Big Data: Tendencias 2016

José Antonio Guerrero

Datrik Intelligence

jaguerrerod@ono.com

BuleBar Café
2 Marzo 2016

Mi etapa profesional en Gestión Sanitaria

Estadística Multivariable vs Aprendizaje Automático

MACHINE LEARNING

Arthur Samuel (1959):

"Field of study that gives computers the ability to learn without being explicitly programmed“

Estadística Multivariable Paramétrica

Hipótesis:

Normalidad
No correlación de errores
Homocedasticidad
No colinealidad

X , Y

Bondad del ajuste:

Grados de libertad
Descomposición de la varianza
Estimaciones puntuales y por IC de errores y coeficientes
Contraste de hipótesis

Debilidades

Asumir hipótesis sobre la distribución de los datos

Mal manejo de la colinealidad (Convergencia y estabilidad de las soluciones)

La limitación en la forma funcional del modelo

Alta sensibilidad a observaciones extremas

Mal manejo de observaciones desconocidas

Problemas de escalabilidad

Mal manejo variables >> casos

Fortalezas

Reproducibles

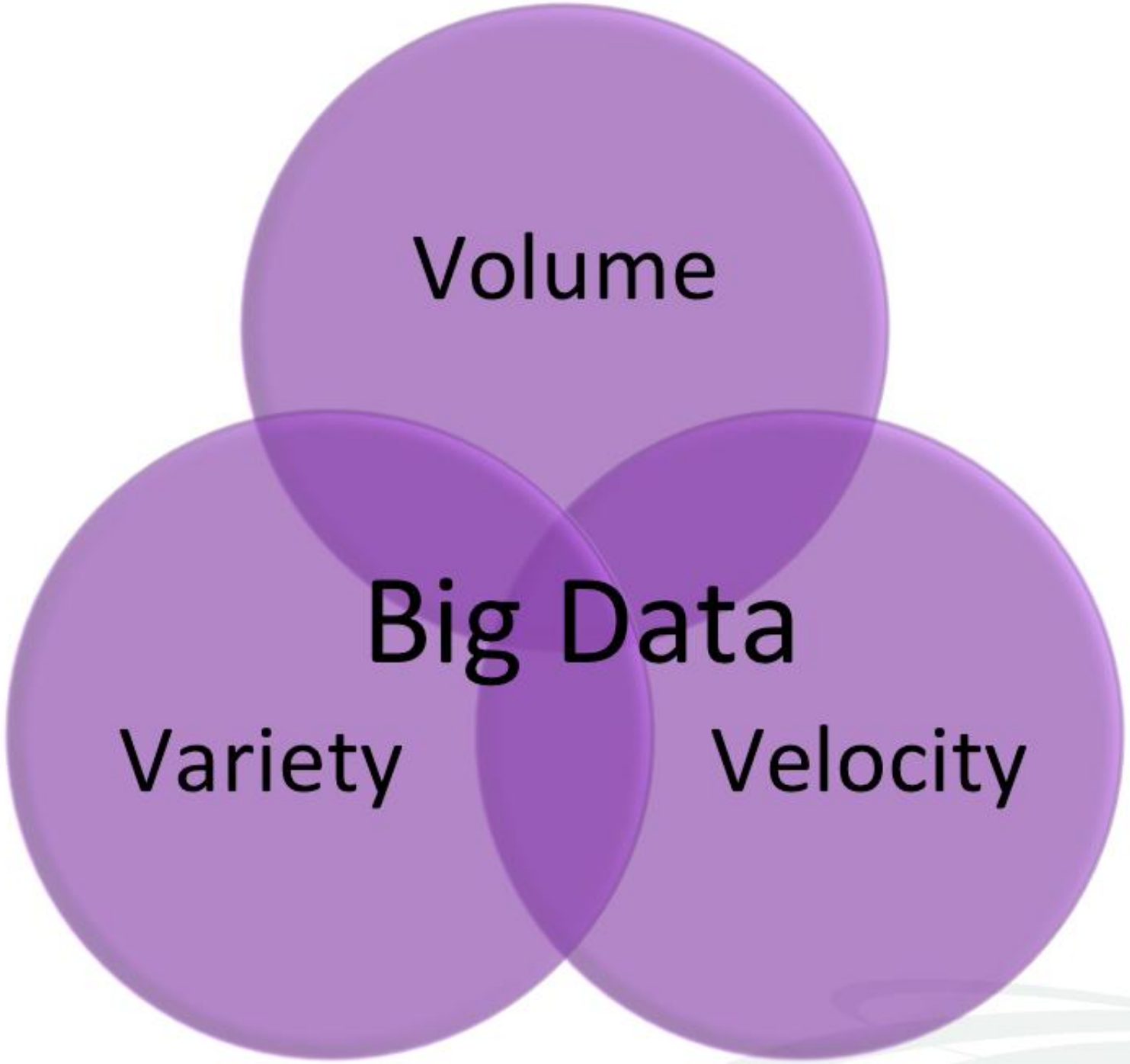
Rápidos de ajustar

Modelos interpretables (expresión analítica)

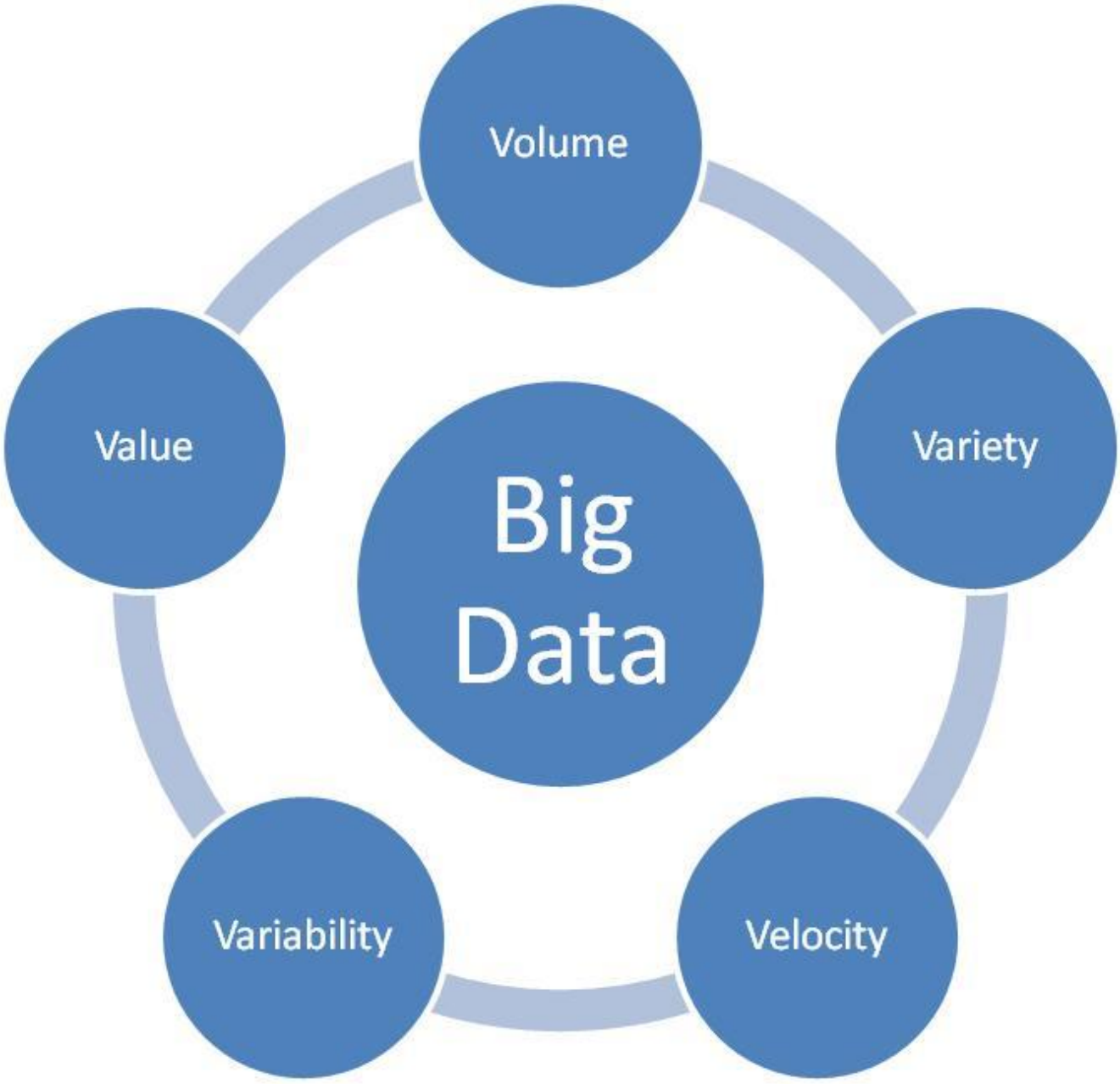
Importancia relativa de variables

Inferencia (bondad de ajuste, coeficientes)


BIG DATA




BIG DATA
VOLUME
DATA SIZE
VELOCITY
SPEED OF CHANGE
VARIETY
DIFFERENT FORMS OF DATA SOURCES
VERACITY
UNCERTAINTY OF DATA


Volume
Value
Variety
Big Data
Variability
Velocity

Therefore, the 3 V's of big data is now 6 V's

Hint 2: Big data should have a clear **business case** to work against

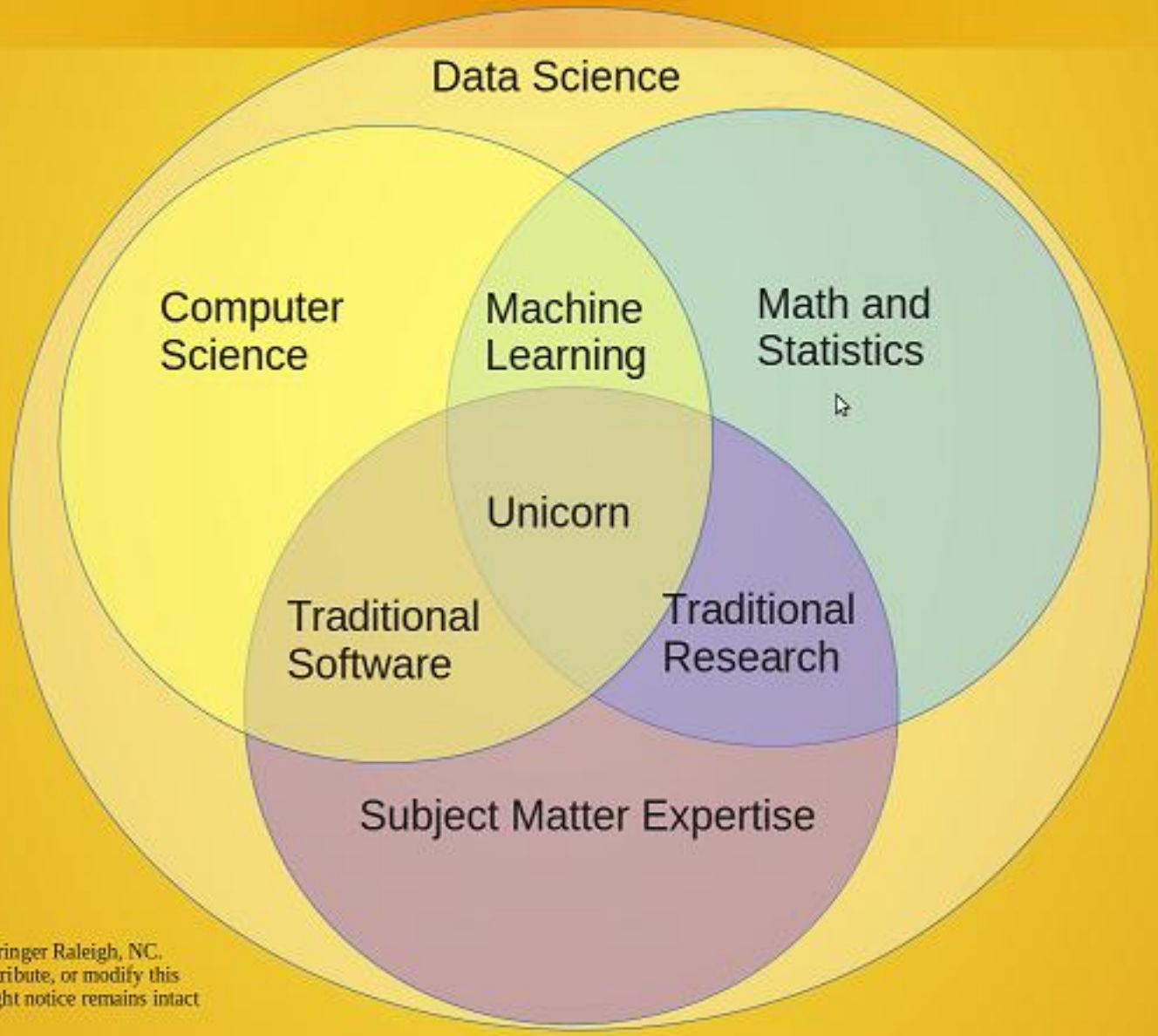
Initially big data was just about having lots of data to play with...

...since then, more attributes have been added to define big data...

..., but from enterprise standpoint the key is in VALUE!

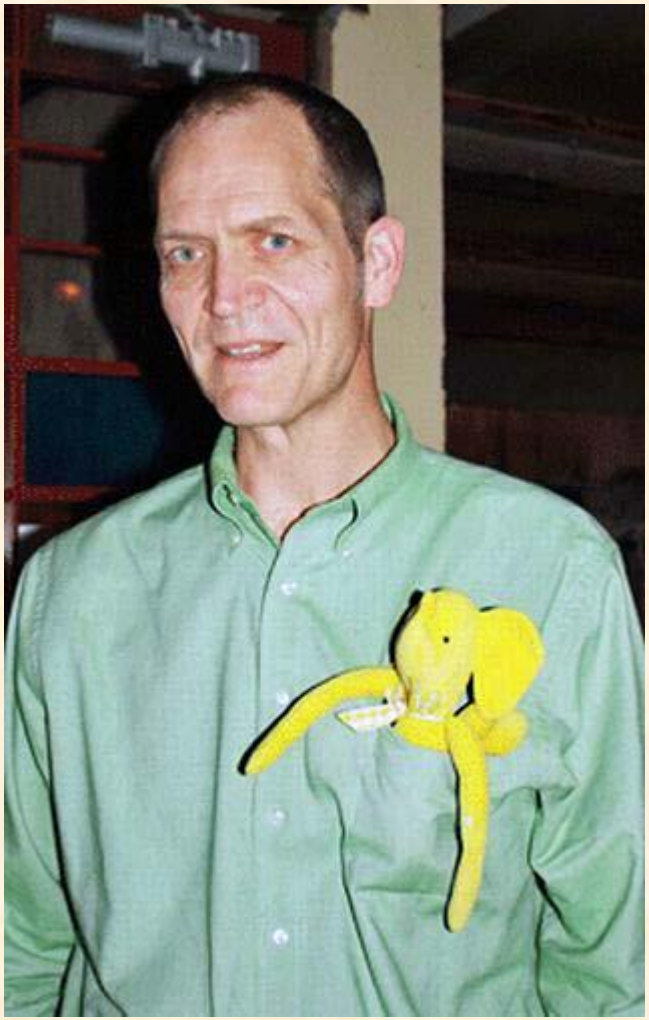
9 de cada 10 Científicos de Datos están buscando palabras con ‘V’ en vez de trabajando en Big Data

Data Science Venn Diagram v2.0



Volumen

ML para Big Data
Aplicaciones distribuidas



Doug Cutting

Volumen

Spark *Lightning-fast cluster computing*

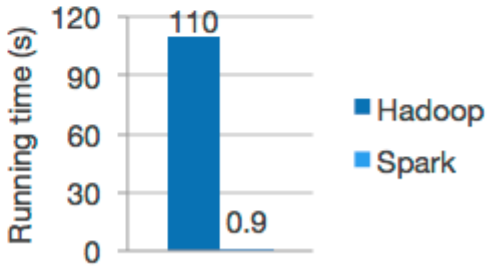
Download | Libraries | Documentation | Examples | Community | FAQ

Apache Spark™ is a fast and general engine for large-scale data processing.

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

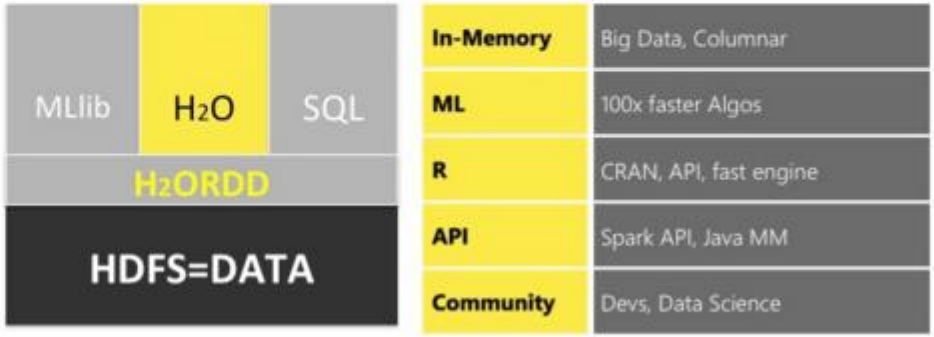
Machine Learning for Spark

Sparkling Water

Spark + H_2O

SPARKLING WATER

H_2O – The Killer-App for Spark



Spark

MLlib

Variedad

Bases de datos noSQL:

Bases documentales:

MongoDB, DynamoDB

Bases de datos orientadas a columna:

Hbase, Cassandra…

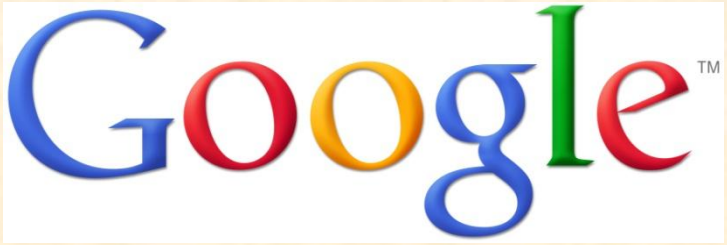
Velocidad

Elmer Fudd
Vorpal Rabbit

John Langford

Velocidad

Sofia – ML

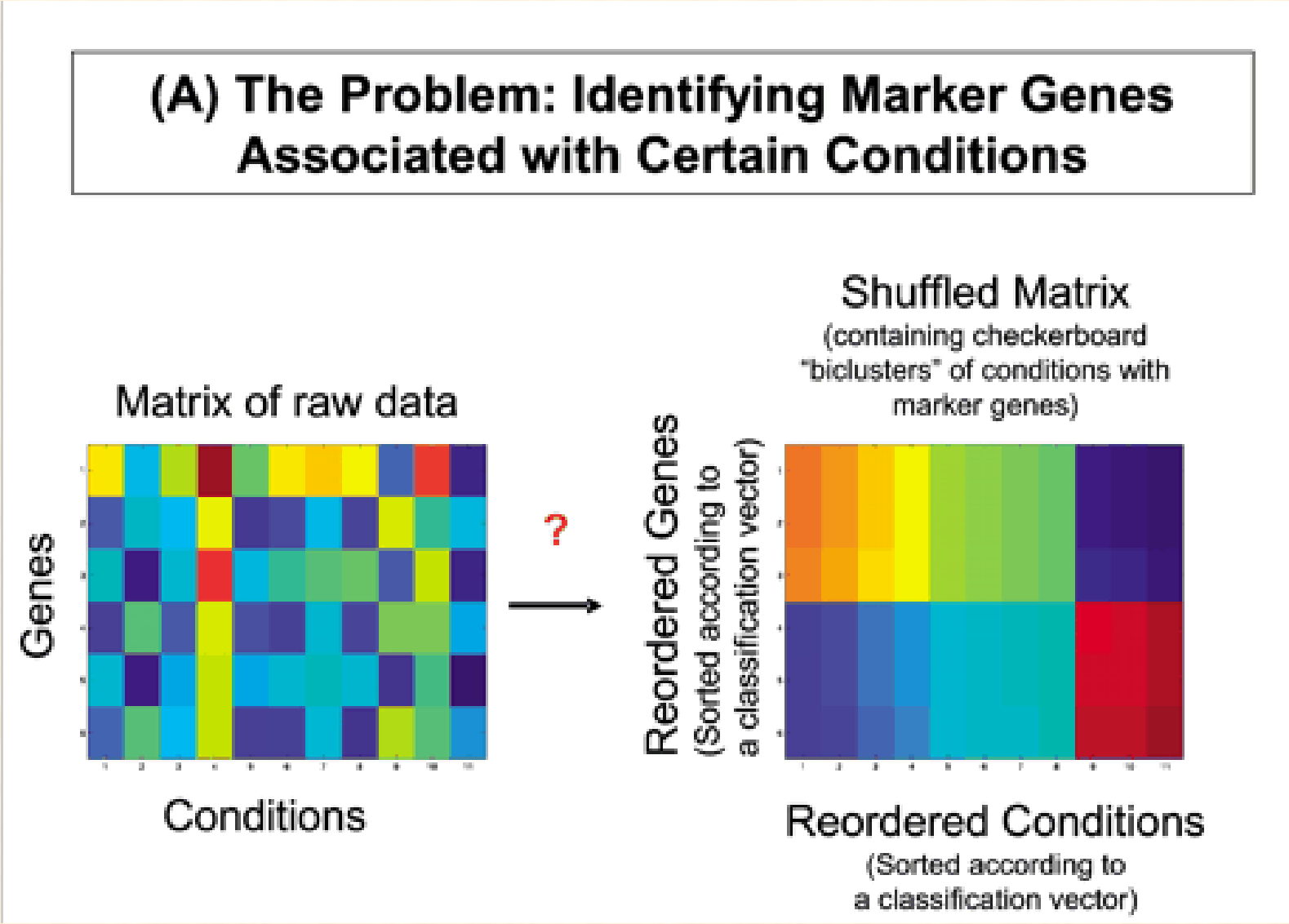


FTRL :
Follow the regularized leader

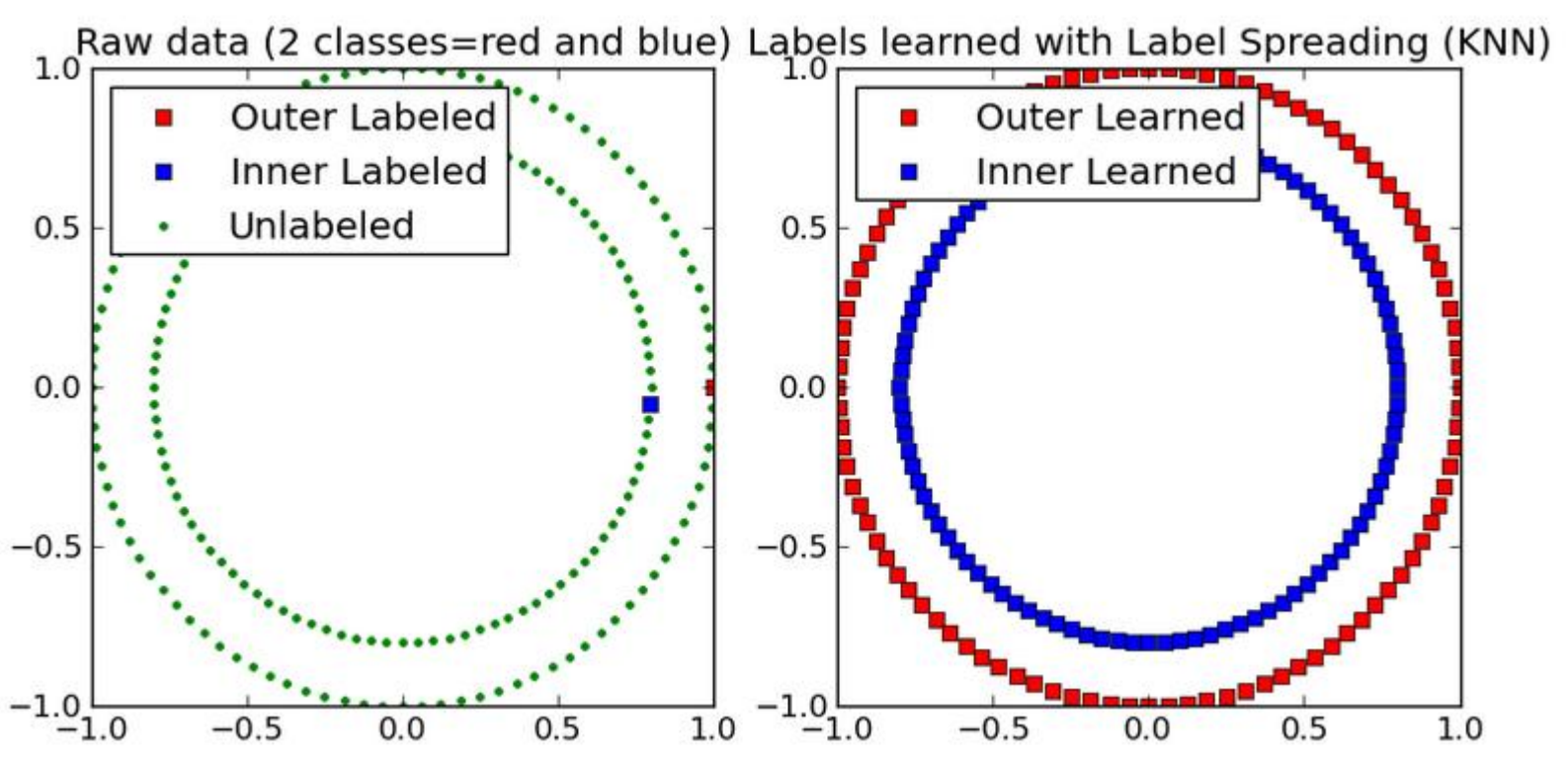
David Sculley

Hashing

Biclustering

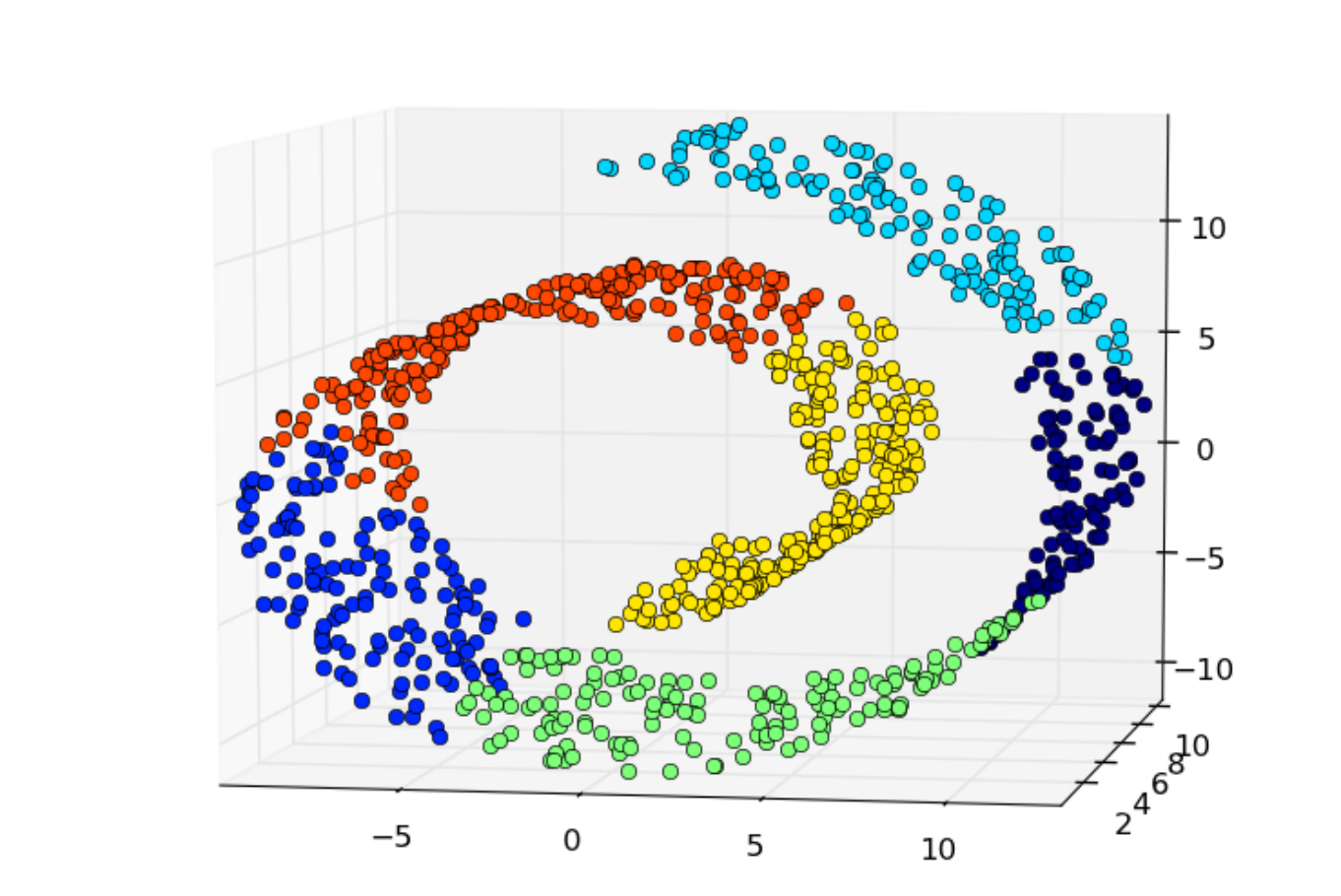


Semisupervised Learning: Label propagation



Semisupervised Learning: Label propagation

Manifold



Selección de variables – Reducción dimensionalidad

PCA (Análisis Componentes Principales)

Stepwise

Regularización: Lasso

Ensembling: Muestreo de variables

T-SNE (t-Distributed Stochastic Neighbor Embedding)