Investigación Reproducible

José Reyes-Valdés

José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Contenido

- Objetivos y pertinencia
- ¿qué es Investigación Reproducible?
- Herramientas
- Ejemplo (integración)

Objetivos y pertinencia

José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

- **Objetivos**.
  - Introducir los conceptos OpenData y OpenScience.
  - Conocer los elementos necesarios para hacer IR.
  - Presentar un enfoque integral de IR.
- **Pertinencia**.
  - El acceso al conocimiento ha tomado una dimensión social y global.
  - Formación de redes de conocimiento es necesaria.
  - Existe un amplio espectro de herramientas informáticas libres poco utilizadas para hacer investigación.

¿Qué es investigación reproducible?

José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Datos abiertos y ciencia abierta

- ***Datos abiertos*** se promueve o exige para realizar investigación o en rendición de cuentas en el servicio público (EOSC, FAIR).
- ***Ciencia abierta (Open Science)*** se refiere, en esencia, a la transformación que la ciencia experimenta debido a la globalización (del conocimiento) (Burgelman et al., 2019).
  - Un repositorio de Ciencia Abierta es Sci-Hub.
  - El artículo referido de Burgelman et al. (2019) se ubica en Open Science.

Investigación reproducible

- Los resultados de una investigación son considerados ***reproducibles***, si se tiene disponible la información suficiente para que otros investigadores, siguiendo los mismos procedimientos, arriben a los mismos hallazgos utilizando nuevos datos (Grandrud, 2015).
- En ciencias computacionales, investigación reproducible significa que con los mismos datos y el mismo código utilizados para llegar a ciertos hallazgos están disponibles, y son suficientes, para que otro investigador pueda recrear estos hallazgos.

Investigación reproducible

- Las herramientas para hacer investigación reproducible tienen un impacto significativo directo en las actividades de investigación.
  - Almacenamiento y acceso a datos (MySQL).
  - Procesamiento de datos (R, Julia, Python).
  - Elaboración de presentaciones, reportes y artículos (R, RStudio, RMarkdown, Quarto, LaTeX).
  - Entornos web de visualización (Frontend) y procesamiento (Backend). (Julia-Genie).

Ejemplo (1/3)

- En una producción cables de acero para acelerador de motocicleta, se especifica que la resistencia media es de $70kg$.
- Se tiene un lote de 5,100 cables.
- Control interno de calidad utiliza una muestra de 35 para verificar si se satisface la especificación.
- Un auditor del comprador realiza también una prueba con 35 cables.
  - Toma la muestra del mismo lote.

Ejemplo (2/3)

Para garantizar la reproducibilidad, se facilita el código utilizado para la estimación de la media y la construcción del intervalo de confianza. Mediante lenguaje *R*, tanto el área de control como el auditor utilizan el mismo código, pero con distintas muestras.

Código:

```
1 data.smpl.1 <- sample(data.src.1, n.smpl)
2 test.1 <- t.test(data.smpl.1, mu = mu.src)
3
4 data.smpl.2 <- sample(data.src.1, n.smpl)
5 test.2 <- t.test(data.smpl.2, mu = mu.src)
```

Ejemplo (3/3)

```
    One Sample t-test

data:  data.smpl.1
t = -0.80285, df = 34, p-value = 0.4276
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 67.64392 71.02167
sample estimates:
mean of x
  69.3328


    One Sample t-test

data:  data.smpl.2
t = -0.57649, df = 34, p-value = 0.5681
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 67.69398 71.28683
sample estimates:
```

Herramientas

José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Sistema operativo

- Los tres principales sistemas operativos son.
  - Windows (comercial). Demanda gran cantidad de recursos. Es vulnerable a ataques al sistema (virus).
  - MacOS (comercial). Estable y robusto (integración hardware-software). Costoso, requiere experiencia para configuraciones. (obsolescencia programada).
  - Linux (gratuito). Manejo óptimo de memoria y recursos. Gran cantidad de recursos disponibles. Diversidad de versiones y entorno amigable.

Desarrollo y procesamiento

- General
  - Python, Julia (Python + R + C)
- Estadística y graficación
  - R (Statistica, Stata, SPSS, SAS)
- Cómputo numérico
  - Octave (Matlab)
- Cómputo simbólico
  - WxMaxima (Mathematica)

Gestores de datos

- Existen diversos gestores de datos, libres y comerciales.
- Gestores libres.
  - MySQL. Versión libre aunque actualmente administrada por la empresa Oracle.
  - MariaDB. Un *Fork* creado por los desarrolladores de MySQL.
- Ambos gestores se instalan de manera local o remota en los sistemas operativos principales.

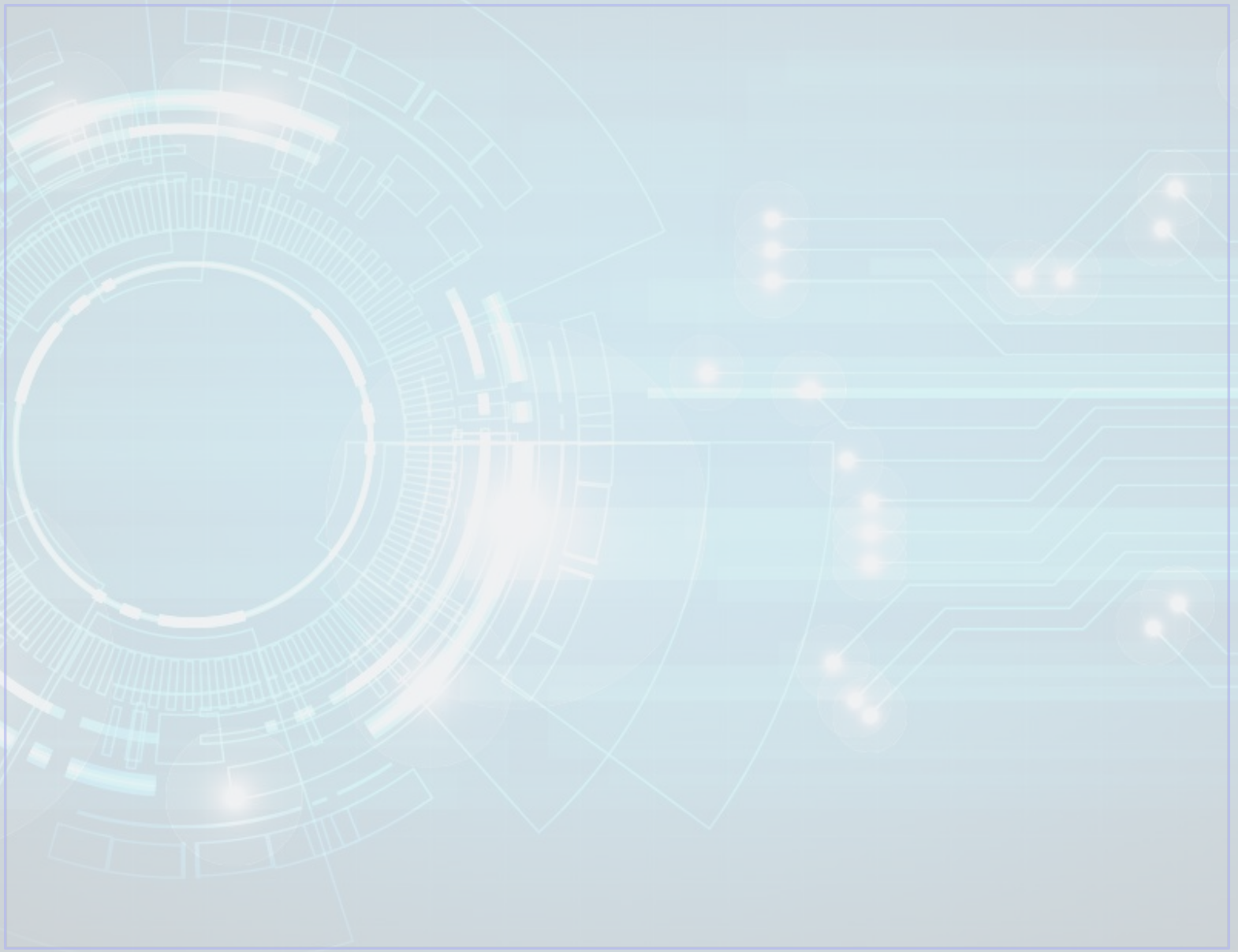
Integración de herramientas

José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Componentes

- Objetivo: crear un documento de notas que genere resultados para compartir y para uso en diversos programas, a través de las heramientas
  - R
  - RStudio
  - MySQL
  - LaTeX
  - JabRef

Estructura de archivos

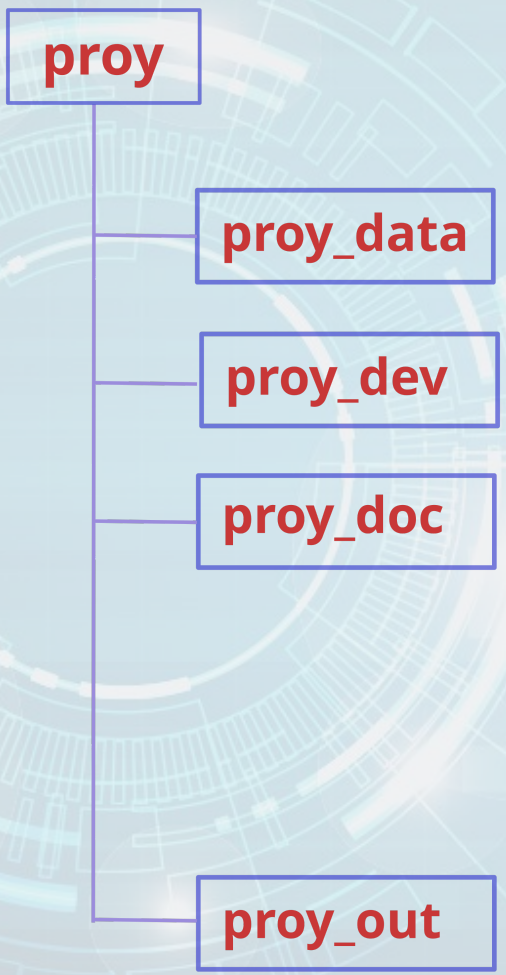


José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos

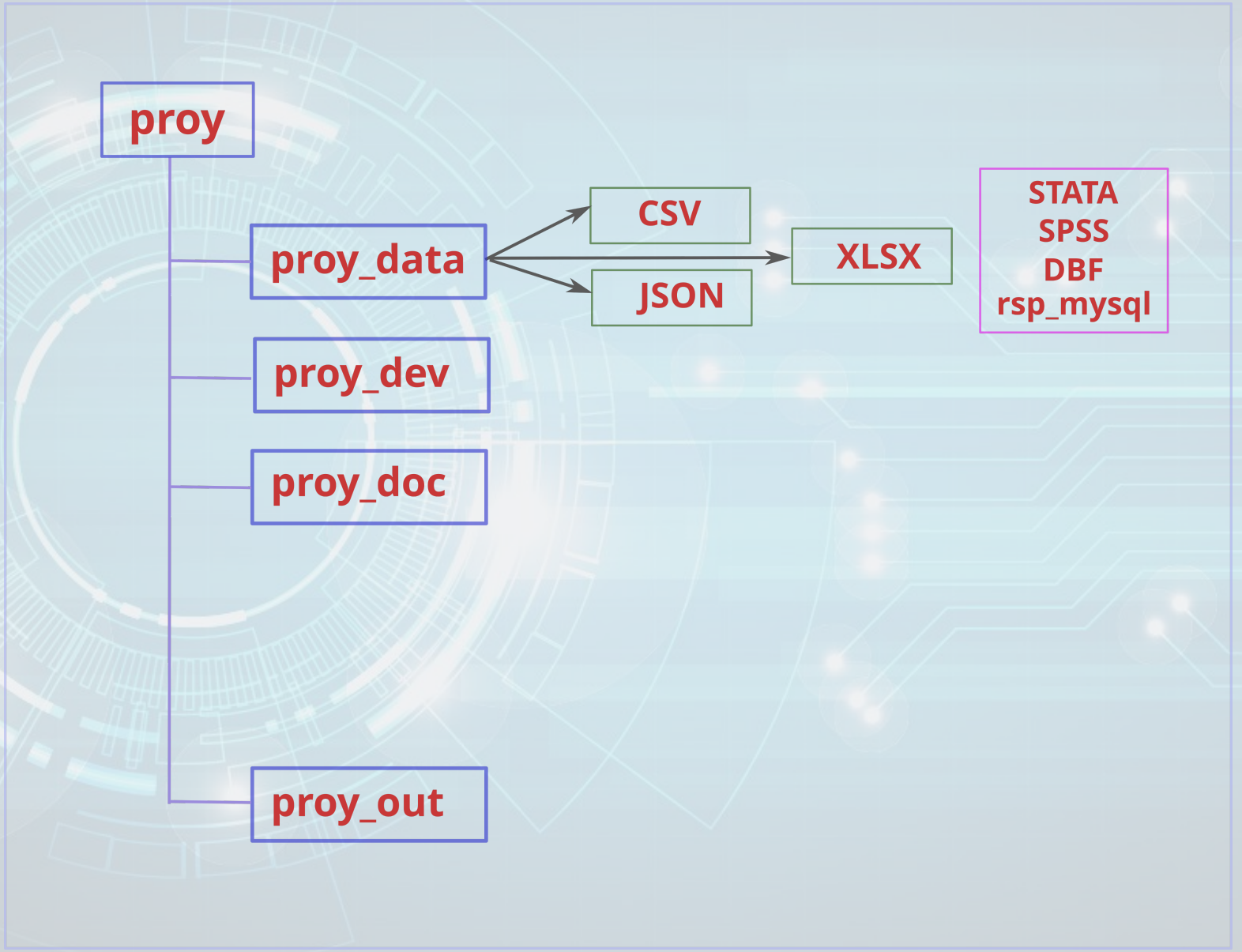
José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos



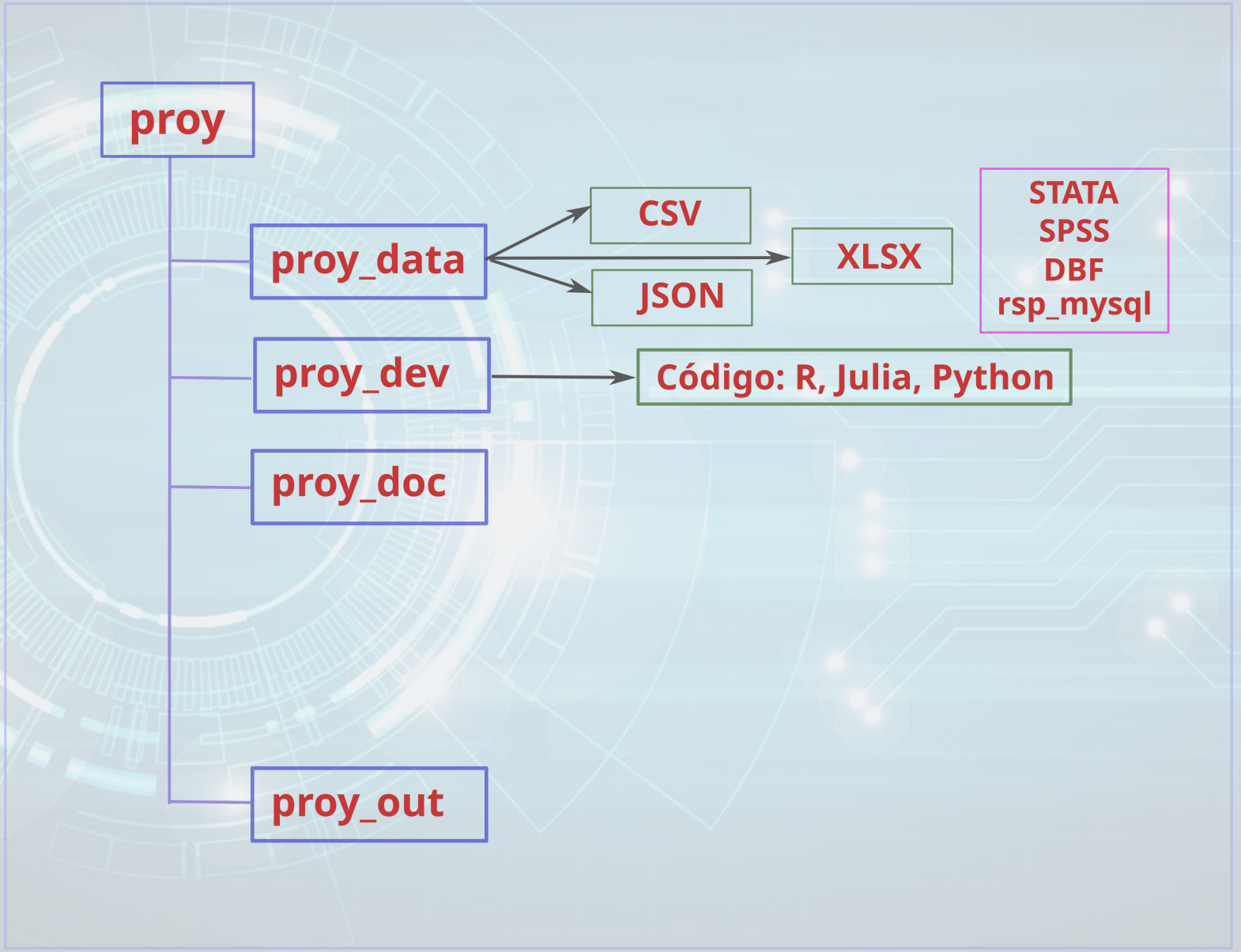
José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos



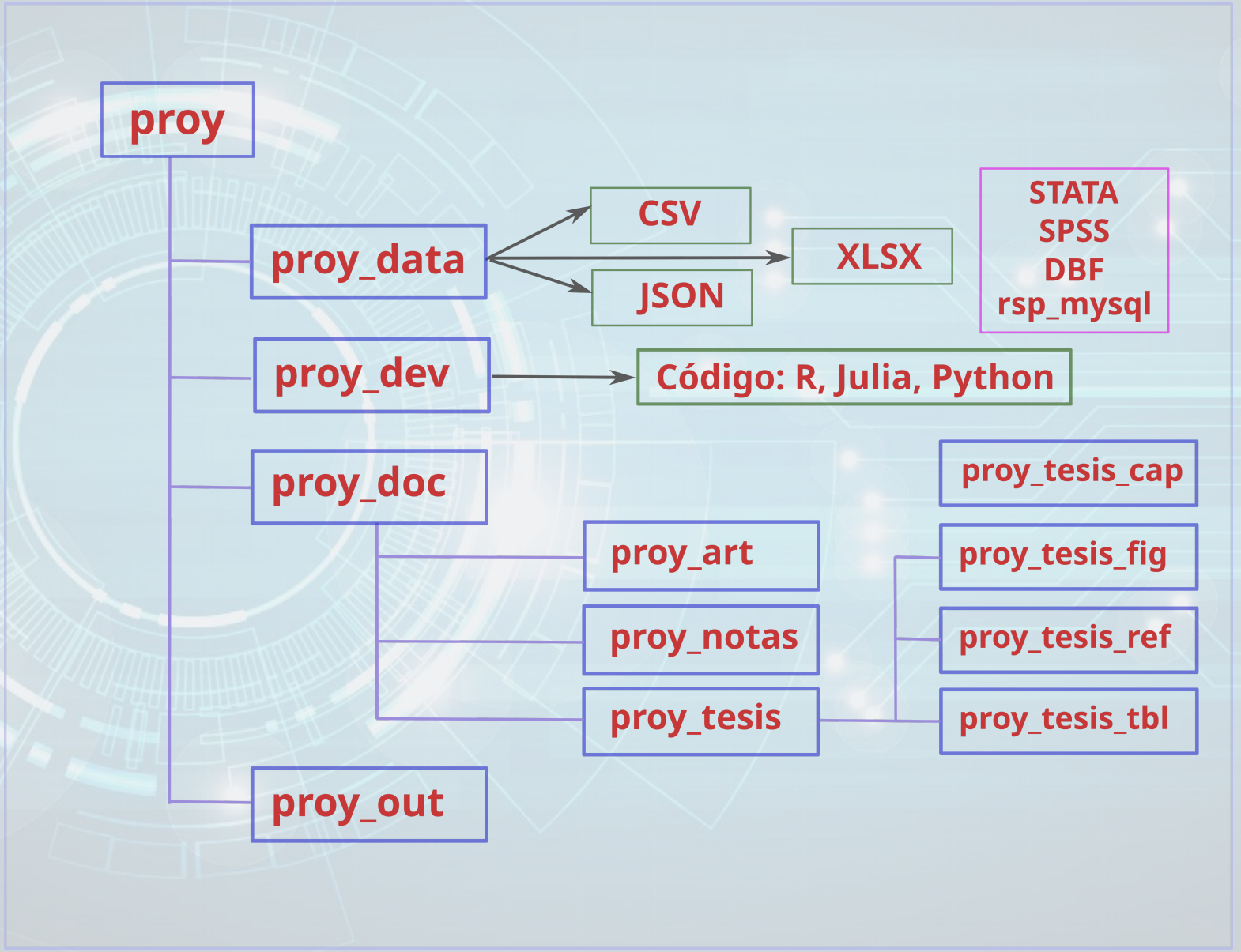
José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos



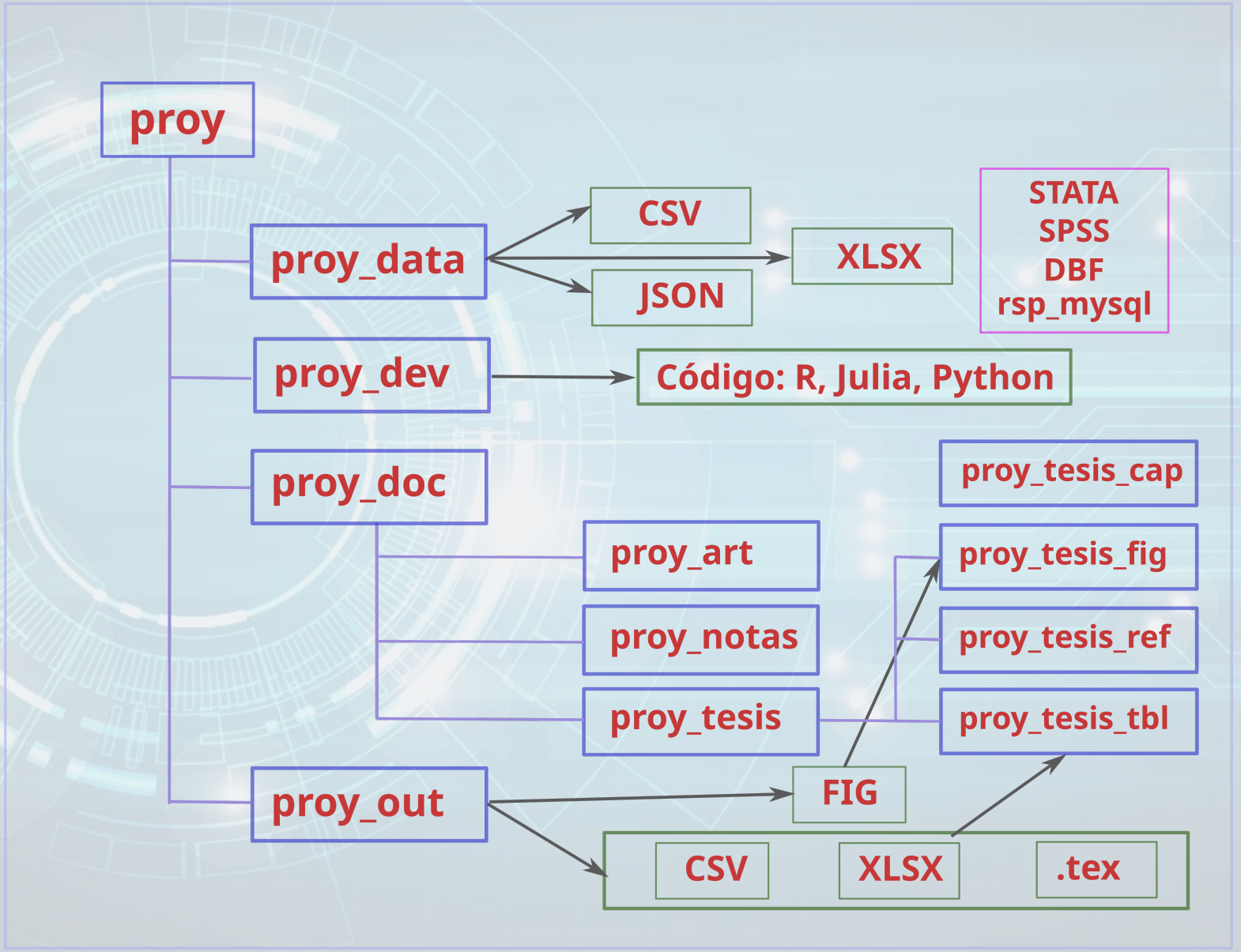
José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos



José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Estructura de archivos



José Reyes-Valdés, CIC40, jose.reyes@cic40.mx

Integración

- Se prepara un documento relacionado con el Índice de Marginación.
- Se hace uso de Inteligencia Artificial para le elaboración del contenido y contexto general.
- Se procesa la información del CENSO 2020 depurada y almacenada previamente en MySQL.
- Se elabora ensayo en formato HTML y en procesador LaTex.

Referencias

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., & Schouppe, M. (2019). Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00043

Grandrud, C. (2015). *Reproducible research with r and RStudio* (J. M. Chambers & T. Hothorn, Eds.). Chapman; Hall/CRC.