

# **Ciencia de datos y problemas sociales**

**Aprendizajes**

# ¿Quién soy?

- Adolfo Javier De Unánue Tiscareño
- Ph.D. en Física Teórica
  - Cosmología y Mecánica Cuántica
- 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 :)
- Transhumanista

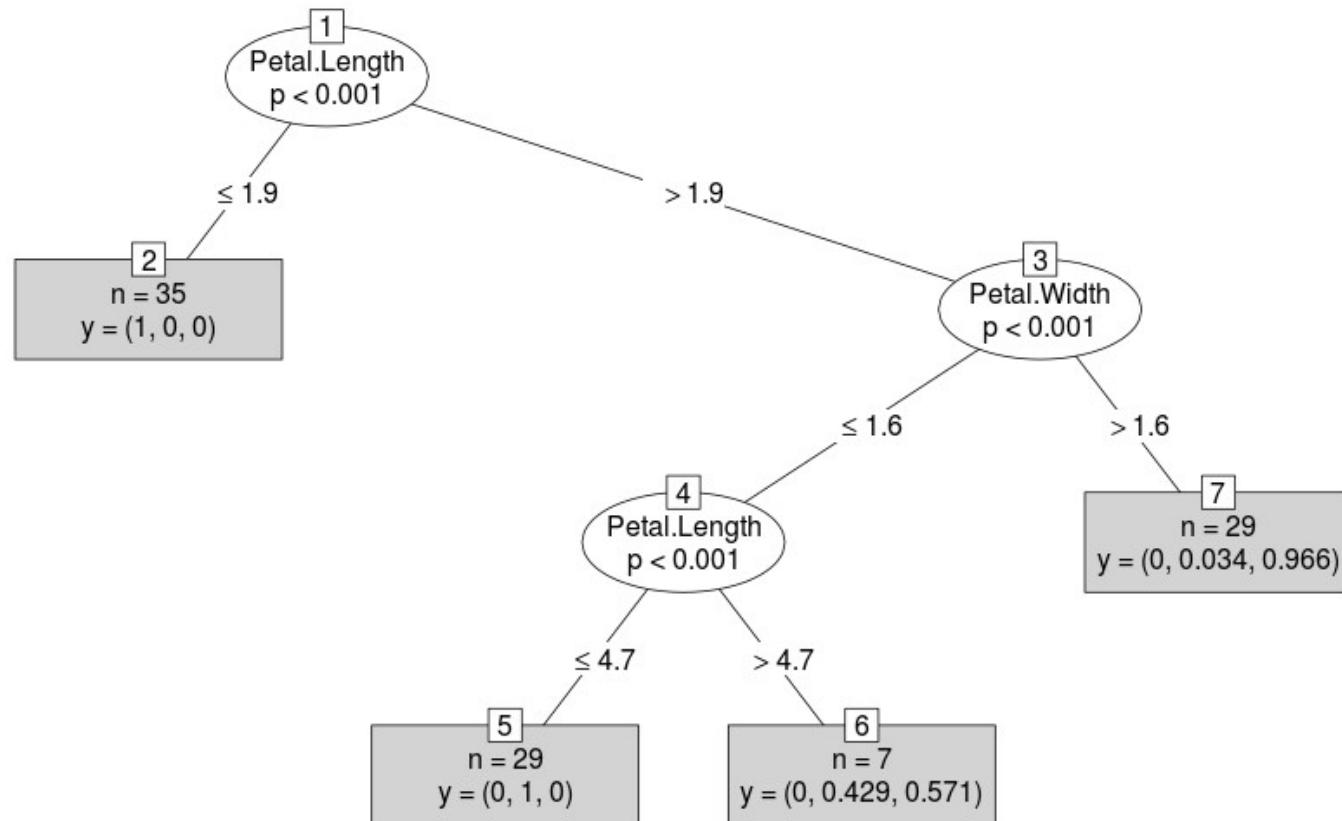
# **¿De qué va esta plática?**

# **Ciencia de datos**

# 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:





# NAILED!

xvii-coneest-2015

22 de Septiembre, 2015

# La mayoría de las personas piensan que hago lo siguiente:

- Bueno, quizá no todas piensen eso y menos (obviamente) a ese nivel de detalle
- Además, no trabajo con el *iris dataset*...
- La realidad es más divertida que eso que acabamos de ver

# En realidad ...

- La ciencia de datos, tiene que ver, con...

# Datos

- i.e. es fenomenológica, empírica
- Por esto las personas la confunden con *Big data*, pero esto lo veremos más adelante

# pero, ¿Para qué?

- Toma de Decisiones Racionales
  - i.e. tomar la mejor decisión basada en la evidencia (datos) disponible.
- Aumento de Inteligencia (AI)
- Aplicar método científico a la toma de decisiones
  - Se ha intentado desde los 40s, al parecer ahora si está teniendo impacto...

# **Sistemas Complejos Adaptativos**

**CAS**

# ¿Por qué ahora?

## Cómputo

Almacenamiento, RAM, CPU

# ¿Qué incluye?

Minería de datos  
Aprendizaje de Máquina  
Big data

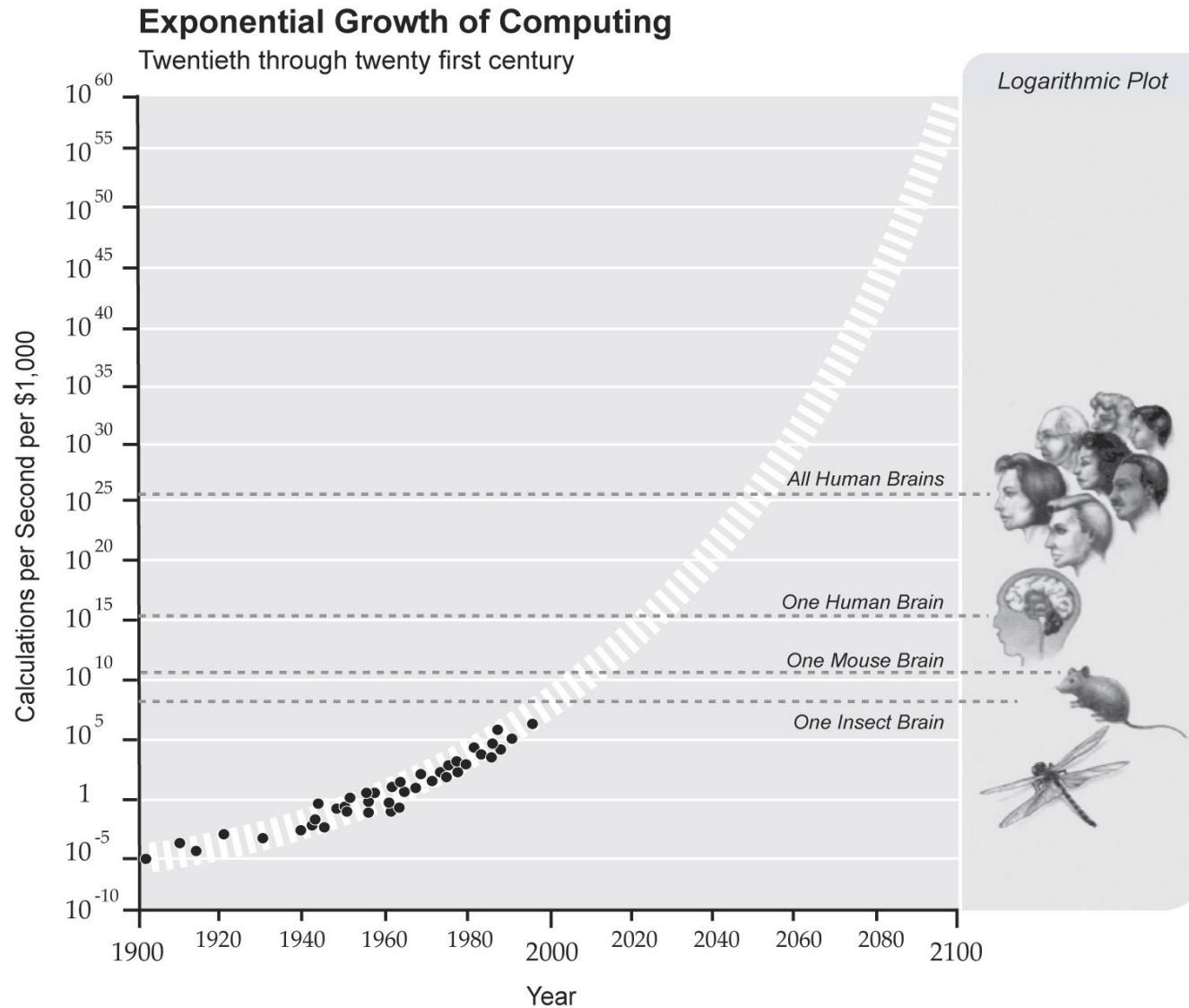
...

(put next buzz word here)

# **En el gran esquema de las cosas**

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

# En el gran esquema de las cosas



# **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*

# Quiero enfatizar esto

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

# Ciencia de datos

- Uno de los principales retos de la ciencia de datos es tratar con la *complejidad de los datos*.
- Su objetivo es crear *Productos de datos*

# Data Complexity

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

NOTA: Tamaño representa la dificultad

# Producto de datos

- ¿Qué es?
- Sistema continuo
  - Todas las partes: reentrenamiento, recalibración, adquisición, movimiento de datos, transformación, limpieza, etc.

# **La segunda parte del título...**

**¿Qué CAS afecta a billones de personas en el mundo\* y  
necesita de toda nuestra capacidad para mejorar?**

**\*otro CAS**

# **Gobierno y Sociedad**

Esto es muy amplio y muy ambiguo

¿A qué me refiero específicamente?

**Gobierno** es aquel sistema encargado de resolver, usando una **teoría** (la cual especifica variables a medir y sus relaciones) **problemas sociales**, mediante la asignación de recursos en áreas geográficas específicas.

No es una idea nueva, se intentó por primera vez en Chile, durante el gobierno de Salvador Allende:

**Cybersyn**

# Breviario Filosófico

Todo sistema interactúa con el mundo.

=> Es perceptible, deja su huella, se “siente”.

No importa (al principio) la calidad de la sensación, lo que importa es la consistencia.

# Problemas que quiero atacar

¿Corrupción?

¿Asignación de Presupuesto?

¿Capacidad de gasto?

Transparencia

Accountability

Medición de ROI

# ¿Por qué no ha funcionado antes?

- Nos hemos 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

# ¿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 siempre 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 loops*
- No se puede ignorar los *delays* en el sistema.

# Ejemplos abundan

- Algunos juegos:
  - *Beer game*
  - El faro
  - *Fishing game*
- 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.

# **The big unsolved problems of the world result from system instabilities**

Dirk Heilberg, ETH Zurich

# ¿Open data?

# ¿Open Government?

(¬\_¬) ... I don't think so

Solos, son pensamiento mágico

# OPI

## **Una aproximación a la solución**

# Ajustando un poco más

- Hay que establecer fases de aproximación
- Aumento de Inteligencia, no Inteligencia Artificial
  - *Human in the loop*: Plantea el problema, usa datos
- Capturar primero todo lo que es open data
  - Muchísimos retos ya en esta fase: ¿Cuál es el universo? ¿Todos son valiosos?
- Lo importante, estar conectados a su generación de datos internos

# Ajustando un poco más

- Queremos que el producto se use para tomar decisiones
- Hay poca actualización de los datos (el problema de la derivada de Euler)

# ¿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?

	Lat	Long	Indicador	...	Implicita la fecha
Obs 1	#	#	#	...	
Obs 2					

	Lugar	Indicador	...	I dem
Obs 1			...	
Obs 2				

	Fecha	Fecha	...	Implicita la variable
Lugar 1			...	
Lugar 2				

	Ind 1	Ind 2	...	Implicita la fecha
Lugar 1			...	
Lugar 2				

	Fecha 1	Fecha 2	...	Implicita la variable
Lugar 1			...	
Indi. 1				
Indi. 2				
Lugar 2				
Indi. 1				
Indi. 2				

	Indicador 1			Indicador 2	
	Tech1	Tech2	...	Tech1	Tech2
Lugar 1					
Lugar 2					

:

	Tran1		Tran2	
	Tech1	Tech2	Tech1	Tech2
Lugar 1				
Ind 1				
Ind 2				
Lugar 2				
Ind 1				
Ind 2				

	Lugar	Key	SubTech1	SubTech2	...
Tech1	Lugar 1	Ind 1			
Tech1	Lugar 1	Ind 2			
Tech1	Lugar 2	Ind 1			
Tech2	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	

	Fech1	Fech2	Fech3
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...

- 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?

# Los datos...

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

# El problema de la tecnología

- No empezar arriba
  - ➔ i.e. Vamos a usar Hadoop en un clúster en GoogleCloud...
- ¿Cuánto esperamos crecer?
- No agregar complejidad creada, a la complejidad implícita del problema.

# 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.

# No olvidar a los usuarios *indirectos*

→ ¿Cómo incorporar a la sociedad?

→ Preguntarle:

→ Encuestas: Muy caras, Mucha certeza estadística, un sólo punto.

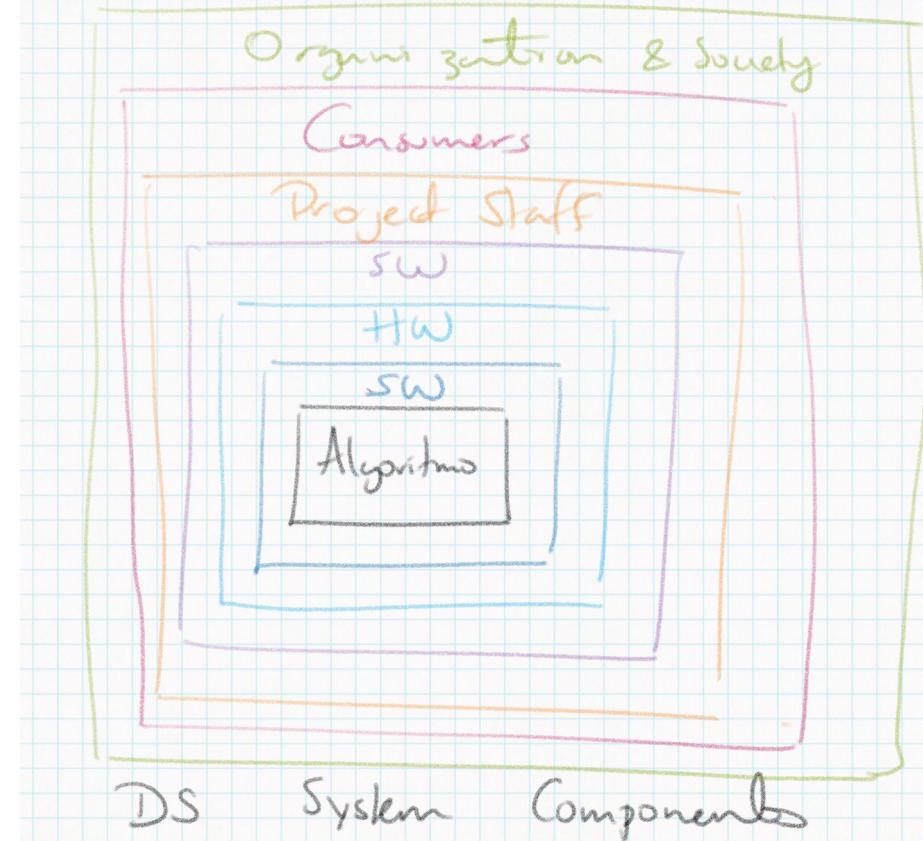
→ Diagnósticos participativos: Se pueden masificar, ejecutar continuamente, baratas.

→ Tenemos una aplicación móvil para esto.

→ Medirla: Sensores en todos lados.

→ Esto trae otros tipos de problemas

Al final el producto también es un CAS...



Función de optimización multiobjetivo

**¿Qué quieres optimizar?**

# ¿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

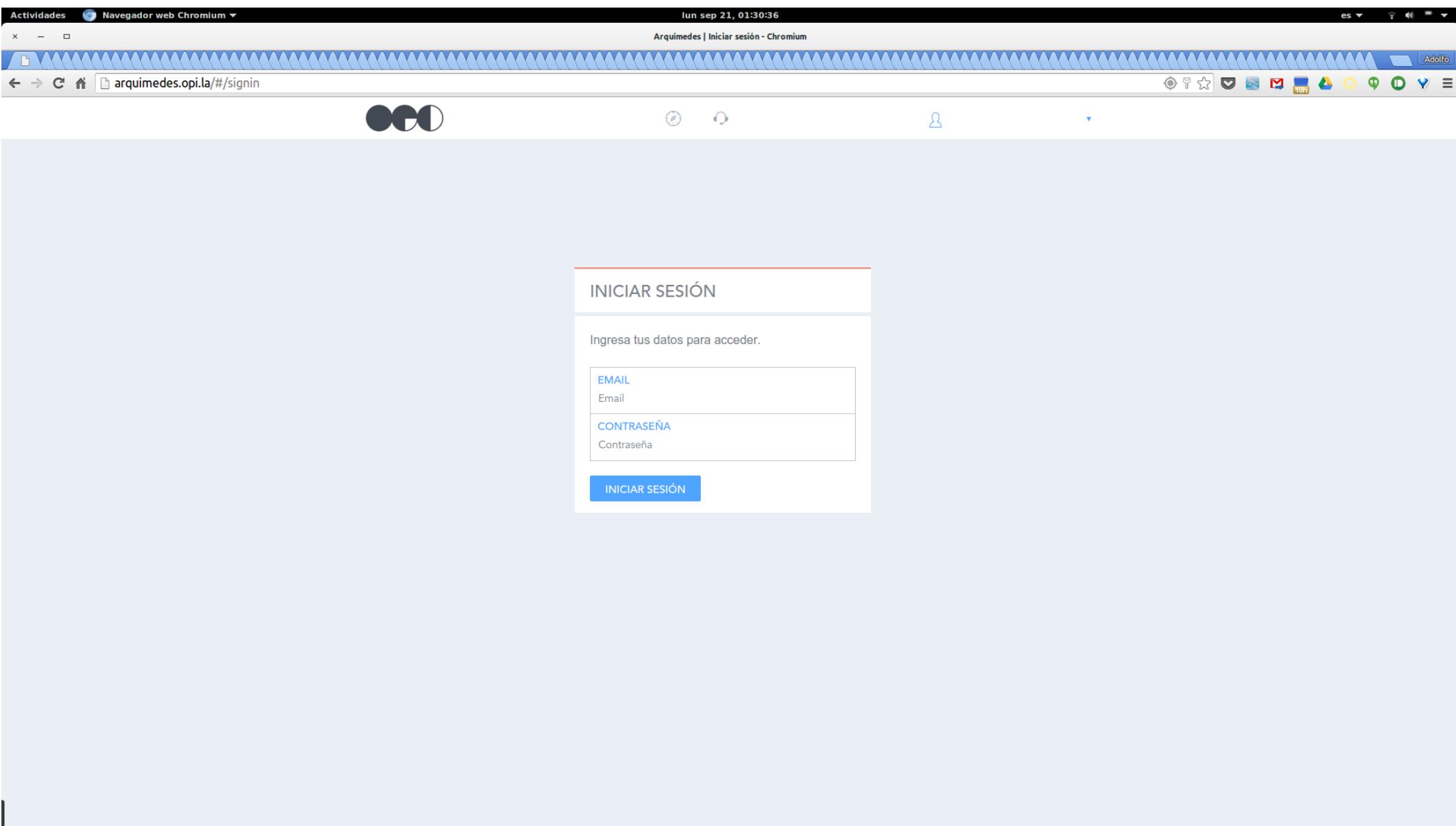
## → Clientes

→ Valor directo, usabilidad, explicabilidad, “accionabilidad”

## → Sociedad

- Valor indirecto

# Arquímedes



# Arquímedes

Actividades Navegador web Chromium es ▾

lun sep 21, 01:27:37

Arquímedes | Explorador - Chromium

arquimedes.opi.la/#/

OEI EXPLORADOR SOPORTE ADOLFO DE UNANUE ▾

Introduce un lugar

The map displays the following labeled locations:

- North America: California, Arizona, New Mexico, Colorado, Kansas, Missouri, Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Georgia, Tennessee, Kentucky, Virginia, North Carolina, South Carolina, West Virginia, Ohio, Indiana, Michigan, Wisconsin, Minnesota, Iowa, Missouri, Kansas, Nebraska, Wyoming, Montana, Idaho, Nevada, Utah, New Mexico, Texas, Oklahoma, Kansas, Missouri, Illinois, Indiana, Michigan, Ohio, Pennsylvania, New Jersey, New York, Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, Maine.
- Caribbean: Nassau, The Bahamas, La Habana, Cuba, George Town, Cockburn Town, Port-au-Prince, Kingston, San Juan, Basseterre, Roseau, Castries, Kingstown, Scarborough, Caracas, Ciudad Bolívar, Georgetown, Paramaribo, Quito.
- Mexico: Baja California, Sonora, Chihuahua, Coahuila de Zaragoza, Durango, Estados Unidos Mexicanos, Zacatecas, San Luis Potosí, Nuevo León, Tamaulipas, Veracruz de Ignacio de la Llave, Jalisco, Guanajuato, Michoacán, Colima, Nayarit, Sinaloa, Tabasco, Chiapas, Belmopán, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, Panama.
- Central America: Provincia de Galápagos.

# Arquímedes

# Arquímedes

Actividades Navegador web Chromium ▾ lun sep 21, 01:28:58 Arquímedes | Explorador - Chromium es ▾ Adolfo

arquimedes.opi.la/#/

 EXPLORADOR SOPORTE ADOLFO DE UNANUE

Aguascalientes 

VER LUGAR

REVISAR CATÁLOGO

produ|

Volumen de la producción de carne en canal de bovino  
INEGI; BIINEGI, 1994 - 2014

Valor de la producción de trigo grano  
INEGI; BIINEGI, 1994 - 2014

Volumen de la producción de leche de bovino  
INEGI; BIINEGI, 1994 - 2014

Valor de la producción de tomate rojo ( jitomate )  
INEGI; BIINEGI, 1994 - 2014

Valor de la producción de pastos  
INEGI; BIINEGI, 1994 - 2014

Valor de la producción de cera en grefña



# Arquímedes

Actividades Navegador web Chromium ▾ Lun sep 21, 01:29:16 Valor de la producción de tomate rojo ( jitomate)... en Aguascalientes - Chromium

arquimedes.opi.la/#/detalle/2/1048619

Adolfo

Valor de la producción de toma... x Aguascalientes x VER DETALLE

Puedes agregar un concepto a tu búsqueda REVISAR CATÁLOGO

UNIDAD: MILES DE PESOS FECHA: 2011

Jerez Ojo Caliente Villa Gonzalez Ortega Ahualulco

Colotán Rincón de Tomor Loreto

Tlaltenango de Sánchez Román Arteaga San Francisco de los Romo Ojuelos de Jalisco

Villa Hidalgo Encarnación de Díaz, Jalisco San Felipe

Calvillo Teocaltiche

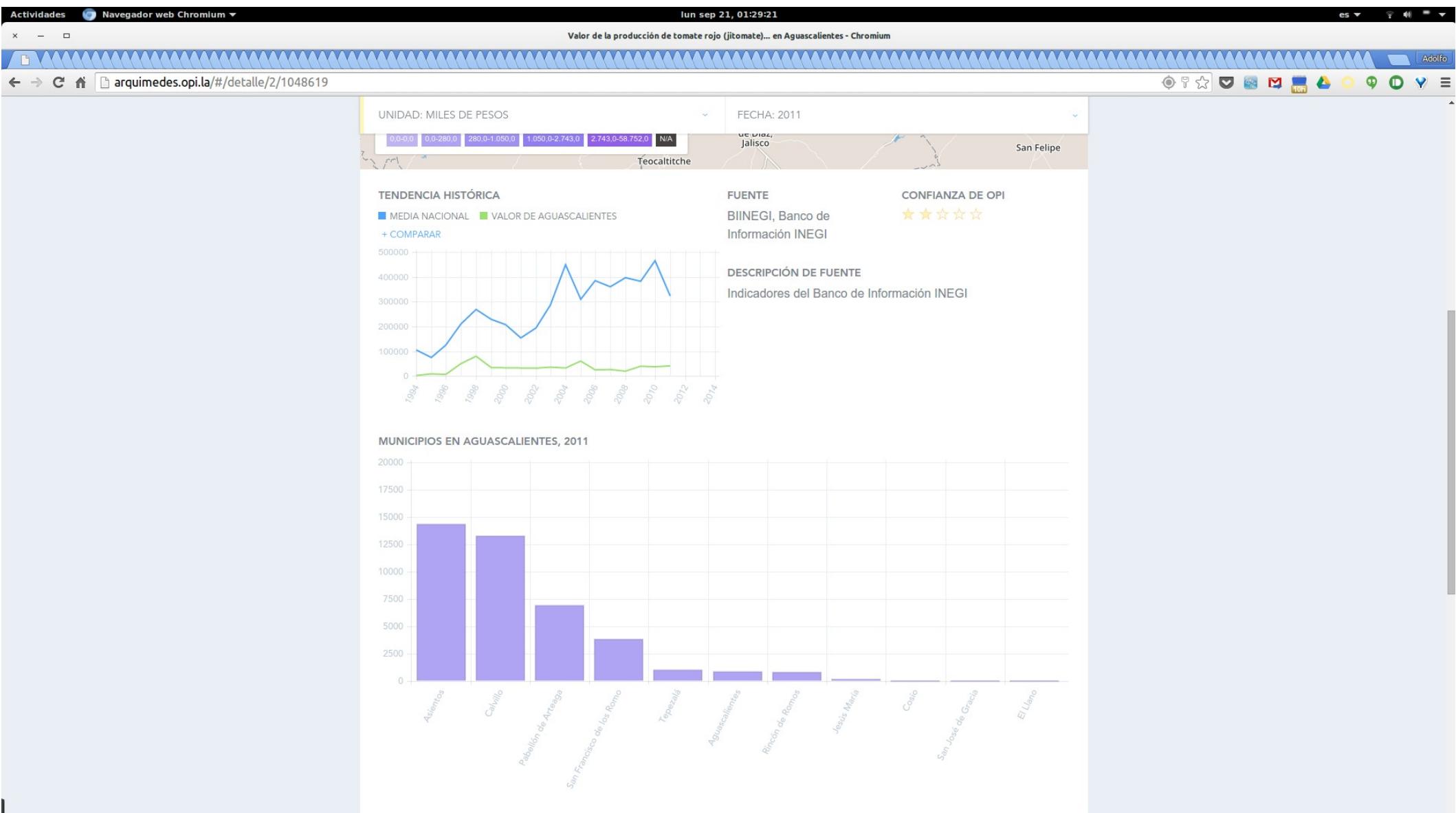
0.0-0.0 0.0-280.0 280.0-1.050.0 1.050.0-2.743.0 2.743.0-58.752.0 N/A

TENDENCIA HISTÓRICA FUENTE CONFIANZA DE OPI

MEDIA NACIONAL VALOR DE AGUASCALIENTES BIINEGI, Banco de Información INEGI 5★

+ COMPARAR DESCRIPCIÓN DE FUENTE Indicadores del Banco de Información INEGI

# Arquímedes



# Arquímedes

Actividades Navegador web Chromium ▾ lun sep 21, 01:30:08  
Cruce de Valor de la producción de tomate rojo (jitoma... y Suma de todos los delitos de modalidad homicidi... en Aguascalientes - Chromium

arquimedes.opi.la/#/cruce/2?var1=1048619&var2=371799

Adolfo

Suma de todos los delitos de m... Valor de la producción de toma... Aguascalientes VER CRUCE

ADOLFO DE UNANUE

Elimina una etiqueta para realizar una búsqueda distinta

MILES DE PESOS 2011 HOMICIDIOS ABRIL, 2015

MUNICIPIOS EN AGUASCALIENTES, MÉXICO

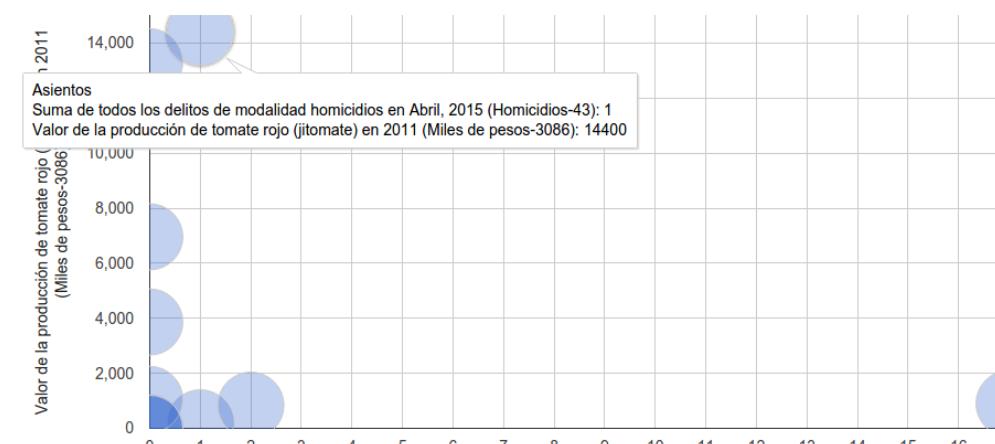
Asientos  
Suma de todos los delitos de modalidad homicidios en Abril, 2015 (Homicidios-43): 1  
Valor de la producción de tomate rojo (jitomate) en 2011 (Miles de pesos-3086): 14400

Valor de la producción de tomate rojo (Miles de pesos-3086)

Suma de todos los delitos de modalidad homicidios en Abril, 2015 (Homicidios-43)

VALOR DE LA PRODUCCIÓN DE TOMATE ROJO ... INEGI, Banco de Información INEGI

SUMA DE TODOS LOS DELITOS DE MODALIDAD... OPI, Incidencia delictiva



# Arquímedes

- Problemas de esta versión
  - No aprovecha las relaciones entre las variables
  - Calcula las variables derivadas al cargar, no al solicitarla
  - Destruye la única relación (la fuente).
  - Difícil de mantener
  - Confunde el *fact* con su representación y su visualización.
- Todos estos problemas se están solucionando para la nueva versión (próximamente).

# Arquímedes

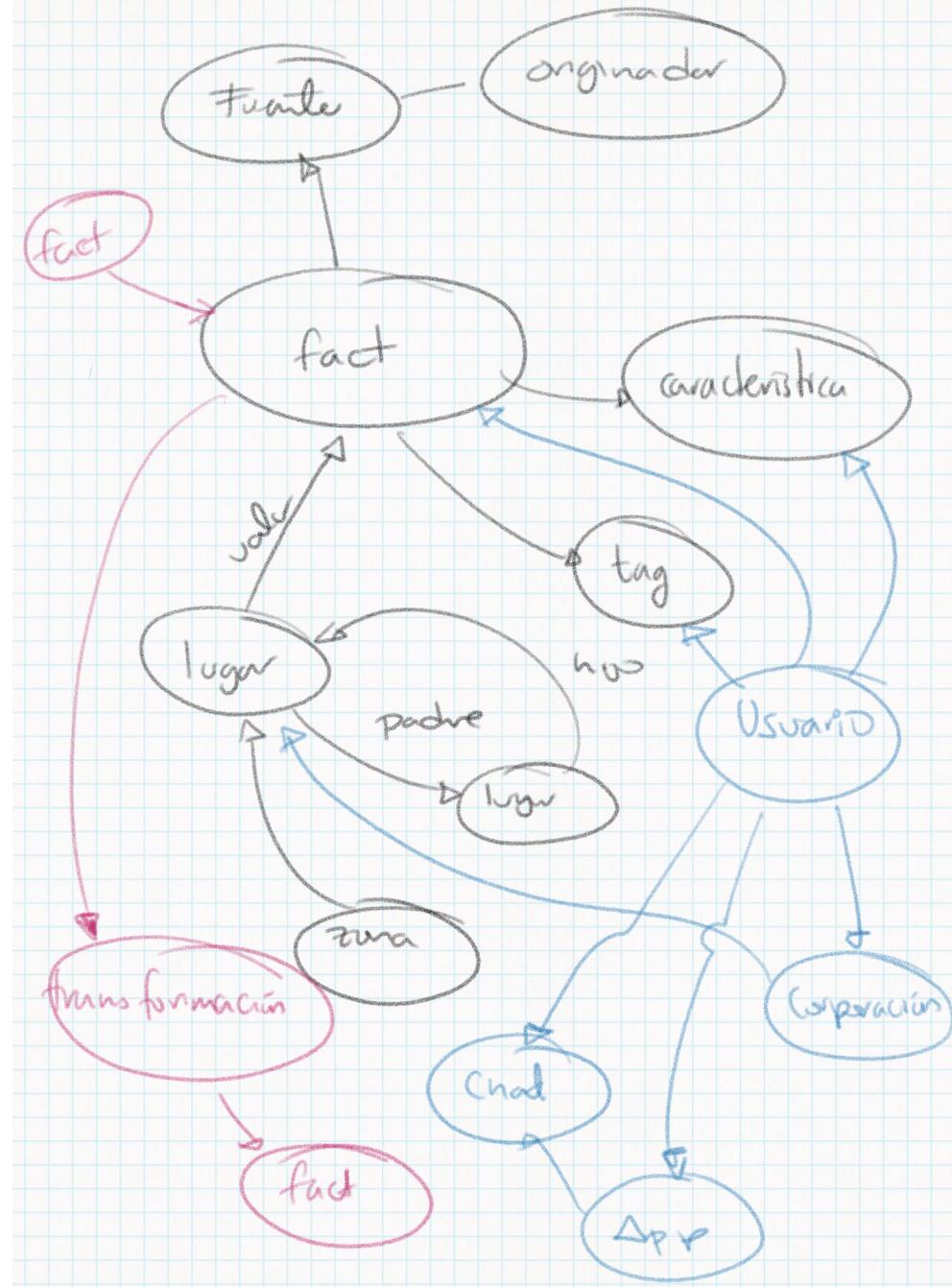
- En las siguientes versiones se agregará
  - Integración con MS Excel.
  - Capacidad de guardar búsquedas, definir proyectos, definir bases de datos propias, compartir descubrimientos.
  - Cruces entre mayor número de variables.
  - Etc.

# Arquímedes

- Resumiendo los principales retos son:
  - El usuario debe de **encontrar**, no buscar.
  - Carga eficiente: Pipeline rastreable, con *data lineage*, actualizable, etc.
  - Eliminar la fragmentación cartesiana de la realidad
  - Eliminar complicación, preservando complejidad
  - Ayudar al usuario con las relaciones, en lugar de destruirlas al cargar.
    - Muy difícil de lograr en una base de datos relacional

# Arquímedes\*

## Grafos



# 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.
- Esta idea se debe de utilizar en el gobierno.
- 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

adolfo.deunanue@itam.mx

@nano\_unanue