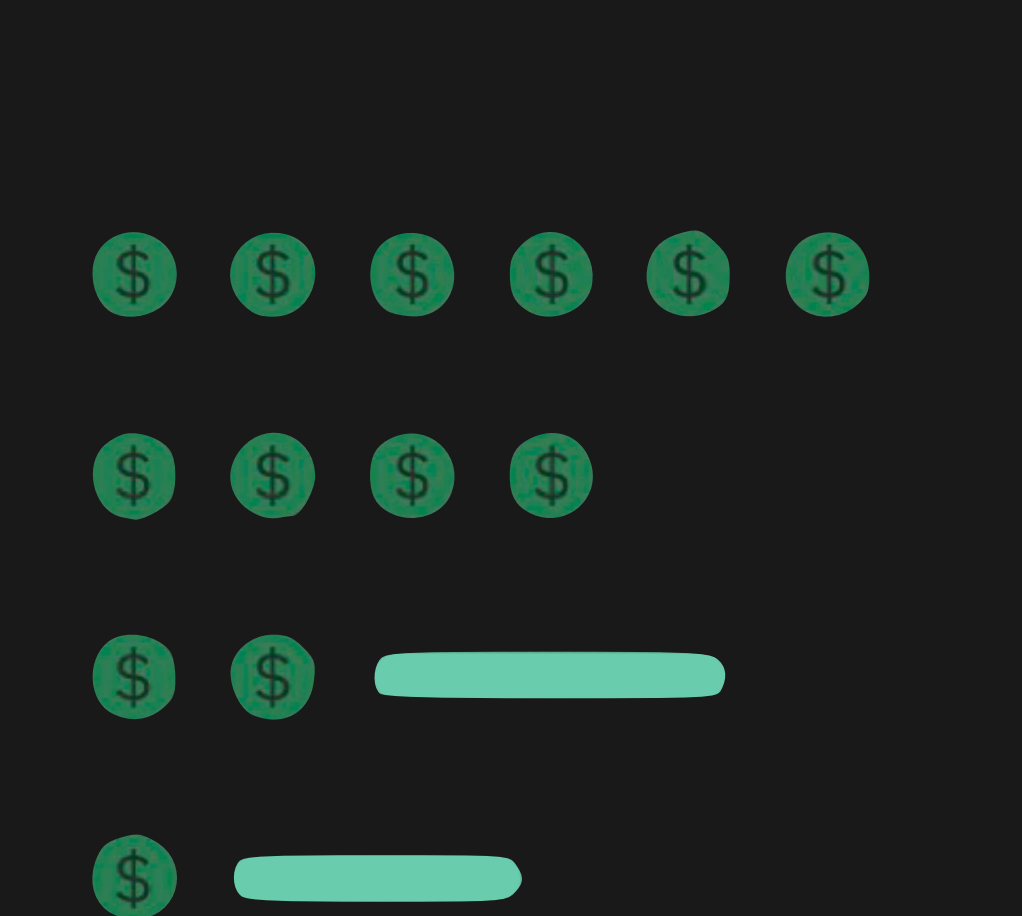


# Corrupción en Chile

# Analizando datos de corrupción



# Bastián Olea Herrera

# Visualizador de corrupción en Chile

- [https://github.com/bastianoolea/corrupcion\\_chile](https://github.com/bastianoolea/corrupcion_chile)

- **Metodología**

- Casos de corrupción donde se involucren recursos públicos

- **Características**

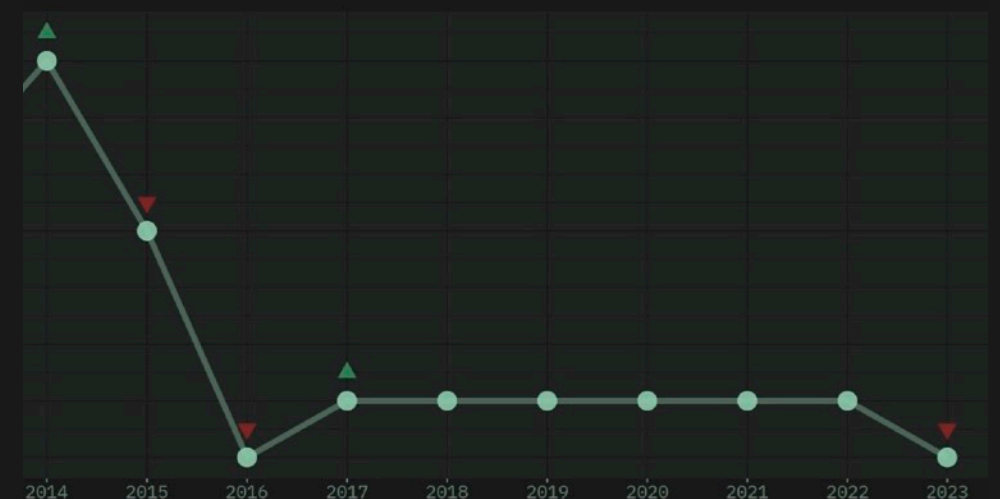
- Serie de gráficos actualizados con datos más recientes
- Énfasis en política



dos usando el *Graficador CEP*, aplicación web para el Centro de Estudios Públicos diseñada y rrera como parte del equipo DataUC.

la **corrupción** de *Transparency International* es un ranking de 180 países sobre corrupción en el sector público, que usa una escala de 0 a 100, donde 0 es

respondiente al año 2023, el índice indica que la **corrupción en Chile aumentó**,



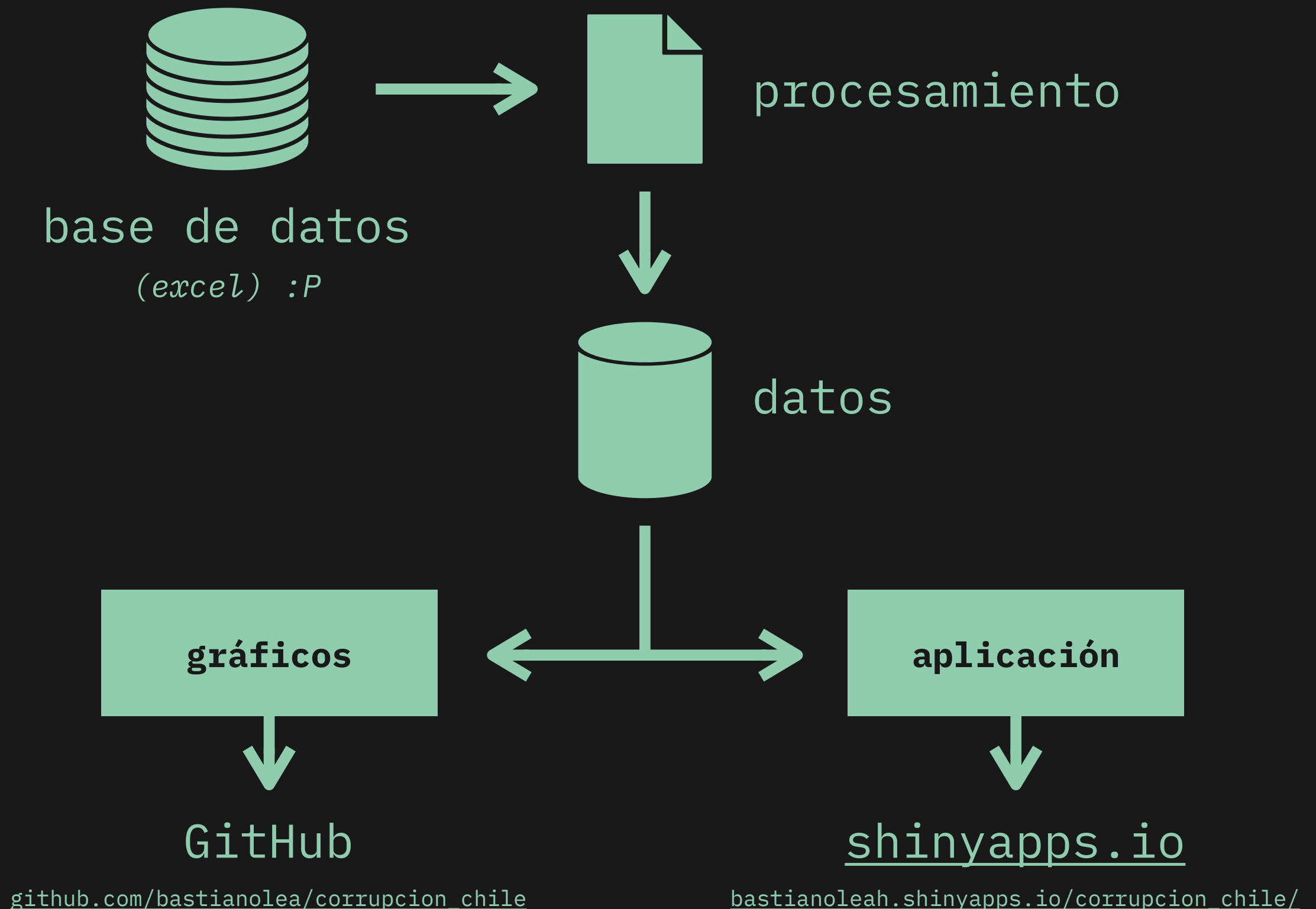
# Stack

## *Tecnologías usadas en la aplicación*

- Aplicación programada con el lenguaje **R**
- Framework de aplicaciones web **Shiny**
- Librería de visualización de datos **{ggplot2}**
- Hosting en la nube [shinyapps.io](https://shinyapps.io)



# Flujo *Proceso de despliegue*



# Transparencia

## *Perspectiva ante una realidad*

- Datos abiertos
- Código abierto
- Tecnologías gratuitas
- Participación social
- Exponer información ofuscada

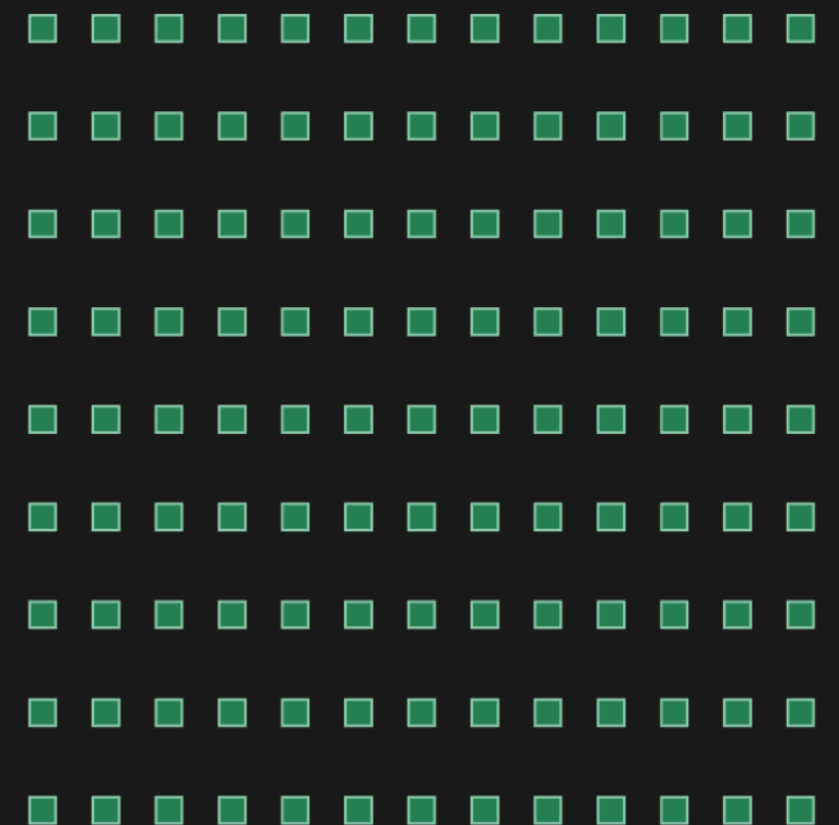
toneladas de palta



El monto de \$3,1  
*Alcaldesa de Mai*  
aproximadamente  
**palta**

precio unitario de re

En esta visualización, **cada punto**  
comprado con la suma de dinero de



# Web scraping

## Extracción de datos desde páginas web

```
E html>
ng="es-ES" dir="ltr"> Desplazar
...</head>
.
/ class="grid-container home"> grid
header class="site-header pt4 pb2 mb4 bb b--transparent ph5 headroom z
d top headroom--not-bottom" role="banner">...</header>
main class="page-main pa4" role="main">
<section class="page-content mw9 center">
  ▼ <div class="flex-l items-center" style="flex-direction: row-reverse
    ▼ <div class="mh4 w-50-l ">
      <h1 class="f2 f1-m f-subheadline-l fw5-ns mv4 lh-solid">Bastián
      </h1> = $0
      <h2 class="f5 fw7 mt0 mb4 ttu tracked">Sociólogo, analista de d
      desarrollador R</h2>
      ▼ <div class="social-icon-links" aria-hidden="true">
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        bastianolea" title="github" target="_blank" rel="me noopener">...
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        profile/bastianolea.rbind.io" title="bluesky" target="_blank" r
        noopener">...</a>
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        bastimapache" title="twitter" target="_blank" rel="me noopener"
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        www.linkedin.com/in/bastianolea/" title="linkedin" target="_bla
        noopener">...</a>
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        www.tiktok.com/@bastimapache" title="tiktok" target="_blank" re
        noopener">...</a>
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        www.instagram.com/raccunnie" title="instagram" target="_blank"
        noopener">...</a>
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="https://
        www.goodreads.com/user/show/53224910-basti-n-olea-herrera" titl
        target="_blank" rel="me noopener">...</a>
        ▶ <a class="link dib h1 w1 ml0 mr2 f6 o-90 glow" href="/contact
        "envelope">...</a>
      </div>
      ▼ <p class="f4 mt4 lh-copy">
        "Este sitio contiene todo tipo de recursos sobre programaciór
        lenguaje R para análisis de datos, con un foco en las ciencia
        En el blog comparto "
        <a href="/blog/">consejos, novedades</a>
        " y "
        <a href="/categories/tutoriales/">tutoriales</a>
        " de R para que otras personas aprendan a programar en este l
      </p>
      <a class="mt4 action text" href="/blog/">blog ></a>
      <a class="mt4 action text" href="https://bastianolea.github.io/
      aplicaciones de datos sociales ></a>
      <a class="mt4 action text" href="https://bastianolea.github.io/
```



# ¿Qué es el web scraping?

*Extracción de datos desde sitios web*

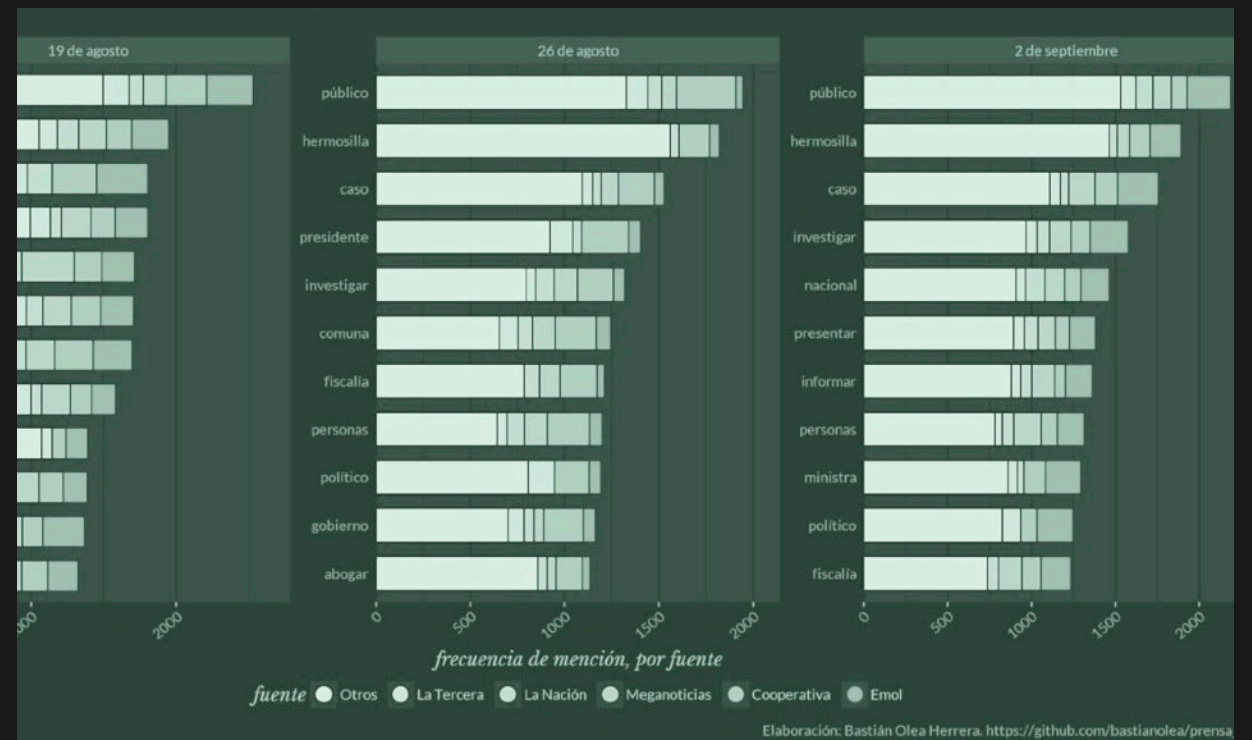
- Se utiliza un programa para obtener el código fuente de un sitio web
- Se usan selectores para identificar elementos de un sitio web
- Extracción de datos y organización tabular de los mismos
- Análisis de los datos

> *Ejemplo práctico*



# Scraping de prensa chilena

## Aplicación de análisis de datos de texto de prensa chilena



### Menciones de un concepto específico

#### hermosilla

iones de una palabra, por semana, y separado por medios de comunicación. Seleccione un concepto, palabra, o nombre para comparar o elegido entre los distintos medios de comunicación escritos. De este modo, es posible identificar si hay ciertos medios que mencionan, o medios que los evitan; o bien, la popularidad de un concepto a través del tiempo, comparada a entre distintos medios.

Comparar

selecciona en el rango de semanas

Cantidad de medios

Destacar un medio

1

5

8

3

10

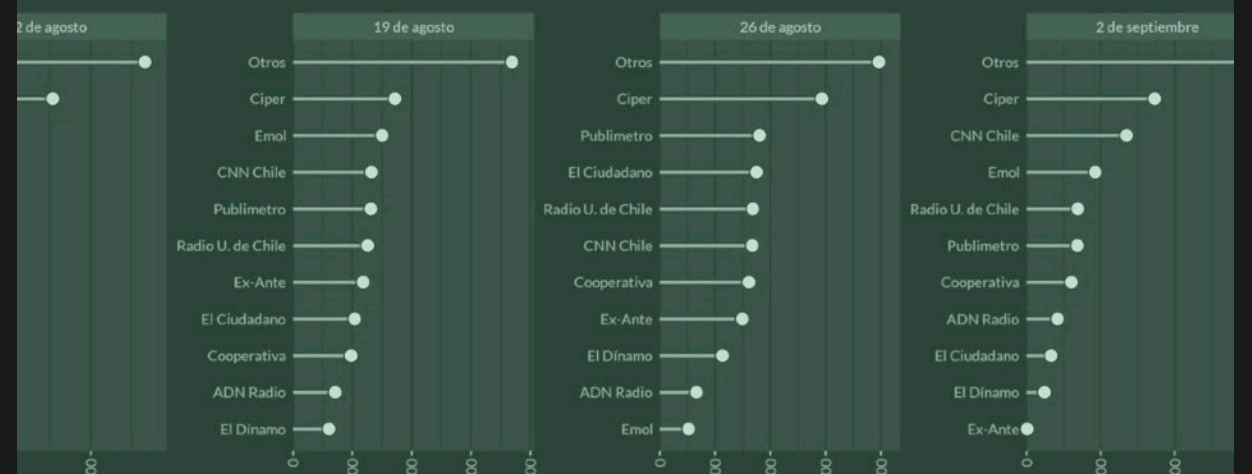
15

Ninguno

Personalice el rango de tiempo que abarcará la visualización. Por defecto, si el rango es muy amplio, se cambia a barras.

Cantidad de medios comunicacionales a identificar. Se mostrarán los nombres de las n fuentes con mayor cantidad de palabras. El resto se agrupará en "Otros".

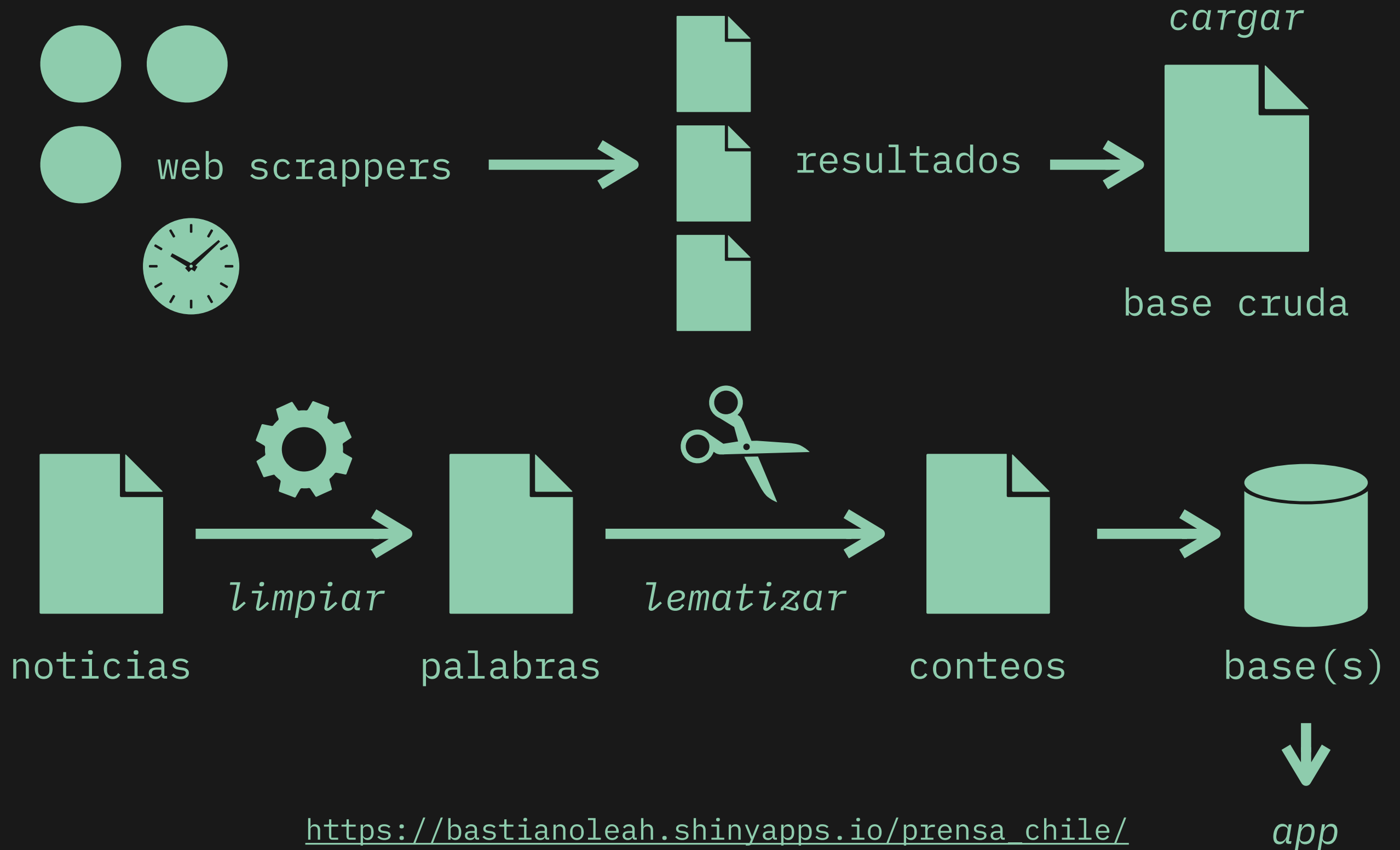
Seleccione un medio de comunicación para destacarlo en el gráfico por sobre el resto de los medios disponibles.



# Scraping de prensa chilena

- Web scraping y análisis de texto sobre un corpus de texto de noticias de la prensa chilena, con nuevos datos obtenidos diariamente, y app actualizada semanalmente
- [https://github.com/bastianoolea/prensa\\_chile](https://github.com/bastianoolea/prensa_chile)

# Flujo *obtención y procesamiento*



# Procesamiento de texto

- Unión y limpieza de noticias
- Tokenizar texto y conteo de frecuencias
- Legalización y conteo de palabras lematizadas
- Cálculo de correlación entre términos
- Modelamiento de tópicos
- Detección de sentimiento
- Generación de resúmenes de noticias
- Actualización de aplicación

# Modelamiento de tópicos

## *Detección de temas emergentes*

- Modelamiento estructural de tópicos (STM)
- Busca modelar tópicos en base a covariantes a nivel de documento
- Puede incluir metadatos
- Permite obtener tópicos latentes desde los datos
- Obtiene el número de tópicos ( $K$ ) que pueden existir en el texto probando modelos

- > *Ejemplo de modelamiento de datos*
- > *Ejemplo de análisis de texto*