# Tres ideas en torno a la reproducibilidad de los procesamientos geodésicos.

Javier José Clavijo

jclavijo@fi.uba.ar

https://github.com/jjclavijo/3ideas2022sirgas

Universidad de Buenos Aires - Facultad de Ingeniería

# Tres ideas en torno a la reproducibilidad de los procesamientos geodésicos.

## Motivación (origen)

- Posiciones, en un MR, $\longleftrightarrow \arg\max_x P(X|\text{datos})$ [1]
- Si resuelvo X por Pedazos (ej: órbitas $\rightarrow X_0$)
- $P(X_1|\text{datos}) = P(X_1|\text{datos}, X_0)$
  - $\mapsto$ condicionar a los parámetros fijos es igual a la marginalizar.
  - $\mapsto$ $X_1$ y $X_0$ son condicionalmente independientes a los datos.
  - No pasa si hay datos compartidos.
- $\mapsto$ Necesito saber si
  $P(X_0, \text{datos}) = P(X_0)P(\text{datos}) \forall \text{modelos}$
- Y necesito sistematizarlo.

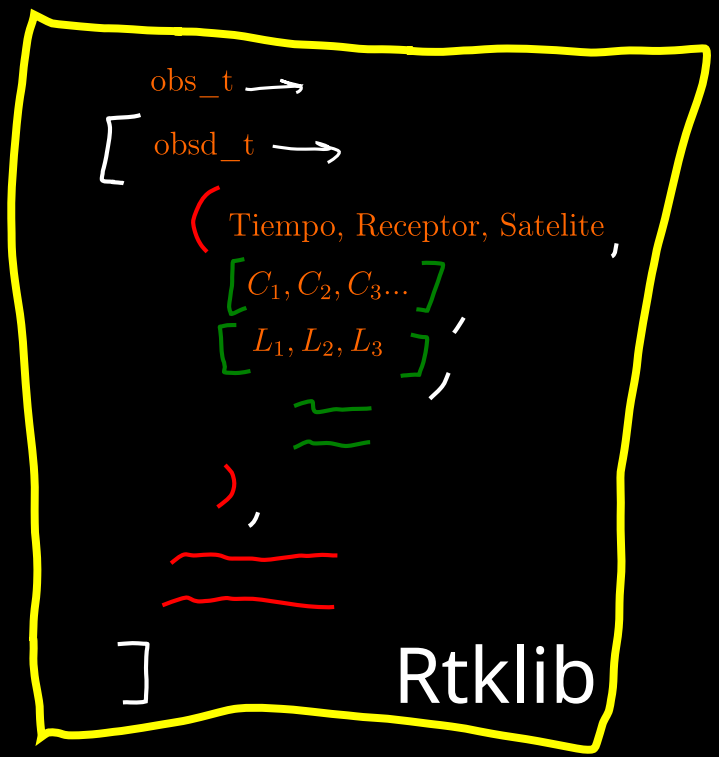
[1] *Marcos de referencia geodésicos: un enfoque Bayesiano; JJ Clavijo, JF Martínez*, en «Seminario de vinculación y transferencia: año 3»**ISBN 978-987-88-4967-6** https://cms.fi.uba.ar/uploads/Libro_SE_Vy_T_2021_VERSION_FINAL_add8815fd9.pdf

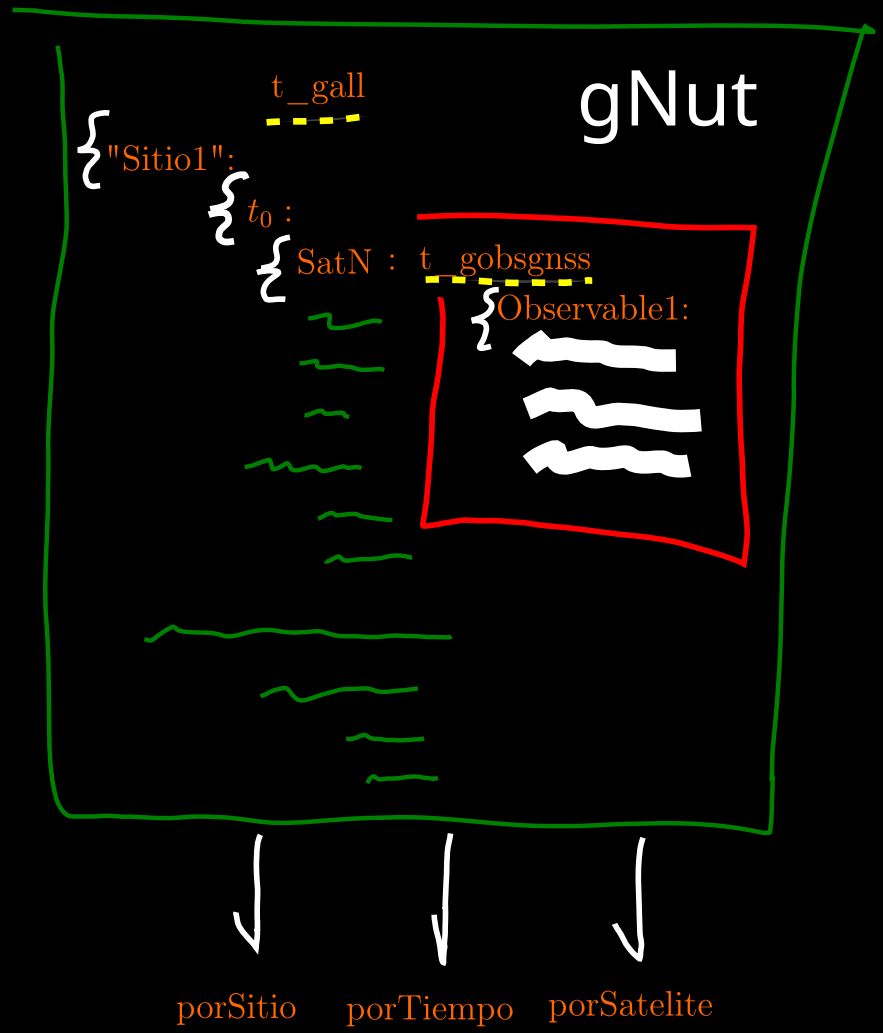
Al analizar trabajos publicados aparece

- Software específico.
- Nomenclatura especial.
- Múltiples Formas de leer y estructurar datos
- Términos y Métodos “heredados”

Puede dificultarse

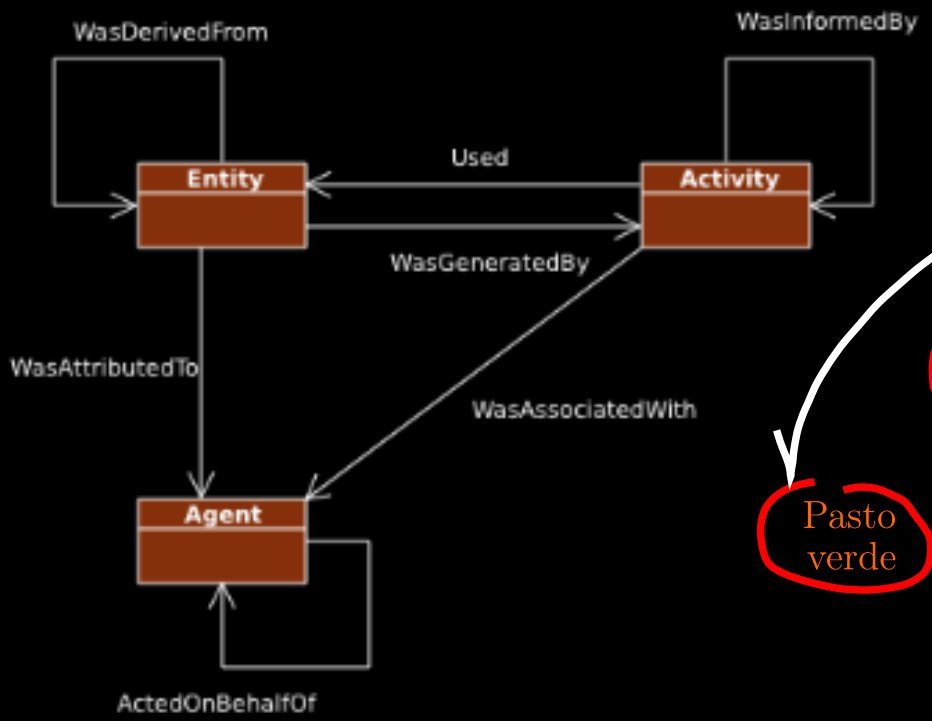
- Aprender por asociación
- Reproducir experimentos
- Validar Resultados

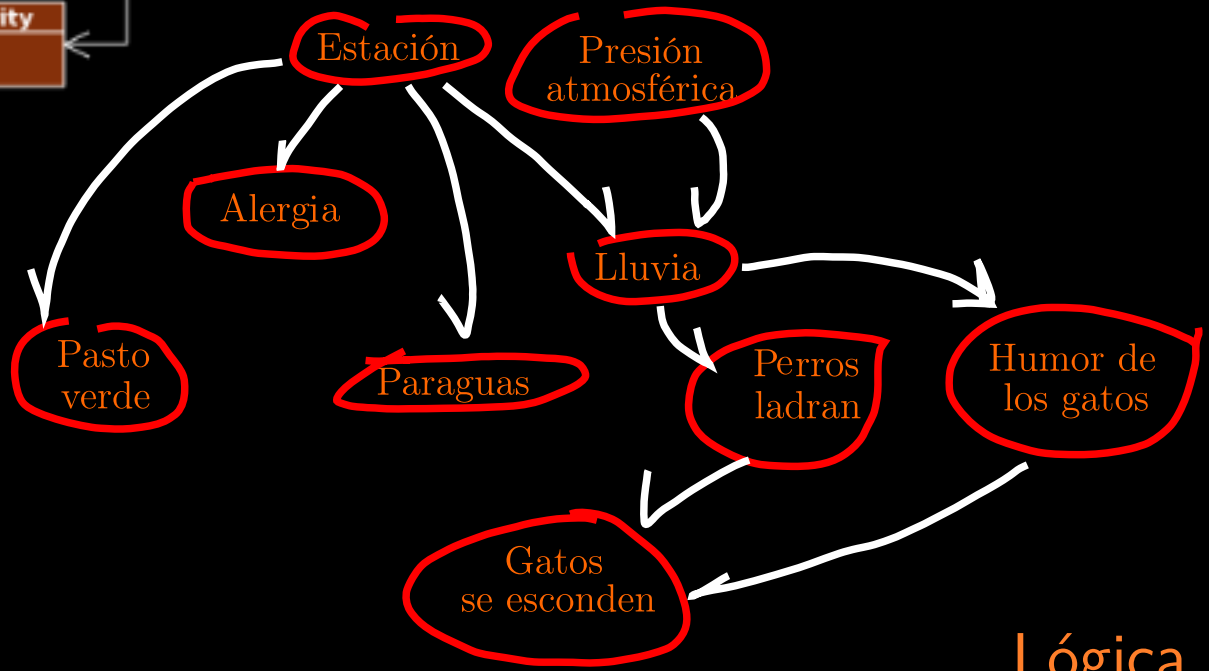


## Premisas

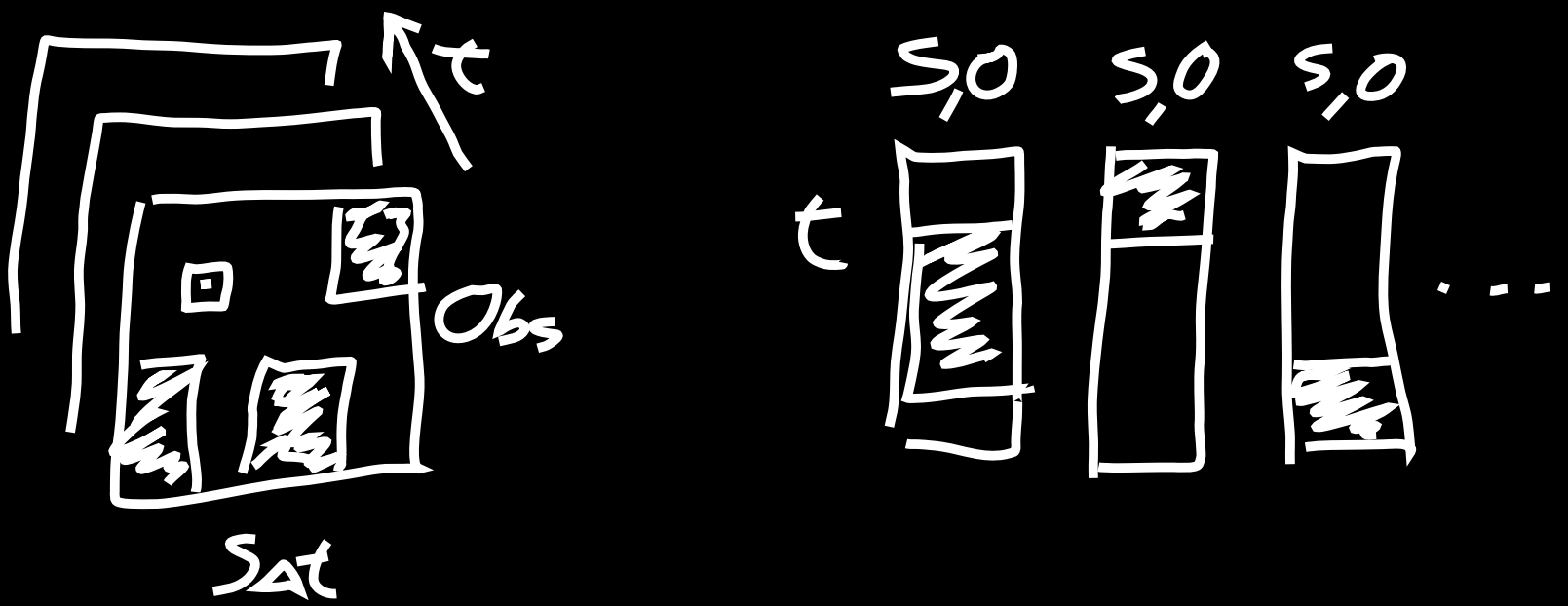
- Formatos estándar de Intercambio $\in$ Geodesia
- Flujos de datos estandarizados $\notin$ Geodesia
- $\exists$ W3C PROV Standard
- $\exists$ Metodos Bayesianos - Redes Bayesianas, etc
- $\exists$ Métodos de validación, Herramientas genéricas de procesamiento y transmisión de datos



## Observando que

- Datos de Observacion $\mapsto$ Tensor Ralo o Columnar Format (índice: t).
- Modelos paramétricos $\mapsto$ Conjuntos de parámetros $\mapsto$ Tensor Denso o Columnar Format (índice: t).
- Resultados $\mapsto \mu, \Sigma$ de una $\mathcal{N}$ (hoy).



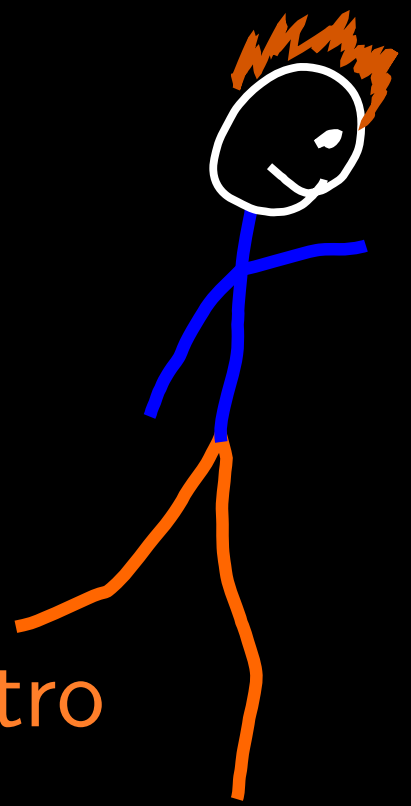
## Lógica

## Podemos plantear

- Segmentación en función del tiempo (con Herramientas Genéricas)
- Validación por bloques (Hash)
- Flujos de trabajo “funcionales” (PROV, Reproducibilidad)
- Identificación y descripción de paramétros y dependencia. (PROV, Redes Bayes)

Bajando

a



nuestro

caso

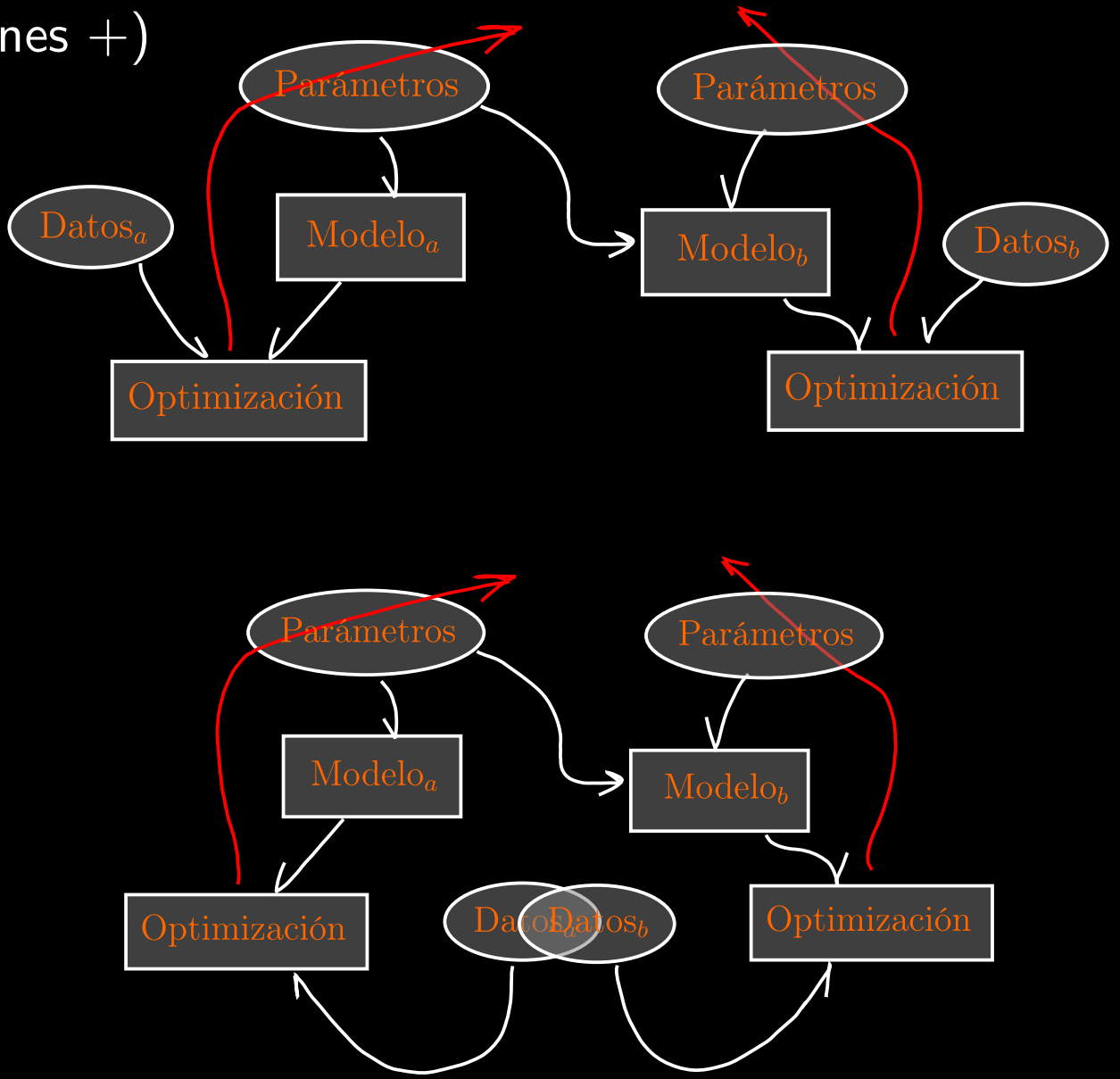
## Datos de entrada

1. $\uparrow\uparrow$ Datos (observaciones $+$)
2. $\uparrow\uparrow$ Procesos
3. $\uparrow\uparrow$ Modelos
   - 3.1 $\uparrow\uparrow$ Parámetros

Provistos

$\neq$

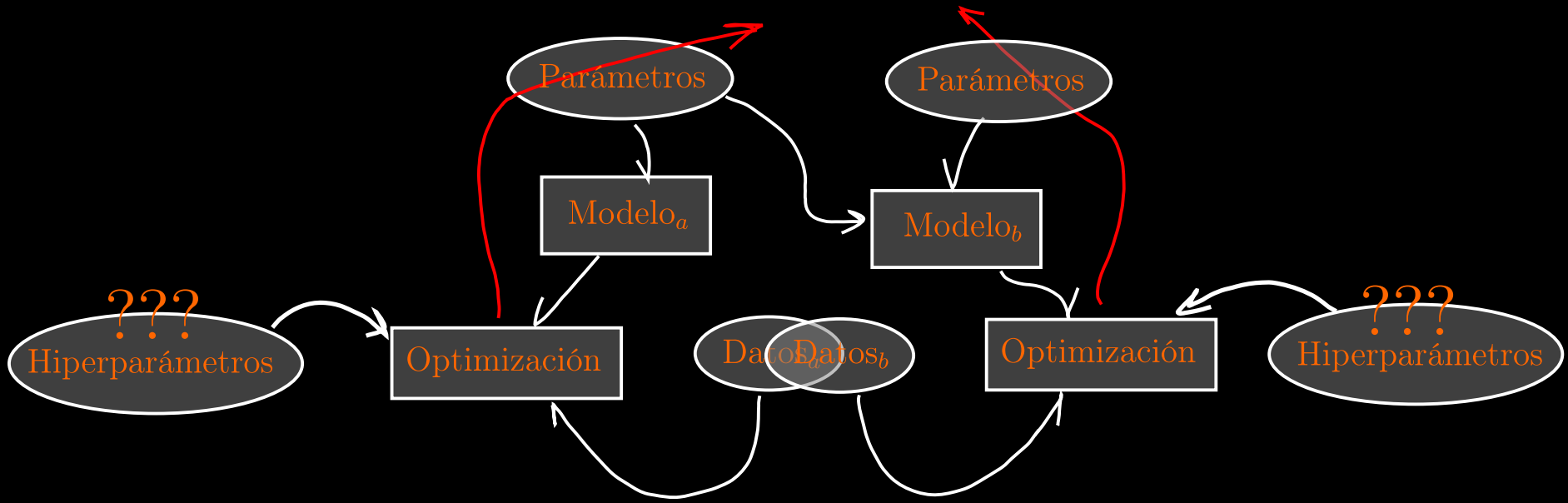
Datos $\rightarrow$
Optimización $\rightarrow$
Ajustados.



## Algoritmos de inferencia

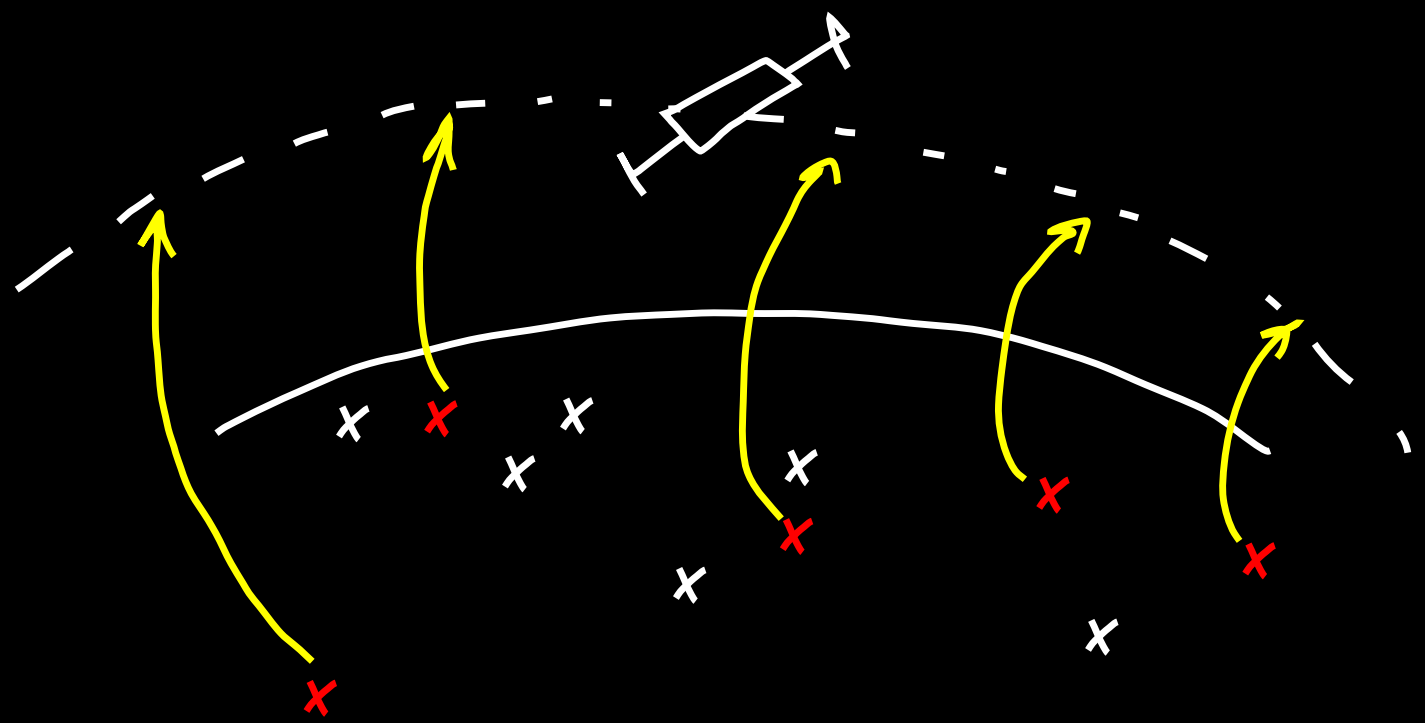
Los métodos que usamos para hallar "resultados", que son valores para los parámetros de un modelo, a partir de "datos".

1. Tienen Hiperparámetros.
2. Pueden variar entre implementaciones (por tener variantes).
3. Pueden tener "imprecisiones" que se deciden en cada implementación.



## Distribución de los datos

1. No es homogénea
2. Difícilmente permite validar con conjuntos independientes
3. Puede haber correlaciones ocultas a través de parámetros "dados por hecho".



# Tres ideas introduciendo

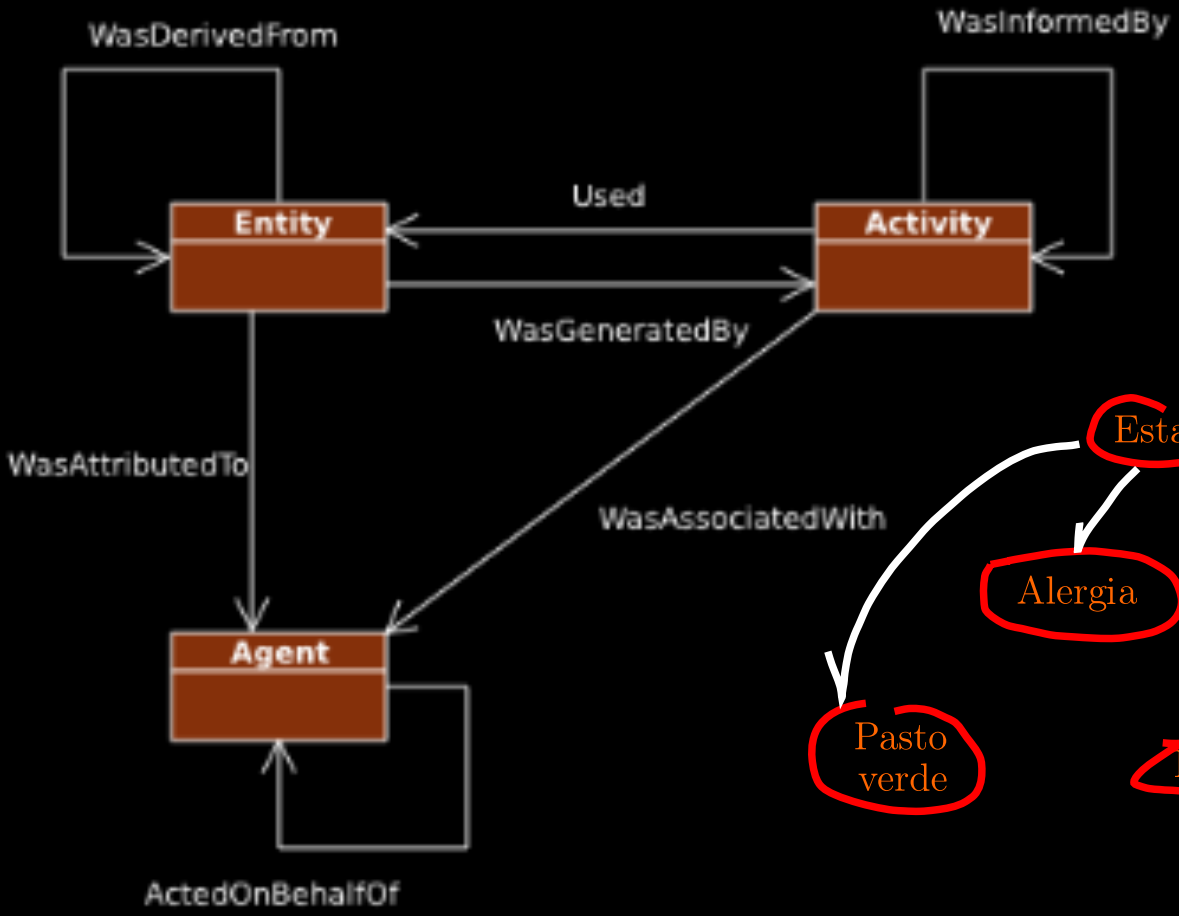
los conceptos de

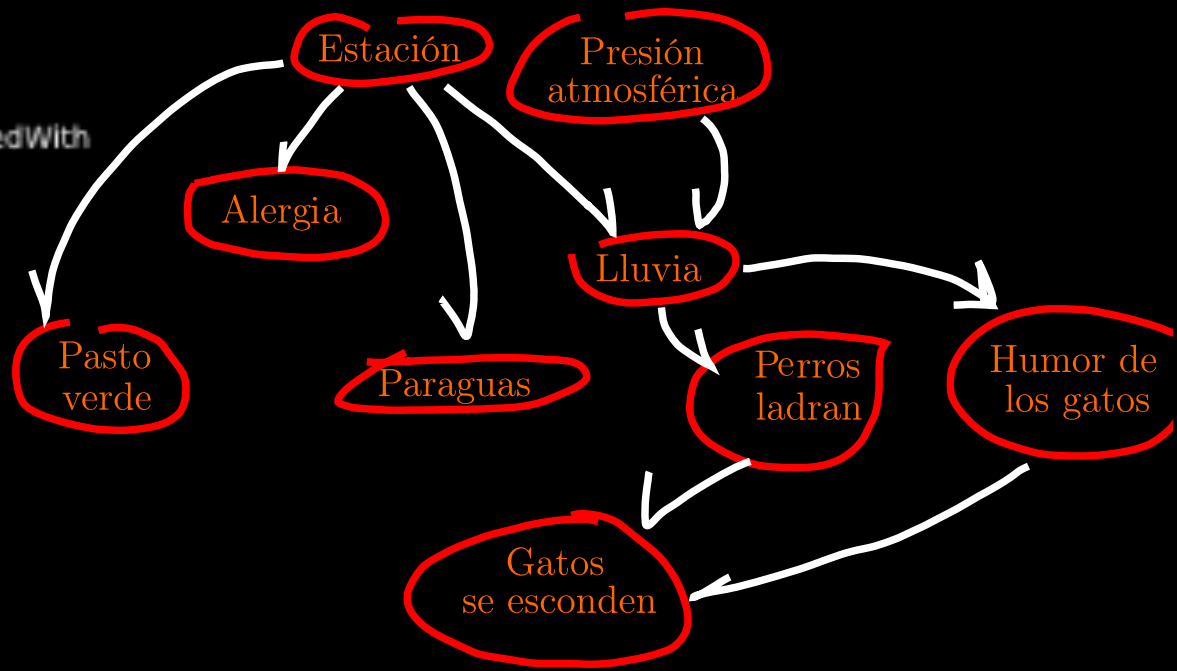
1. Funcionalidad
2. Replicabilidad.
3. Reproducibilidad.
4. Robustez.
5. Validación.
6. Rastreo de origen. (data provenance)

# Idea 1 Aplicar el estándar PORV-O

Agente - Entidad - Acción

- Esquema funcional (= entradas $\mapsto$ = salidas)
- Red Bayesiana



Ejemplo: Analizar algoritmos de detección de discontinuidad

Leer 150 papers para saber

- ¿En qué MR trabajan? (como hiperparámetro al menos)
- ¿Qué tipo de modelo aplica? (Lineal?)
- ¿Hasta qué distancia "detectan"?

Ej de provecho: Modelos lineales no dependen de cambios lineales del MR

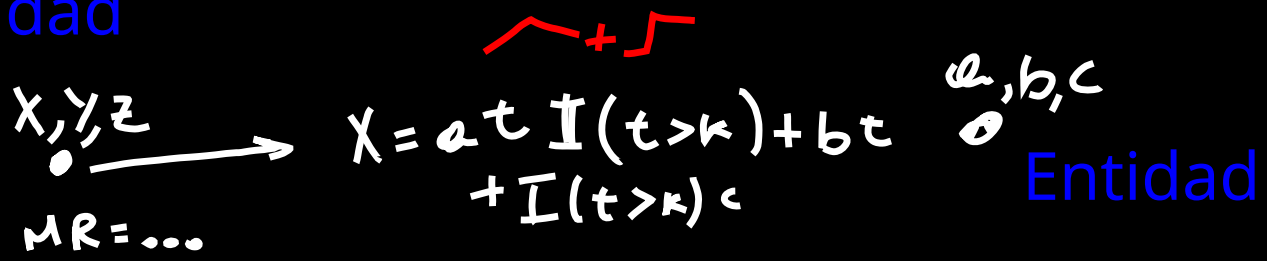
Mapping of PROV core concepts to types and relations

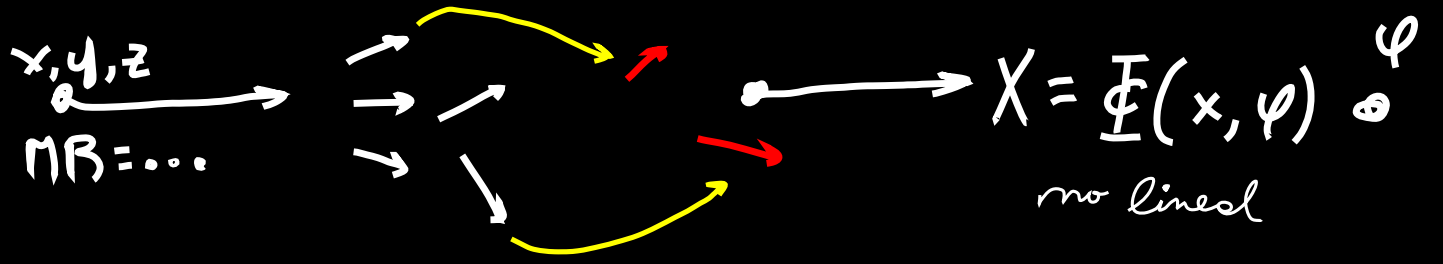
| **PROV Concepts** | **PROV-DM types or relations** | **Name** |
|---|---|---|
| Entity | | Entity |
| Activity | PROV-DM Types | Activity |
| Agent | | Agent |
| Generation | | WasGeneratedBy |
| Usage | | Used |
| Communication | | WasInformedBy |
| Derivation | PROV-DM Relations | WasDerivedFrom |
| Attribution | | WasAttributedTo |
| Association | | WasAssociatedWith |
| Delegation | | ActedOnBehalfOf |

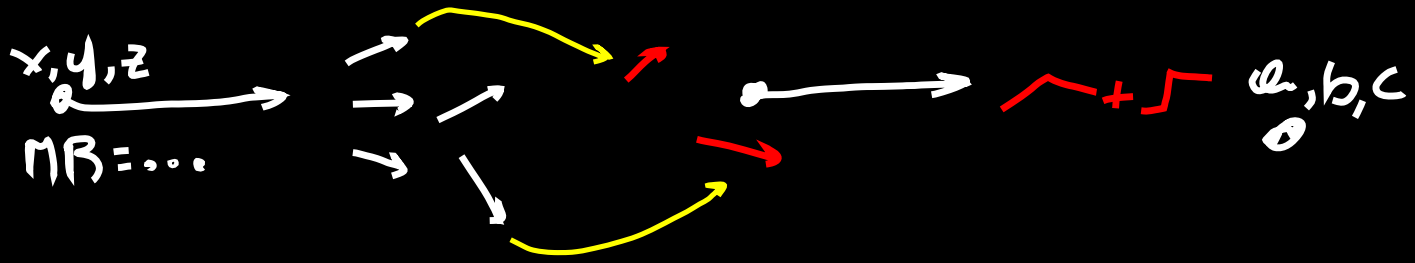
Aplicar el estándar PORV-O

Actividad

Entidad




El Marco de Referencia puede ser una propiedad de la entidad.

Ser lineal/no lineal puede ser una propiedad de una actividad (como transformación)

# Idea 2

## Planteo

El formato RINEX es texto $\mapsto$

- No define tipo de dato estandarizado para cada variable.
- Depende de un parser que puede ser mas o menos tolerante.
- Espacios y ceros pueden “cambiar” el archivo sin cambiar el dato.

Datos transmitidos por NTRIP También son (otro) texto.

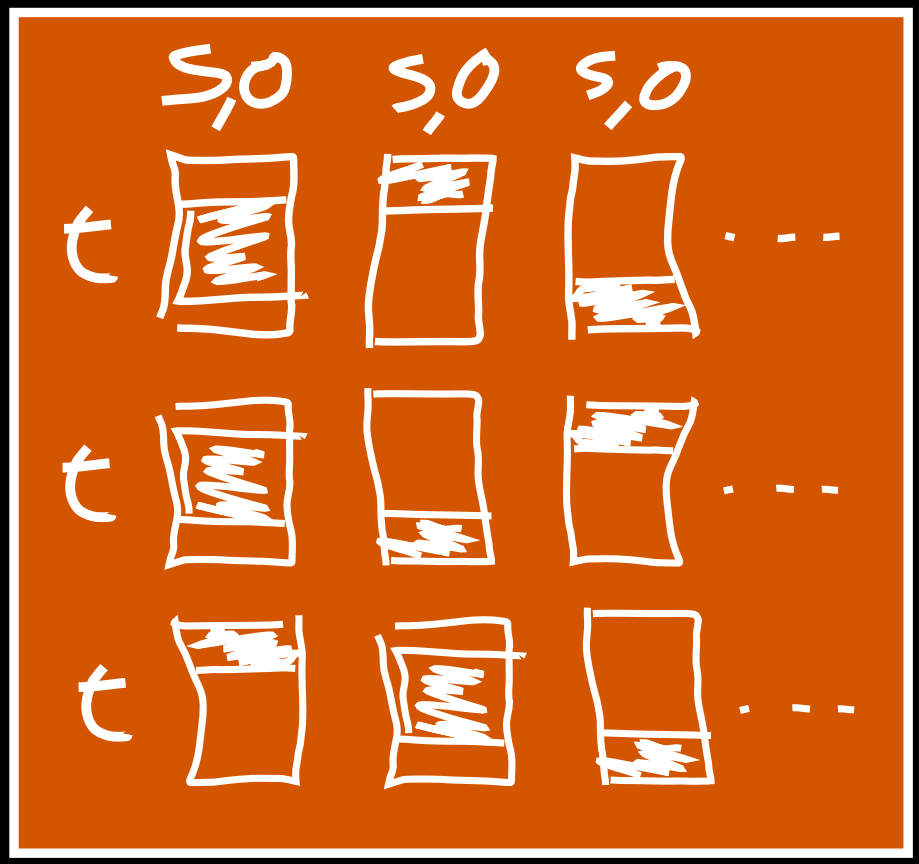
## BINEX

- Mejora todo esto, pero no es Columnar y los tipos de dato no son estándares de computadora (sigue habiendo algo de nicho)...

## Planteo

La naturaleza del dato es Observaciones de una variedad de indicadores indexados por el tiempo

- Se puede estandarizar en columnas con tipo de dato determinado



## Utilidades

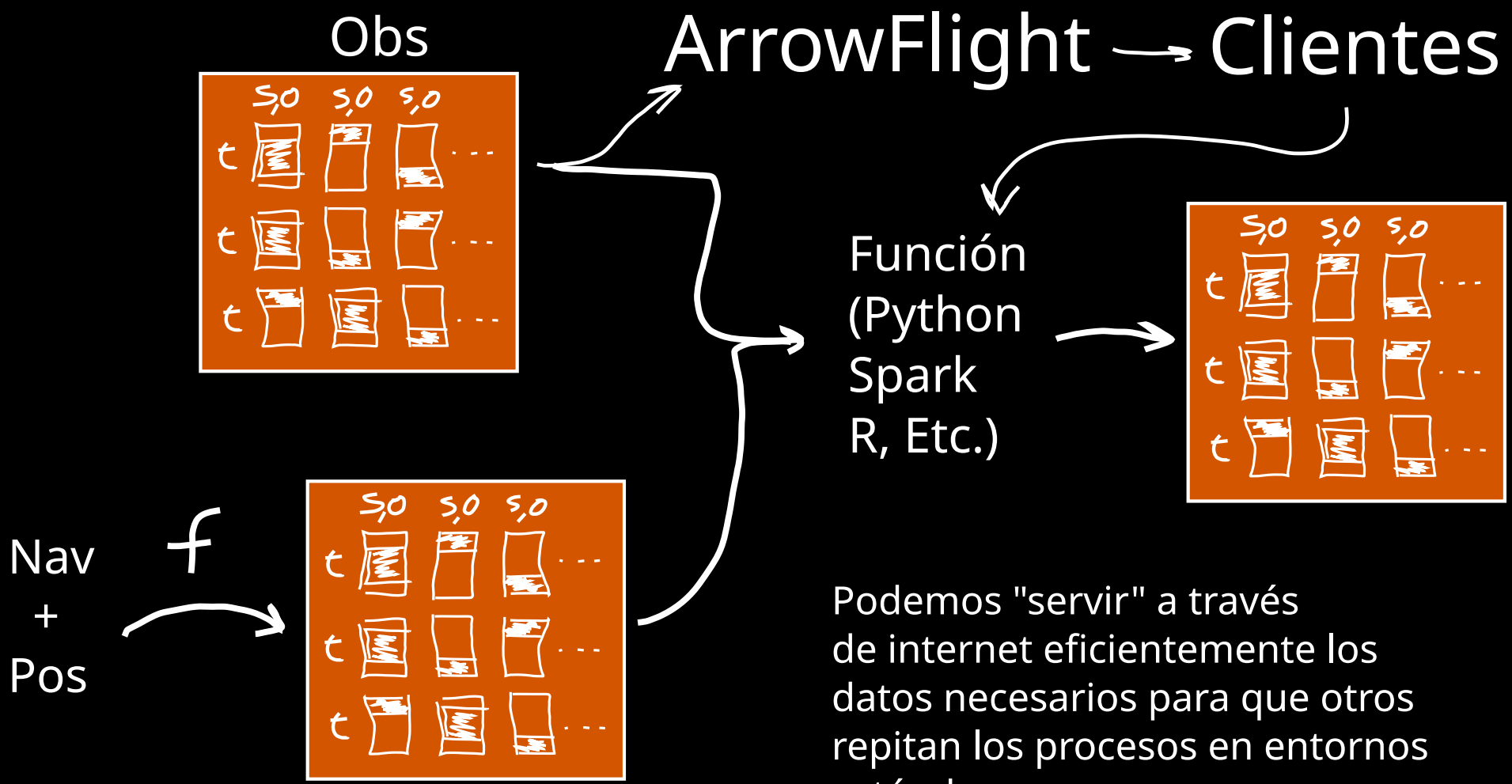
- Al fijar el tipo de dato se puede comparar datos (no archivos) con hashes.
  - (BINEX podría permitirlo, aunque mezcla muchos indicadores "extra")
- Se podría partir y transmitir los datos eficientemente. (aplicaciones RT) (y validarlos)
- Se puede usar entornos como spark, arrow-flight, etc.

## Hashes

https://github.com/jjclavijo/sirgas2022hashes

# Columnar Format: Flujos de trabajo

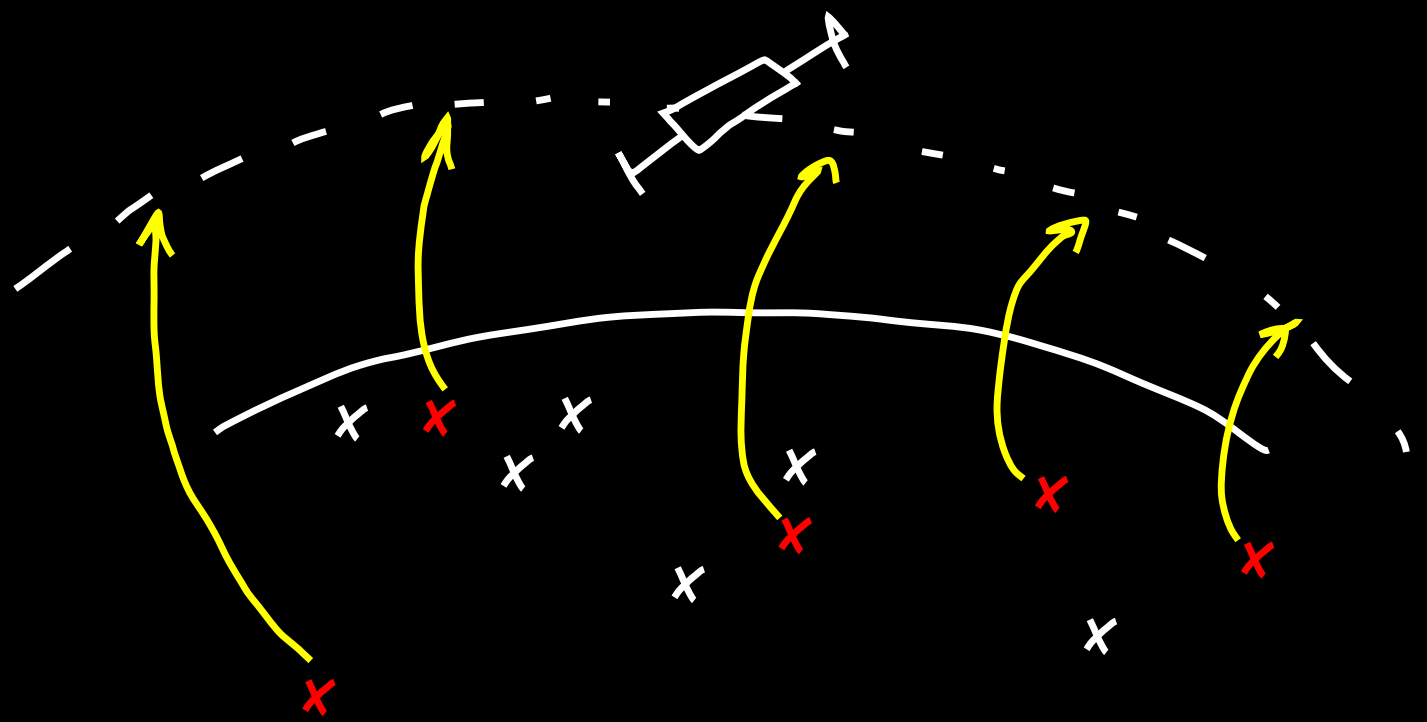
https://github.com/jjclavijo/sirgas2022arrow




Podemos "servir" a través de internet eficientemente los datos necesarios para que otros repitan los procesos en entornos estándar.
(python pandas, R tidyverse, etc.)

## Planteo

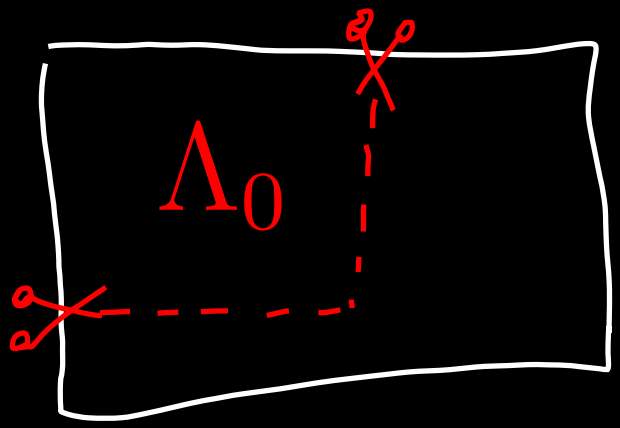
Es muy difícil garantizar la independencia real de procesamientos, aún si separáramos redes que no se conectan.

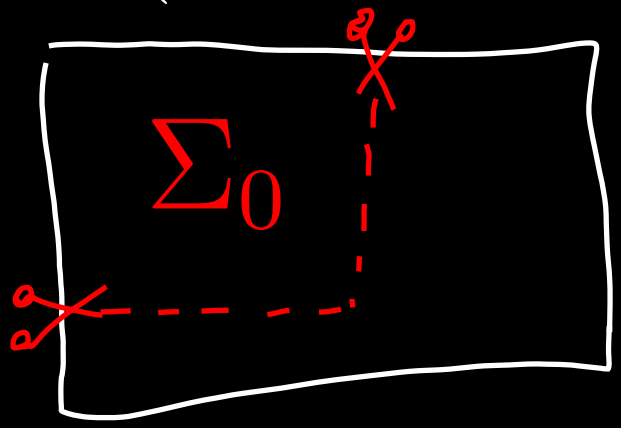


## Observación

Los productos publicados no describen su propio origen.
Ejemplo: Los SINEX pueden ser subsets: condicionales $X_1 = X|X_0$ o marginales $X_1 \ni P(X_1) \times P(X_0) \neq 0$ Los sp3 dan parámetros ($\theta$) de órbitas pero no describen $P(\theta, x_0)$
Si nuestros parámetros ($X_2$) son tales que $X_0 \cap X_2 \neq \emptyset$ estamos perdiendo información, porque no sabemos nada de P(X_0).

$$\Lambda = A^T P A \qquad \Sigma = (A^T P A + \Gamma)^{-1}$$

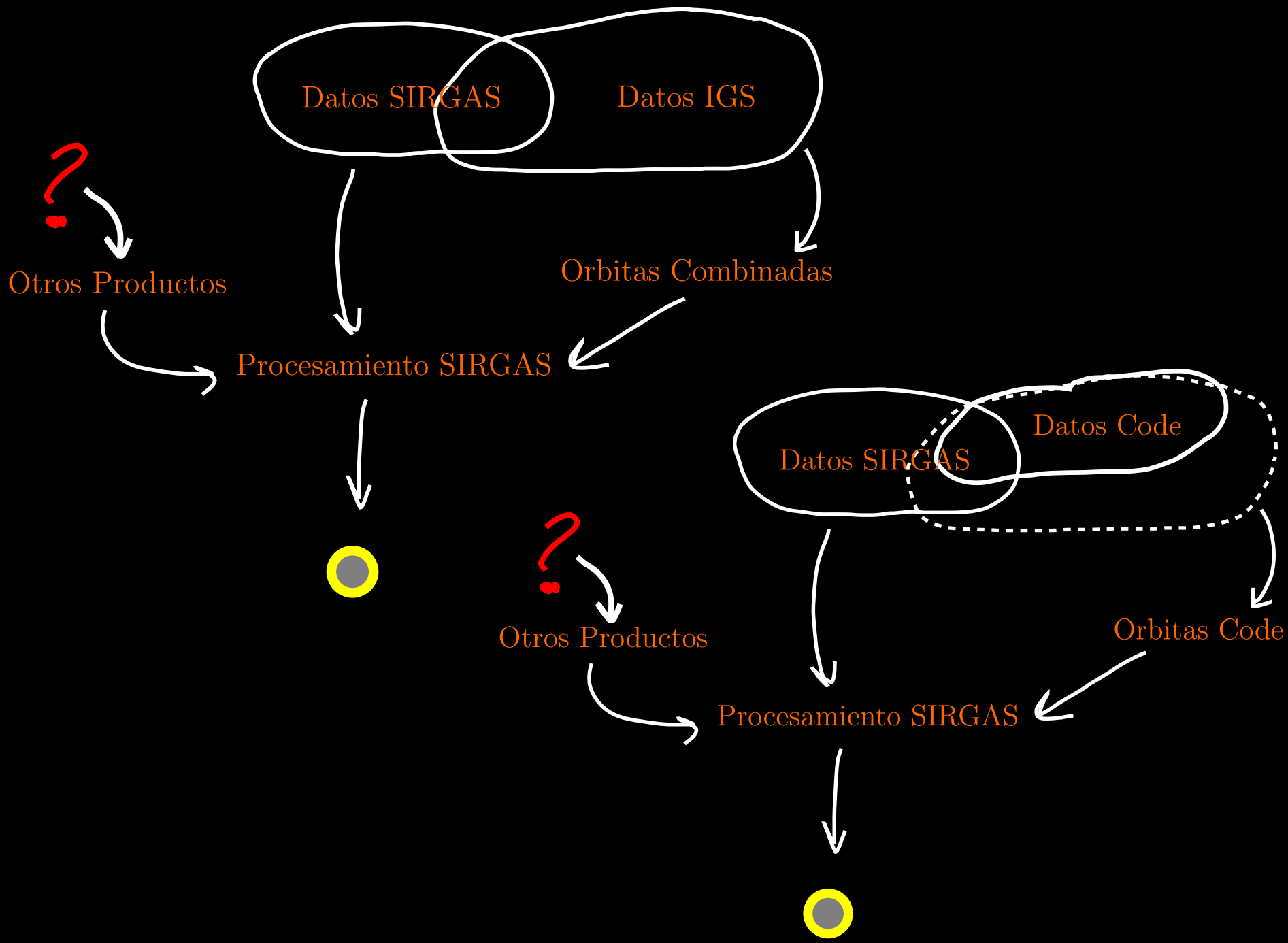



$$(\Lambda_0 + \Gamma)^{-1} = \Sigma_{X_0|X_1} \qquad \Sigma_0 = \Sigma_{X_0}$$

## Observación

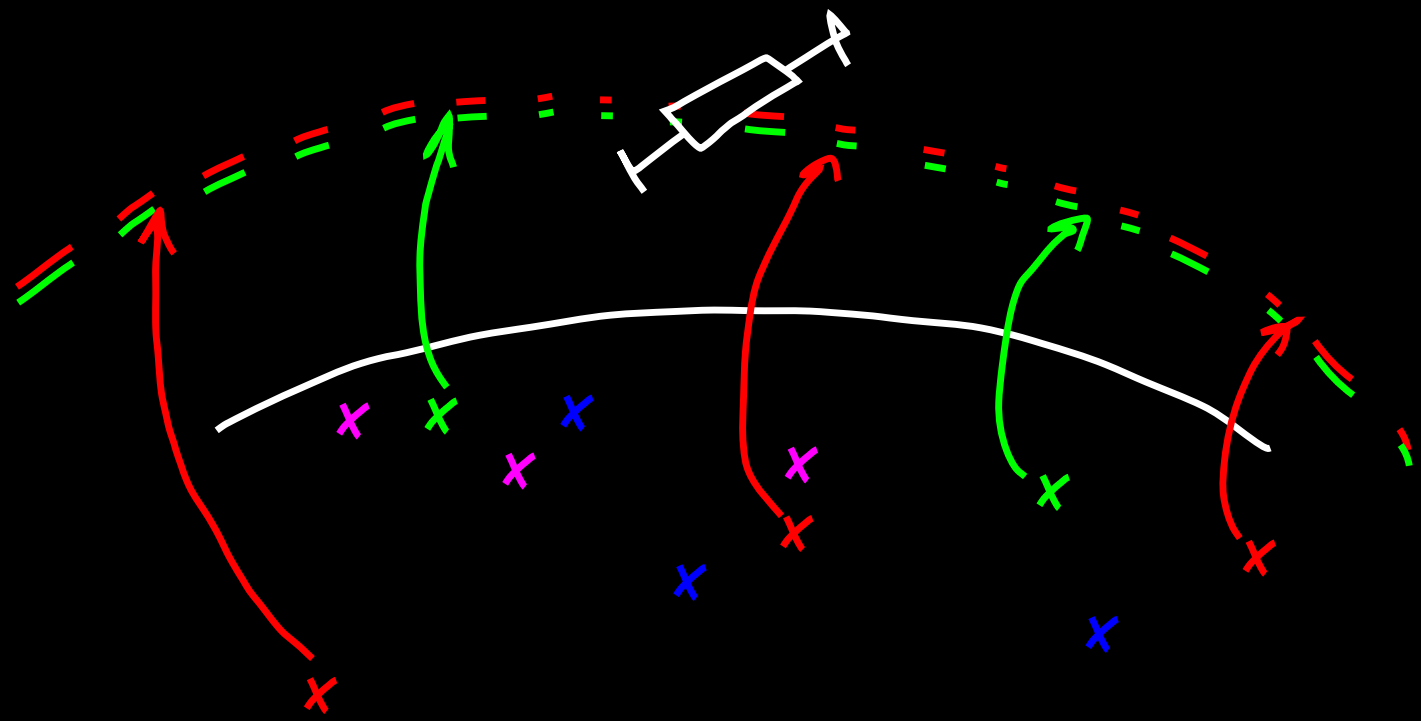
- El SINEX de SIRGAS no describe su dependencia del SINEX de óbitas de IGS, de hecho lo ignora porque usa los sp3 para calcular.
- Rastreando el origen de datos podría rastrearse esa dependencia (desde el SINEX de IGS)
- Esto permitiría diseñar la combinación de forma distinta usando una red bayesiana y Beleaf propagation


Datos SIRGAS
Datos IGS
Otros Productos
Orbitas Combinadas
Procesamiento SIRGAS
Datos Code
Datos SIRGAS
Otros Productos
Orbitas Code
Procesamiento SIRGAS

# Validación total (?)

Una idea intereseante:

- Tener conjuntos de órbitas independientes.
- Validar procesamientos con verdadera independencia modelos / datos.



¡Muchas Gracias!