# Sueña la gente con algoritmos eléctricos?

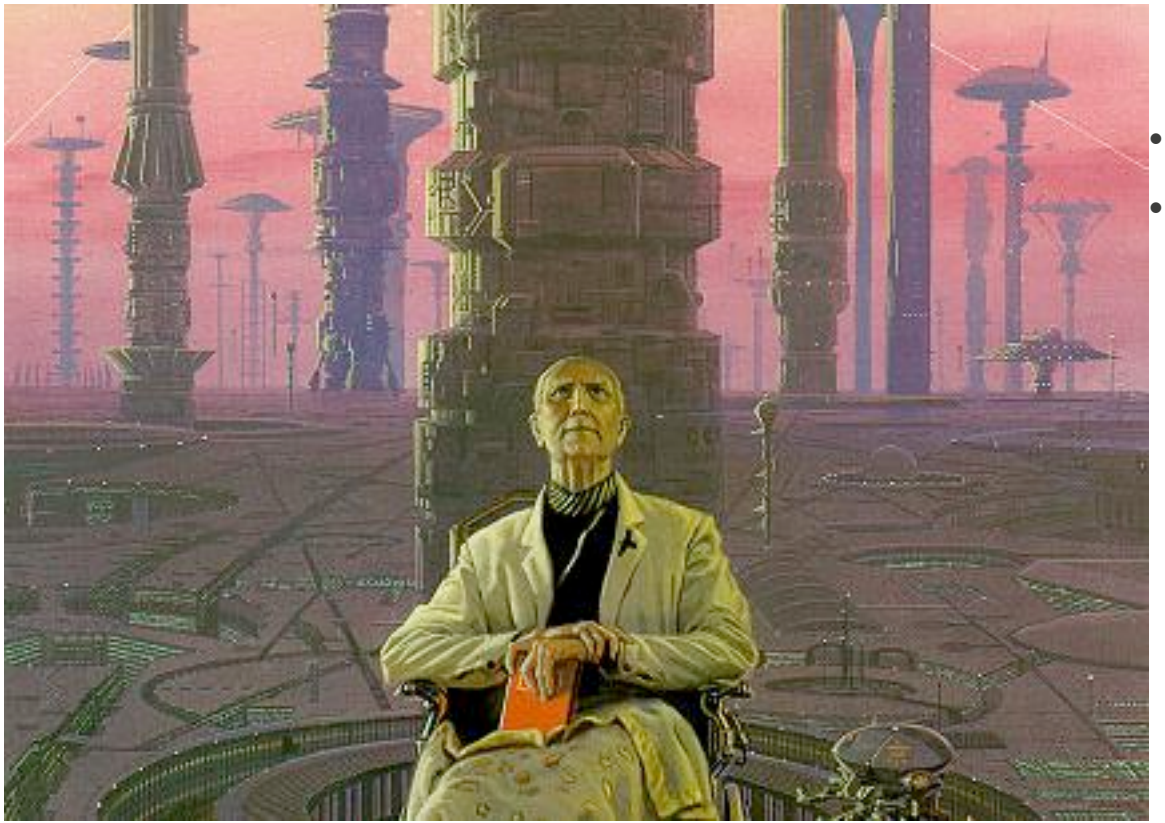
Daniel Collico Savio
Teradata
@dcollico

# Índice

1. Isaac Asimov, Hari Seldon y la Psicohistoria
2. "Esto yo ya lo vi, esto ya lo escuché".
3. Definición del problema: insights vs interpretaciones
4. Autos, jets, aviones, barcos (y Telcos): IoT
5. Machine learning
6. Cómo Neux procesó la data de #PanamaPapers

# Psicohistoria

# Isaac Asimov: “Fundación e Imperio”



- Hari Seldon
- Apariciones fugaces en una bóveda cada tantos siglos, contando “cómo van las cosas”.

# Big Data según Asimov

1. Se necesita mucha data para tener buenos resultados
2. Necesidad de altísimo proder de procesamiento.
3. Los modelos predictivos simples pueden ser refinados.
4. Uso de porcentajes para indicar precisión del modelo.
5. Intervalos de confidencia
6. Predicciones sobre individuos son inciertas.
7. Predicciones sobre el futuro lejano son difusas.

# Hari Seldon sobre los ciclos de la historia

» Durante siglos, la civilización Galáctica se ha estancado y ha declinado, aunque sólo unos pocos se dieron cuenta de ello. Pero ahora, al fin, la Periferia se está desligando y la unidad política del imperio se ha quebrantado. En algún punto de estos cincuenta años pasados, los historiadores del futuro trazarán una línea imaginaria y dirán:

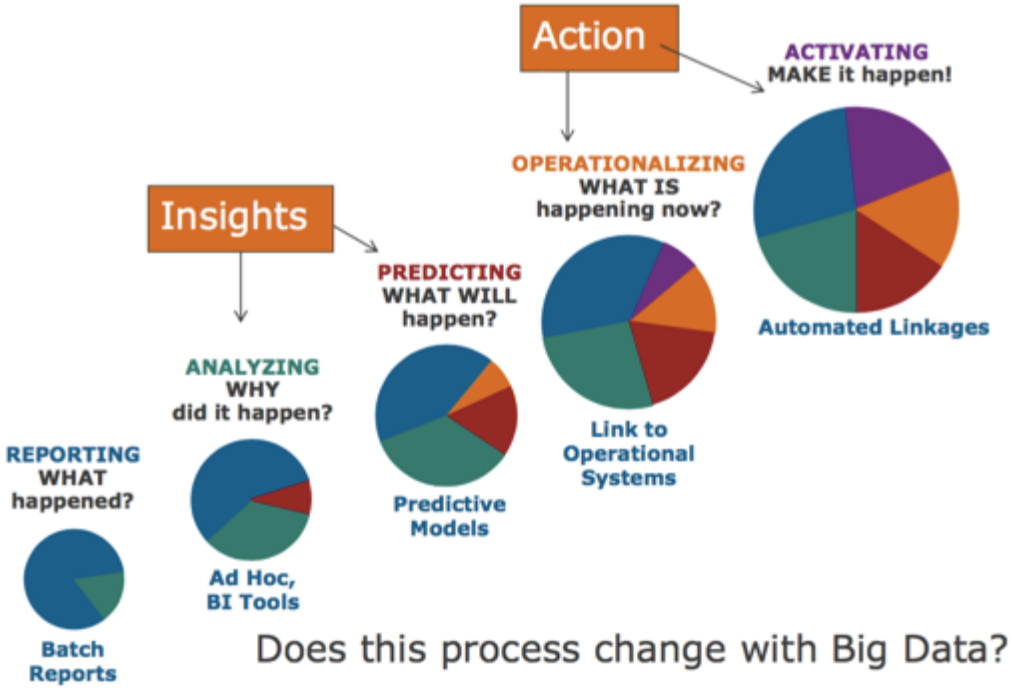
"Esto señala la Caída del imperio galáctico."

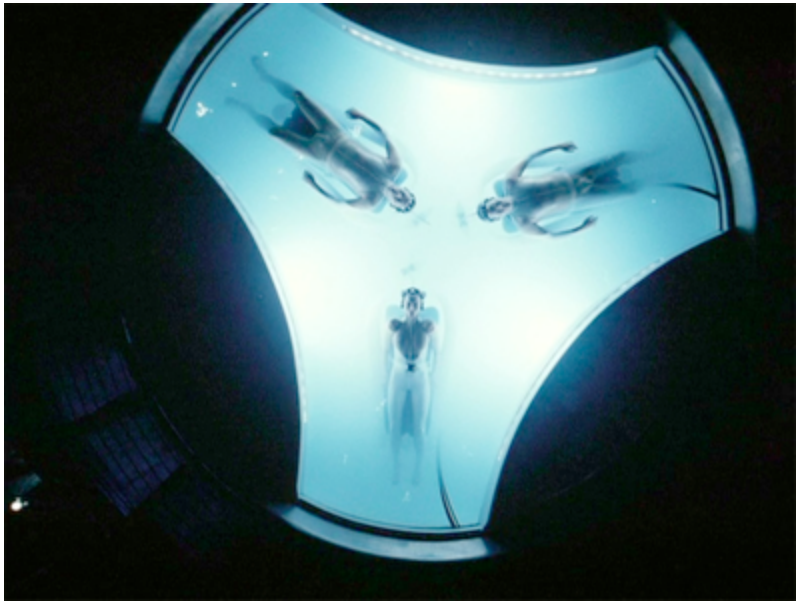
» Y tendrán razón, aunque casi ninguno reconocerá esta Caída durante muchos siglos.

» Y después de la Caída sobrevendrá la inevitable barbarie, un período que, según dice nuestra psicohistoria, debería durar, bajo circunstancias normales, otros treinta mil años. No podemos detener la Caída. No deseamos hacerlo, pues la cultura del imperio ha perdido toda la vitalidad y valor que había tenido. Pero podemos acortar el período de barbarie que debe seguir reduciéndolo hasta sólo un millar de años.

https://www.balabit.com/blog/7-data-science-principles-introduced-in-asimovs-foundation/

# Predicción => acción: PK Dick y los precogs



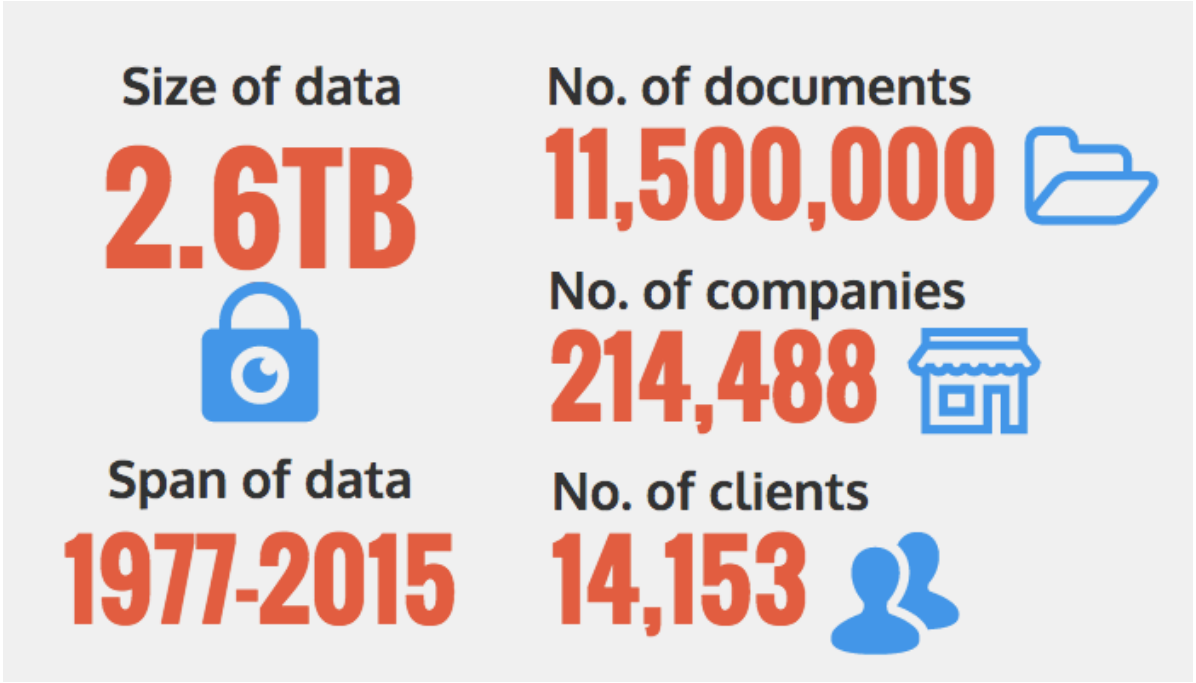



Does this process change with Big Data? No.

# Esto yo ya lo ví.


BIG
DATA
Valor
Volume
Velocity
Variety

# Es realmente #PanamaPapers un asunto de Big Data?




- Falta el requisito de "velocidad".
- Es más bien un gran proyecto de Data Science.
- Sin duda agregó la dimensión del "valor".



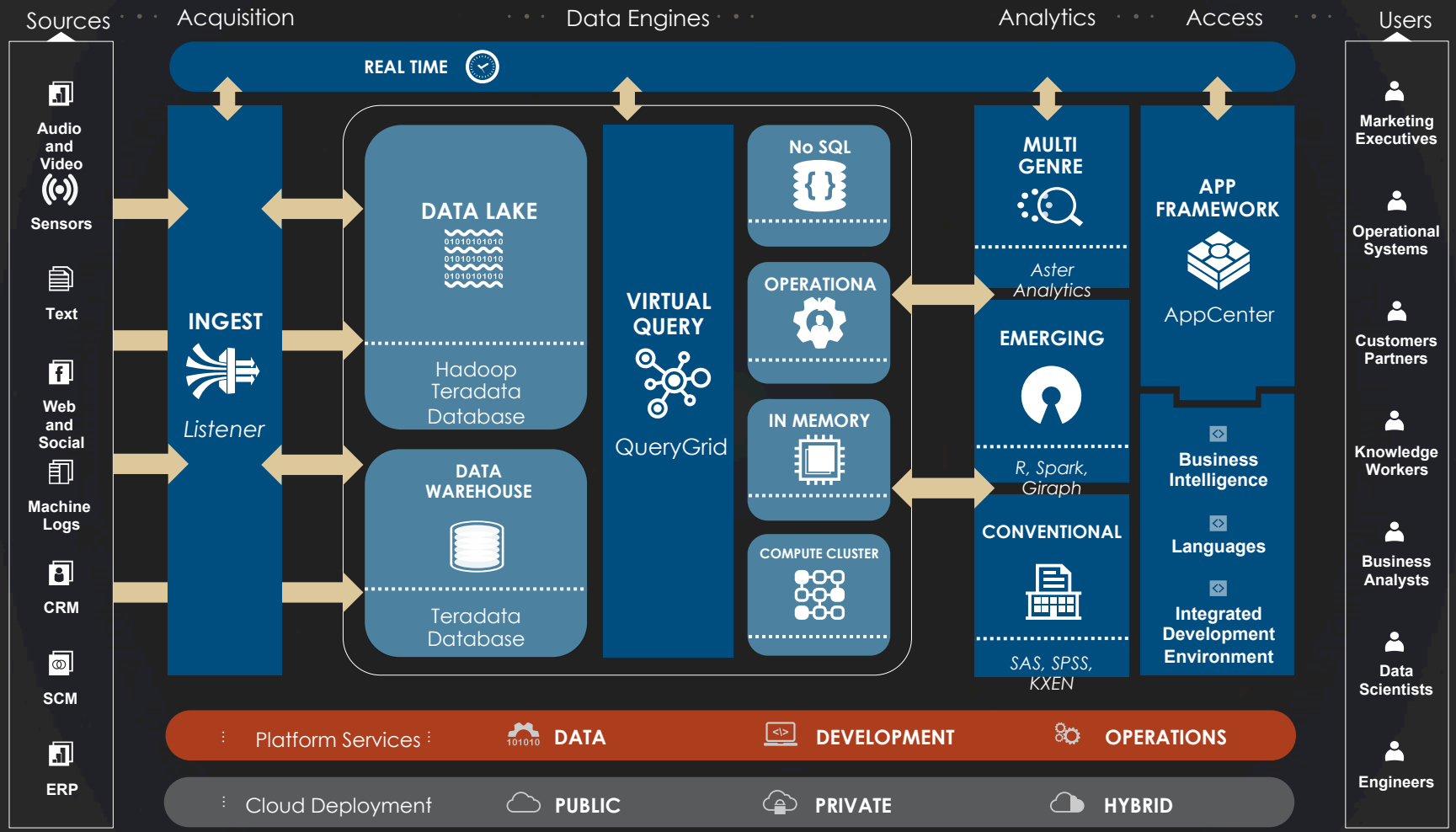
TERADATA.

# El “state of the art” en analytics

1. No se habla de “Big Data” o de “BI”.
2. Analytics es todo.
   Multigénero, a escala, data integrada, estructurada o no.
   Statistical, Predictive, Text, Graph, Pattern, Path Analysis.

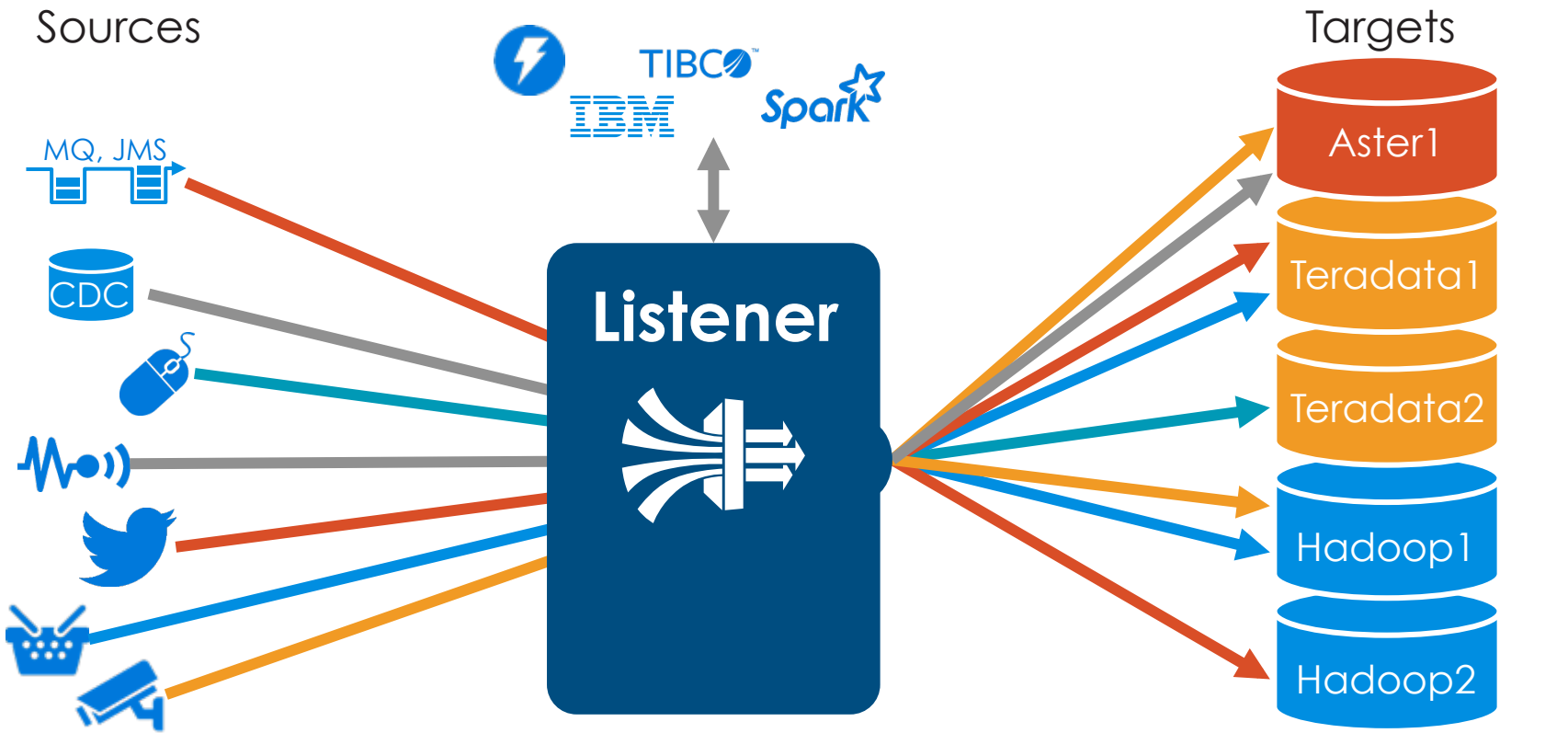
3. Ya no hay marcas sino un “ecosistema” (todo sirve)
4. Analytics as a Service
5. Data lakes: “vast, raw, unestructured info”.
6. La parte oscura: el modelo lógico o LDM.
7. Obtener rápido los “insights”.

# Múltiples alianzas en torno al analytics


Sources
Acquisition
Data Engines
Analytics
Access
Users
REAL TIME
Audio and Video
Sensors
Text
Web and Social
Machine Logs
CRM
SCM
ERP
INGEST
Listener
DATA LAKE
Hadoop
Teradata
Database
DATA WAREHOUSE
Teradata
Database
VIRTUAL QUERY
QueryGrid
No SQL
OPERATIONA
IN MEMORY
COMPUTE CLUSTER
MULTI GENRE
Aster Analytics
EMERGING
R, Spark, Giraph
CONVENTIONAL
SAS, SPSS, KXEN
APP FRAMEWORK
AppCenter
Business Intelligence
Languages
Integrated Development Environment
Marketing Executives
Operational Systems
Customers Partners
Knowledge Workers
Business Analysts
Data Scientists
Engineers
Platform Services
DATA
DEVELOPMENT
OPERATIONS
Cloud Deployment
PUBLIC
PRIVATE
HYBRID

# Nuevas cosas: Listener, Kafka, Cassandra

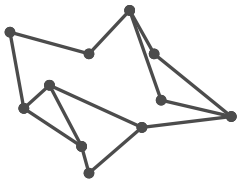
## Foco absoluto en "real time"






TERADATA.

What is the change in risk profiles by age group over the past 6 months?



**SQL ANALYTICS**

What is the typical path to purchase for a policy with increased deductions?

**PATH / TIME SERIES ANALYTICS**

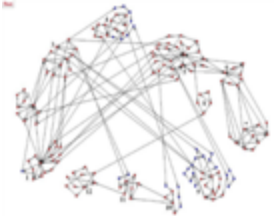
What can text based service forms tell us about potentially larger safety issues?


**TEXT ANALYTICS**

How many customers that called Customer Service expressed a frustrated tone of voice?

**RICH MEDIA ANALYTICS**

Which customers are highly influential on social media and regularly post about our claims service?



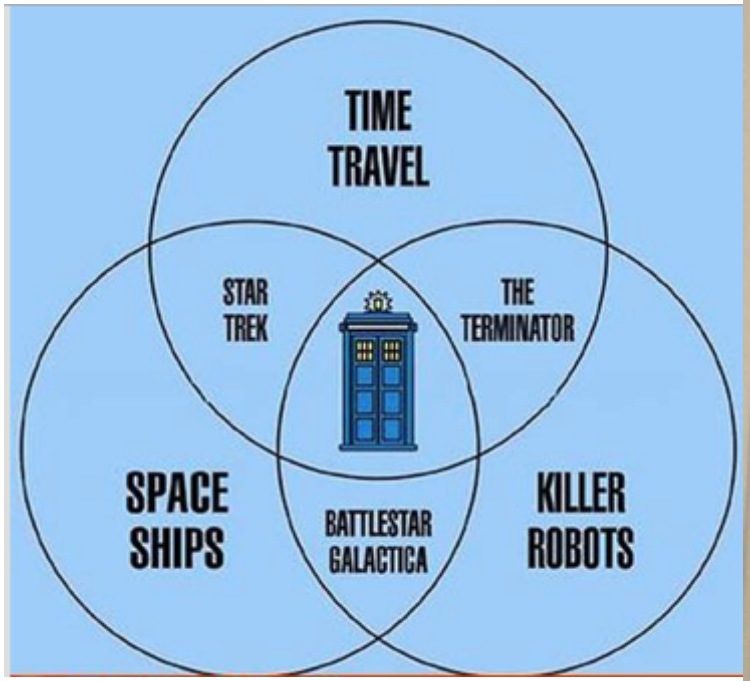
**GRAPH ANALYTICS**



TERADATA.

# Insights

# Insights

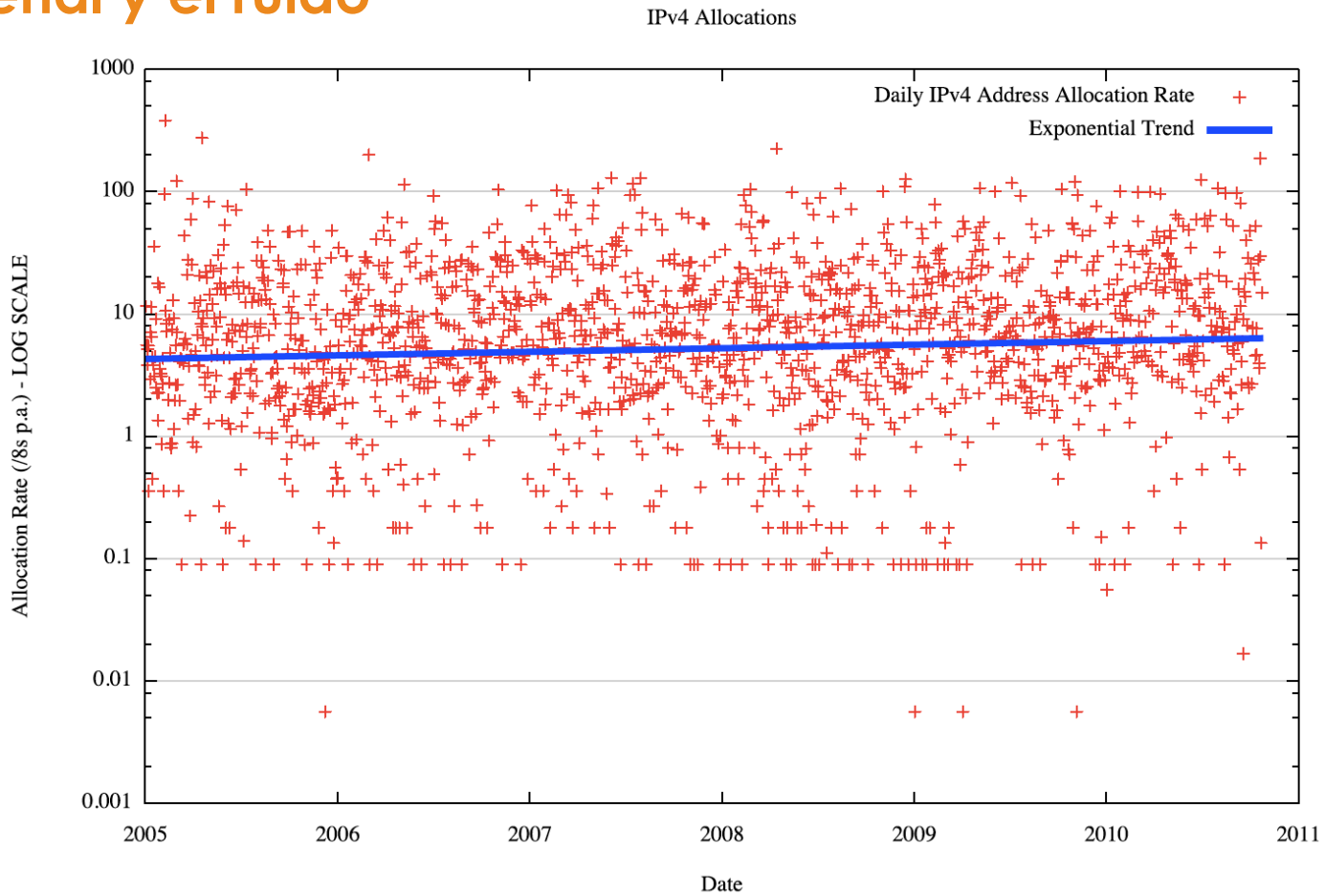






TERADATA.

# La señal y el ruido



# No hay Big Data sin insights. No hay insights si no se conoce la industria.

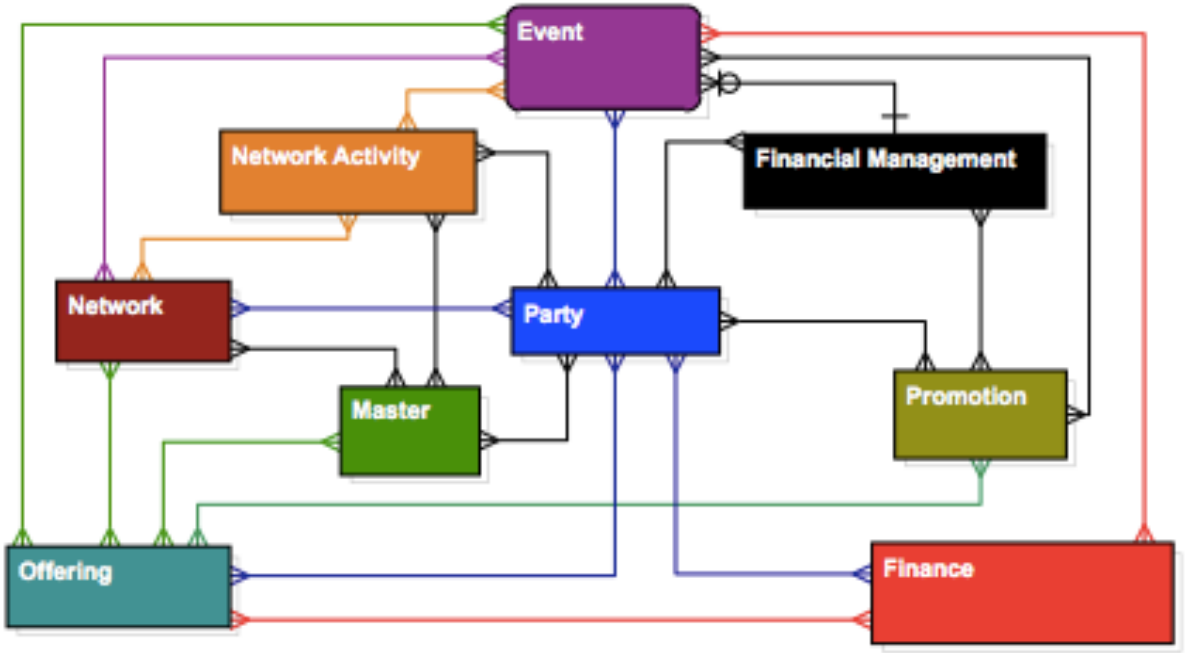
# Comparemos…

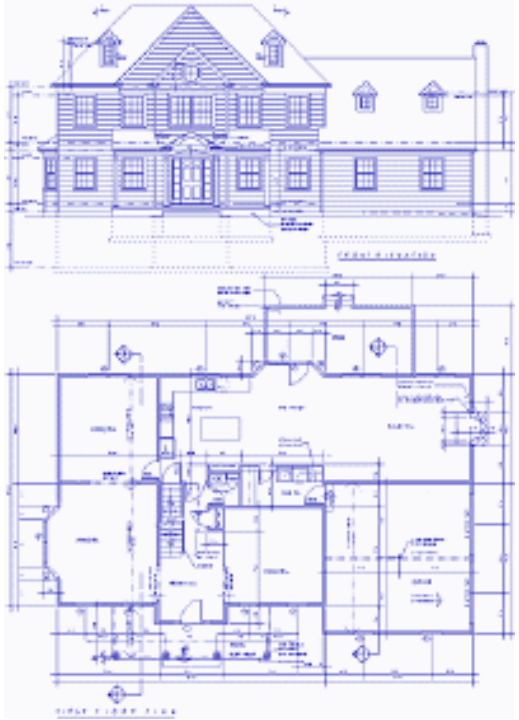
**Data**
muestra que los clientes tuvieron 2,000 sesiones en un web site durante el último mes

**Analytics**
revela cuántas sesiones ocurrieron usando un iPhone en Vicente López en ese mismo mes.

**Insight** indica cuáles de estos clientes que usan están en el 20% más probable a comprar un dado servicio

# El modelo lógico (LDM) significa conocer la industria.



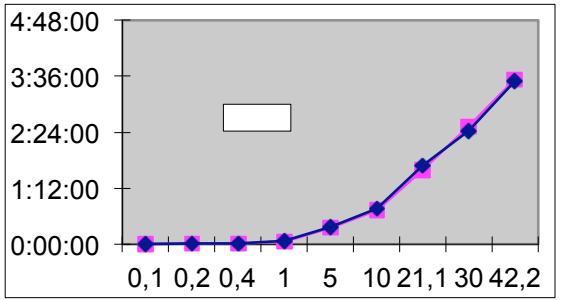





TERADATA.

# Ejemplo 1: Marathones

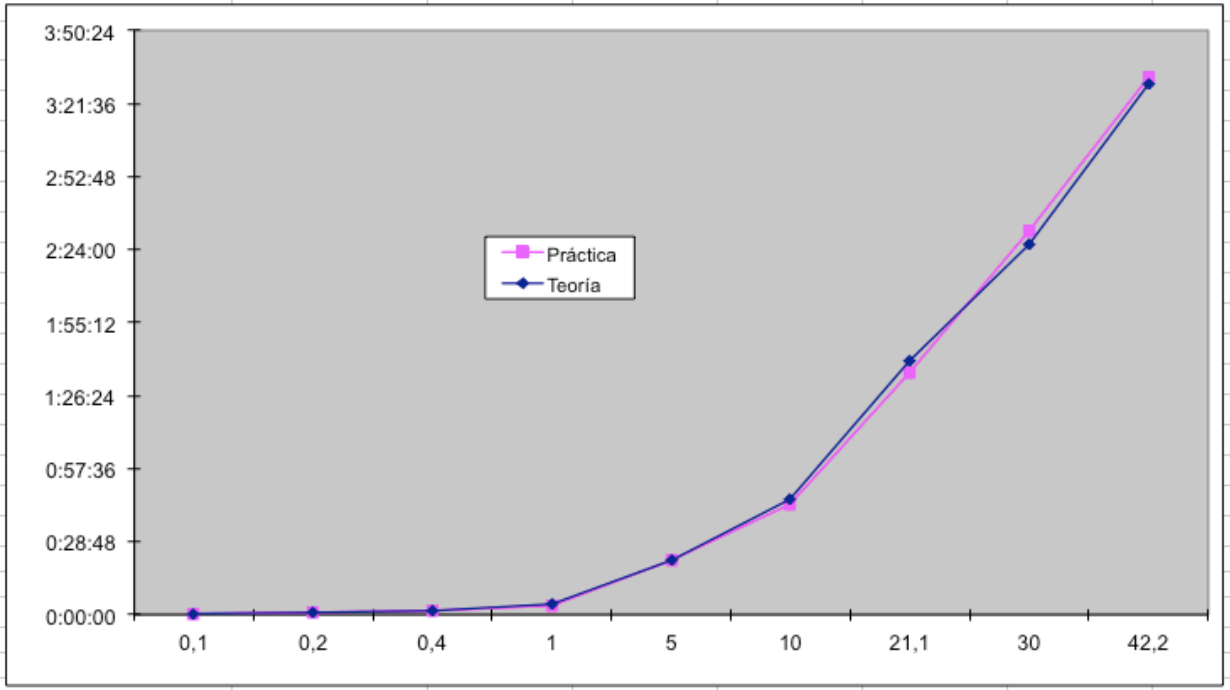

- "Rule of thumb" 1:- Tiempo de 10K * 4,66
- "Rule of thumb" 2: 2*(Tiempo de 21K) +10m
- Experimentar el "muro" = enriquecer el algoritmo.
- Mejor predictor:



# Prediciendo tiempos de marathones:
# (buen ejemplo de cómo enriquecer un algoritmo)

**T2 = T1 x (D2/D1)^1.06**

| | Práctica | Teoría | Paso teórico |
|---|---|---|---|
| 0,1 | 0:00:17 | 0:00:21 | 0:03:27 |
| 0,2 | 0:00:39 | 0:00:43 | 0:03:36 |
| 0,4 | 0:01:30 | 0:01:30 | 0:03:45 |
| 1 | 0:03:40 | 0:03:58 | 0:03:58 |
| 5 | 0:21:10 | 0:21:49 | 0:04:22 |
| 10 | 0:43:20 | 0:45:29 | 0:04:33 |
| 21,1 | 1:35:30 | 1:40:23 | 0:04:45 |
| 30 | 2:31:30 | 2:25:46 | 0:04:52 |
| 42,2 | 3:31:55 | 3:29:17 | 0:04:58 |

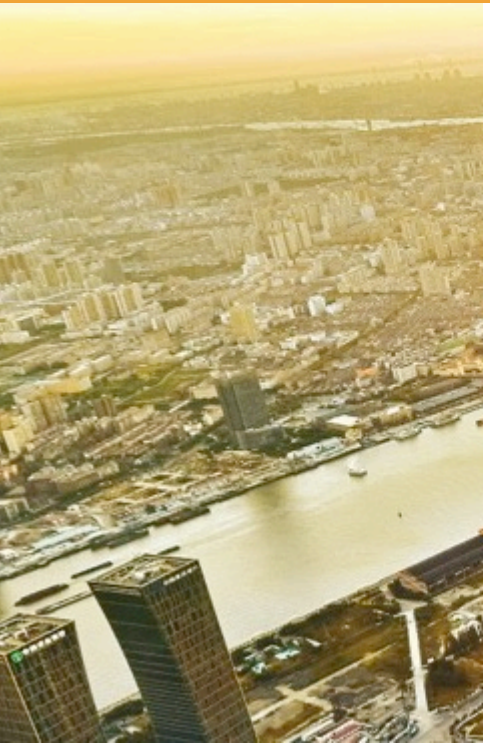


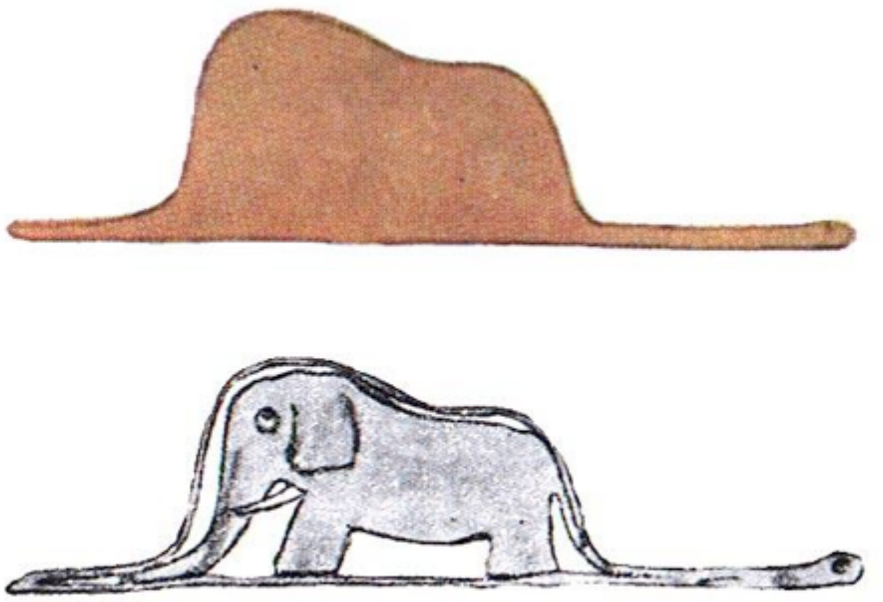
# Entonces… qué es modelo, algoritmo y qué es ML?

- Modelo general: explicación de cómo funciona el mundo.
- Modelo en el ejemplo: suponemos que corremos un marathon de un modo "lineal". O sea, que **corremos a una velocidad constante**.
- El algoritmo sería sencillo: velocidad= espacio / tiempo. **El viejo MRU**!
- Alteración de modelo: la "pared" de los 30K, cae nuestro rendimiento.
- Este modelo vincula inputs y outputs. Espacios y tiempos.
- Otros modelos (ML) pueden simplemente **reconocer patrones**, entrenando pares de inputs/outputs.
- Refinando el modelo podemos **predecir un nuevo output**: en el ejemplo, cuál va a ser nuestro tiempo en los 42K.

# El famoso “Churn” = los clientes que se van

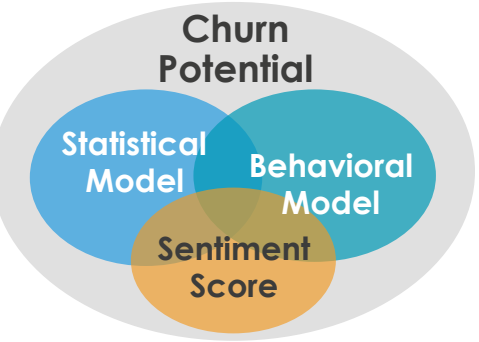
# El problema de las Telcos: qué hacen los clientes





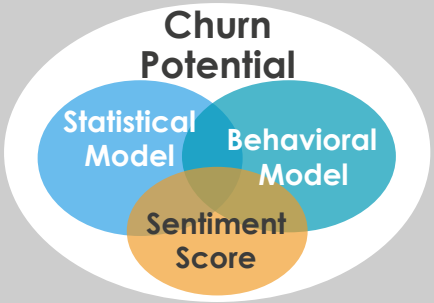
# Introducción: atacar y entender el churn desde distintos frentes

Statistical Analysis + Multi Channel Behavioral Path Analysis + Sentiment Analysis =

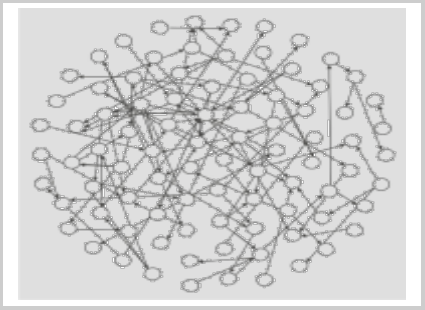



**Best in Class Churn Analytics**

**Qué pasaría si se pudiera visualizar todo al mismo tiempo?**




+



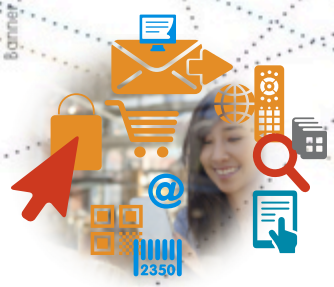
= **Insights**

# Entendiendo el “viaje del Cliente”

Un universo de interacciones cruzadas entre distintos canales y distintos sistemas.


Social Media

The Customer



Marketing Channels

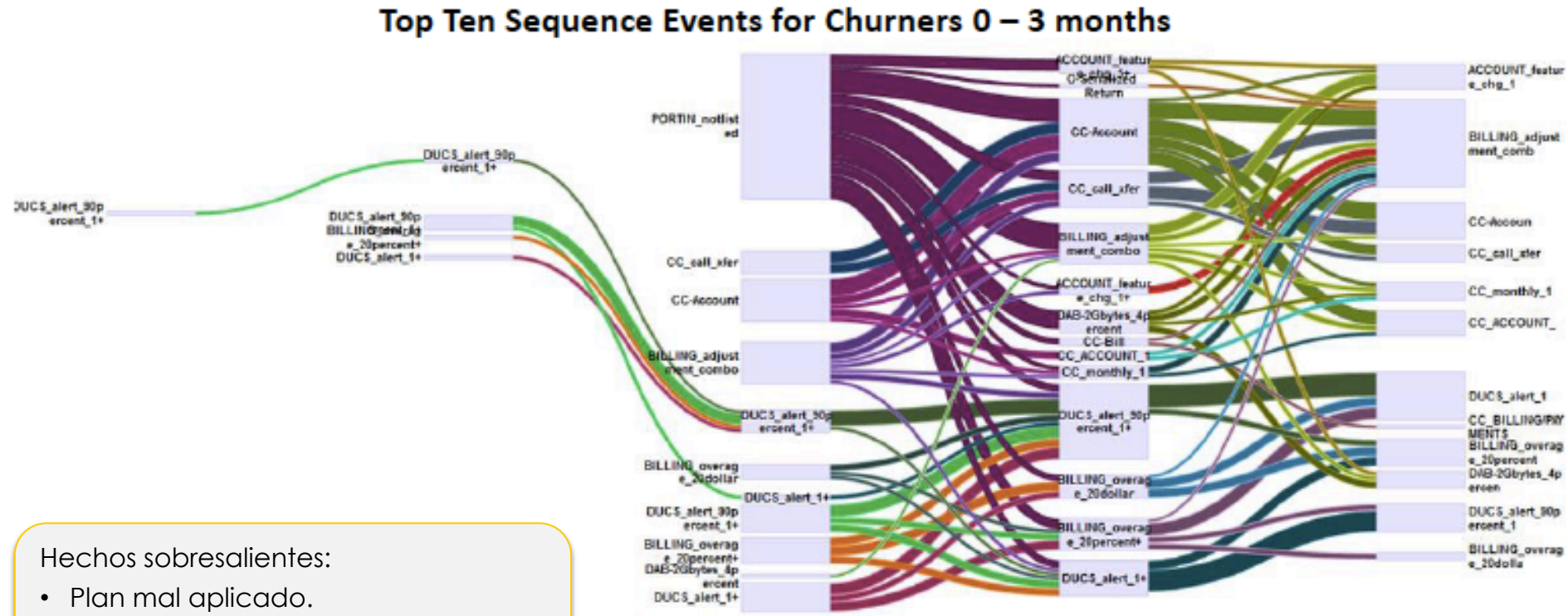
Mobile Apps

Devices & Form-factors



TERADATA.

# Ejemplo 2, Telco de USA: Visualización del churn




Hechos sobresalientes:

- Plan mal aplicado.
- Llamadas al Call no resueltas.
- Ajustes de facturación

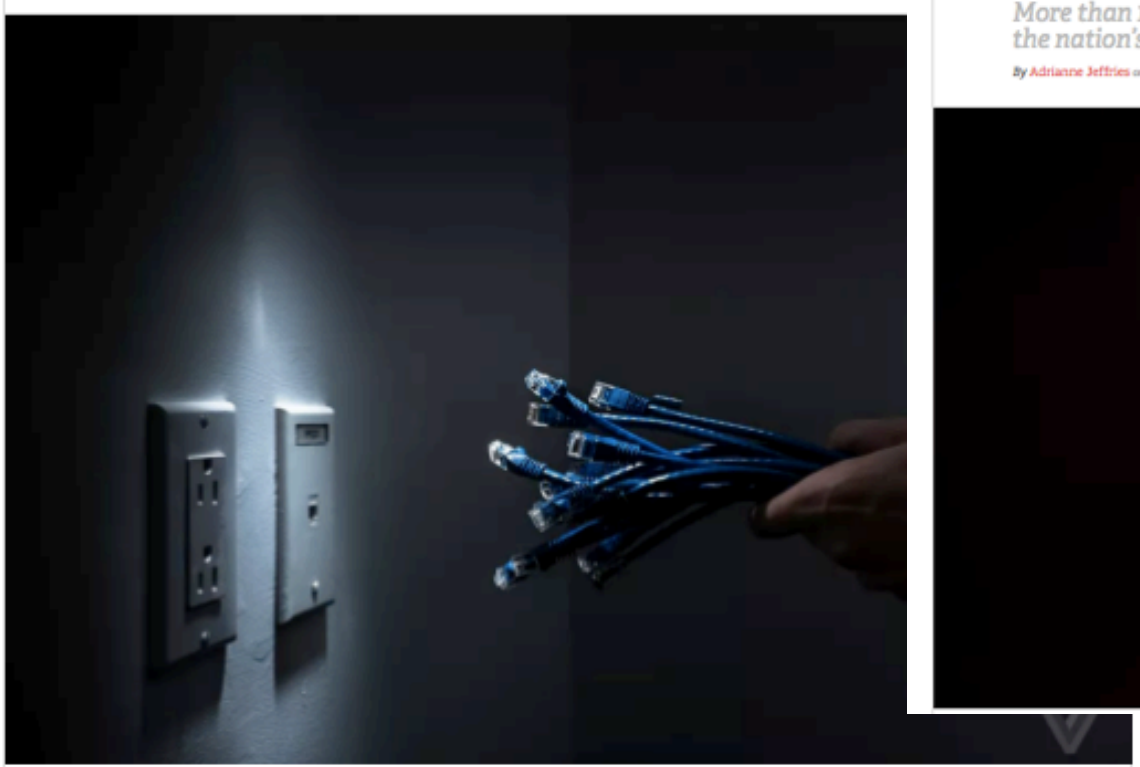
TERADATA.

COMMENTS

# Comcast Confessions: growing pains of a Goliath

*There is no one Comcast*

By Adrianne Jeffries on August 11, 2014 08:36 am



COMCAST

THE VERGE LONGFORM REVIEWS VIDEO TECH CIRCUIT BREAKER SCIENCE ENTERTAINMENT CARS TL;DR BUSINESS FOR

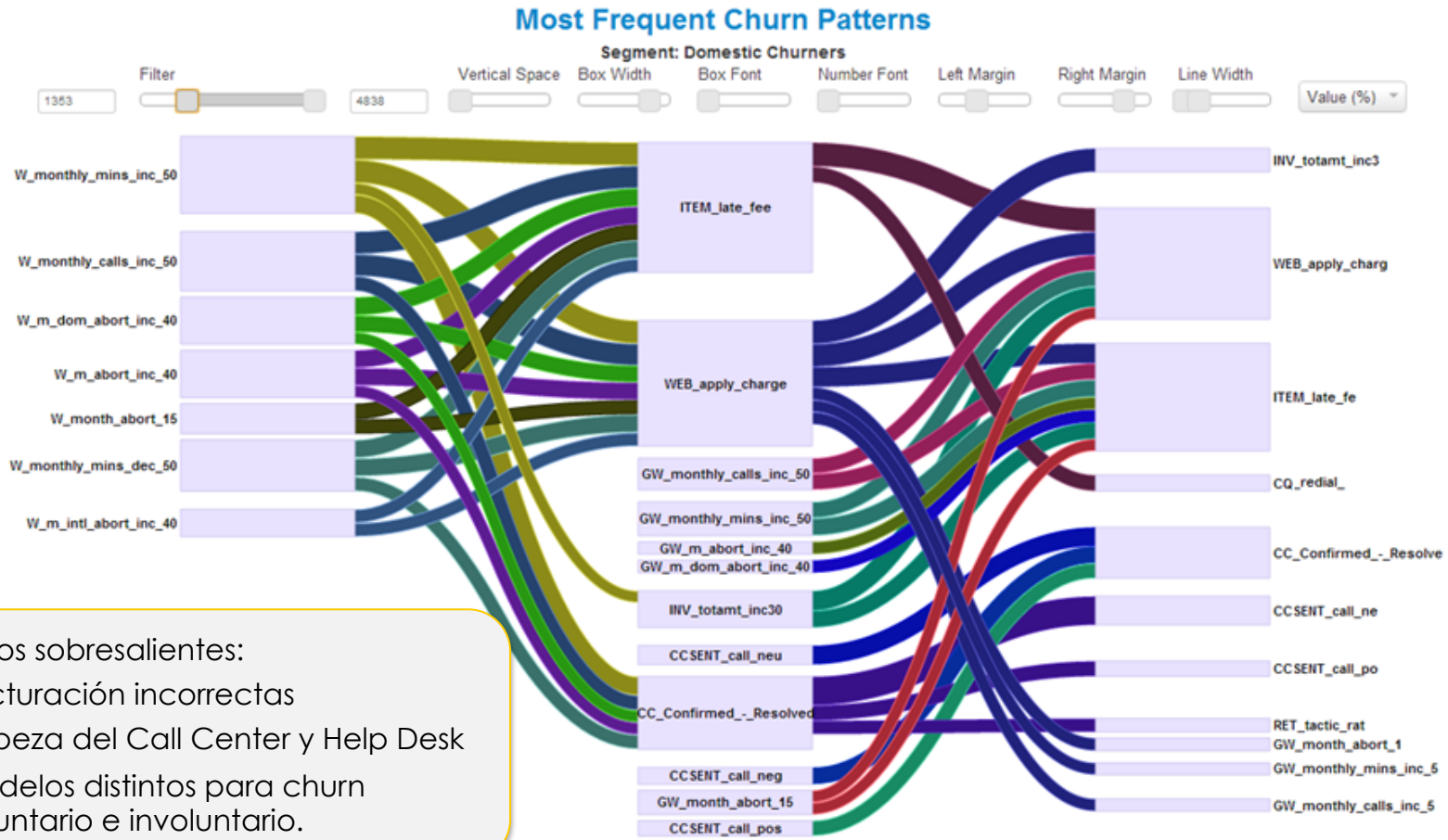
# Comcast Confessions: when every call is a sales call

*More than 100 Comcast employees spoke to The Verge about life inside the nation's largest cable and broadband company*

By Adrianne Jeffries on July 28, 2014 09:00 am

TERADATA

# Ejemplo 3, cablera de USA: tarifas mal aplicadas




Hechos sobresalientes:

- Facturación incorrectas
- Torpeza del Call Center y Help Desk
- Modelos distintos para churn voluntario e involuntario.

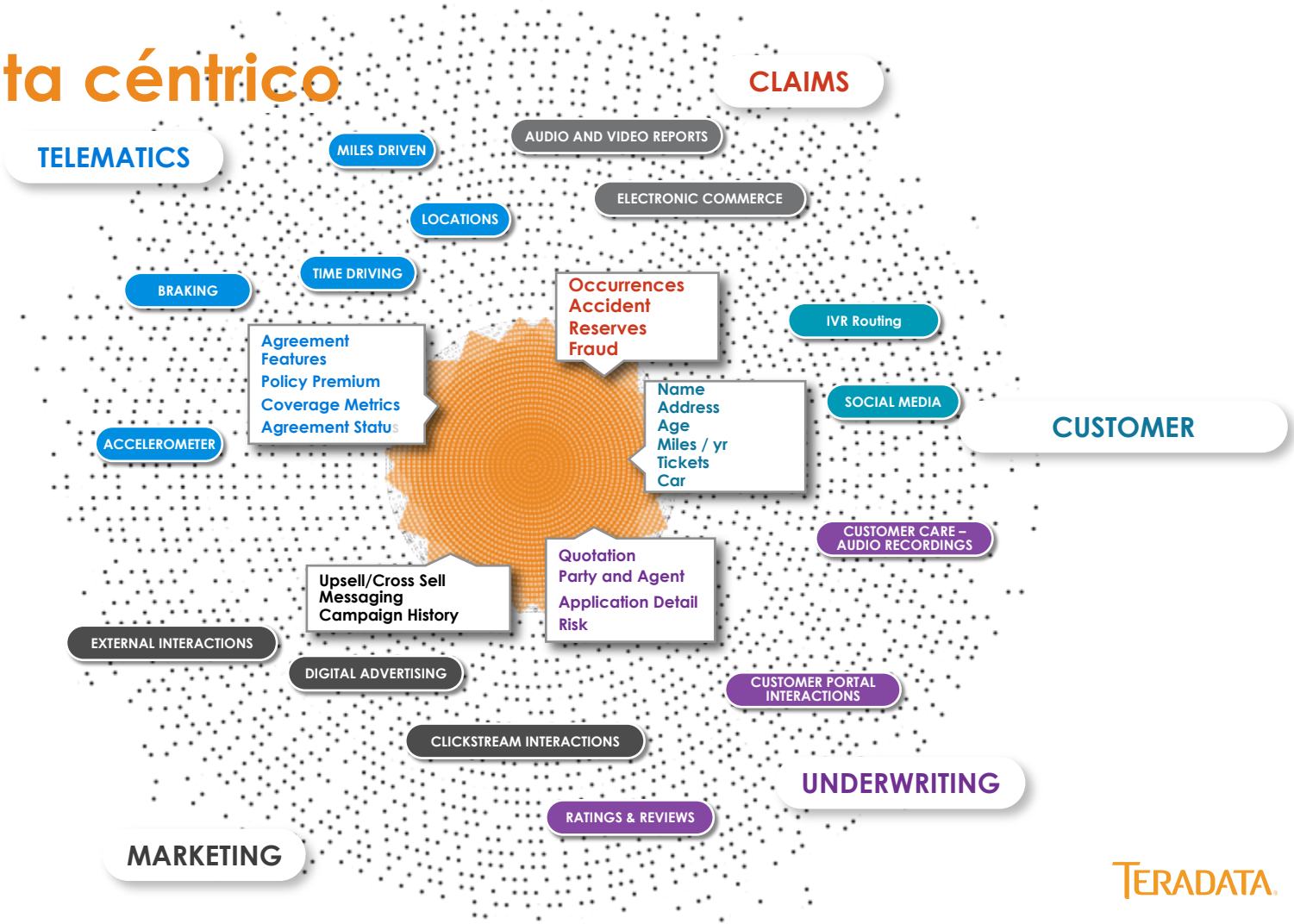
# Autos, jets, … = IoT

Pero los insights hacen la diferencia.


Sensores


No se habla tanto de Volumen, Velocidad y Variedad.
**Todo se refiere a integrar la data y el analytics.**

# Ser data céntrico






TERADATA.

# Ejemplo 4: Metadata de contenidos





TERADATA.

# Metadata: imposibilidad de categorizar

- "...los animales se dividen en
  (a) pertenecientes al Emperador / (b) embalsamados
  (c) amaestrados / (d) lechones / (e) sirenas
  (f) fabulosos / (g) perros sueltos
  (h) incluídos en esta clasificación
  (i) que se agitan como locos / (j) innumerables
  (k) dibujados con un pincel finísimo de pelo de camello
  (l) etcétera
  (m) que acaban de romper el jarrón
  (n) que de lejos parecen moscas."
- "No hay clasificación del universo que no sea arbitraria y conjetural. La razón es muy simple: no sabemos qué cosa es el universo."

Jorge luis Borges, "El idioma analítico de John Wilkins" (Otras Inquisiciones, 1952)"

# Machine learning y Watson

# Qué es Apache Spark?


- Proyecto de Open Source de Apache
  - Middleware paralelo para clusters de servidores
  - Spark.apache.org (2014)


- Dearrollado por AMPlab de UC Berkeley's
  - Soportado por DataBricks


- En qué se lo usa:
  - SQL en Hadoop
  - **Machine learning**
  - Streameado "mini-batches" de información

# Pero qué es Machine Learning y qué es Watson?

- **Machine learning** es dejar que la máquina aprenda sola.
  - Middleware paralelo para clusters de servidores
  - Spark.apache.org (2014)

- **Watson** es la marca de ML para IBM
  - Muchísima prensa!
  - Jeopardy!

- Hemos hablado poco sobre **algoritmos**
  - (ML **puentea** el algoritmo, usa "patrones")

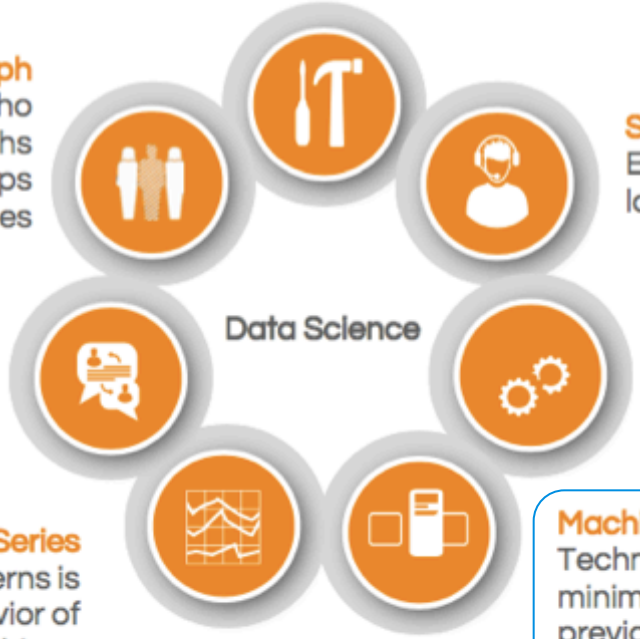
IBM's Vice President for Watson Analytics and business intelligence, Marc Altshuller, explains "With a cognitive system like Watson you just bring your question – or if you don't have a question you just upload your data and Watson can look at it and infer what you might want to know.

# Contexto

**At Scale**
If it can't be done *at scale*, you're not doing Big Analytics

**Graph**
Not just a visual graph of who knows who, but analytical graphs of the strength of relationships based on attributes

**SQL**
Ease of data manipulation using the language of business tools




**Text**
Uncover sentiment, word clouds, compliance infractions in communications and documents of all kinds

**Statistics**
Leverage statistical modeling techniques that have been proven over time

**Path/Pattern and Time Series**
Insight on interaction patterns is critical whether it is the behavior of people, products or things

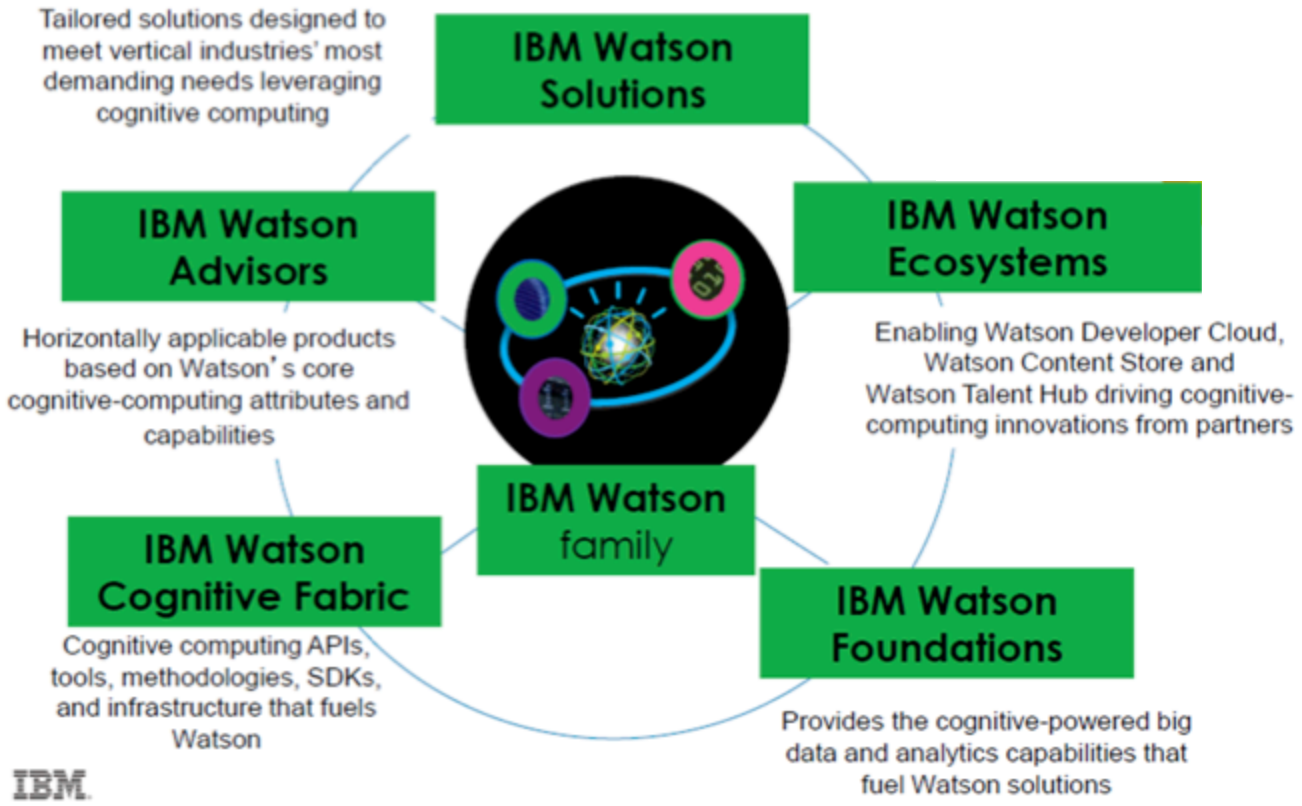
**Machine Learning**
Techniques to sift through data with minimal human input to gain new insights previously undetected

@natbusa | linkedin.com: Natalino Busa

# Principales algoritmos usados en ML

- **Modeling predictivo /Clasificación – "Aprendizaje supervisado"**
  - Linear / Logistic Regression (el "MRU" del ejemplo del marathon)
  - Decision Trees
  - Random Forests
  - Support Vector Machines
  - k-nearest-neighbour
  - Naïve Bayes
  - Neural Networks
- **Segmentación/Clustering – "Aprendizaje no supervisado"**
  - k-means
  - hierarchical clustering
  - affinity propagation
  - self-organizing maps
  - …

# IBM Watson: el branding al poder




- El "branding" de Watson Brand: conjunto de tech y productos
- Montado arriba del éxito de "Jeopardy".
- "Framework cognitivo" donde se extrae automáticamente el insight de big data a gran escala
- **"smart machines will require substantial custom work and trial and error. We are a long way away from …"plug and play" panaceas."** - Gartner



TERADATA.

# Jeopardy"! (2011)




TERADATA.

# Desafíos de “Jeopardy!”

- Entendimiento de ironía y referencias oscuras.
- Velocidad en la respuesta.
- Confianza (%) en la propia respuesta.

**“Hard times,” indeed! A giant quake struck New Madrid, Mo., on Feb. 7, 1812, the day this author struck England".**

**“According to C.S. Lewis, it was bordered on the east by the Eastern Ocean and on the north by the River Shribble".**




TERADATA.

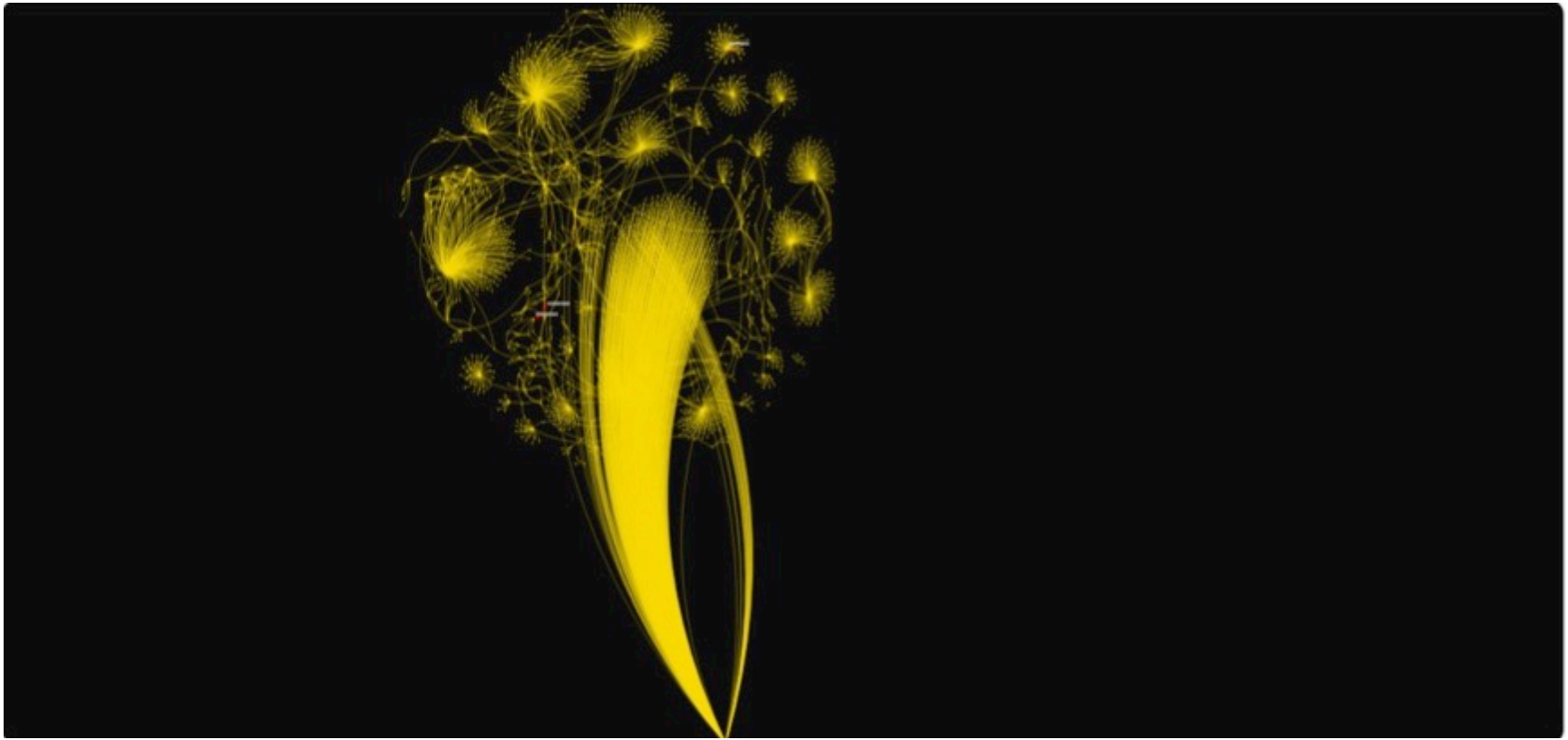
# Aplicaciones de Machine Learning (I)

- **Spam Detection**: Given email in an inbox, identify those email messages that are spam and those that are not. Having a model of this problem would allow a program to leave Nonspam emails in the inbox and move spam emails to a spam folder. We should all be familiar with this example.
- **Credit Card Fraud Detection**: Given credit card transactions for a customer in a month, identify those transactions that were made by the customer and those that were not. A program with a model of this decision could refund those transactions that were fraudulent.
- **Digit Recognition**: Given zip codes hand written on envelops, identify the digit for each handwritten character. A model of this problem would allow a computer program to read and understand handwritten zip codes and sort envelops by geographic region.
- **Speech Understanding**: Given an utterance from a user, identify the specific request made by the user. A model of this problem would allow a program to understand and make an attempt to fulfil that request. The iPhone with Siri has this capability.
- **Face Detection**: Given a digital photo album of many hundreds of digital photographs, identify those photos that include a given person. A model of this decision process would allow a program to organize photos by person. Some digital cameras and software like iPhoto has this capability.

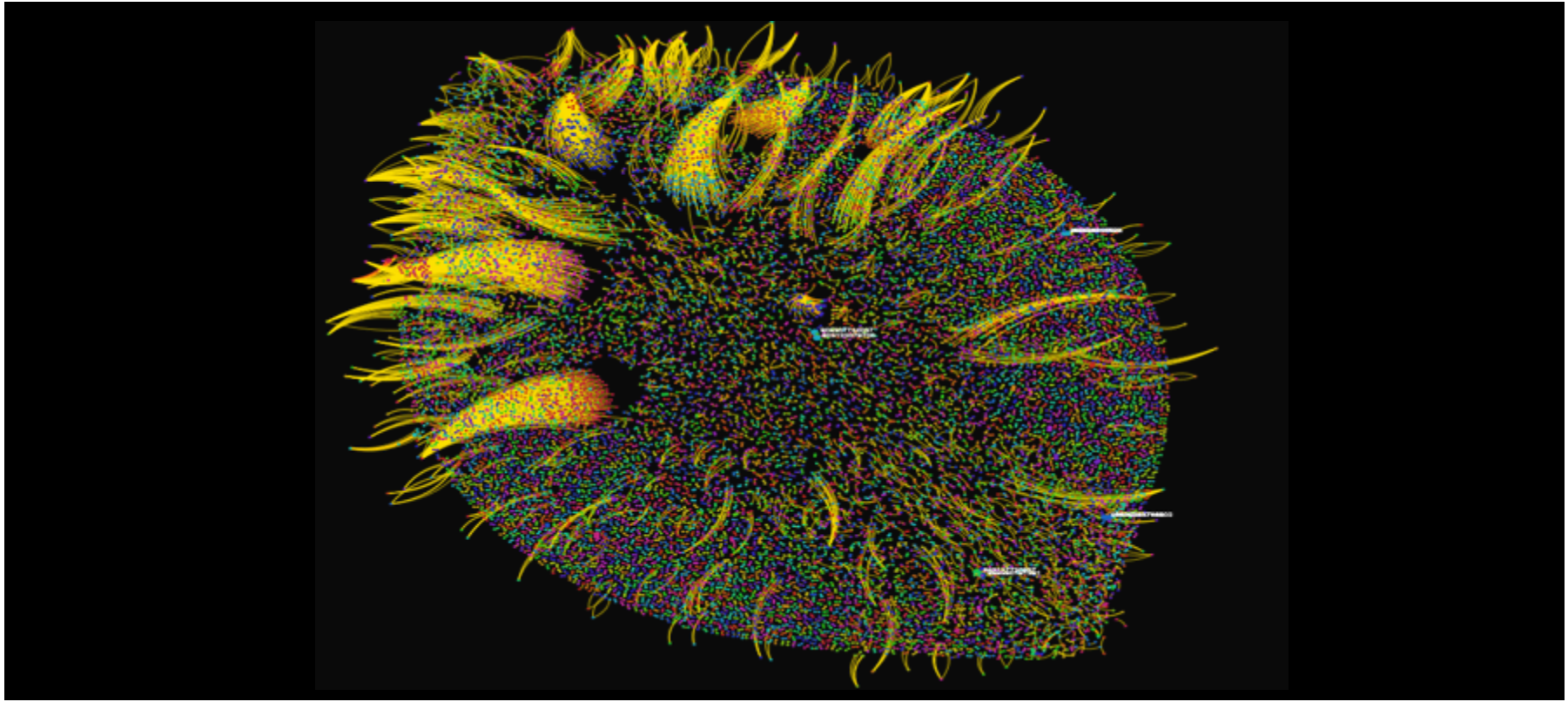
# Aplicaciones de Machine Learning (II)

- **Product Recommendation**: Given the purchase history for a customer and a large inventory of products, identify those products in which that customer will be interested and likely to purchase. A model of this decision process would allow a program to make recommendations to a customer and motivate product purchases. Amazon has this capability. Also think of Facebook, Google+ and Facebook that recommend users for you to connect with.
- **Medical Diagnosis**: Given the symptoms exhibited in a patient and a database of anonymized patient records, predict whether the patient is likely to have an illness. A model of this decision problem could be used by a program to provide decision support to medical professionals.
- **Stock Trading**: Given the current and past price movements for a stock, determine whether the stock should be bought, held or sold. A model of this decision problem could provide decision support to financial analysts.
- **Customer Segmentation**: Given the pattern of behaviour by a user during a trial period and the past behaviours of all users, identify those users that will convert to the paid version of the product and those that will not. A model of this decision problem would allow a program to trigger customer interventions to persuade the customer to covert early or better engage in a limited trial.
- **Shape Detection**: Given a user hand drawing a shape on a touch screen and a database of known shapes, determine which shape the user was trying to draw. A model of this decision would allow a program to show the platonic version of that shape the user drew to make crisp diagrams. The Instaviz iPhone app does this.

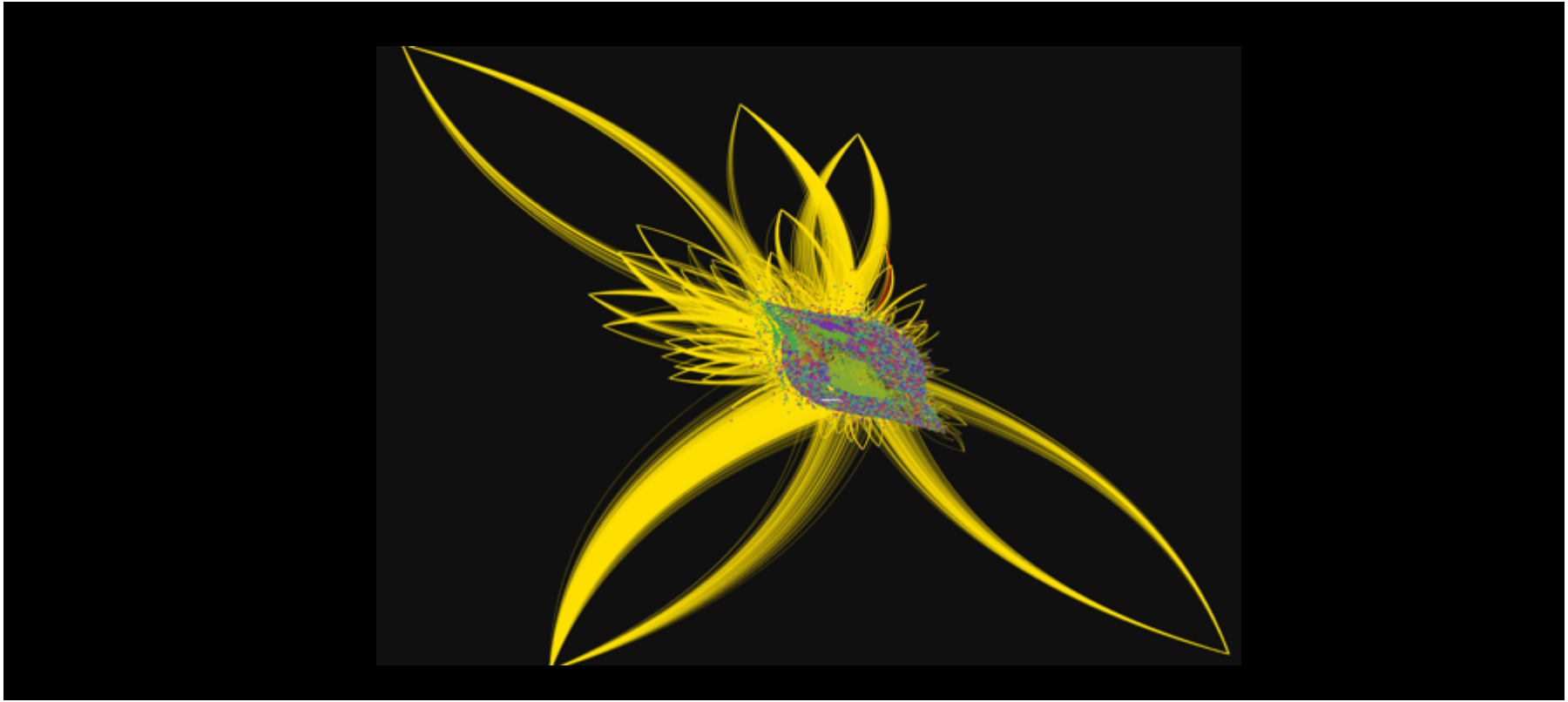
# The Art of Analytics: Flujo de fondos / supply chain



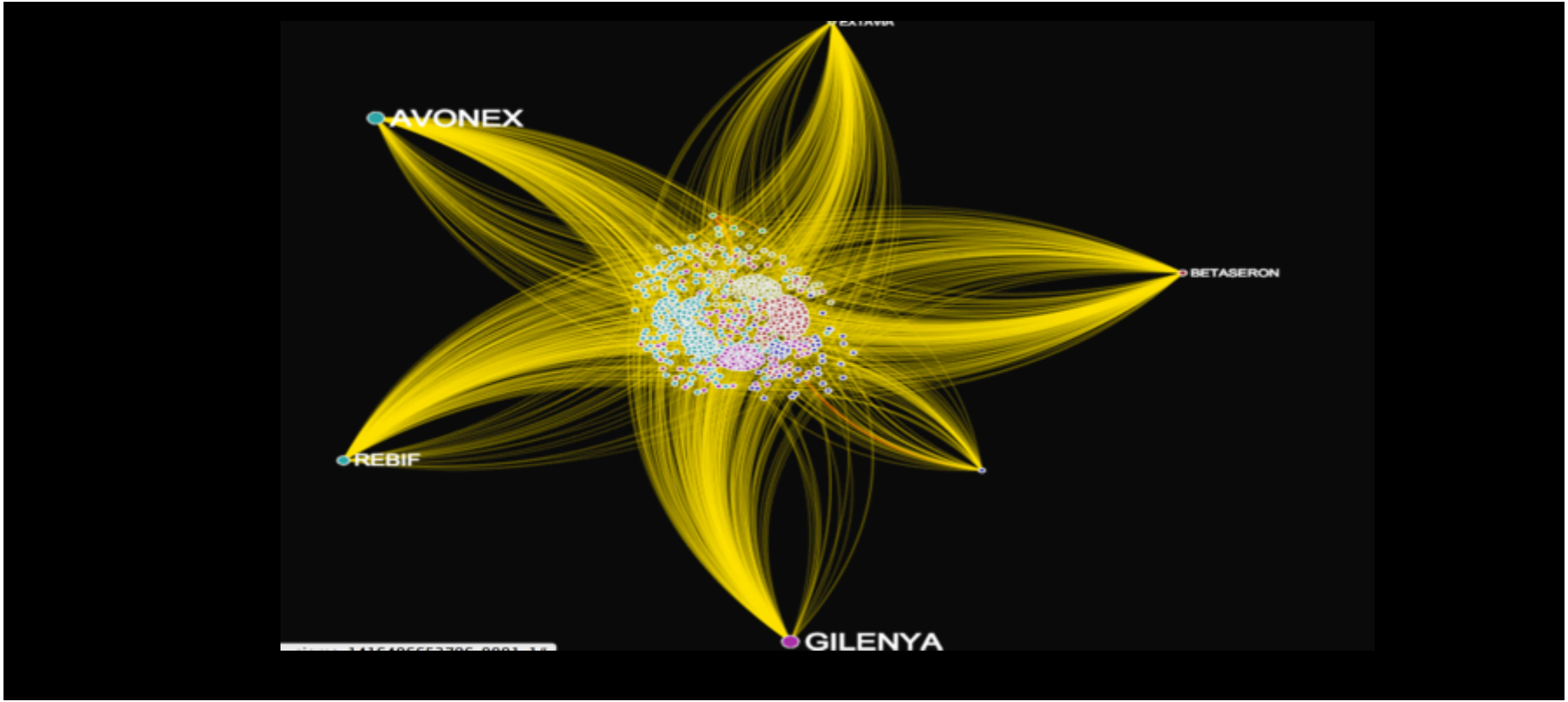
# The Art of Analytics: Clusters de llamadas en Telcos



# The Art of Analytics: influencia en Twitter



# The Art of Analytics: drogas y efectos colaterales



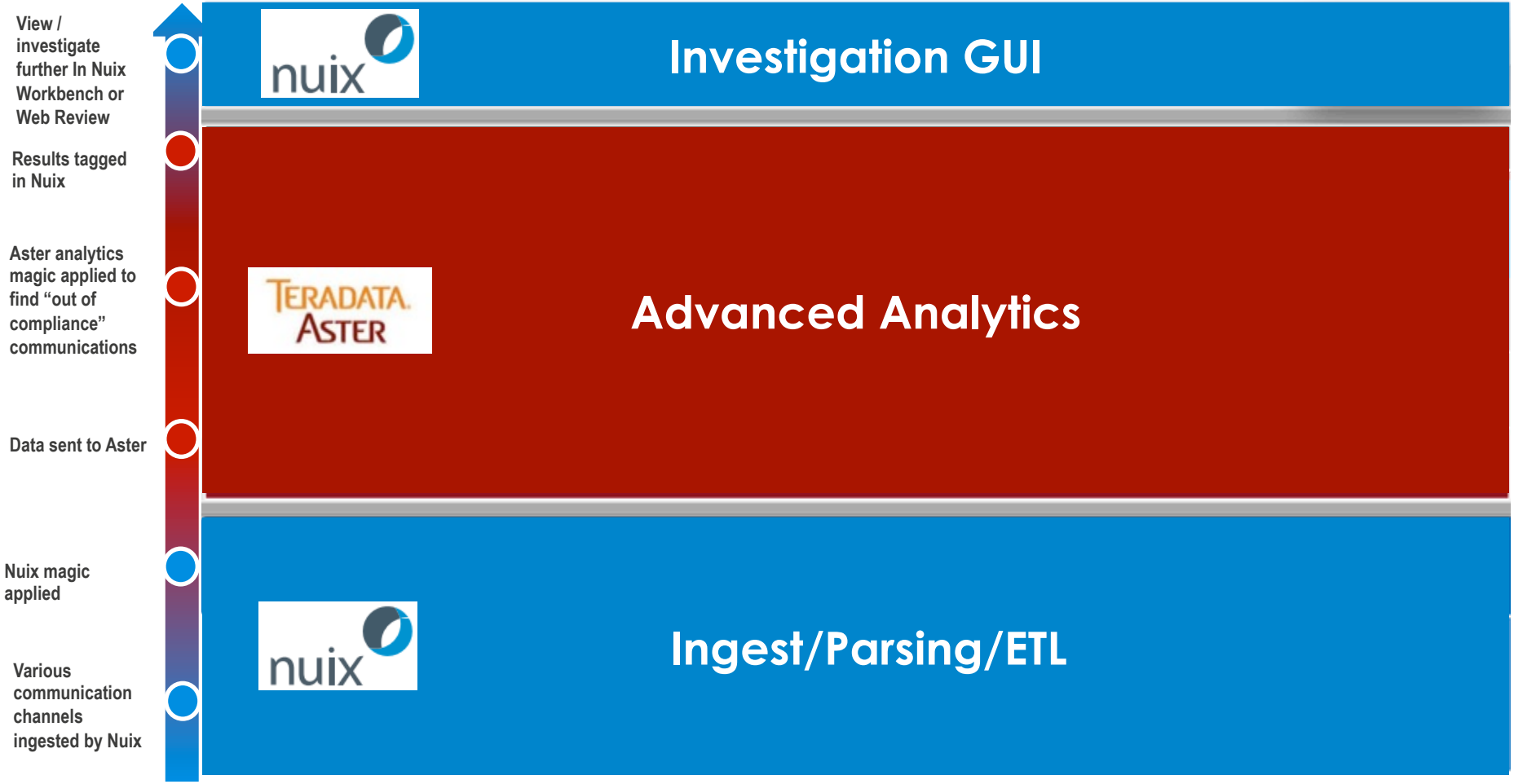



TERADATA.

# #PanamaPapers


SZ → nuix → TERADATA

# Aster Analytics and Nuix Investigative Engine



# Aster Analytics and Nuix Investigative Engine

View / investigate further In Nuix Workbench or Web Review

Results tagged in Nuix

Aster analytics magic applied to find "out of compliance" communications

Data sent to Aster

Nuix magic applied

Various communication channels ingested by Nuix


Teradata Aster AppCenter

Deep Learning

Time Series

Machine Learning

MULTI GENRE

Aster

Statistics

Cluster

Path, Pattern and Time Series

Association

Text

Graph

Data Transformation

Naïve Bayes

Decision Tree

Location Services

TERADATA ASTER

100+ Pre-built Analytic Functions

Teradata Aster Discovery Platforms

nuix

Single point of data collection, Pre-processed, filtered data

(at scale, speed, 100s of containers/file types)

Network Info.

Browser Data

Images

E-Mail & Docs.

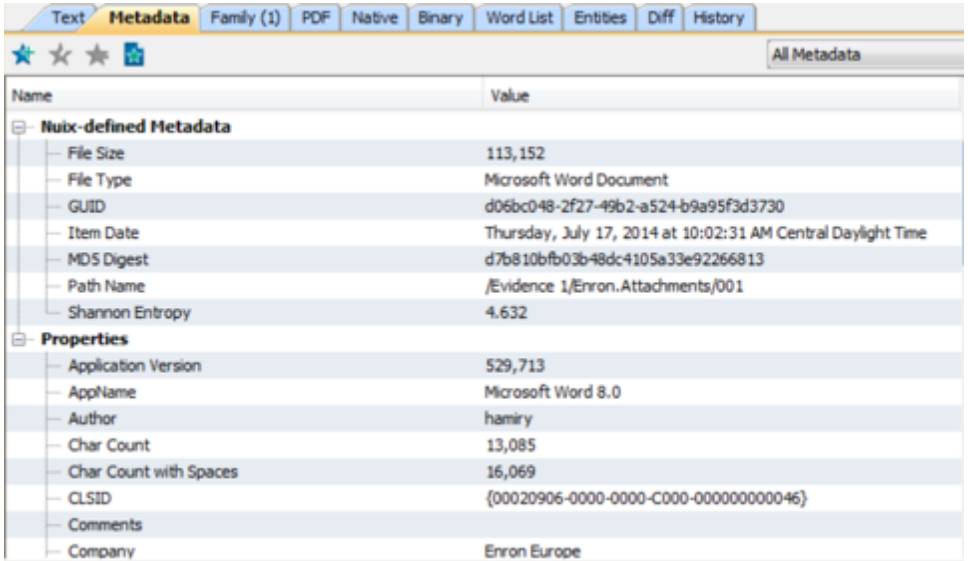
Device Data

skype

Office 365

Bloomberg

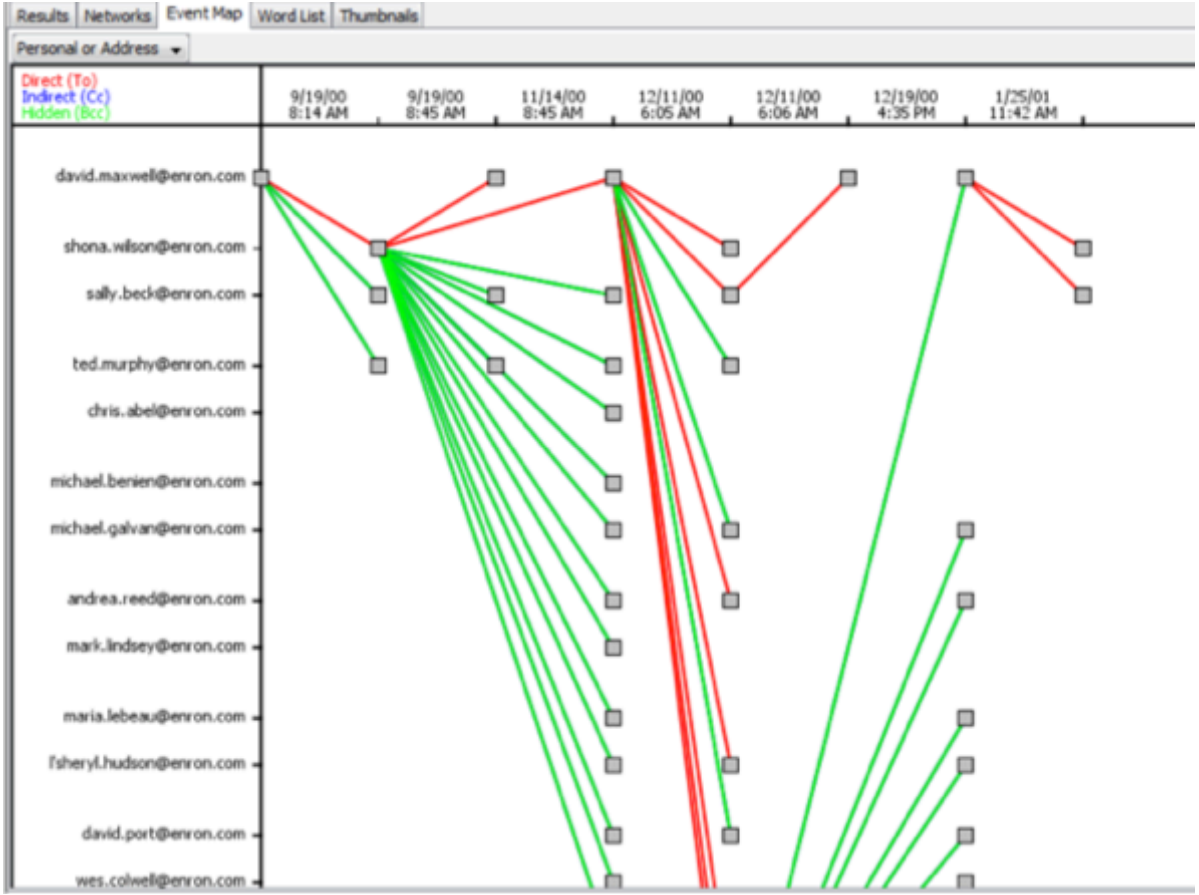
# Investigación y Contexto– Text & Metadata




TERADATA.

# Investigación y contexto – Email thread



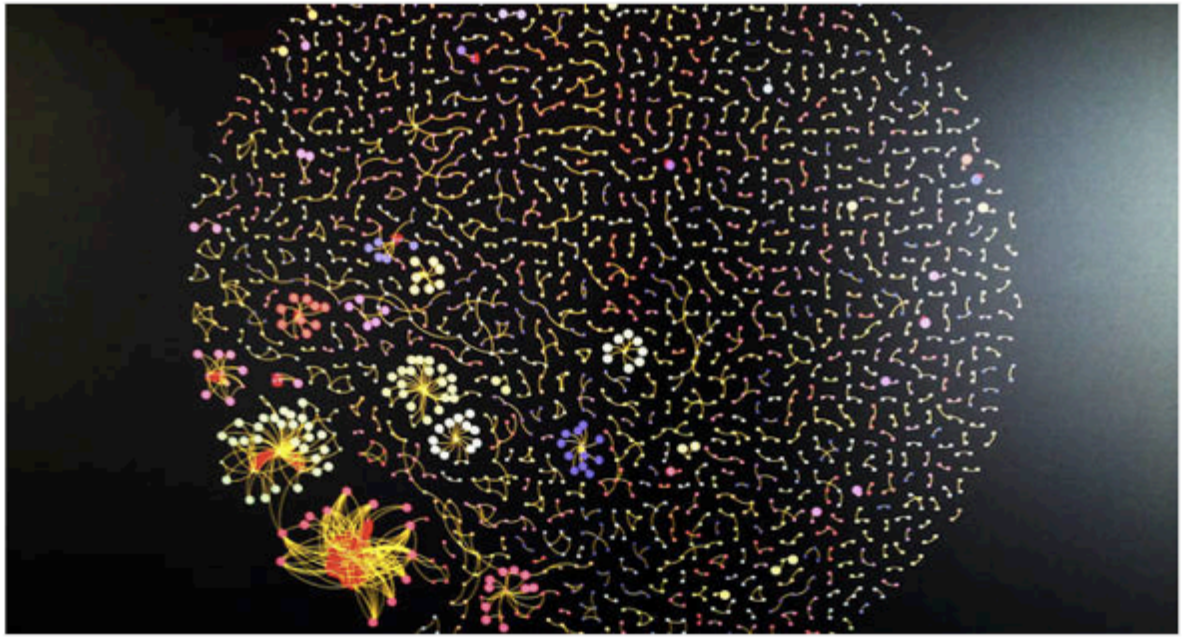



TERADATA.

# Final feliz

## Teradata and Nuix are dancing in the dark

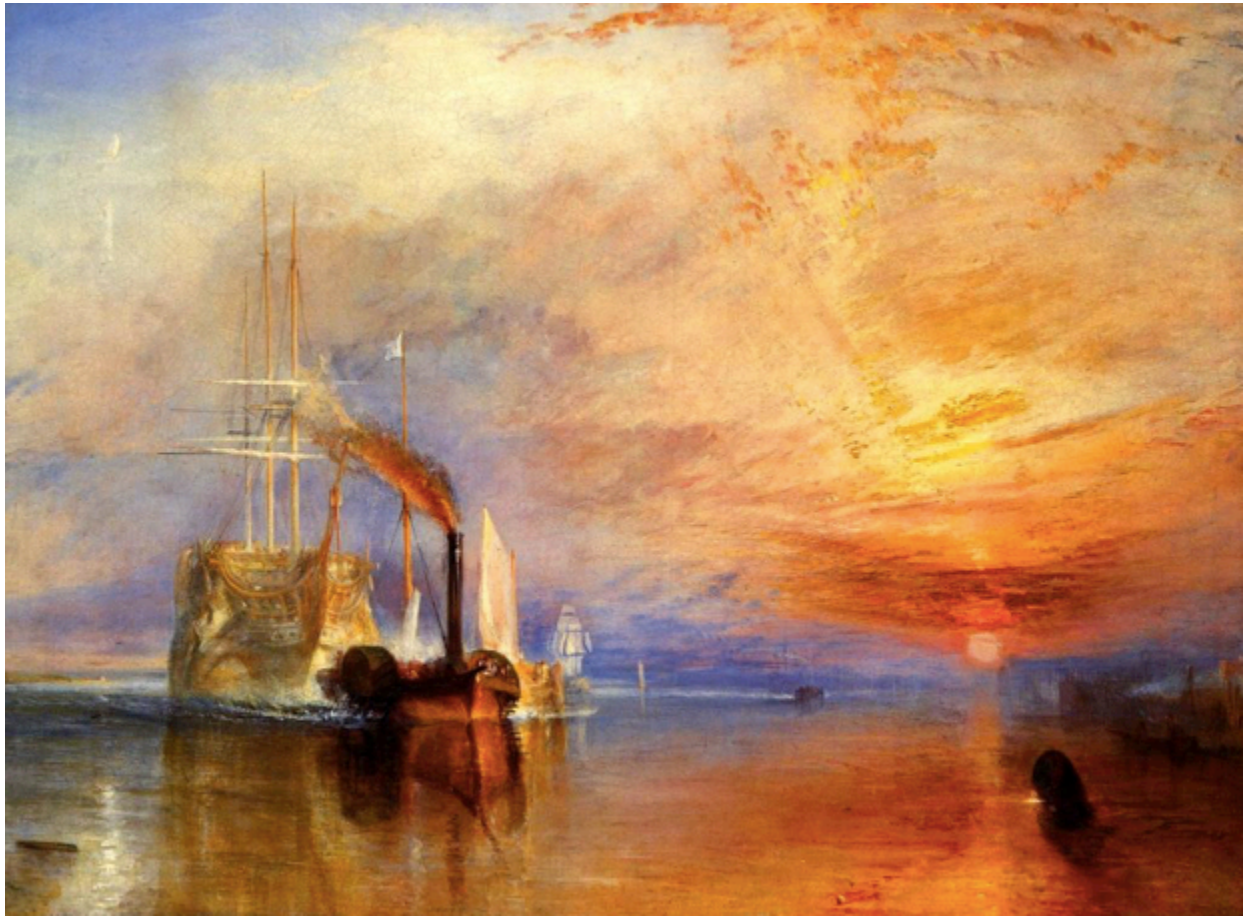
Where there is dark data let there be...umm...light data



Aster Analytics fraud detection graph

http://www.theregister.co.uk/2016/07/20/teradata_nuix_partnership/?mt=1475012474803

# Preguntas?



Daniel Collico Savio
Teradata
@dcollico