

# Productos de Datos

# ¿Quién soy?

- Adolfo Javier De Unánue Tiscareño
- Ph.D. en Física Teórica
- Cofundador y CTO de OPI
- Director académico de la MCDatos en el ITAM, México
- [@nano\\_unanue](https://twitter.com/nano_unanue), [adolfo.deunanue@itam.mx](mailto:adolfo.deunanue@itam.mx), [adolfo@opi.la](mailto:adolfo@opi.la)

# Antes de empezar

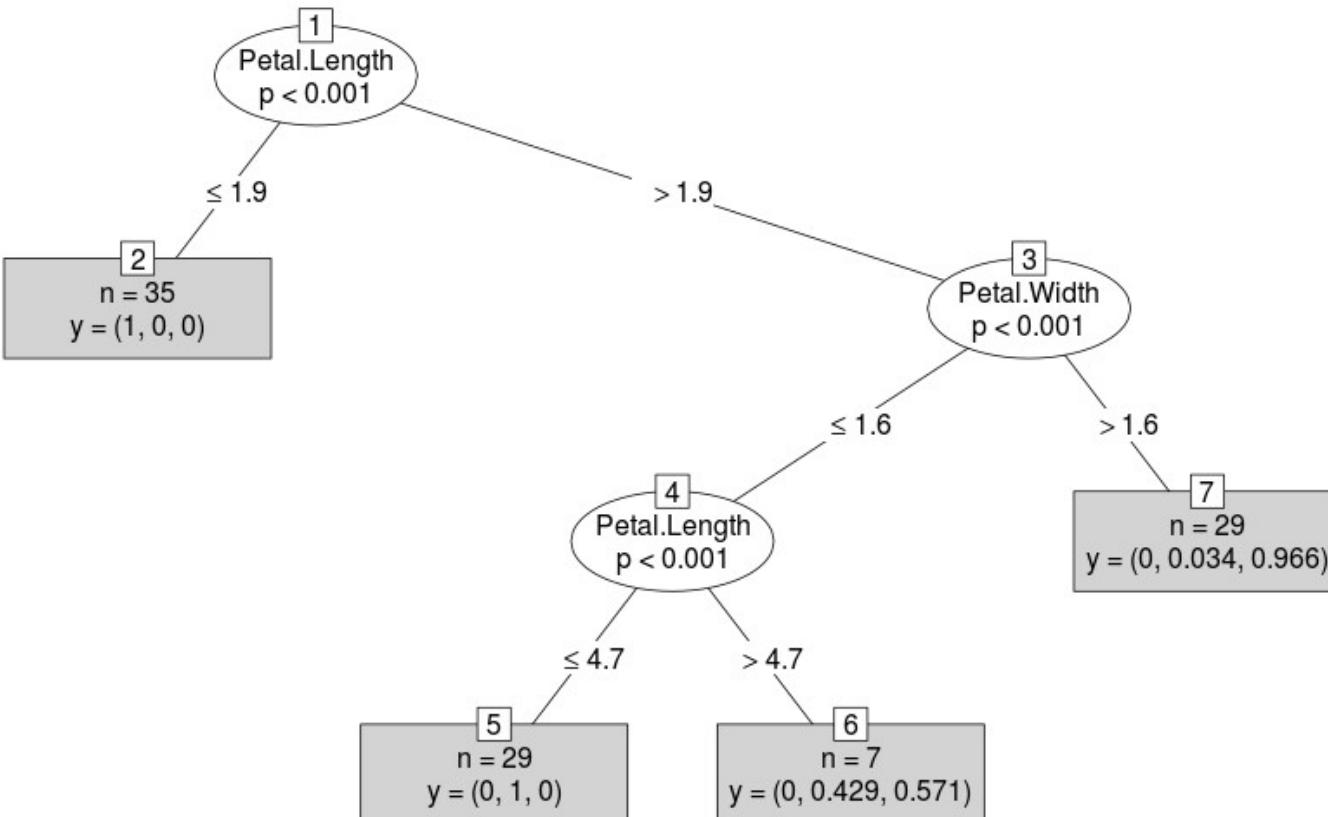
- Ciencia de datos NO es *Big data*
  - *Big data* es un conjunto de técnicas y tecnología para tratar con datos.
  - Aunque la incluye
- Tampoco es Aprendizaje de Máquina
  - Aunque la incluye
- Ni mucho menos Inteligencia de Negocios
  - Aunque la incluye

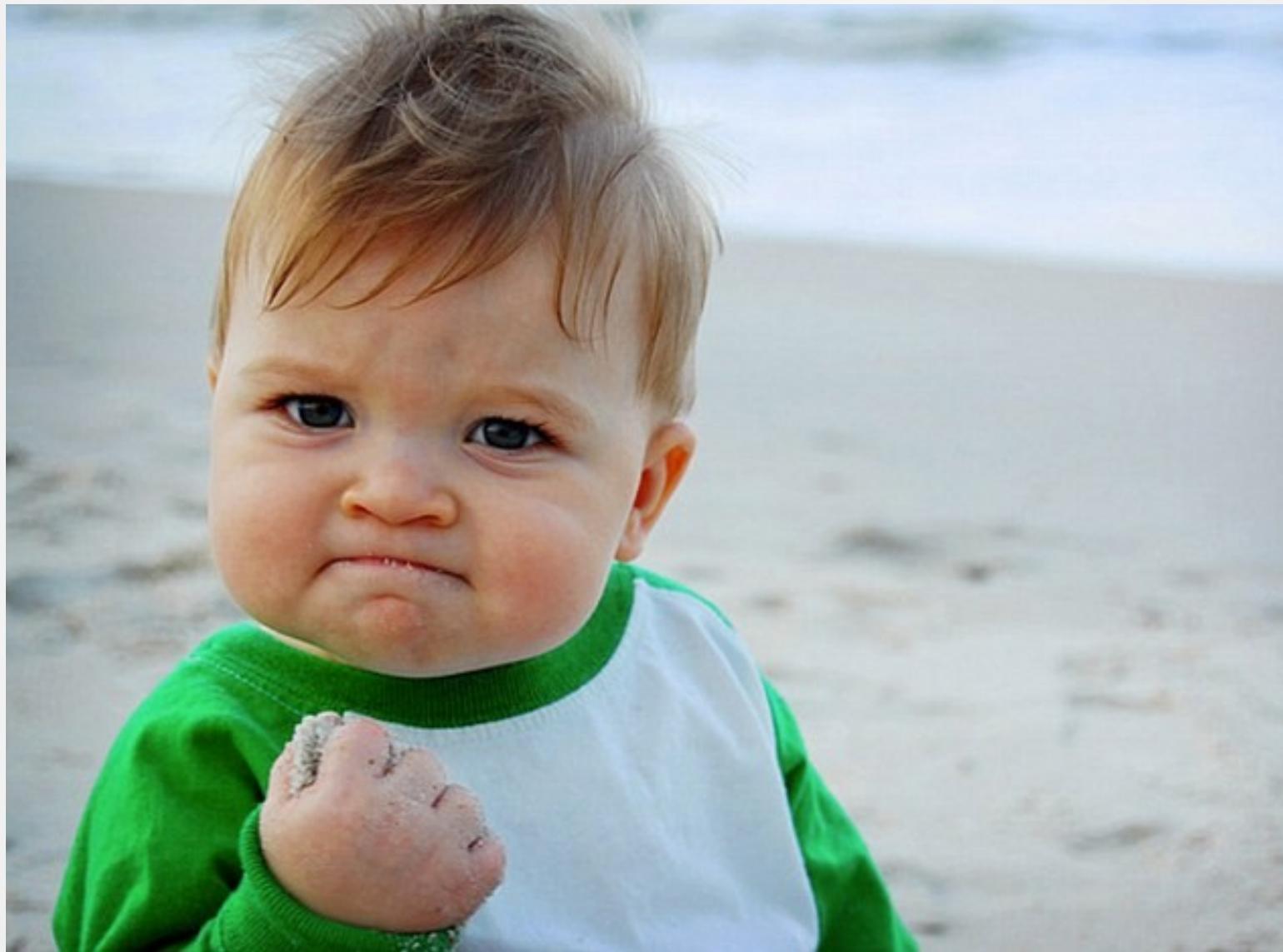
La mayoría de las personas piensan que hago lo siguiente:

```
> target <- Species ~ .  
  
> train <- sample(nrow(iris), size = 100)  
  
> iris_train <- iris[train,]  
> iris_test <- iris[-train,]  
  
> cdt <- ctree(target, iris_train)  
  
> table(predict(cdt, new_data=iris_test), iris_test$Species)
```

	setosa	versicolor	virginica
setosa	15	0	0
versicolor	0	15	0
virginica	0	2	18

La mayoría de las personas piensan que hago lo siguiente:





# En realidad . . .

→ La ciencia de datos, tiene que ver, con, bueno...

## Datos

→ i.e. es fenomenológica, empírica

# Sistemas Complejos Adaptativos

Sistemas  
Complejos  
**Adaptativos**

# **Personas de la tercera edad**

# **Defraudadores**

# **Crimen y violencia**

# Gobierno

# Productos de Datos

Brazalete para vigilancia de  
personas de la tercera edad

(2010)

# Sistema de prevención de fraude bancario

(2011-2012)

# Modelo explicativo para prevención de crimen y violencia

(2013)

Captura, procesamiento y  
ontología de datos abiertos  
de manera automatizada

(2015)

# ¿Para qué?

## → **Toma de Decisiones Racionales**

- i.e. tomar la mejor decisión basada en la evidencia (datos) disponible.

## • **Aumento de Inteligencia (AI)**

- *Human in the loop*: Plantea el problema, usa datos

## → **Aplicar método científico a la toma de decisiones**

- Se ha intentado desde los 40s, al parecer ahora si está teniendo impacto
- Cibernetica, O.R. Control Theory, Scientific Management
- Cybersyn

# Retos

# Retos de la Ciencia de datos

- No ignorar la complejidad y no linealidad del fenómeno:  
No “desbloquear” negativamente
- Uno de los principales retos de la ciencia de datos es tratar con la complejidad de los datos.
- Desarrollar Productos de datos
  - También es un CAS.
  - Cuya optimización es multiobjetivo...

# Complejidad

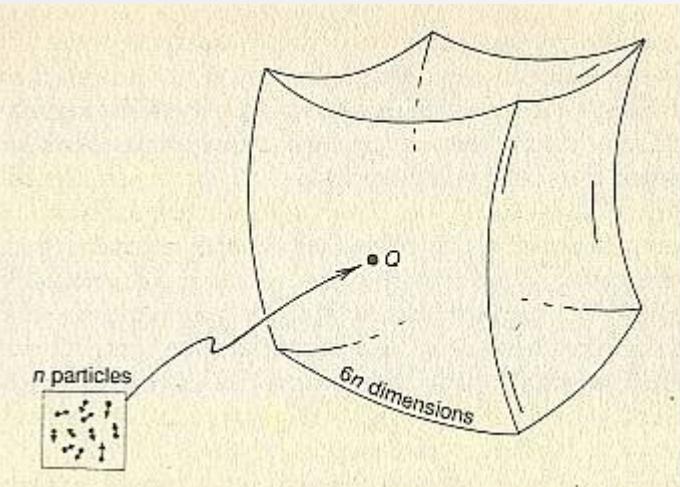
The big unsolved problems of  
the world result from system  
instabilities

Dirk Heilberg, ETH Zurich

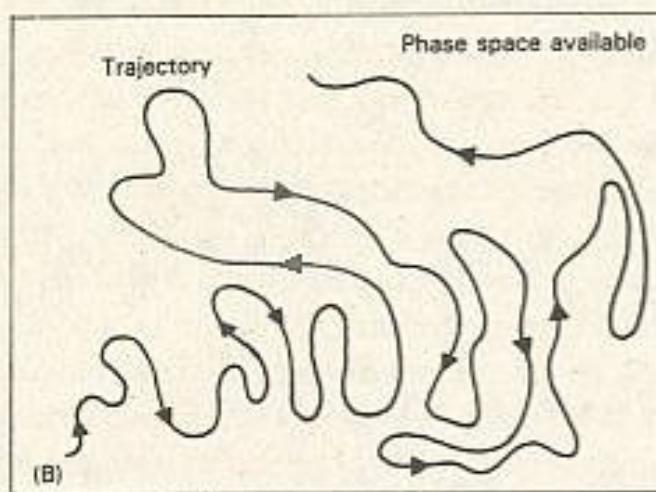
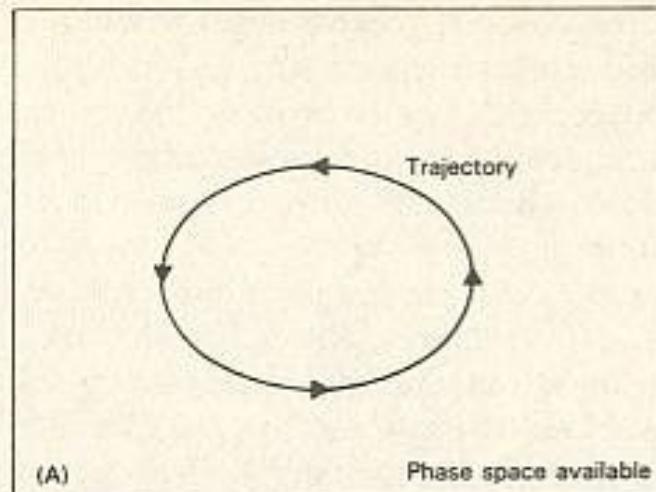
# ¿Por qué no había funcionado?

- Nos habíamos conformado con los pocos datos que podíamos obtener (medir) y basado en eso reducíamos la dimensionalidad a unos pocos indicadores
- No teníamos datos para hacer frente a la complejidad de la realidad, y jugábamos a la segura.
- Esta situación ya **no** es la actual

# DESBLOQUEO



Penrose, The Emperors New Mind



Coveney & Highfield, The Arrow of Time

NOTA: Con fines ilustrativos únicamente...

# ¿Por qué no funcionaría ahora?

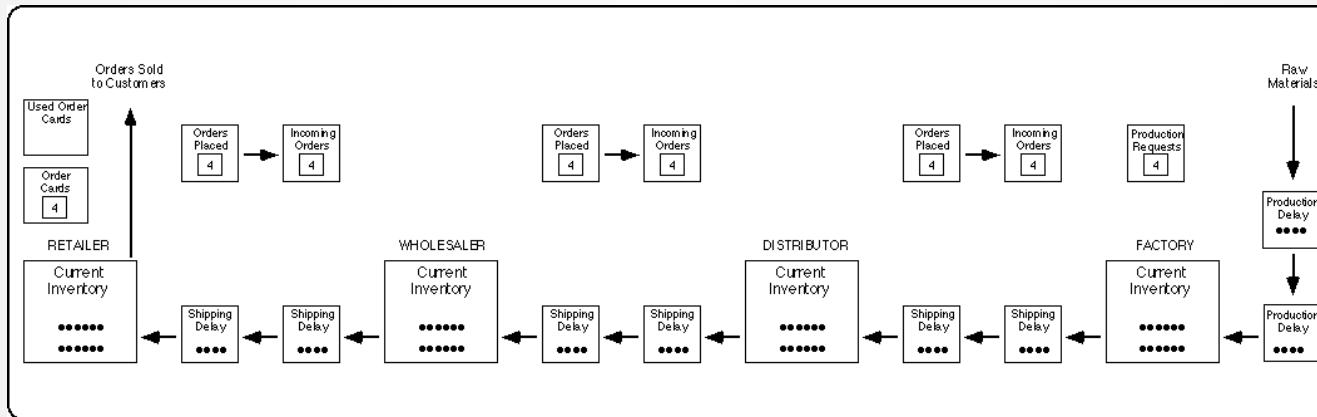
- Pensar cartesianamente en un mundo no lineal, es el mejor de los casos temerario, en el peor catastrófico.
- El cerebro humano piensa en causa/efecto lineal y coloca las causas fuera del sistema siempre que puede, ignorando las relaciones.
- No se pueden ignorar los ciclos de retroalimentación.
  - *Feedback/forward loops*
- No se puede ignorar los *delays* en el sistema.

# Ejemplos

- Algunos juegos:
  - *Beer game*
  - *El farol problem*
  - *Fishbank game*
  - *Friday Night at the ER*
- En todos ellos (a pesar de lo simples que son) el caos emerge, debido a que se ignoran las bucles de retroalimentación y los retrasos.

# Beer distribution game

(Sterman 1992)

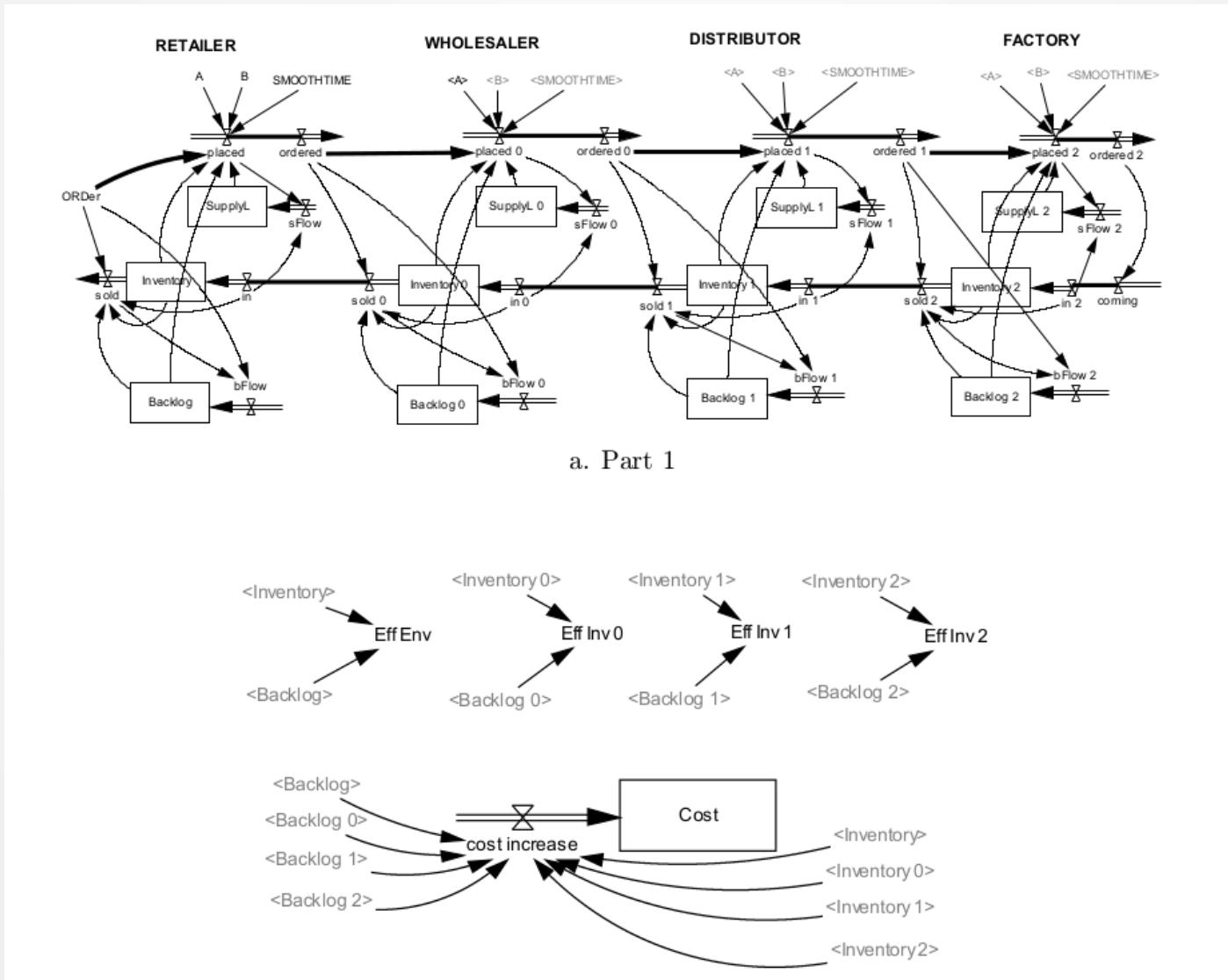


<http://jasss.soc.surrey.ac.uk/17/4/2.html>

- Retroalimentación y *delays*
- Información imperfecta → predicción → no linealidades → *Bullwhip effect*
- Existe una variante con información perfecta y aún aparece el *bullwhip effect*.

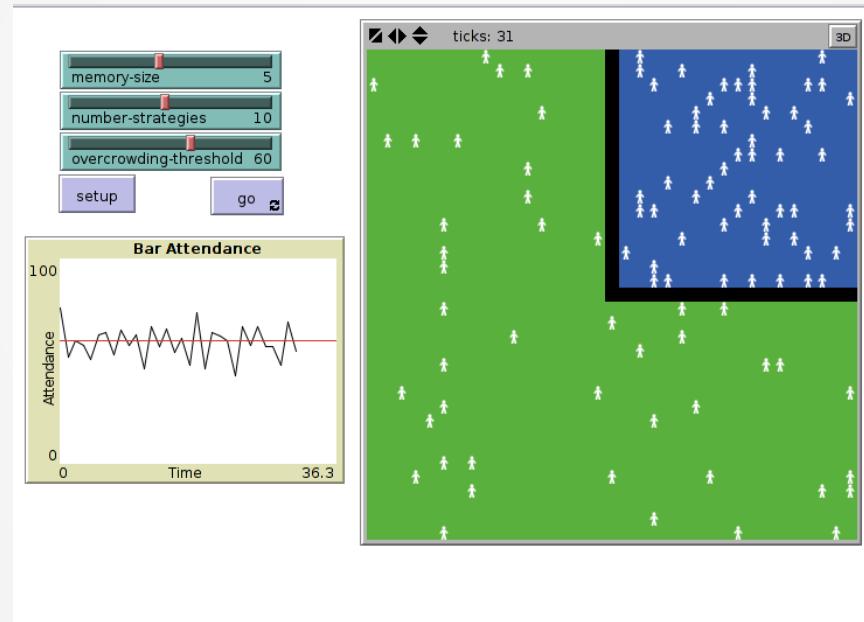
# Beer distribution game

(Sterman 1992)



# El farol problem

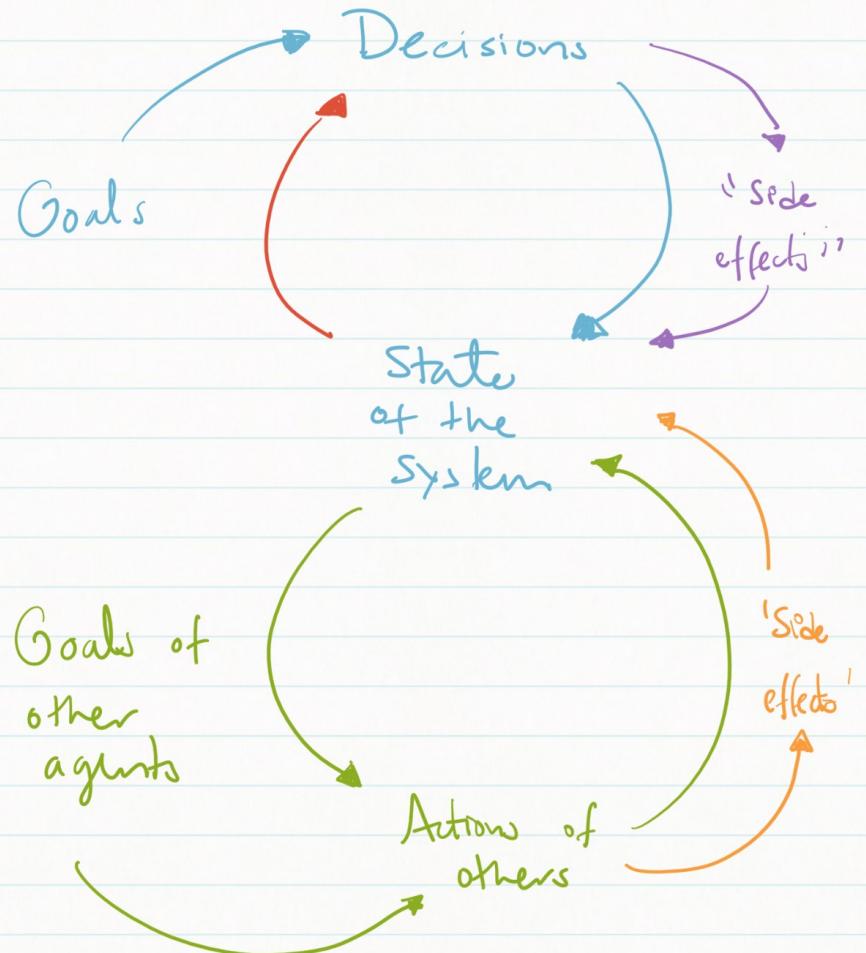
(Brain Arthur, 1994)



- No hay solución 'racionalmente deductiva' al problema
- Múltiples agentes haciendo predicciones
- Ya ocurre en **Waze**:

AdamRFisher

Please develop an app that simulates Waze recommendations to assess which routes will open up based on everybody else follow Waze. 13/05/2015 07:12



No sólo hay bucles nuestros, hay otros actores compitiendo y una única realidad.

# “Flight Simulators”

Importante ya que sólo hay una realidad

# Complejidad de los datos

# Data Complexity

- Volumen
- Semi-estructurado/No estructurado
- Variedad
- Conectividad

NOTA: Tamaño representa la dificultad

# ¿Cómo son los datos?

- Mediciones en un lugar y en un tiempo
- También hay datos transaccionales
- Existen “hechos”
  - Alcalde gobernante en el año xxxx en el lugar xxxx
  - Variables categóricas
  - En DWH se les conoce como *factless facts*
- Descripciones de objetos
- Datos relacionales o con conexiones
  - Importación, migración, redes de contactos, redes temporales, etc.

¿Qué tan difícil puede ser?

	Cat	long	Indicador	...
Obs1	#	#	#	Implicita la fecha

	Lugar	Indicador	...
Obs1			
Obs2			

	Fecha	Tech1	...
Lugar1			
Lugar2			

	Ind1	Ind2	...
Lugar1			
Lugar2			

	Fecha1	Fecha2	...
Lugar1			
Ind1			
Ind2			
Lugar2			
Ind1			
Ind2			

Indicador 1			Indicador 2		
Fecha1	Fecha2	...	Fecha1	Fecha2	
Lugar1					
Lugar2					

Tran1		Tran2	
Tech1	Tech2	Tech1	Tech2
Lugar1			
Ind1			
Ind2			
Lugar2			
Ind1			
Ind2			

	Lugar	key	SubTech1	SubTech2	...
Fech1	Lugar1	Ind1			
Tech1	Lugar1	Ind2			
Fech1	Lugar2	Ind1			
Fech2	Lugar1	Ind1			

		Key	
Fecha 1	Lugar 1	Ind 1	
Fecha 1	Lugar 1	Ind 2	
Fecha 2	Lugar 2	Ind 1	
Fecha 2	Lugar 2	Ind 2	

	Fecha 1	Fecha 2	Fecha 3
Lugar 2			
Lugar 3			
Lugar 4			
...			

Implicado el lugar ↗

Además hay que tomar en cuenta el formato, si es abierto o no, etc.

Si se puede manipular, etc

# Los datos...

- Una perspectiva funcional ayuda muchísimo al conceptualizar el repositorio de datos
- No hay variables, hay *values*
  - Son inmutables, i.e. su valor está ligado a una posición espacio-temporal y no puede ser cambiada.

# Los datos...

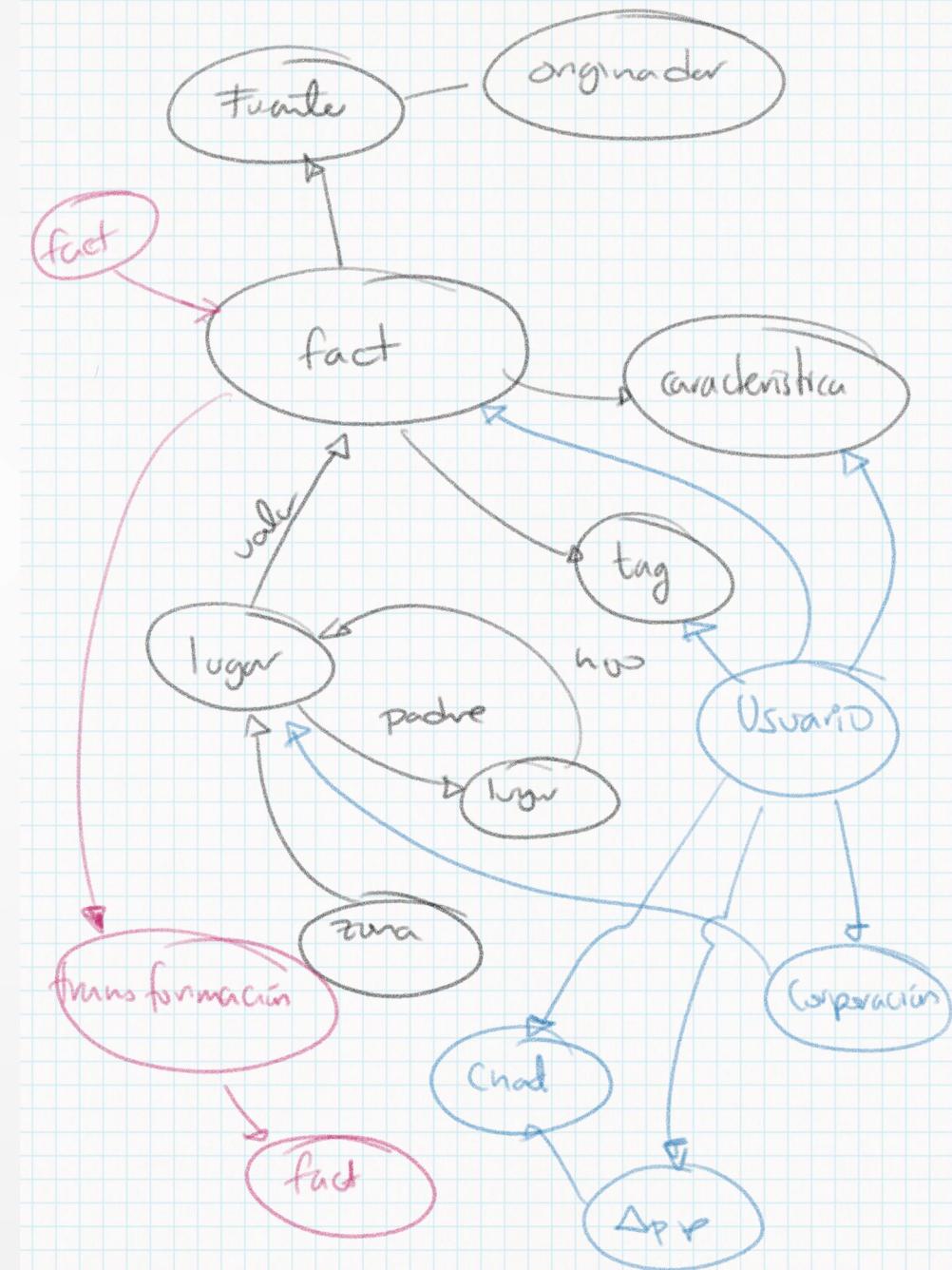
- Diferentes formatos y estructuras de datos
- Decidir qué automatizar
- ¿Estandarizar el input al pipeline?



# Los datos . . .

- ¿Cómo guardarlos?
- ¿Dónde guardarlos?
- Las variables derivadas ¿Dónde crearlas?
  - ¿En el pipeline de tal manera que queden precalculadas?
  - ¿A la hora que el usuario las solicite?

# Datos en Grafos



# Producto y Procesos de datos

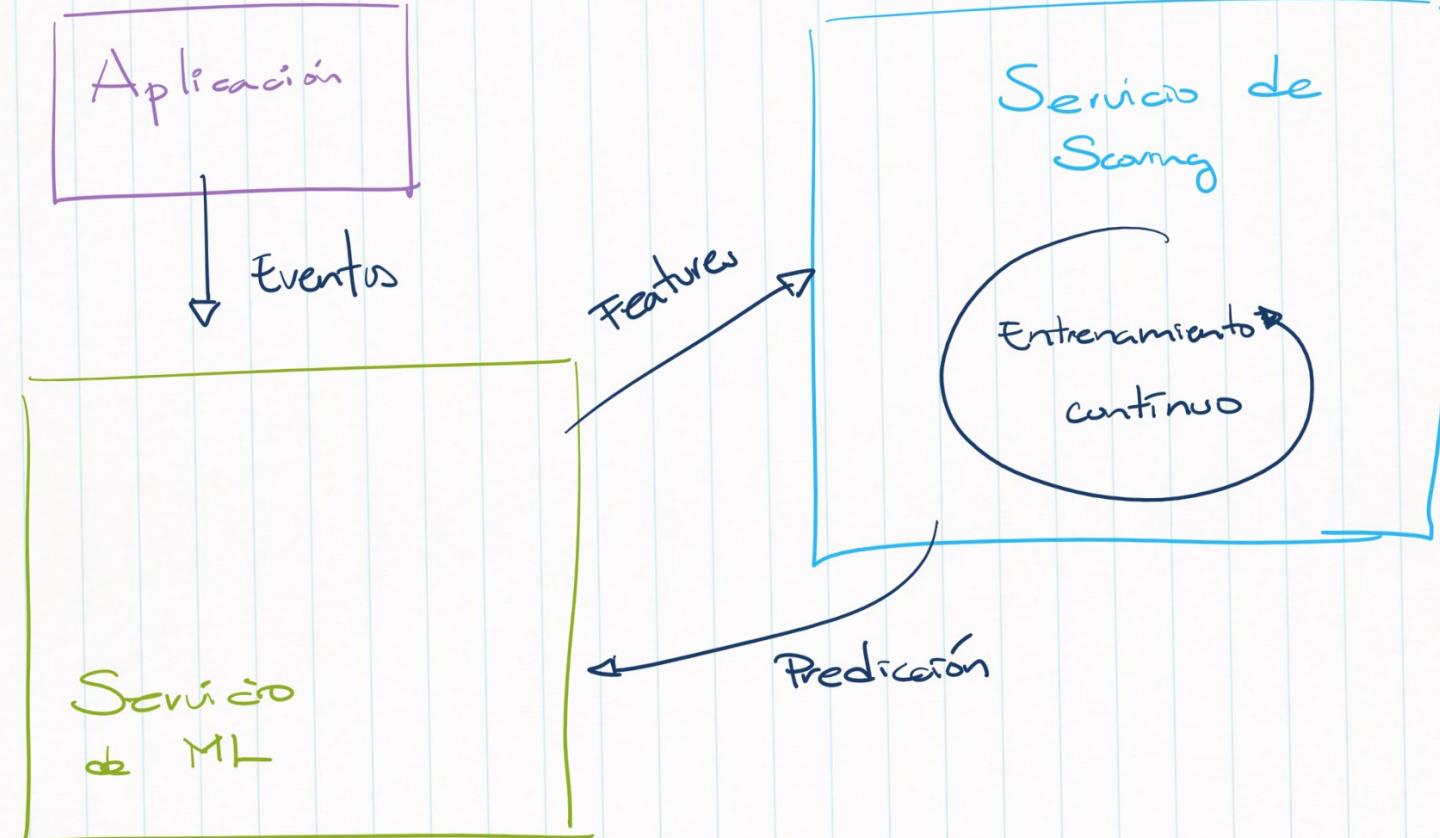
# Producto de datos

- Debe de ser un sistema continuo
  - Recuerda que es un CAS
- Todas las partes: reentrenamiento, recalibración, adquisición, movimiento de datos, transformación, limpieza, etc.
  - i.e. el Proceso

# Procesos Vitales

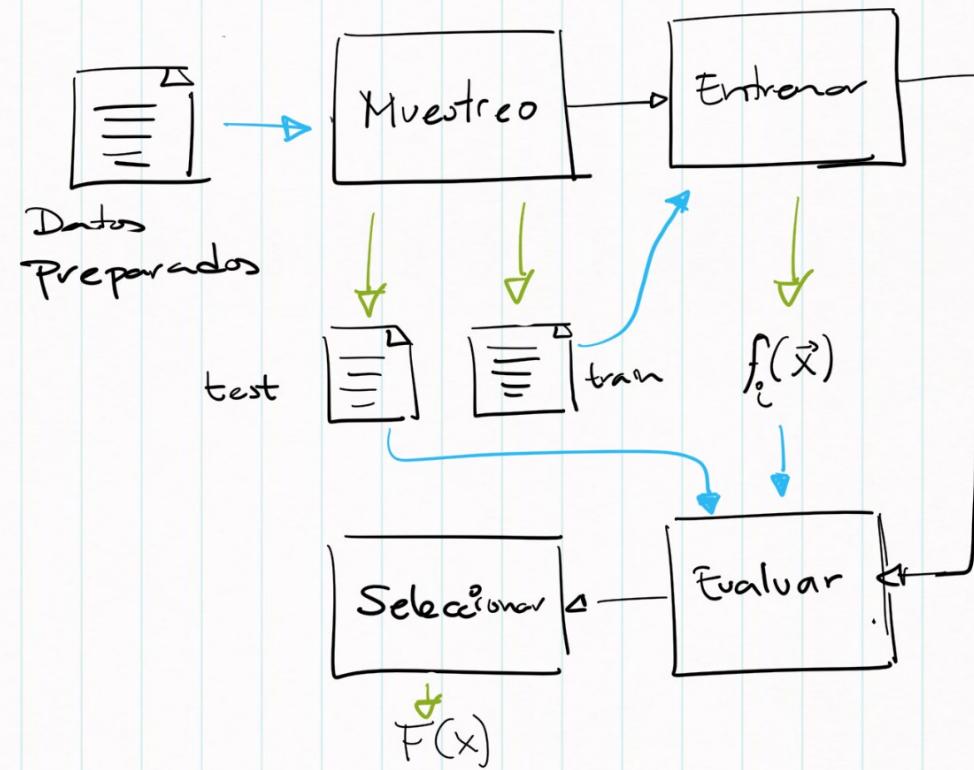
- Regularmente existen varios pasos de procesamiento para preparar los datos.
  - Extraer los datos (desde una carpeta, el internet, una base de datos) e importarlos al *data lake*.
  - Validar los datos.
  - Transformarlos a un formato más adecuado.
  - Ejecutar agregaciones y generación de variables.
- Y pasos para preparar el modelo
  - Entrenar, validar y seleccionar modelos.
  - Poner en producción el modelo seleccionado

# ¿Cómo se ve un producto de datos? (uno de muchos posibles)



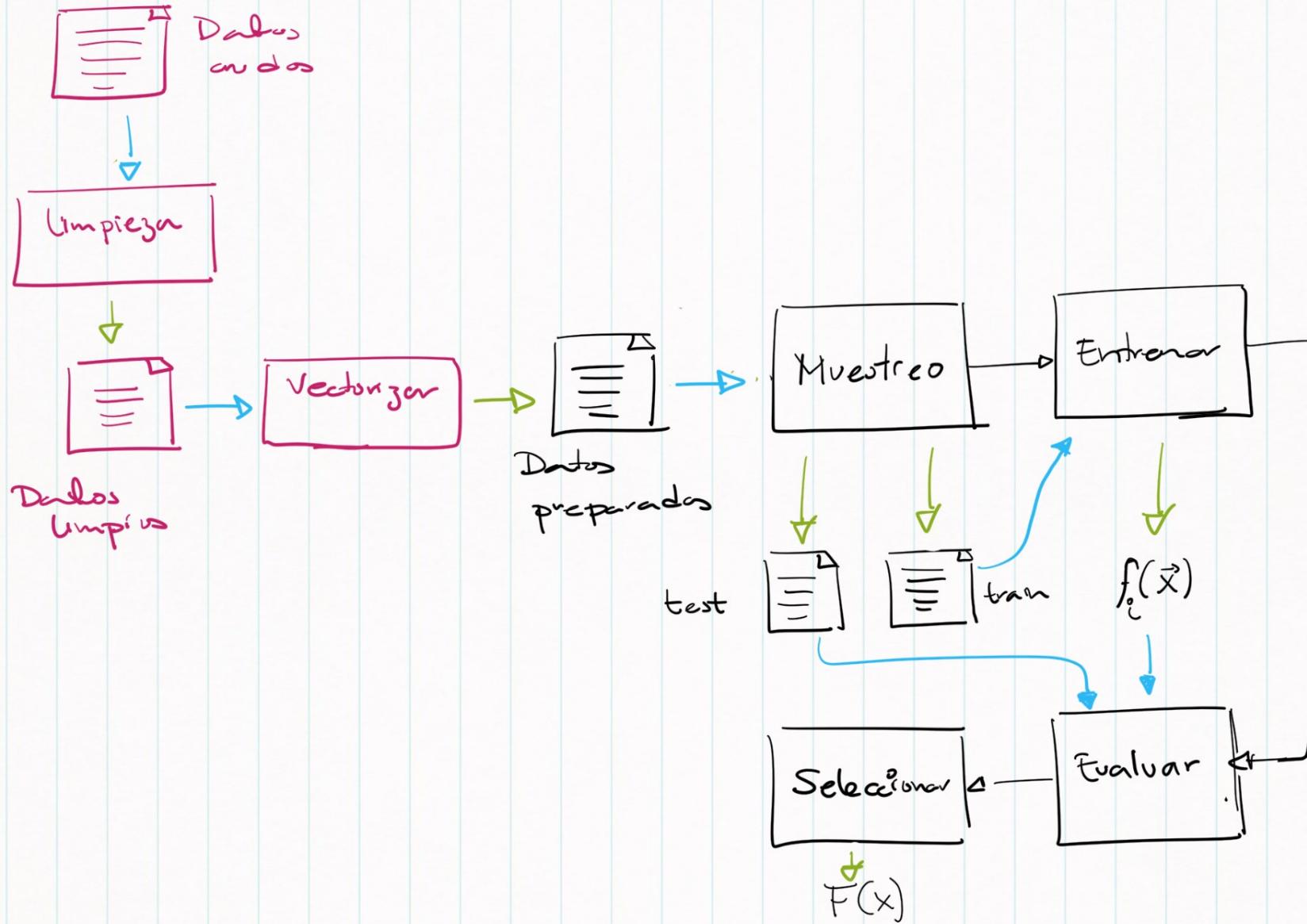
# El proceso de modelar

(regularmente se hace a mano)

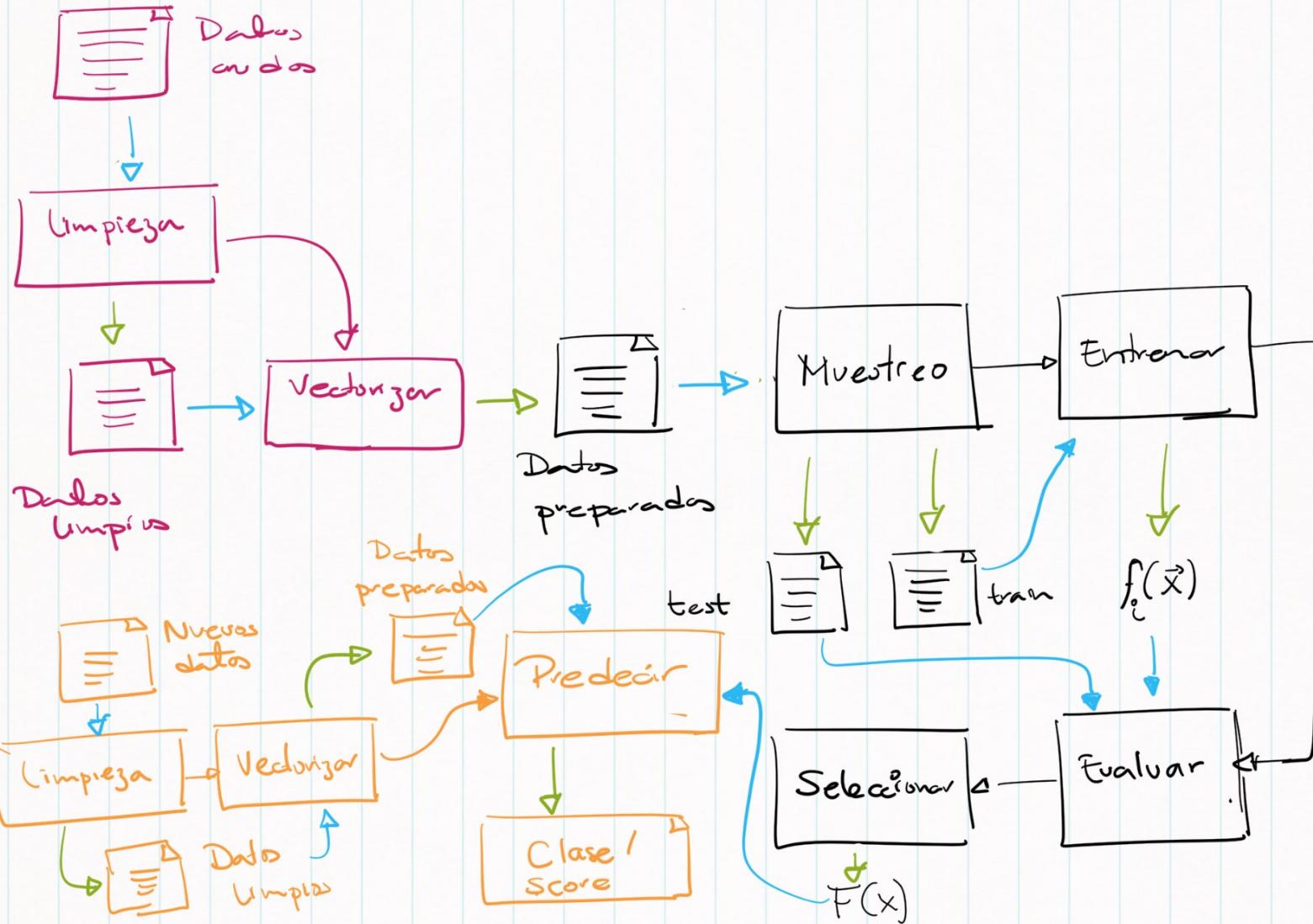


# Lo que no queremos hacer...

(pero hay que hacer)



# El significado de un producto de datos



# No olvidar a los usuarios

- Aquí empieza la multiplicación de los procesos de ciencia de datos
  - Más procesos
  - Más algoritmos
  - Más pipelines
- Recomendaciones basadas en comportamiento y/u otros usuarios.
- Necesitamos ver qué hacen los usuarios

# No olvidar a los usuarios

- Es muy importante tener los logs funcionando en cuanto antes
- Servicio al cliente
- Aumenta la inteligencia interna
  - Organizacional
  - De la “máquina”
- La captura de datos no es lo único, si se proveé de Flight simulators, se generan nuevos datos.

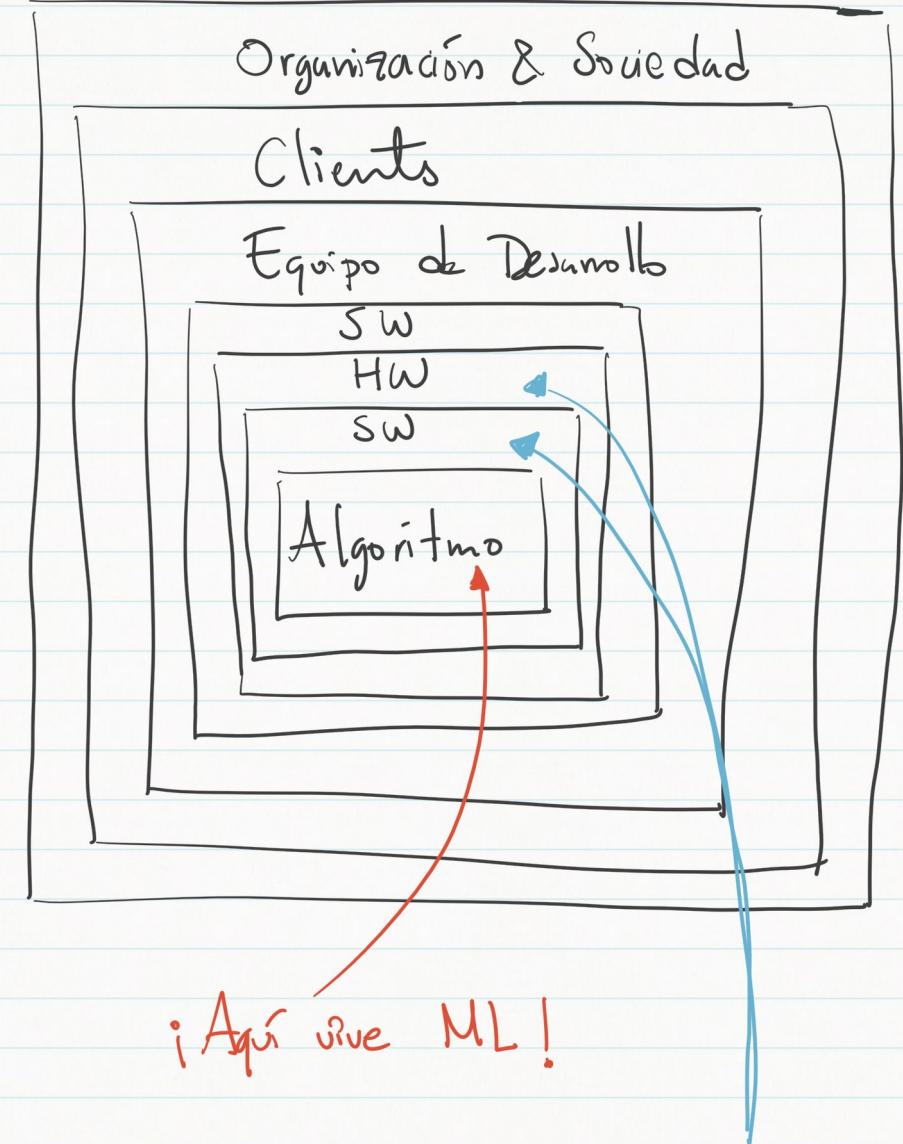
Al final el  
producto  
también es un  
CAS...

Algoritmo

Al final el producto también es un CAS...



Al final el producto también es un CAS...



# Función de optimización multiobjetivo

# ¿Qué quieres optimizar?

- **Algoritmo**
  - *Learning rate, convexity, error bound, etc.*
- **SW/HW**
  - RAM, Disco, CPU, tiempo en aprender, tiempo en predecir
- **Recursos Humanos**
  - Tiempo para implantar, mantenibilidad, *reliability*, recursos/costos
- **Clientes**
  - Valor directo, usabilidad, explicabilidad, “accionabilidad”
- **Sociedad**
  - Valor indirecto

# Ideas para llevar

- Ciencia de datos, es una herramienta para analizar CAS.
- El método científico y la tecnología son vitales para resolver nuestros grandes problemas.
- Desarrollar un producto de datos es muy complejo y tiene muchos retos muy interesantes de representación, modelado, comunicación, etc.
- Traté de dar una visión general de los problemas a los que me enfrento en la empresa

# ¿Preguntas?

y quizá posibles respuestas...

¡Gracias por su  
tiempo!

[adolfo@opi.la](mailto:adolfo@opi.la)

[adolfo.deunanue@itam.mx](mailto:adolfo.deunanue@itam.mx)

[@nano\\_unanue](https://twitter.com/nano_unanue)