Retos de la Ciencia de datos

¿Quién soy?

- Adolfo Javier De Unánue Tiscareño
- Ph.D. en Física Teórica
- Cofundador y CTO de OPI
  - Aquí hago ciencia de datos
  - Estamos contratando :)
- Director académico de la MCDatos en el ITAM, México
  - Aquí también hago ciencia de datos
  - Inscripciones en Agosto :)

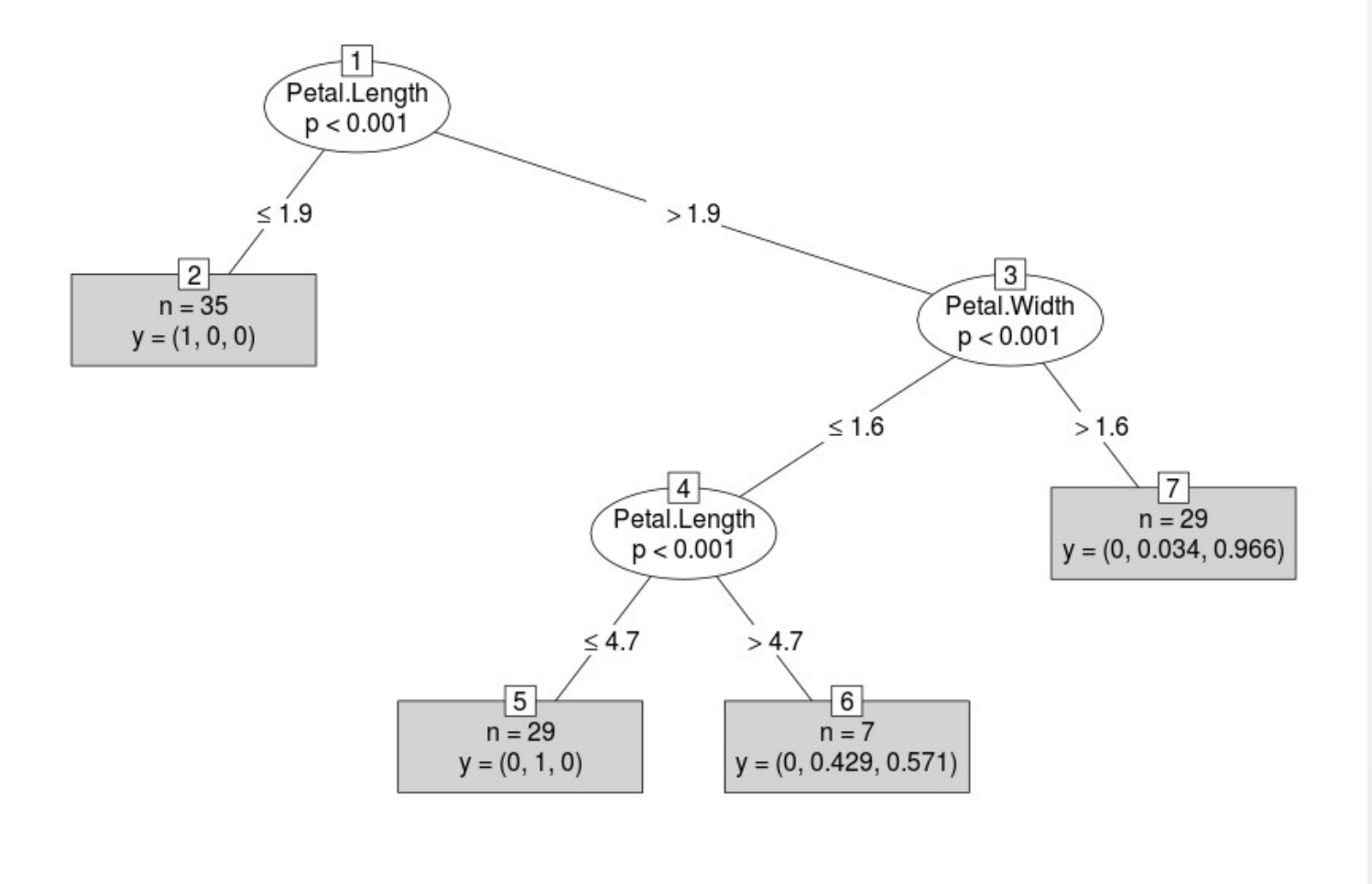
Antes de empezar

- 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

Complejidad

Urgencia para resolver

¿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**

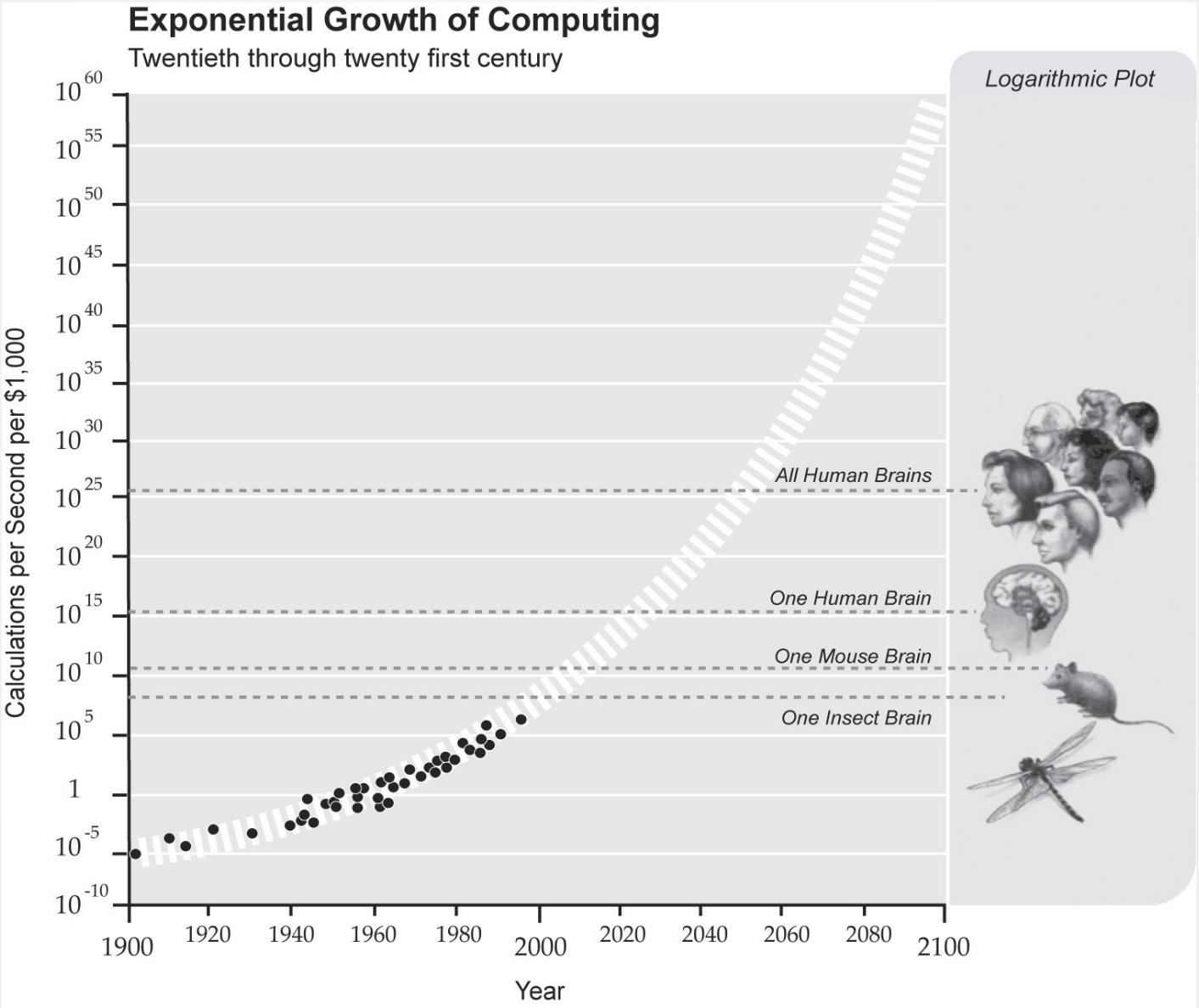
Cómputo

Almacenamiento, RAM, CPU

En el gran esquema de las cosas

Machine Learning → **Artificial Narrow Intelligence**
Data Science → **Intelligence Augmentation**
Singularidad → **Artificial General Intelligence**

En el gran esquema de las cosas



http://www.singularity.com/images/charts/ExponentialGrowthofComputing.jpg

¡Estamos
construyendo el futuro!

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

¿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

¿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.

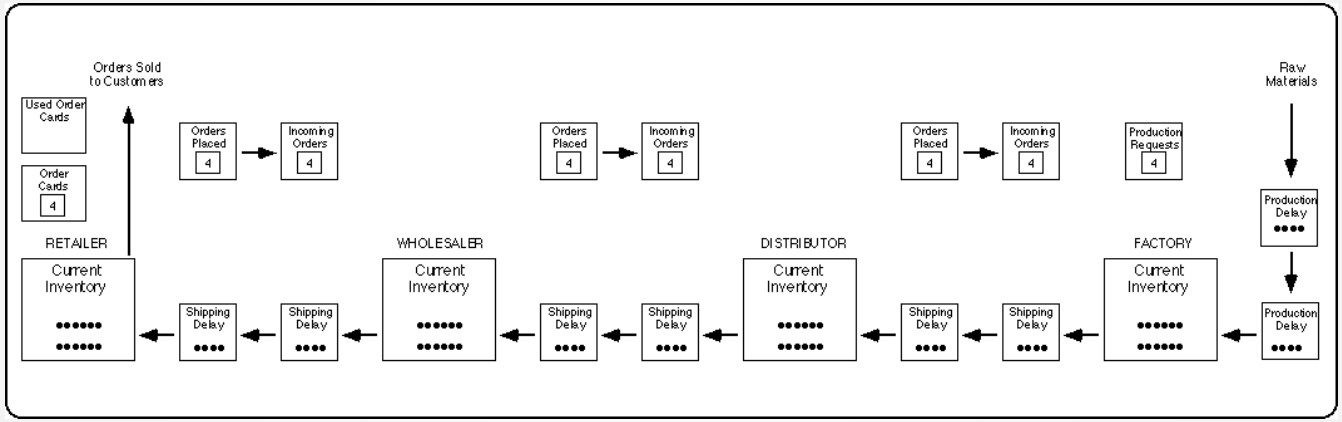
Ejemplos

- 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.
- *Flight simulators:*
  - Hay una sola realidad

Beer distribution game

(Sterman 1992)

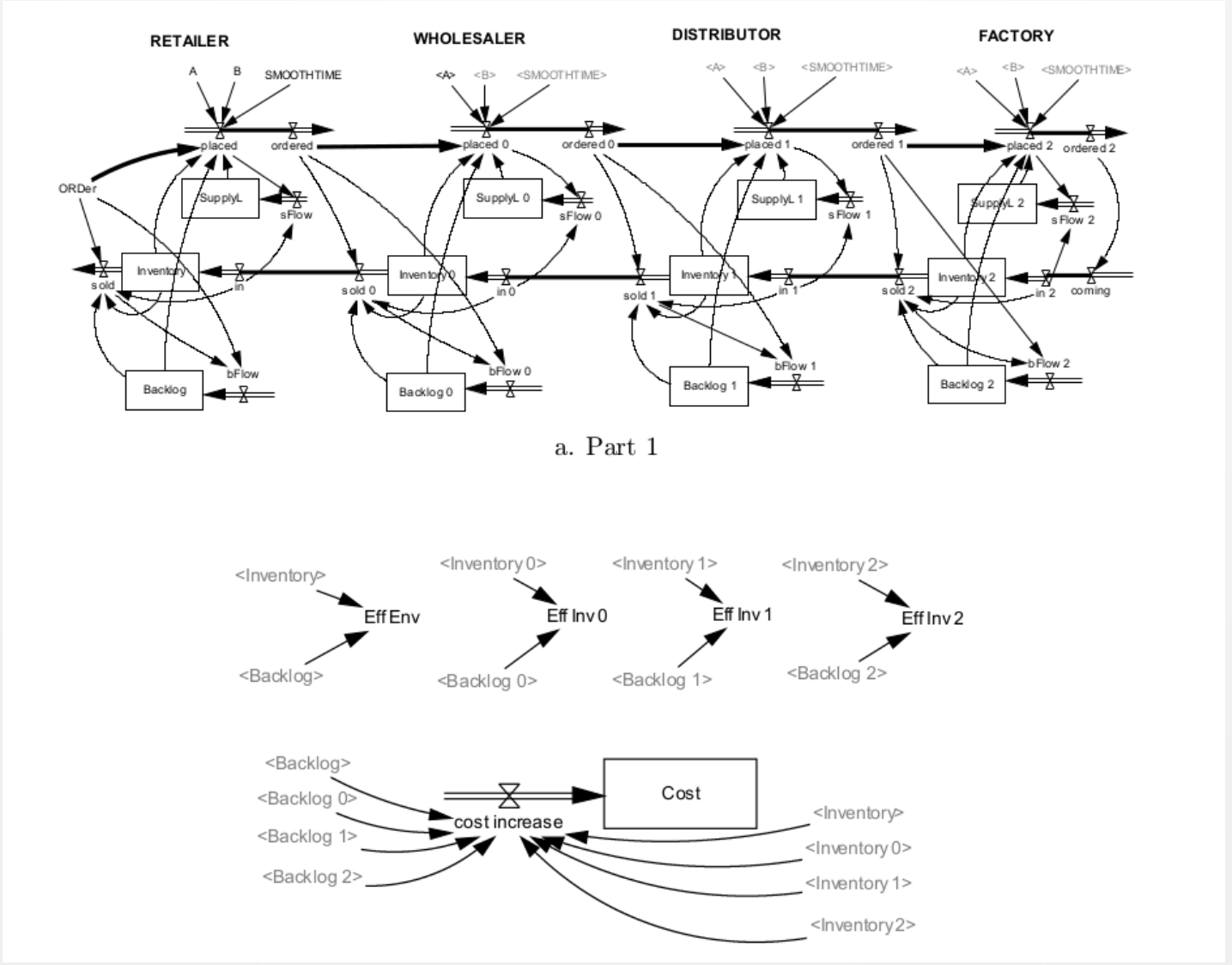


http://jasss.soc.surrey.ac.uk/17/4/2.html

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

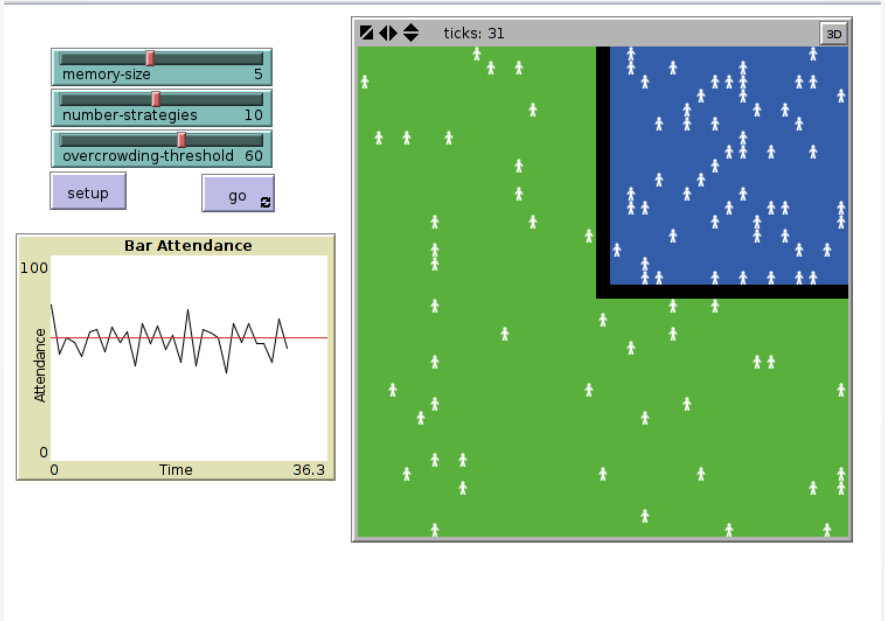
Beer distribution game

(Sterman 1992)



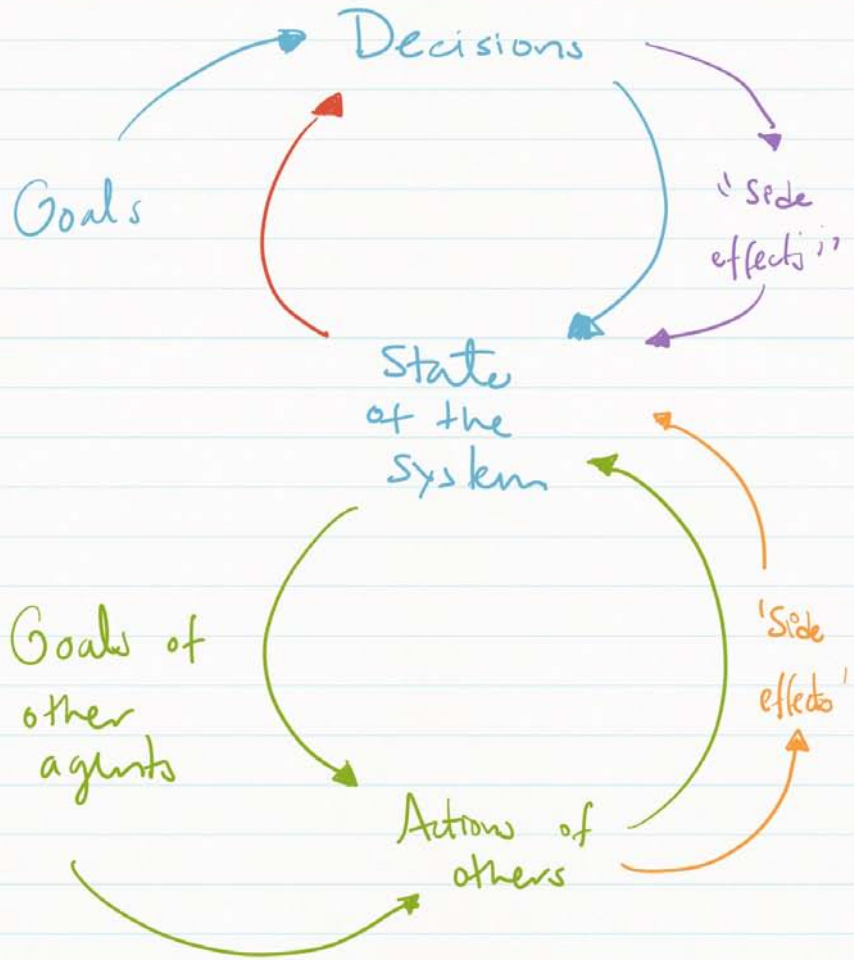
El farol problem

(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



Business Dynamics, Sterman

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

The big unsolved problems of the world result form system instabilities

Dirk Heilberg, ETH Zurich

Complejidad de los datos

Data Complexity

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

¿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 pede ser?

	Lat	Long	Indicador	...
Obs 1	#	#	#	

Implicita la fecha

	lugar	indicador	...
obs 1			
obs 2			

Idem

	Fecha	Fecha	...
lugar 1			
lugar 2			

Implicita la variable

	Ind 1	Ind 2	...
lugar 1			
lugar 2			

Implicita la fecha

	Fecha 1	Fecha 2	...
lugar 1			
Ind 1			
ind 2			
lugar 2			
Ind 1			
ind 2			

Implicita la variable

	Indicador 1		Indicador 2	
	Fecha 1	Fecha 2 ...	Fecha 1	Fecha 2
lugar 1				
lugar 2				

⋮

	Trans 1		Trans 2	
	Fecha 1	Fecha 2	Fecha 1	Fecha 2
lugar 1				
Ind 1				
Ind 2				
lugar 2				
Ind 1				
ind 2				

	lugar	Key	SubFecha 1	SubFecha 2 ...
Fecha 1	lugar 1	ind 1		
Fecha 1	lugar 1	ind 2		
Fecha 1	lugar 2	ind 1		
Fecha 2	lugar 1	ind 1		

		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			

...

⋮

Implícito el lugar

Además hay que tomar en cuenta el formato, si es abierto ono, 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.

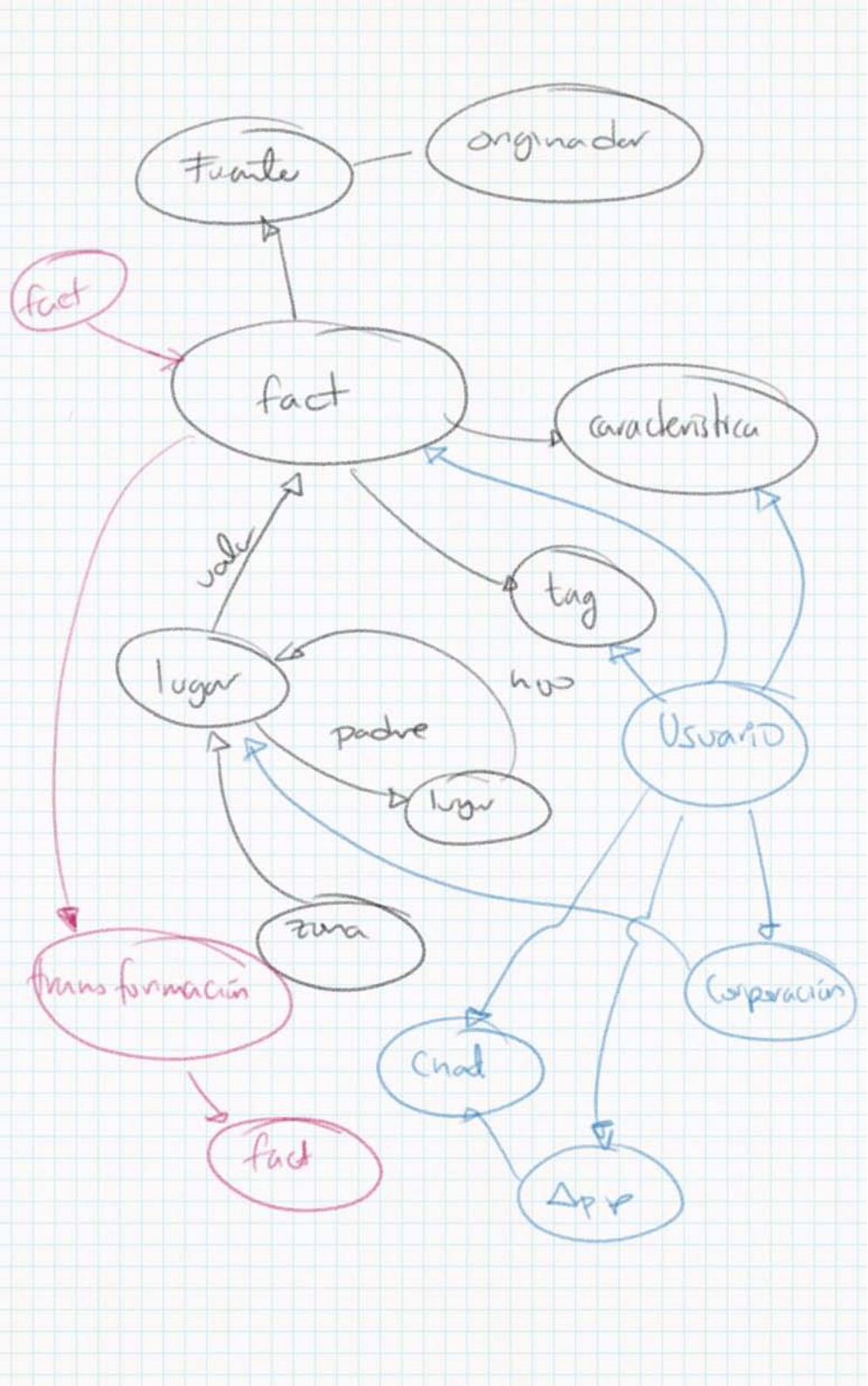
Los datos…

- 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**

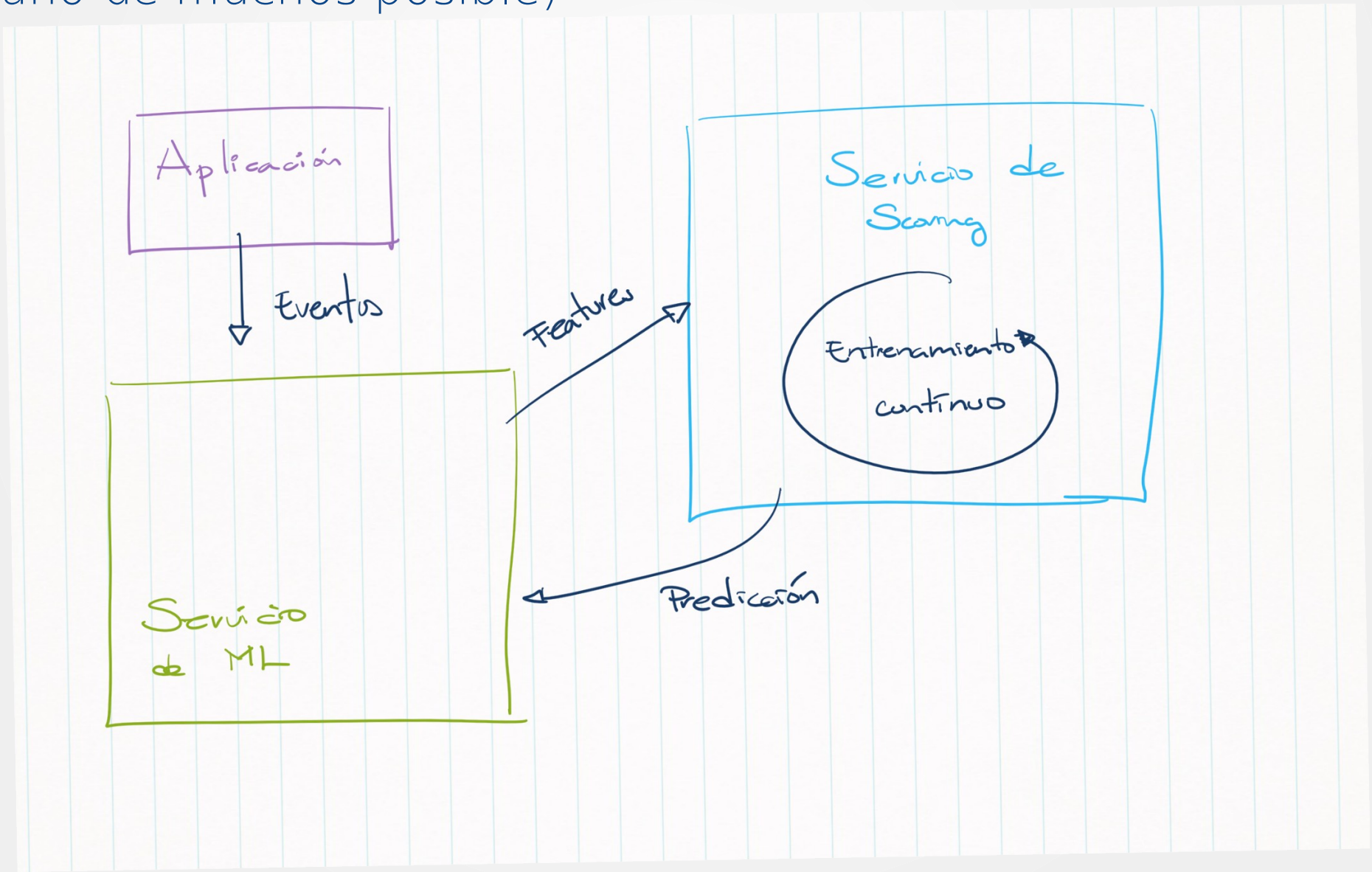
Procesos Vitales

- 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

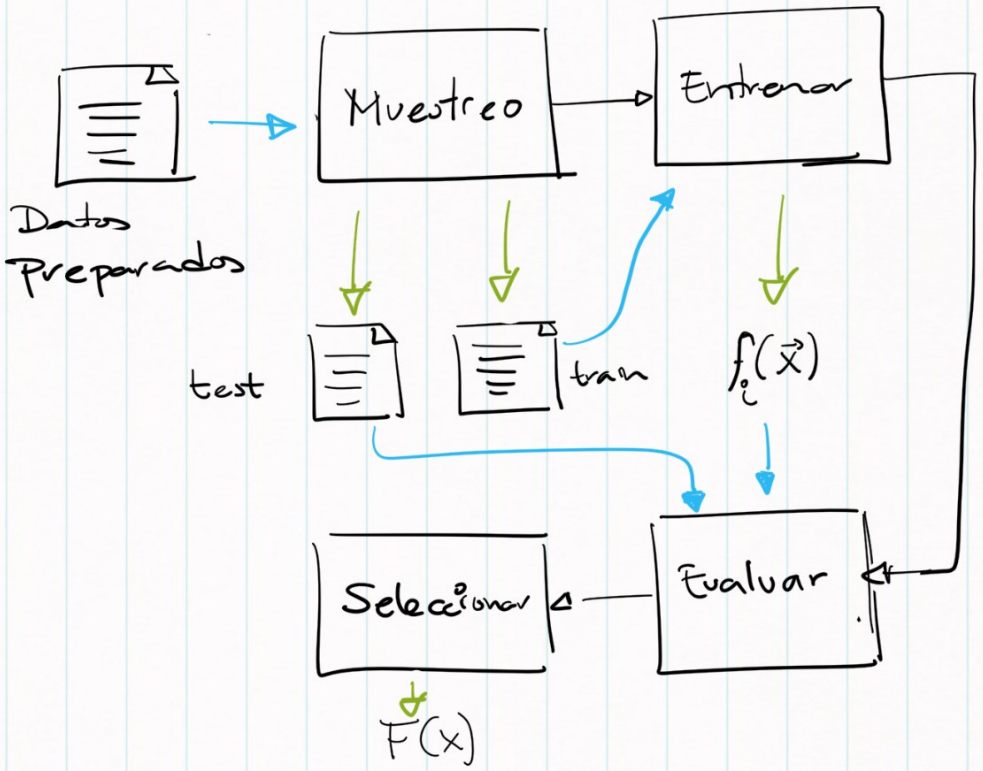
¿Cómo se ve un producto de datos?

(uno de muchos posible)



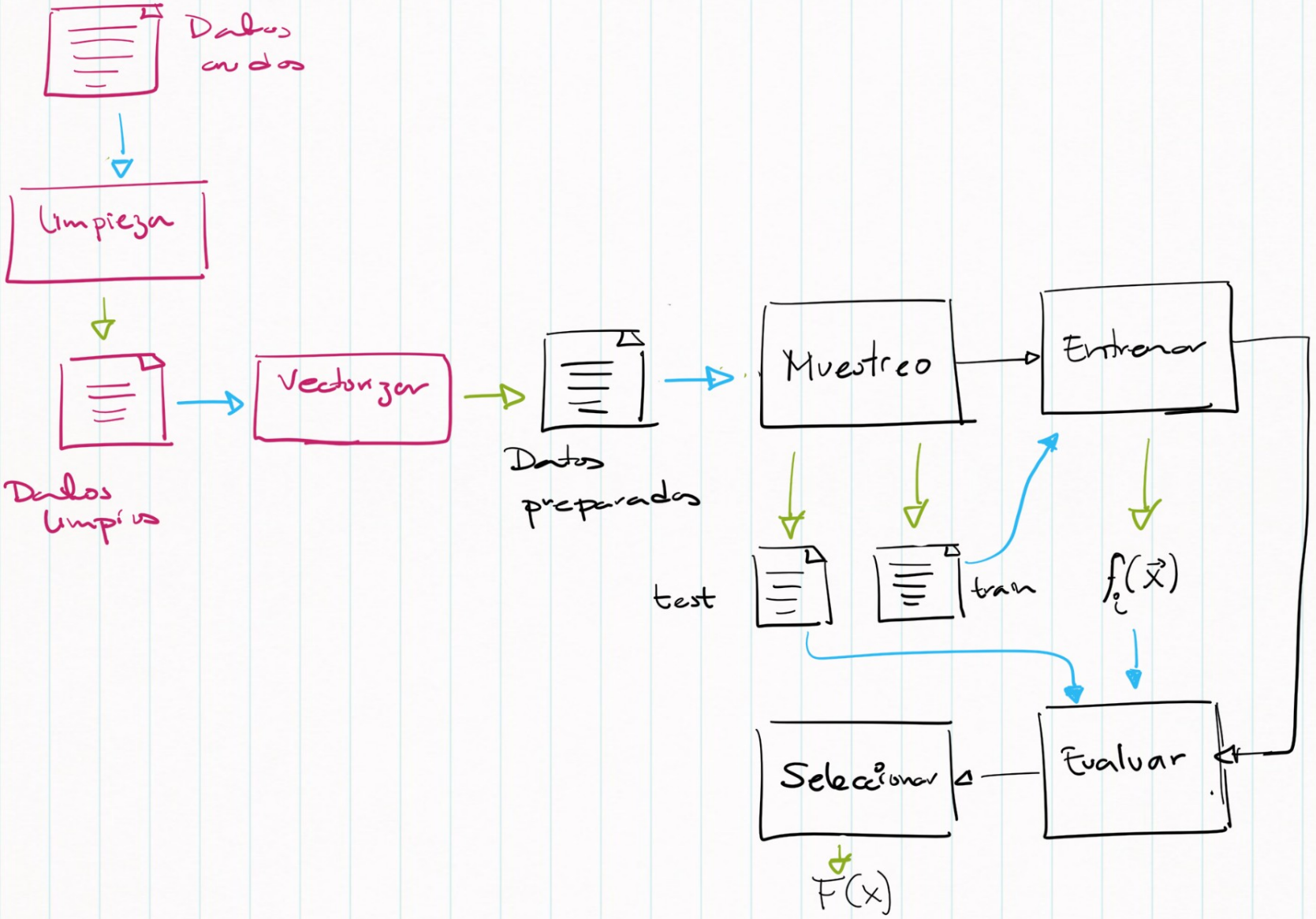
El proceso de modelar

(regularmente se hace a mano)

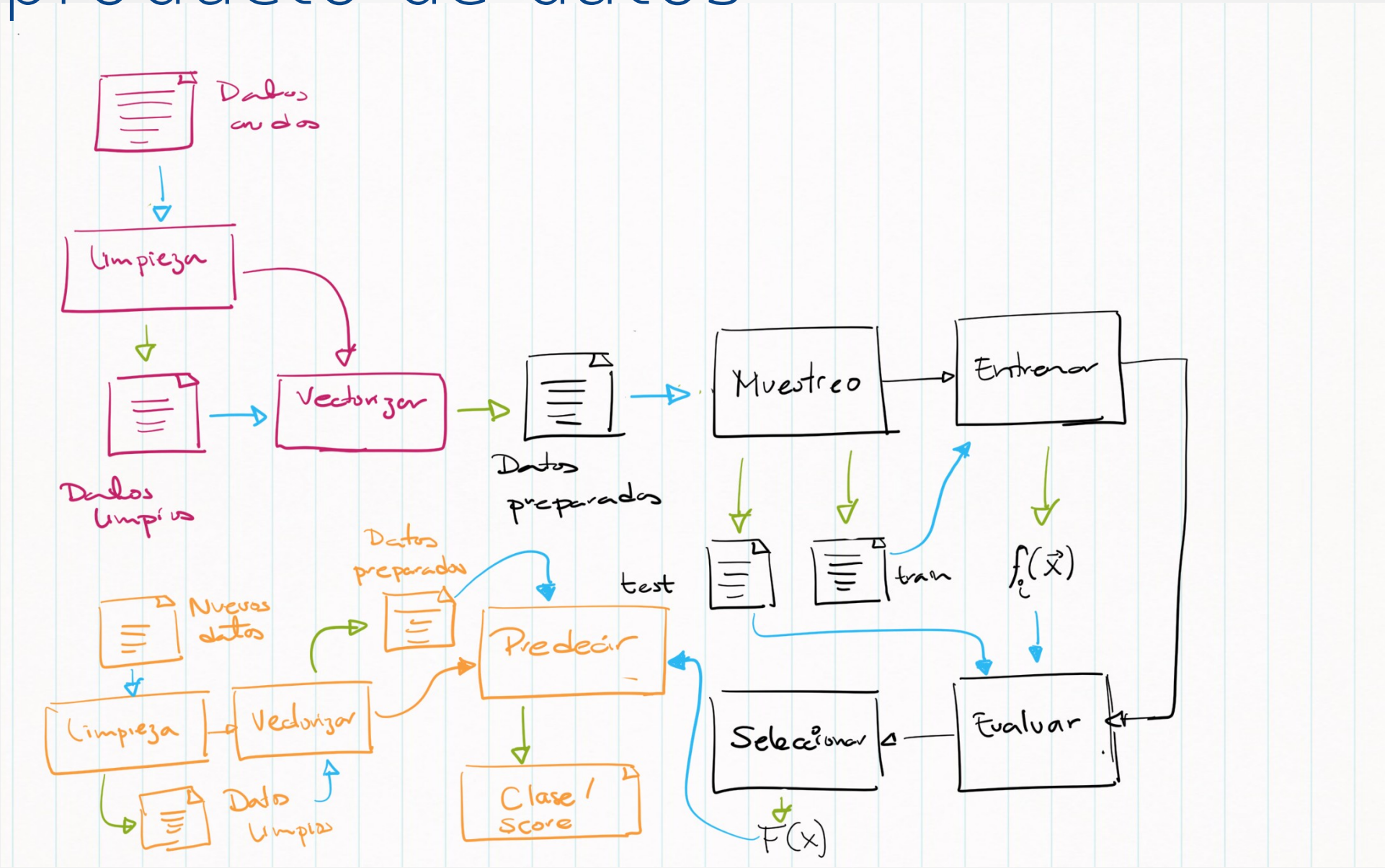


Lo que no queremos hacer…

(pero hay que hacer)



El significado de un producto de datos



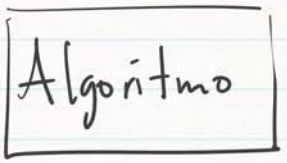
No olvidar a los usuarios

- 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

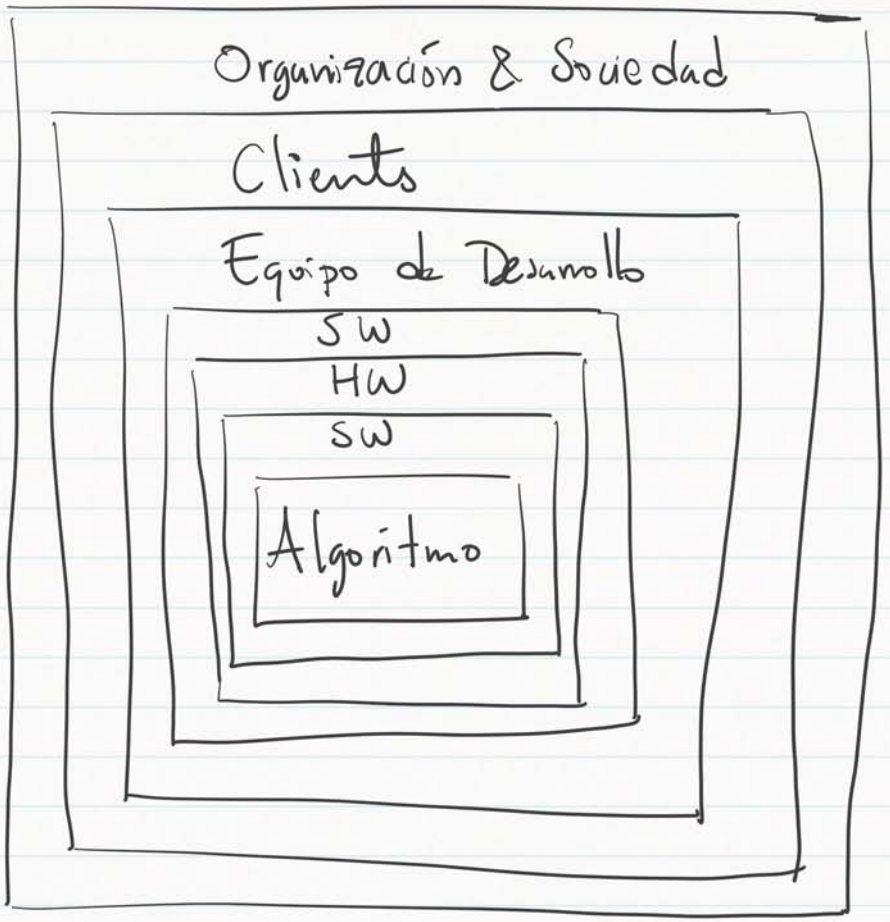
- Es muy importante tener los logs funcionando en cuanto antes
  - Servicio al cliente
  - Aumenta la inteligencia interna
    - Organizacional
    - De la "máquina"
- La captura de datos no es lo único, si se proveé de *Flight simulators*, se generan nuevos datos.

Al final el producto también es un CAS…



A Systems View of ML – Josh Bloom
https://youtu.be/i-1UmCYyzi4

Al final el producto también es un CAS…



A Systems View of ML – Josh Bloom
https://youtu.be/i-1UmCYyzi4

AI final el producto también es un CAS…

A Systems View of ML – Josh Bloom
https://youtu.be/i-1UmCYyzi4

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

Reality Check

- Estructurales
  - CAS dentro de CAS
  - CAS formados por **humanos**

Reality Check

- Procesos
- Generar datos
- Capturar datos
- Procesar datos
  - *Estructurados, no estructurados, chicos, grandes, variedad, velocidad*
- Explicar datos
  - *No provienen de experimentos, valor, comprobación, hipótesis, bayes, realidad*
- Analizar datos
- Presentar datos
  - *Modelado de los datos, Representación de conocimiento, "desbloqueo"*

Reality Check

- Técnicos
- ¿Dónde hago esto?
  - *Infraestructura, Software, ¿Hardware o nube?, Algoritmos*
- ¿Qué hago con esto?
  - *Storytelling*,
  - Simulación (AB, Discrete),
  - *System Dynamics*,
  - Network theory

¿Preguntas?

y quizá posibles respuestas...

¡Gracias por su tiempo!

adolfo@opi.la

adolfo.deunanue@itam.mx

@nano_unanue