 75% (\$)	electores (cant)	CAMBIEMOS (%)	FPV / UC (%)	PJ no CFK (%)	Resto (%)
00655A	6090	9870	14210	54400	22.6	54.1	15.2	8.1
00654	5510	9930	14210	31900	21.1	53	14.6	11.3
00652A	6400	10260	14750	139000	24.6	51.1	16.4	8
00653	6310	10330	14960	68000	22.7	51.9	15.5	9.9
00652	7570	12210	18000	67700	36.6	35.3	16.3	11.7
00655	8140	12720	18100	53200	39.7	35.7	16.4	8.2

Mapa:



Numeros de votos. 2015-17:

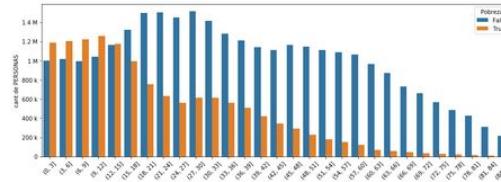
votos por circuito. (distritos con mas de 5000 electores)

Dashboards

codigo -> html



Pobreza por grupo etario



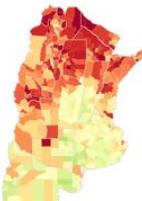
Ingresos, pobreza e indigencia por niveles educativos

Bachiller	Mediana Ingresos	Pobreza (%)	Indigencia (%)
No asistió	108000.0	25.8	3.8
Primaria	114000.0	29.2	5.0
P. completa	128000.0	27.4	4.0
Secundaria	158000.0	26.0	3.4
S. completa	197000.0	14.2	1.6
Terciario	19700.0	11.6	1.4
T. completo	240000.0	4.4	0.4
Universidad	238000.0	5.4	0.8
U. completa	356000.0	1.2	0.2
Postgrado	338000.0	1.2	0.2
P. completo	509000.0	0.4	0.2

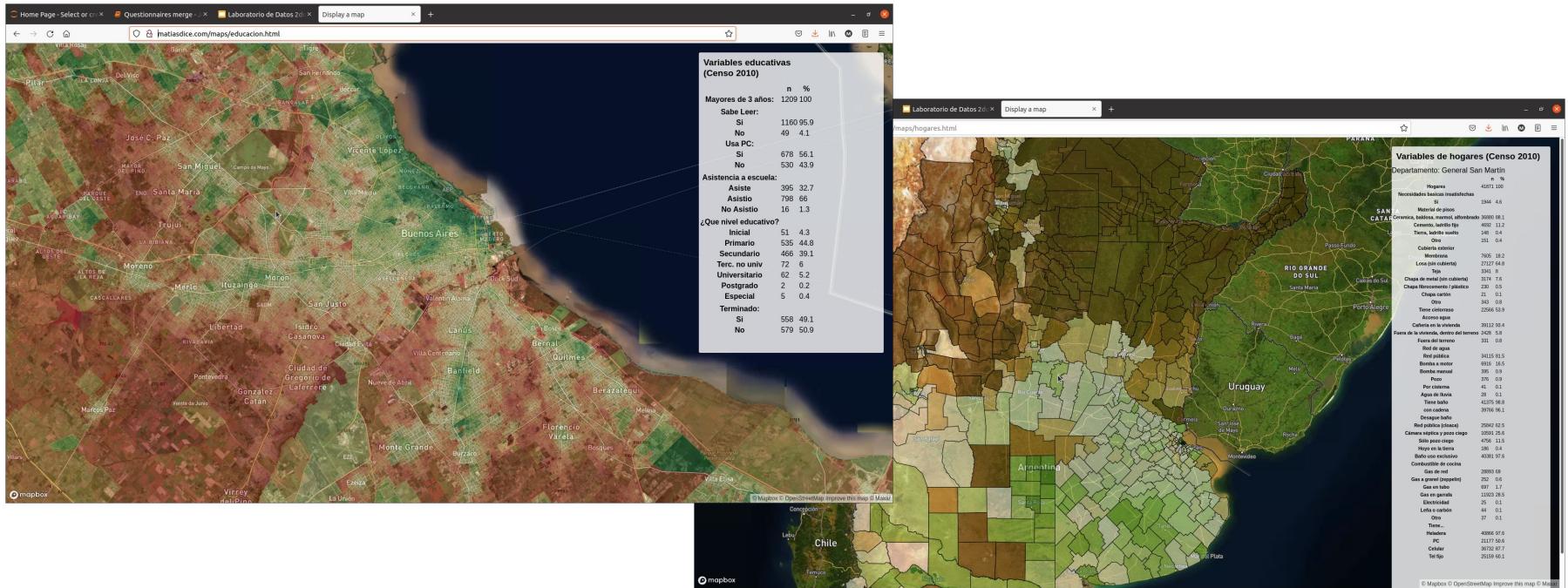
*Mayores de 24 años

Dependencia geográfica

Porcentaje de personas en pobreza por departamento.



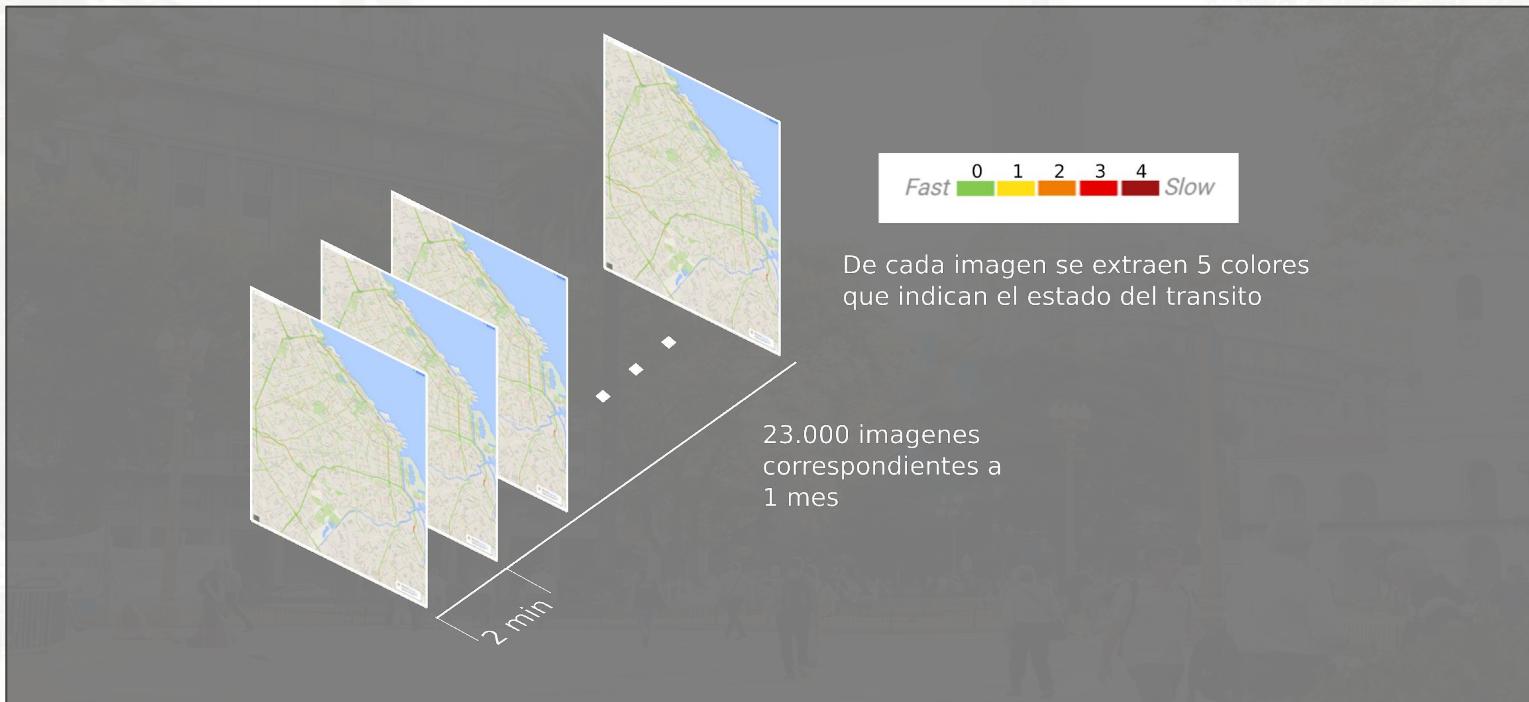
Mapeo interactivo info geográfica + (Mapbox/ArcGIS) + html



3

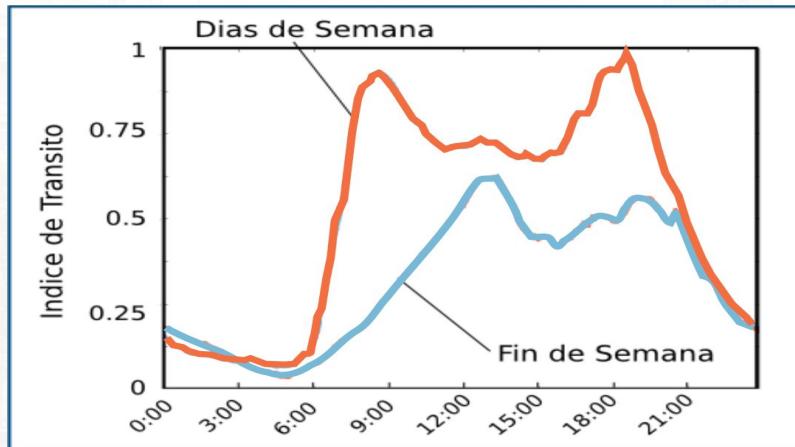
Tránsito

Monitoreo con Google Maps



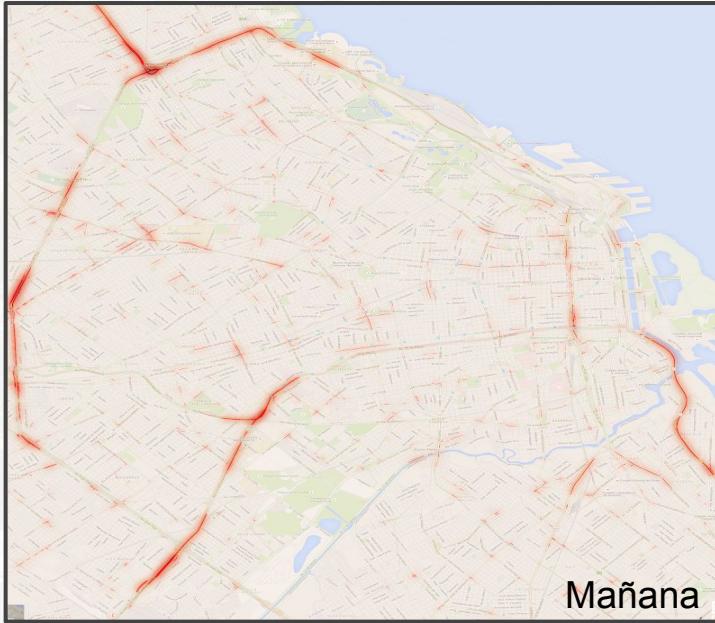
Tránsito

Monitoreo con Google Maps



Tránsito

Monitoreo con Google Maps



Tránsito

Monitoreo con Google Maps



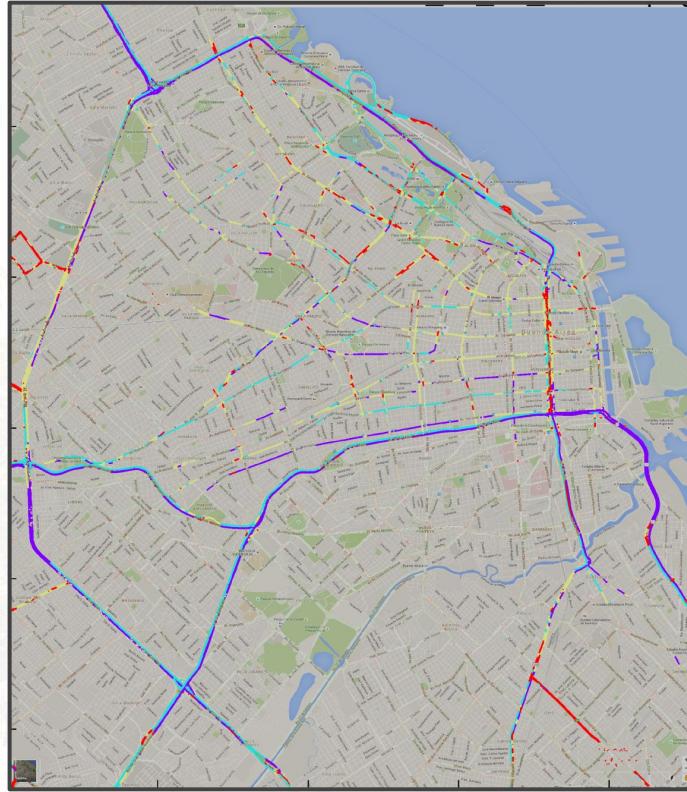
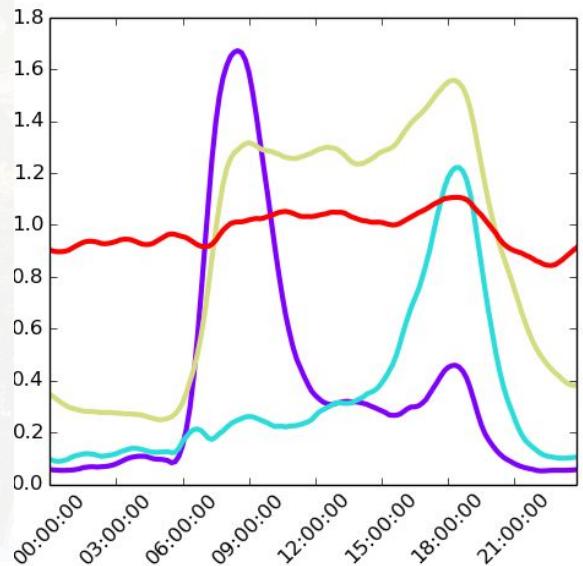
Noche

Clasificación de corredores

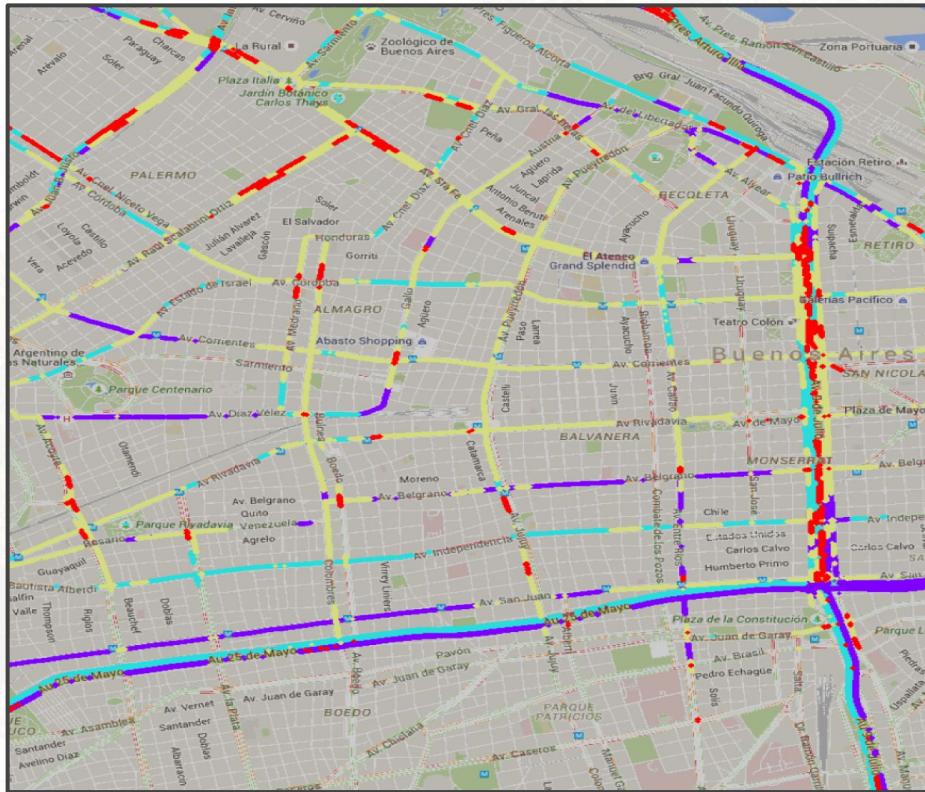
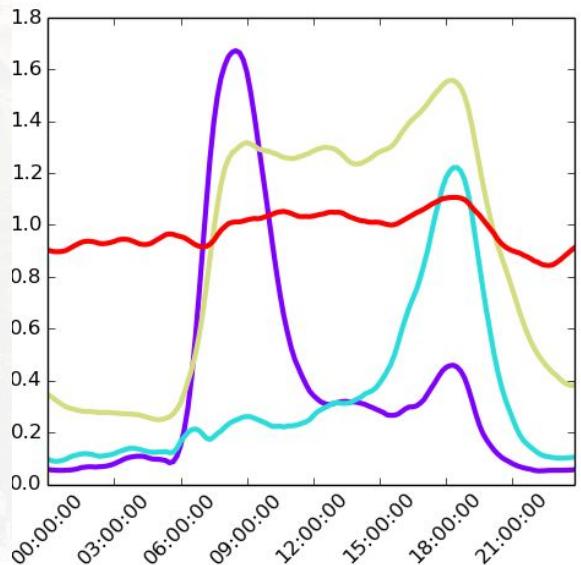
Clasificación de corredores por patrones similares



Clusters



Clusters



Recapitulando...

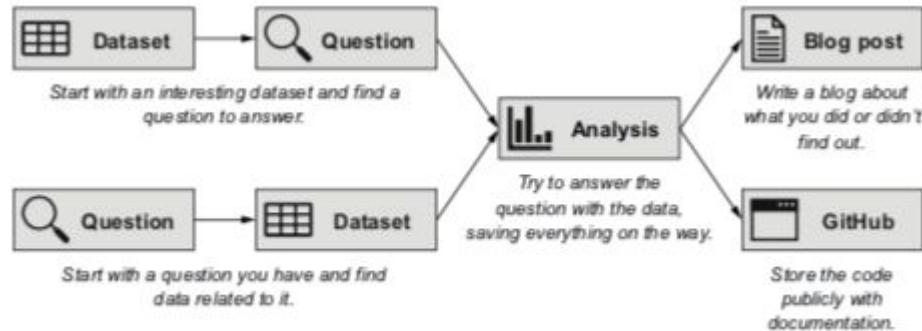
Al terminar la materia van a tener herramientas para empezar un proyecto personal de ciencia de datos.

Elegir una pregunta que les interese!

Conseguir un dataset

Explorar/Aprender de los datos

Comunicar / visualizar



Fuentes de datos:

- APIs (NYT, Yelp, Twitter, ...) algunos con paquetes de R
- WebScraping
- Noticias: FiveThirtyEight.com (encuestas USA, deportes, ...)
- Datos gubernamentales: <https://datos.gob.ar/>
- Datos propios