Introducción al Machine Learning

Caso Práctico: Predicción de Demanda

Septiembre 2016

20k+
REGISTERED CUSTOMERS

7.8M+
TAREAS

720k+
DATASETS

4.6M+
MODELOS

15
DESARROLLADORES

- Fundada en **Enero 2011** para automatizar el machine learning.
- Pionera en **MLAAS.**
- **12 aplicaciones de patentes**
- **API-first** con una UI sencilla y cuidada.
- Implementaciones privadas **cloud-based** y **on-premise** para empresas

Gartner 2015
CoolVendor

Co-founder and
Chief Scientist
BigML, Inc

Machine Learning Community

1980 1990 2000 2010 **2016**

El primer workshop de Machine Learning tuvo lugar en Pittsburgh en 1980.

Agenda

1	¿Qué es el Machine Learning?
2	Origen y Evolución
3	Aplicación del Machine Learning
4	Caso Práctico – Predicción de Demanda

¿Por qué Machine Learning?

Cliente	Minutos	SMS	Compras	Datos	Edad	Cancelación
cliente_1	148	72	0	33,6	50	Verdadero
cliente_2	85	66	0	26,6	31	Falso
cliente_3	183	64	0	23,3	32	Verdadero
cliente_4	89	66	94	28,1	21	Falso
cliente_5	115	0	0	35,3	29	Falso
cliente_6	166	72	175	25,8	51	Verdadero
cliente_7	100	0	0	30	32	Verdadero
cliente_8	118	84	230	45,8	31	Verdadero
cliente_9	171	110	240	45,4	54	Verdadero
cliente_10	159	64	0	27,4	40	Falso

¿Puedes encontrar un patrón que te ayude a predecir la cancelación de un cliente?

ML vs Capacidad humana

- Importancia de las variables en la variable objetivo:
  1) **SMS:** 55%
  2) **Edad:** 45%

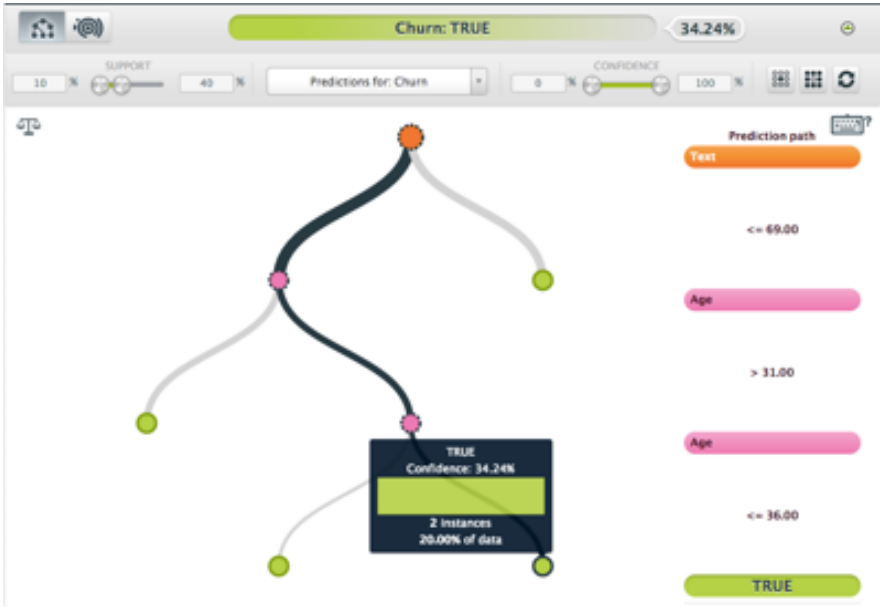
- Resumen de las reglas:

Cancelan:

- (SMS < 69) y (31 < Edad < =36)

No cancelan:

- (SMS < 69) y (Edad < 31)
- (SMS > 69) y (36 < Edad)



Otra manera de verlo

Too Small

Too Big

Definición

"Hacer que los ordenadores aprendan sin que se les programe de forma explícita"

Arthur Samuel, 1959

Otra definición

Utilizar datos para encontrar patrones que se repetirán en el futuro y poder hacer mejores decisiones en el presente

¿Cómo funciona?

- No requiere relaciones causa-consecuencia conocidas. Explota correlaciones.
- No necesita conocer todas las variables que influyen para alcanzar un gran performance.
- No es preciso un % de acierto alto para aportar mucho valor al negocio. Basta con mejorar el método sin ML, a veces incluso por poca diferencia.

Agenda

1	¿Qué es el Machine Learning?
2	Origen y Evolución
3	Aplicación del Machine Learning
4	Caso Práctico – Predicción de Demanda

Orígenes del ML

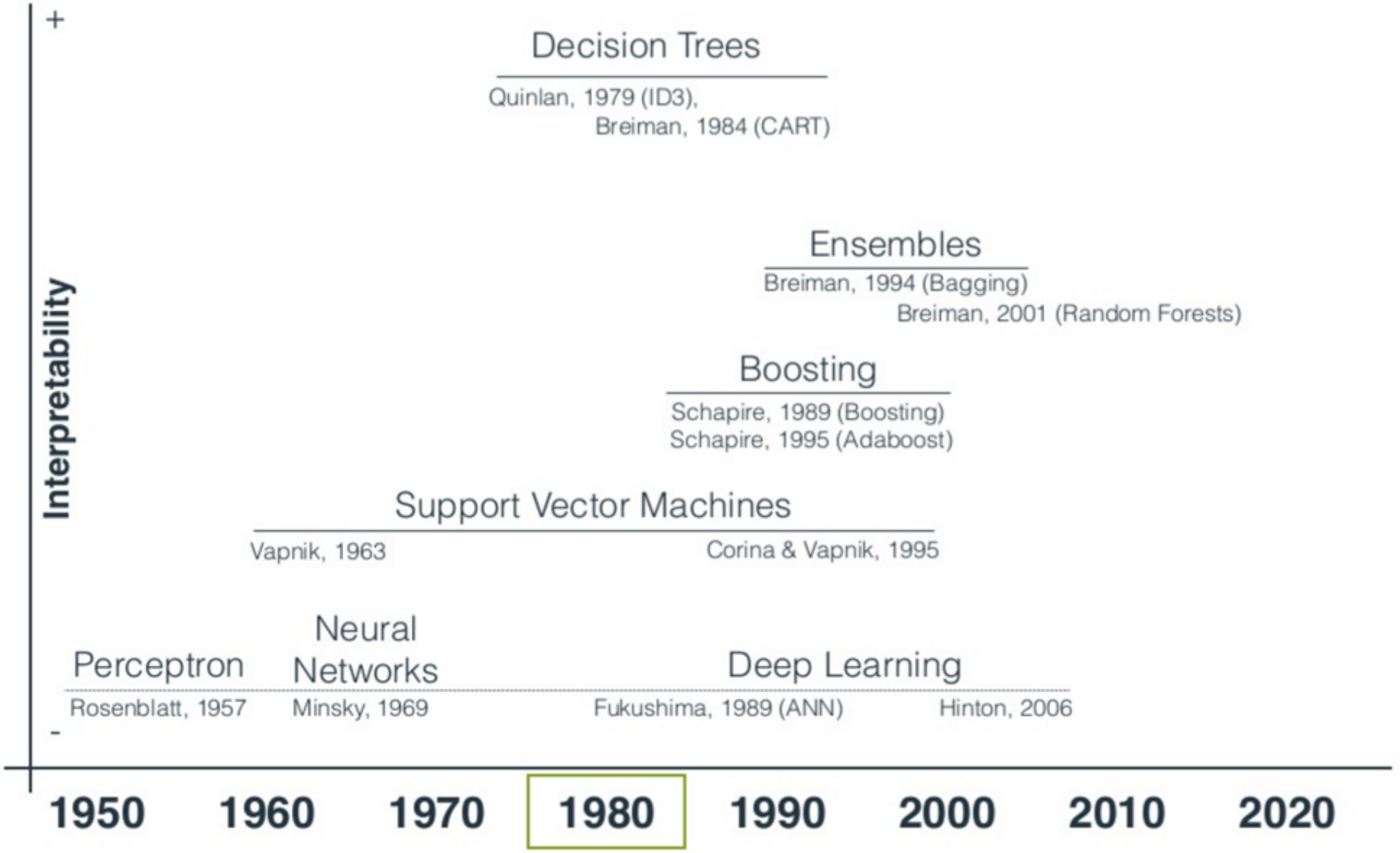
”A field of study that gives computers the ability to learn without being explicitly programmed”

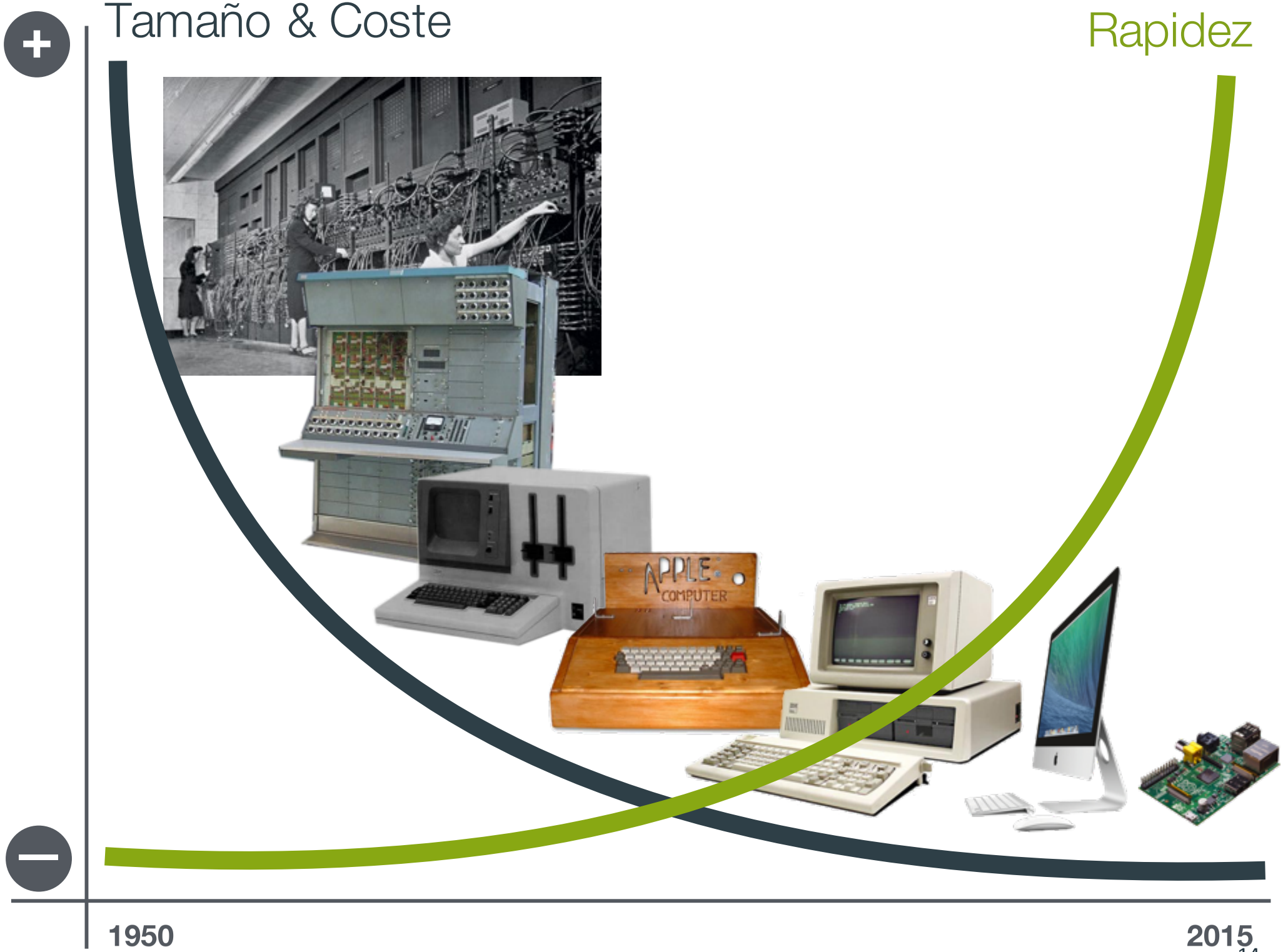
Professor Arthur Samuel

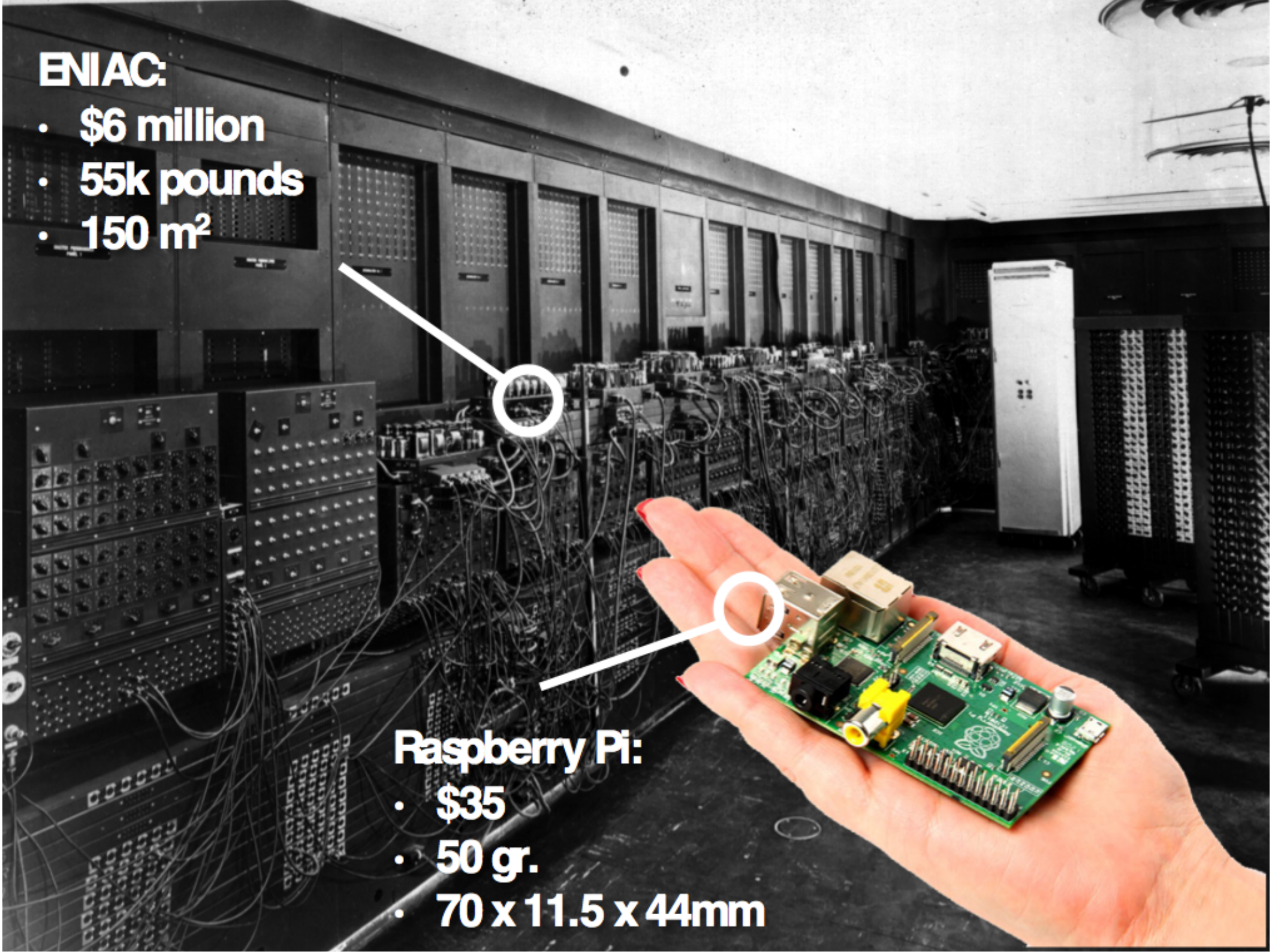
- El primer programa de auto-aprendizaje del mundo servía para jugar a las damas y fue desarrollado para **IBM** por el Profesor **Arthur Samuel** en 1952.

- Thomas J. **Watson** Sr., el fundador y Presidente de IBM, predijo que la demostración de Samuel iba a provocar un incremento de 15 puntos en la acción de IBM. Acertó.

Orígenes de la Algoritmia




Tamaño & Coste
Rapidez
+
−
APPLE
COMPUTER
1950
2015


ENIAC:
• $6 million
• 55k pounds
• 150 m²
Raspberry Pi:
• $35
• 50 gr.
• 70 x 11.5 x 44mm

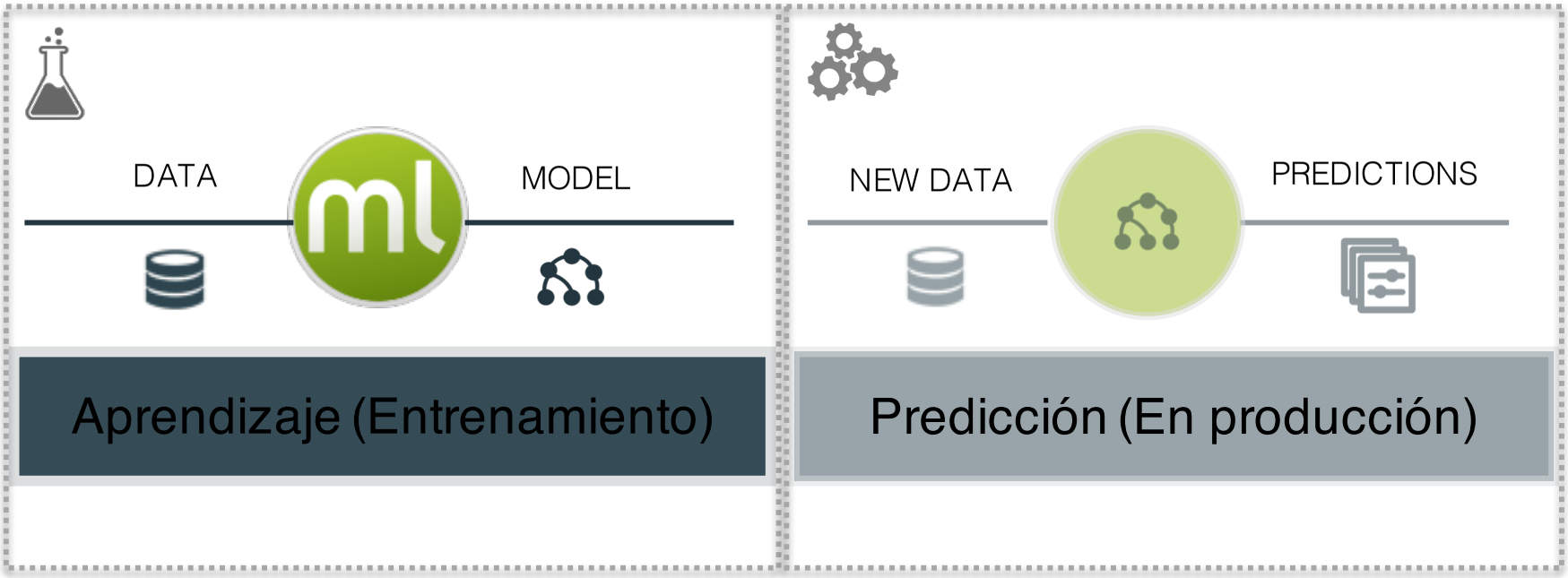
Hitos para el gran público

2016
AlphaGo

2011
Watson

1997
Deep Blue

El Machine Learning hasta hoy

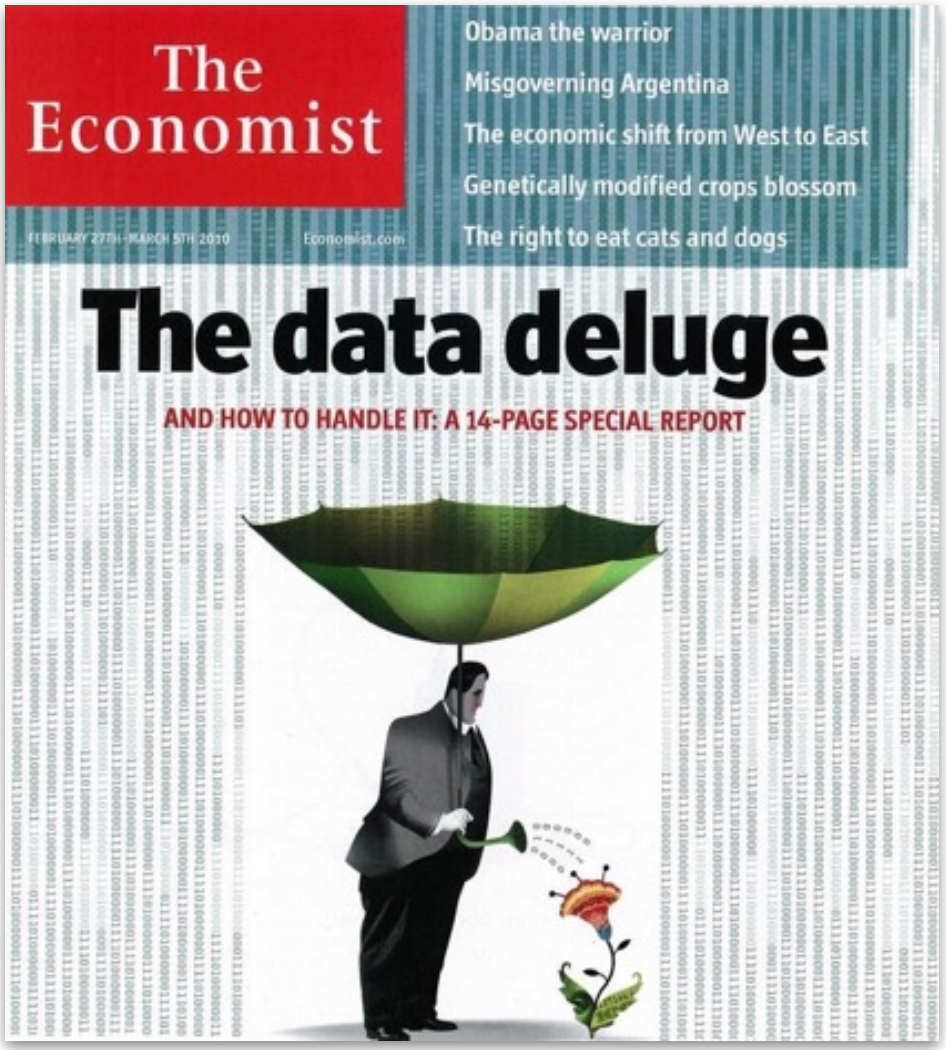


- La mayoría de herramientas se han centrado en entrenar modelos para **usos académicos** y **científicos** "en laboratorio"
- Hasta la actualidad el ML no se ha utilizado para resolver los **problemas del mundo real** en entornos de producción

Herramientas de ML del pasado

- De científicos (con un Ph.D.) para científicos (con un Ph.D.).
- Exceso de algoritmos
- Aplicaciones de escritorio sólo para grandes datasets
- **Demasiado complicado** para el trabajador común
- **Demasiado simple** para problemas del mundo real
- **Pobremente desarrollado** para casos de mundo real
- Las herramientas comerciales (SPSS, SAS) no sólo heredan esos problemas sino que además son tremendamente caras.

El origen de BigML



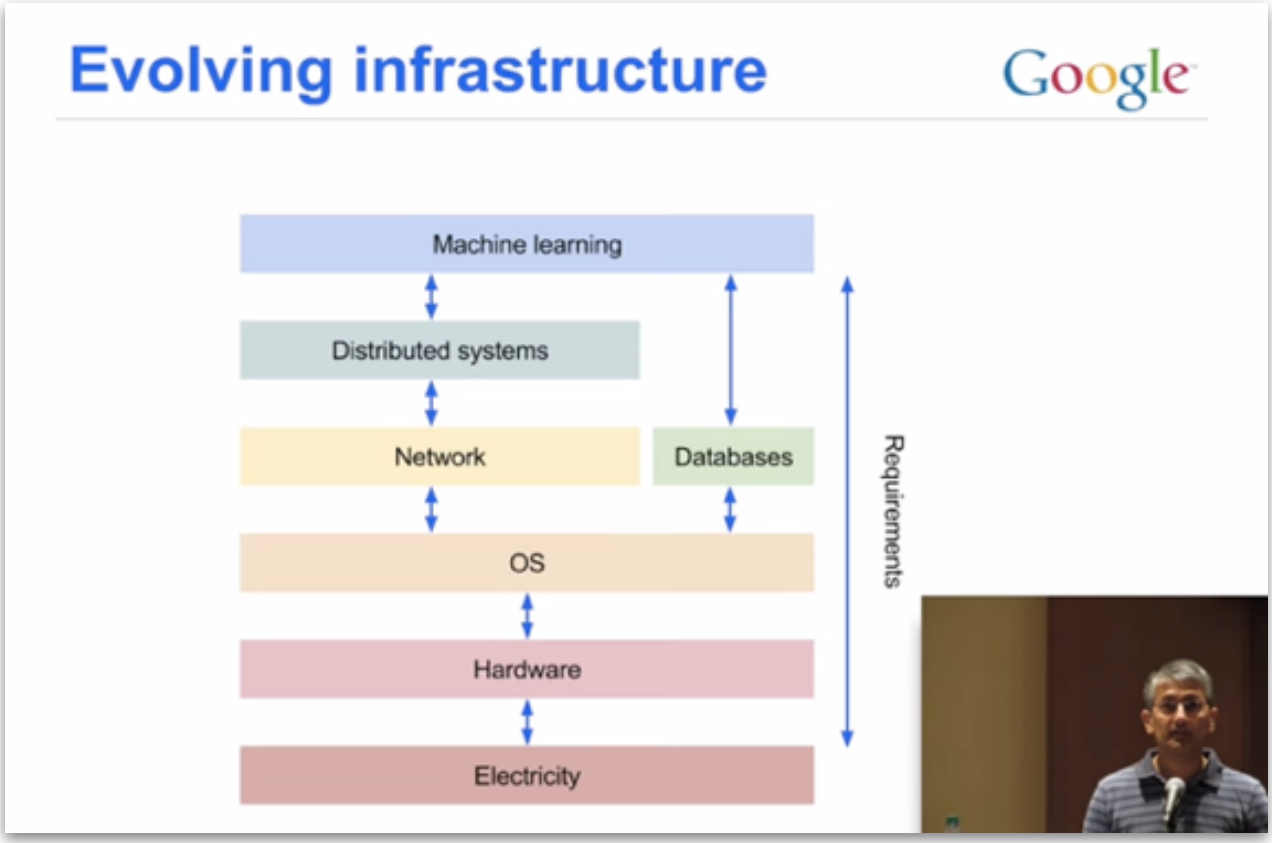
Febrero del 2010

El Machine Learning por su gran complejidad sólo está al alcance de grandes compañías tecnológicas que se pueden permitir muchos especialistas o bien de aquellas que se pueden permitir el elevado coste de soluciones a medida

Idea:

¡Crear una plataforma que democratice el Machine Learning!

ML – una nueva capa

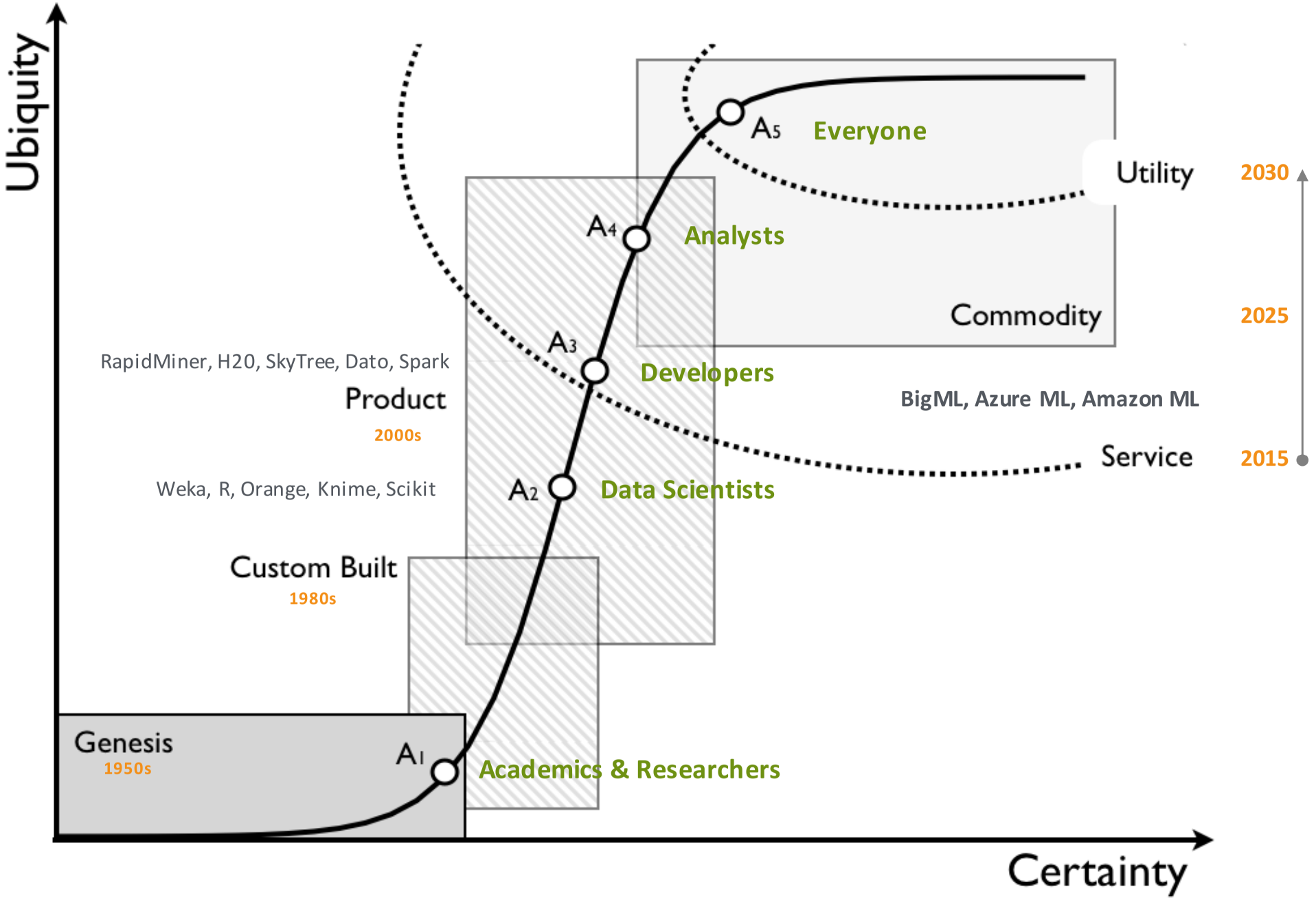


Tushar Chandra, Google

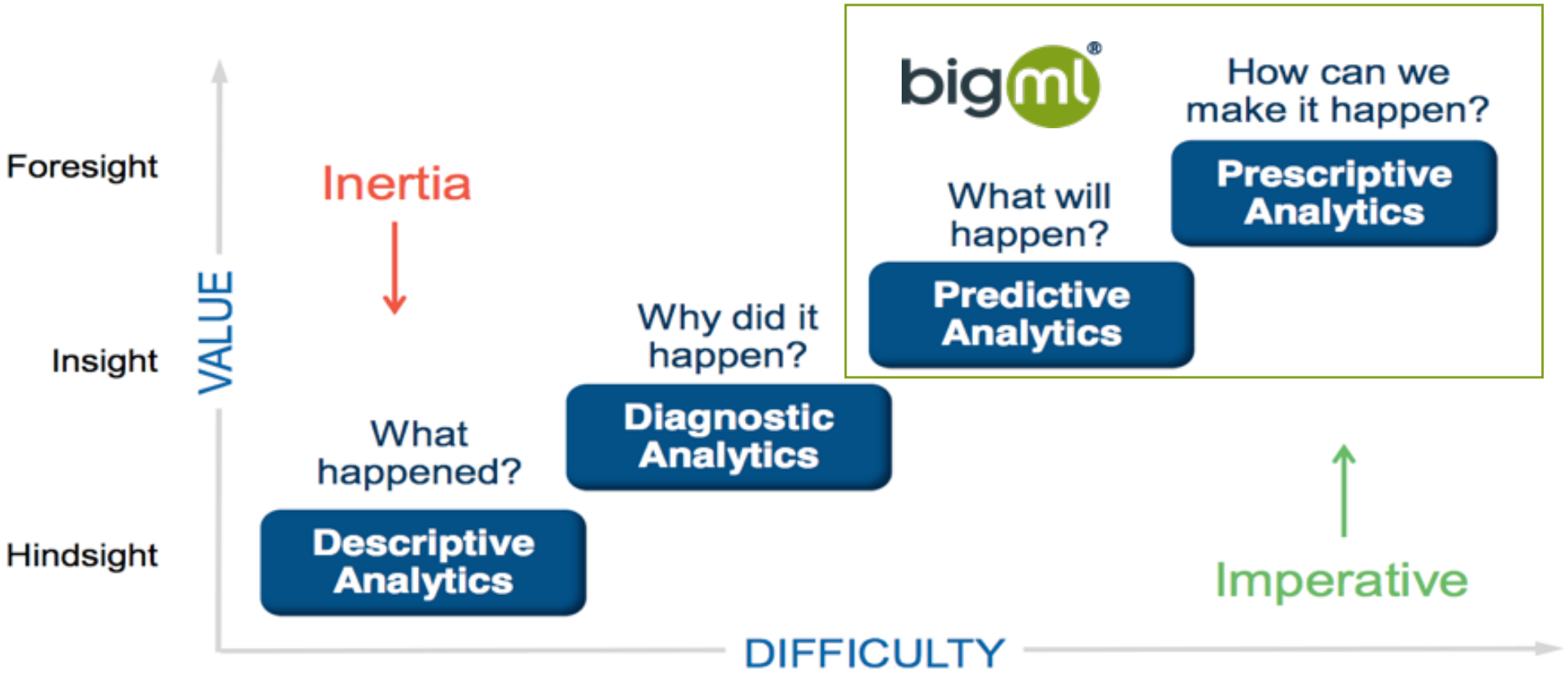
- Machine Learning se está convirtiendo en una nueva capa de abstracción de la infraestructura de computación.

- Un desarrollador de aplicaciones de empresa espera tener acceso a una plataforma de Machine Learning.

Democratización del ML



Evolución de la Analítica



Agenda

1	¿Qué es el Machine Learning?
2	Origen y Evolución
3	Aplicación del Machine Learning
4	Caso Práctico – Predicción de Demanda

¿Quién usa BigML?

Más de 20.000 usuarios de todos los sectores que han creado entorno a 4,6M de modelos predictivos

Empresas

Startups y empresas de software

Innovar en los negocios existentes

PYMEs

Aplicar ML para expandir y mejorar los análisis tradicionales

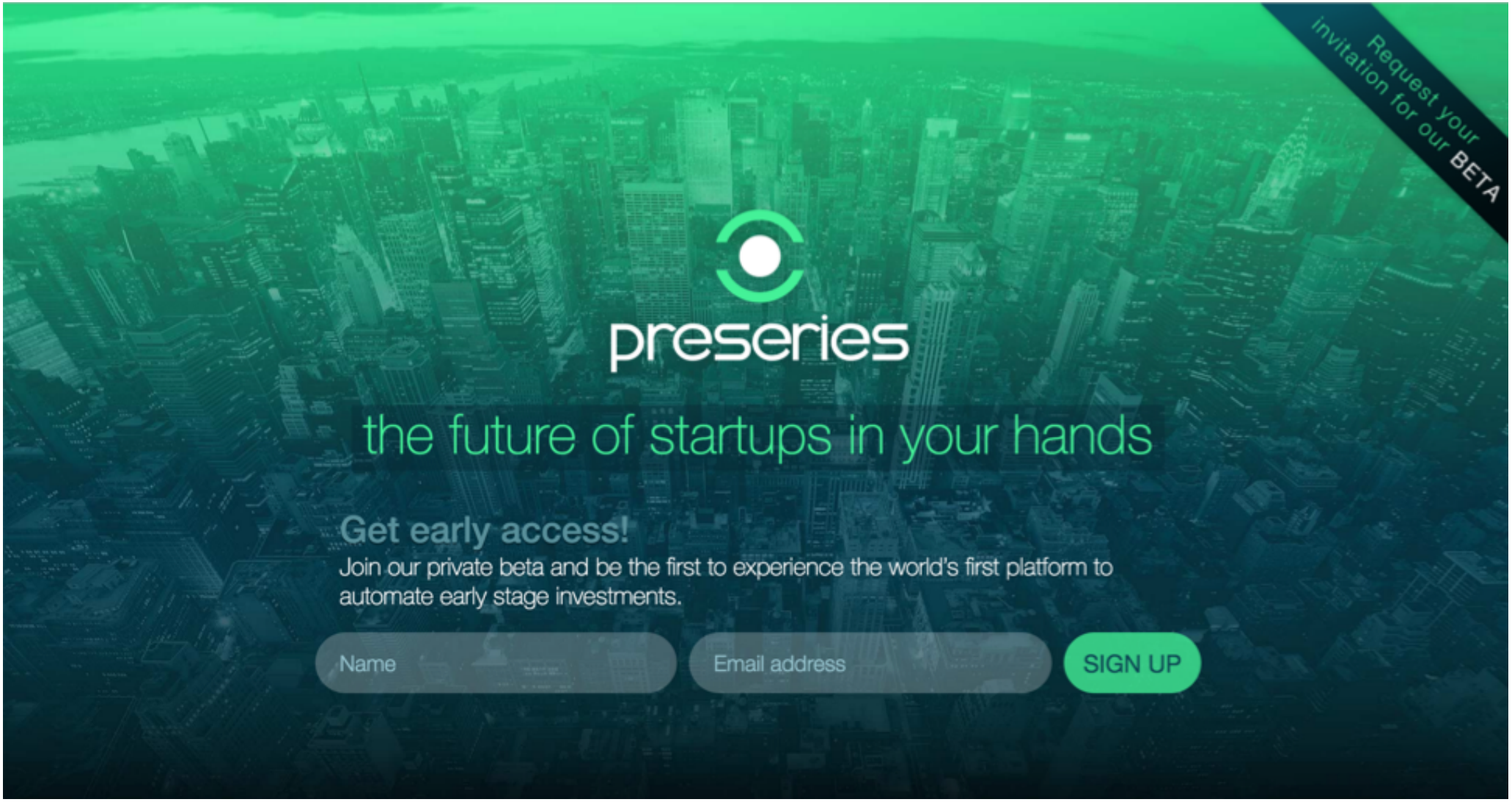
Grandes compañías

Aumentar o reemplazar sus actuales procesos de ML

Ejemplos

- **Marketing**: priorización de potenciales clientes, análisis de cancelaciones, retención de clientes
- **Salud y biología**: diagnósticos de pacientes, administración de hospitales, I+D para medicamentos
- **Recursos Humanos**: ranking y priorización de candidatos, bajas de empleados
- **Publicidad**: análisis de campañas, contenido dirigido, análisis de sentimiento
- **Operaciones**: análisis de fraude, renovaciones de licencias, morosidad

Predecir el éxito de Startups

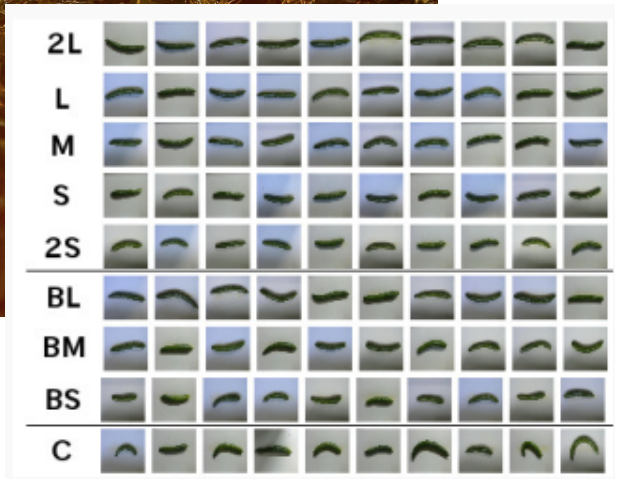


https://www.youtube.com/watch?v=IWmTs0-K-3E

Crear el Trailer de una Película

https://www.youtube.com/watch?v=gJEzuYynaiw

Clasificar Pepinos



http://qz.com/771921/the-ultimate-promise-of-artificial-intelligence-lies-in-sorting-cucumbers/

Neveras Inteligentes

https://techcrunch.com/2016/09/02/microsoft-is-putting-cortana-machine-learning-in-a-fridge/

El Dataset

features → | label

instances ↓

brand	power	age	duty	temp	humidity	FAIL?
koala	45	338	1	16	0.03	FALSE
otter	15	140	1	27	0.27	FALSE
koala	15	315	1	19	0.37	TRUE
otter	45	338	1	29	0.27	TRUE
koala	45	211	1	23	0.85	TRUE
otter	15	328	1	17	0.56	FALSE
koala	15	318	2	22	0.45	TRUE

Aprendizaje Supervisado

Classification

animal	state	…	proximity	action
tiger	hungry	…	close	run
elephant	happy	…	far	take picture

label

Regression

animal	state	…	proximity	min_kmh
tiger	hungry	…	close	70
hippo	angry	…	far	10

Multi-Label Classification

animal	state	…	proximity	action1	action2
tiger	hungry	…	close	run	look untasty
elephant	happy	…	far	take picture	call friends

Aprendizaje No Supervisado

Clustering

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
The	Sally	6788	sign	food	26339	51

similar

Anomaly Detection

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
The	Sally	6788	sign	food	26339	51

unusual

Aprendizaje No Supervisado

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
The	Sally	6788	sign	food	26339	51

Rules:

{customer = Bob, account = 3421} → zip = 46140

{class = gas} → amount > 80

Algoritmos

Aprendizaje supervisado

Clasificación

Árboles de decisión, Ensembles

- Riesgo de cancelación
- Análisis de crédito
- Análisis de riesgo
- Análisis de sentimiento

Regresión logística

- Análisis de campañas
- Mantenimiento predictivo
- Next Best Offer (“NBO”)
- Priorización de contenido

Detector de anomalías

- Diagnóstico de pacientes
- Análisis de retención
- Reclutamiento por objetivos
- Análisis de fraude

Regresión

- Lifetime Value
- Publicidad predictiva

- Optimización de precios
- Estimación de ventas

Sólo para problemas de clasificación

Aprendizaje no supervisado

Clusters

- Análisis de fraude
- Segmentación de mercado
- Segmentación de clientes
- Gestión de portfolios

Detector de anomalías

- Análisis de fraude
- Limpieza de datos
- Detección de intrusos
- Autenticación

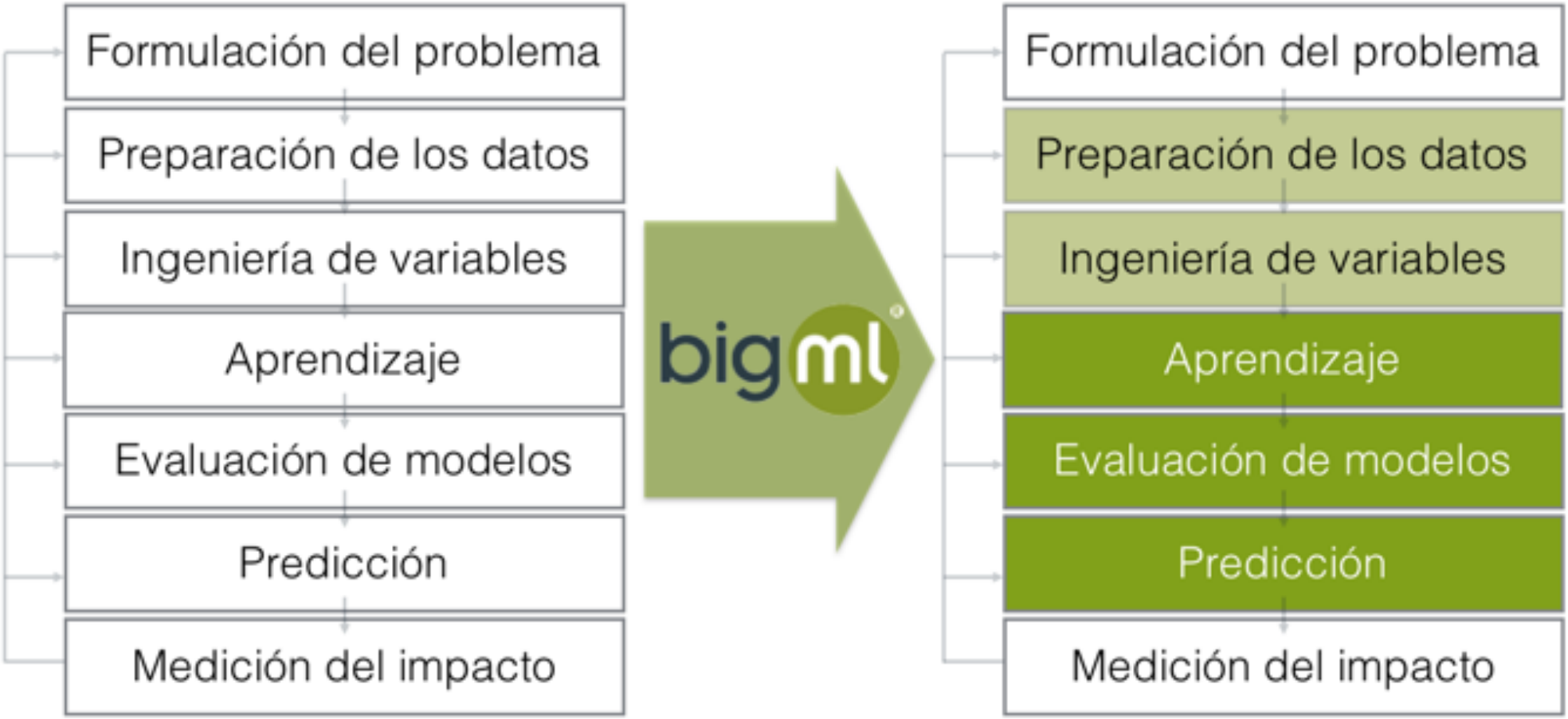
Asociaciones

- Market-Basket Analysis
- Patrones de UX
- Bioinformatica
- Detección de incidentes y Digital Forensics

Agenda

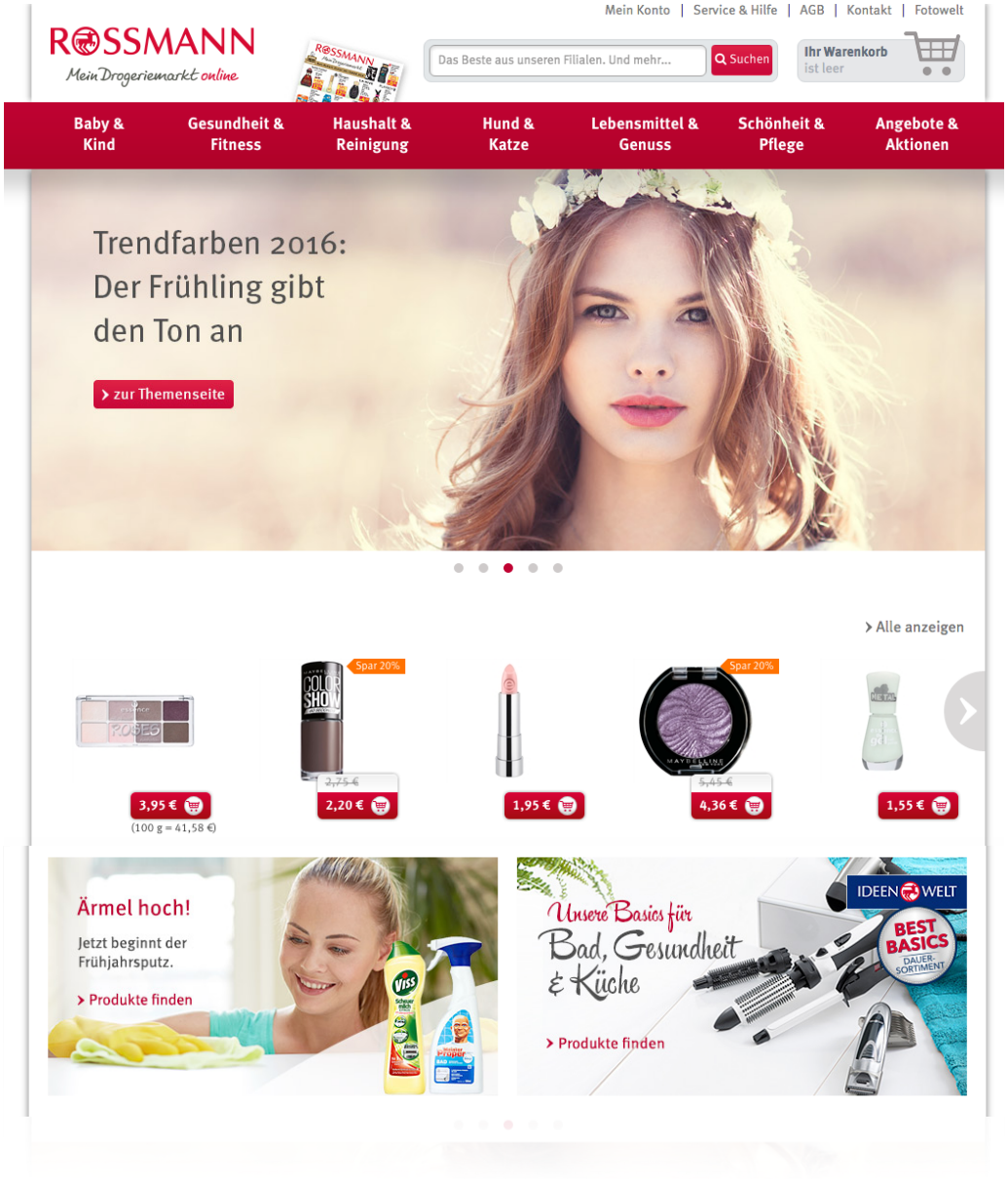
1	¿Qué es el Machine Learning?
2	Origen y Evolución
3	Aplicación del Machine Learning
4	Caso Práctico – Predicción de Demanda

Fases del proceso predictivo

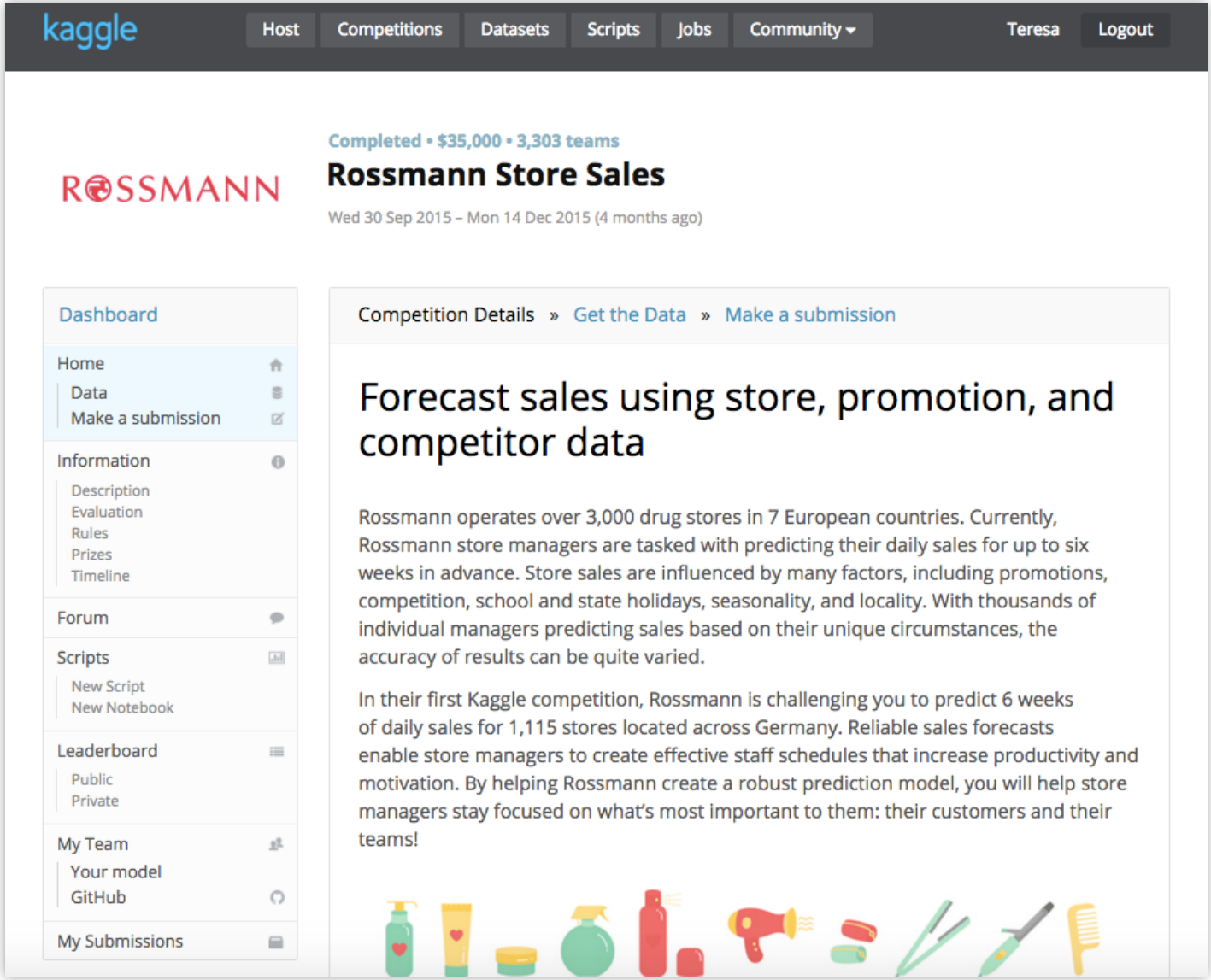


Rossmann

- Segunda cadena más importante de droguería y perfumería de **Alemania**
- 3,000 tiendas en 7 países europeos



El Caso



https://www.kaggle.com/c/rossmann-store-sales/

El Dataset

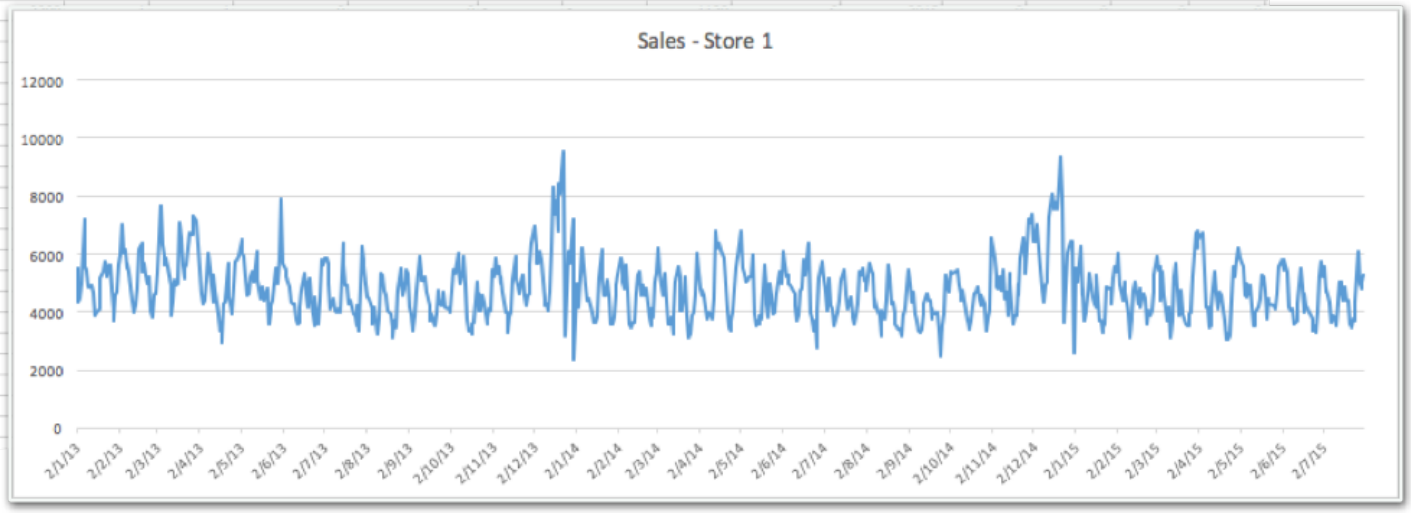
Histórico de **ventas** diarias durante **+2.5 años** (desde enero de 2013 hasta julio de 2015) de **1,115 tiendas** de Alemania

Media de ventas / día: 6.480€

Variables

Instancias

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
131	3	01/07/15	5977	570	1	1	0	1	c	a	920	7	2015	0	0	0	0
304	3	01/07/15	9370	1341	1	1	0	0	a	a	1950	7	2015	0	0	0	0
403	3	01/07/15	8905	788	1	1	0	1	a	a	4970	7	2015	0	0	0	0
859	3	01/07/15	9235	776	1	1	0	0	c	a	21770	7	2015	0	0	0	0
944	3	01/07/15	8854	1177	1	1	0	0	c	a	1670	7	2015	0	0	0	0
1053	3	01/07/15	12617	1290	1	1	0	0	a	a	1710	7	2015	0	0	0	0
269	1	01/06/15	17960	1935	1	1	0	0	a	c	60	6	2015	0	0	0	0
496	1	01/06/15	14251	1140	1	1	0	0	d	c	2780	6	2015	0	0	0	0
550	1	01/06/15	9645	653	1	1	0	0	d	c	50	6	2015	0	0	0	0
595	1	01/06/15	23971														
718	1	01/06/15	10765														
225	5	01/05/15	2401														
5	3	01/04/15	7720														
286	3	01/04/15	9218														
630	3	01/04/15	10916														
770	3	01/04/15	10145														
837	3	01/04/15	5574														
918	3	01/04/15	9079														
1034	3	01/04/15	10870														
37	1	01/12/14	10616														
828	1	01/12/14	9807														
878	1	01/12/14	15554														
556	6	01/11/14	7300														
599	6	01/11/14	11571														
8	3	01/10/14	6826														
95	3	01/10/14	9734														
138	3	01/10/14	9181														
819	3	01/10/14	7867														
249	1	01/09/14	8525														
84	5	01/08/14	13569														



Variables

Variables	Tipo	Descripción
Store	id	identificador único para cada tienda
DayOfWeek	entero	día de la semana
Date	dd/mm/yy	día, mes y año en que se realizó la venta
Sales	real	facturación total durante un día dado
Customers	entero	número total de clientes durante un día dado
Open	categórico	si la tienda estaba abierta (1) o cerrada (0) durante ese día
Promo	categórico	si la tienda estaba haciendo alguna promoción ese día (1) o no (0)
StateHoliday	categórico	si había alguna fiesta a nivel nacional (a=fiesta nacional, b= pascuas, c= navidad, 0=no fiesta)
SchoolHoliday	categórico	si ese día era festivo escolar (1) o no (0)
StoreType	categórico	diferencia entre 4 tipos distintos de tiendas (a, b, c, d)
Assortment	categórico	diferencia el nivel de diversidad de producto de la tienda (a=básico, b=extra, c=extendido)
CompetitionDistance	entero	la distancia en metros del competidor más cercano
CompetitionOpenSinceMonth	entero	mes aproximado de apertura del competidor más cercano
CompetitionOpenSinceYear	entero	año aproximado de apertura del competidor más cercano
Promo2	categórico	es una promoción continuada que poseen algunas tiendas (1=participa, 0=no participa)
Promo2SinceWeek	entero	en qué semana del año empezó la tienda a participar en la Promo2
Promo2SinceYear	entero	en qué año empezó la tienda a participar en la Promo2
PromoInterval	categórico	describe los intervalos consecutivos en los que la Promo2 empieza (por ejemplo: Feb,May,Aug,Nov)

Resumen de Variables

Para cada día:

- Perfil de Tienda: Tipo de tienda, Surtido, Grado de Competencia.
- Actividad Comercial: Abierto/Cerrado, # Customers, Venta.
- Actividades Promocionales
- Vacaciones oficiales y escolares

Feature Engineering

1. Eliminar instancias si Open=0, o Sales<=0.

2. Creación automática de variables a partir de fechas.

3. Reemplazar *CompetitionOpenSinceMonth* y *CompetitonOpenSinceYear* por el número de meses que la competencia lleva abierta en relación al mes en el que se realiza la venta (CompetitionAge).

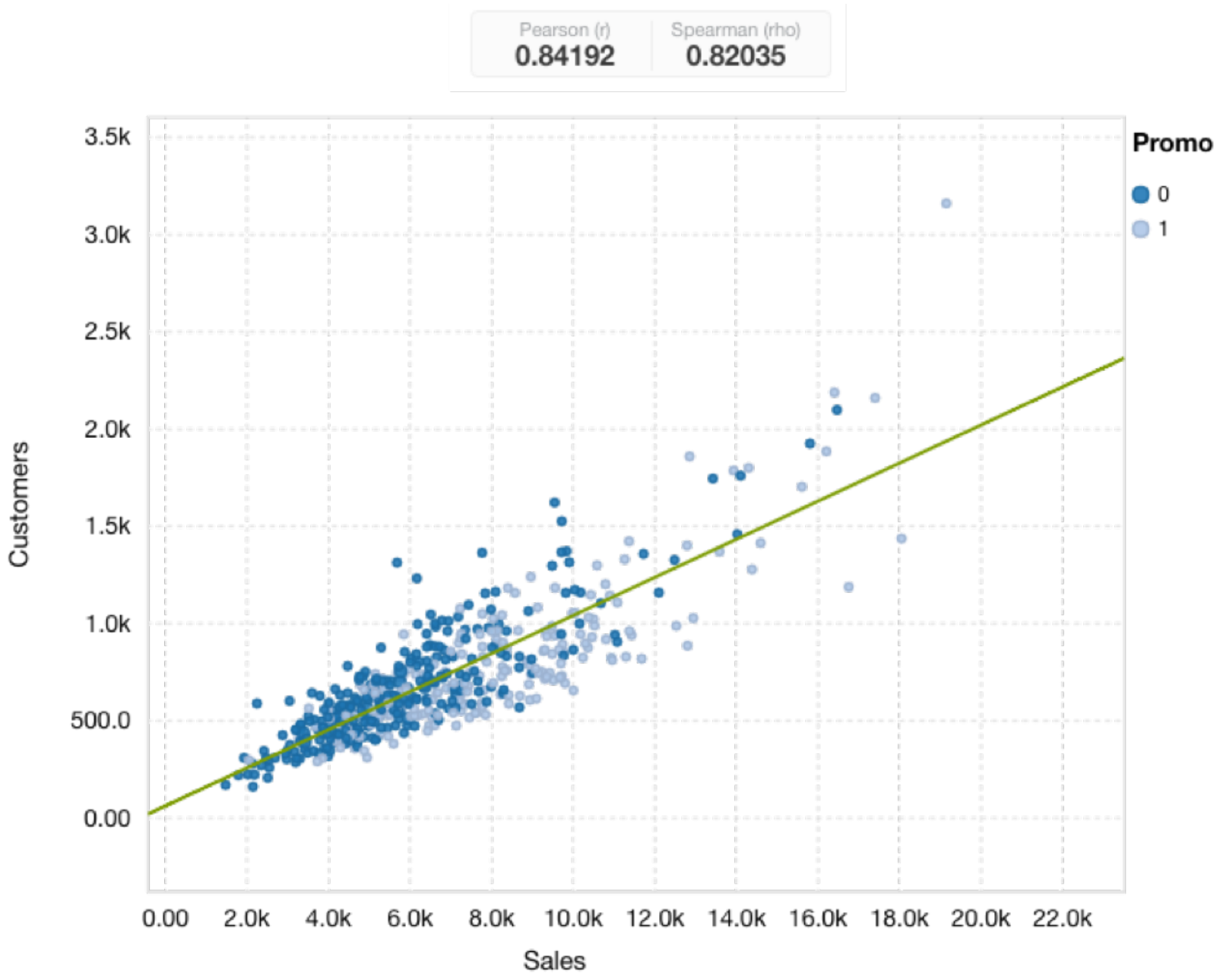
5. Reemplazar *CompetitionDistance* por la misma variable pero que tenga en cuenta si la competencia estaba abierta en el día en que se realiza la venta. En caso que no hubiera competencia se sustituye la distancia por un "N/A" (CompetitionDistance2).

6. Reemplazar *Promo2SinceWeek* y *Promo2SinceYear* por una variable categórica que indique si en el día en que se realiza la venta, la *Promo2* ya estaba vigente o no (Promo2Started).

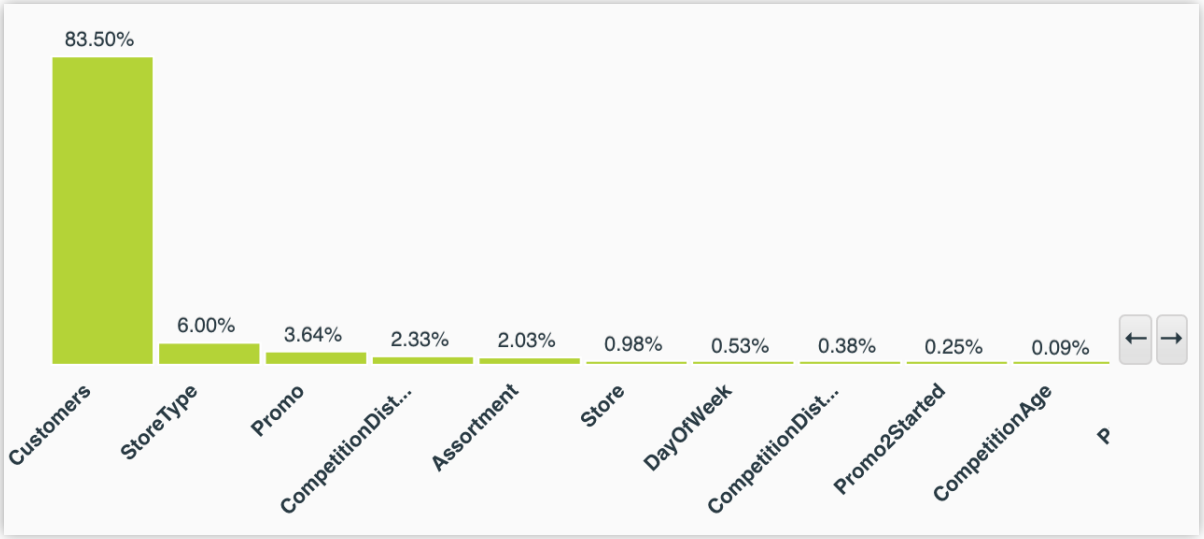
7. Reemplazar *PromoInterval* por otra variable categórica que indique si en el mes en que se realiza la venta estaba realizándose la *Promo2* o no (Promo2Running).

Proceso Rápido

Oops...


Pearson (r)
0.84192
Spearman (rho)
0.82035
Promo
0
1
Customers
3.5k
3.0k
2.5k
2.0k
1.5k
1.0k
500.0
0.00
0.00
2.0k
4.0k
6.0k
8.0k
10.0k
12.0k
14.0k
16.0k
18.0k
20.0k
22.0k
Sales

Aterrizaje

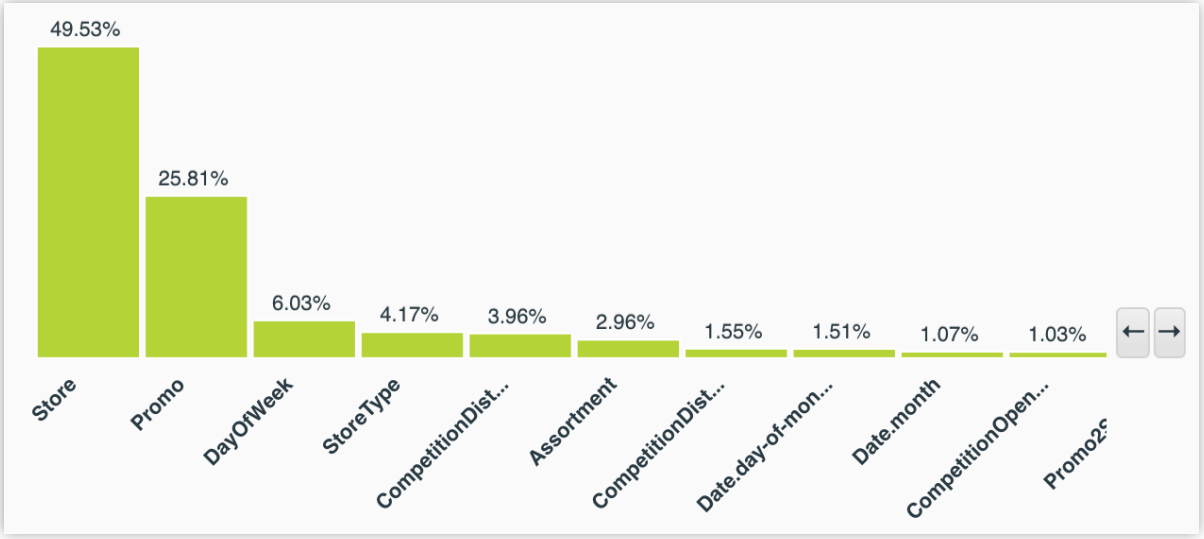


MODEL

MEAN ABSOLUTE ERROR
835.92

MEAN SQUARED ERROR
1,263,744.16

R SQUARED
0.86



MODEL

MEAN ABSOLUTE ERROR
1,581.51

MEAN SQUARED ERROR
4,329,084.46

R SQUARED
0.53

Más Feature Engineering

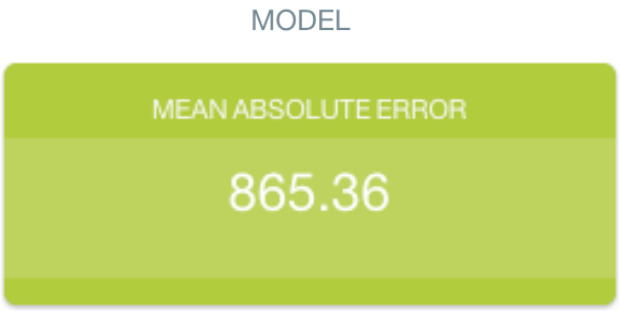
Para cada día:

- Perfil de Tienda: Tipo, Surtido, Grado de Competencia…
- Actividad Comercial: Abierto/Cerrado, # Customers, Ventas…
- Actividades Promocionales
- Vacaciones oficiales y escolares
- Ventas y Customers: Día previo / Semana previa / Últimos 15 días
- Ticket Medio: Última semana / Último mes
- Customers y Ticket Medio: Media histórica para días con mismas promociones

Resultados (I)

Modelo simple sin Customers

Modelo simple, sin Customers, con features adicionales



Resultados (II)

Modelo simple sin Customers

MODEL

MEAN ABSOLUTE ERROR

1,581.51

Modelo simple, sin Customers, con features adicionales

MODEL

MEAN ABSOLUTE ERROR

865.36

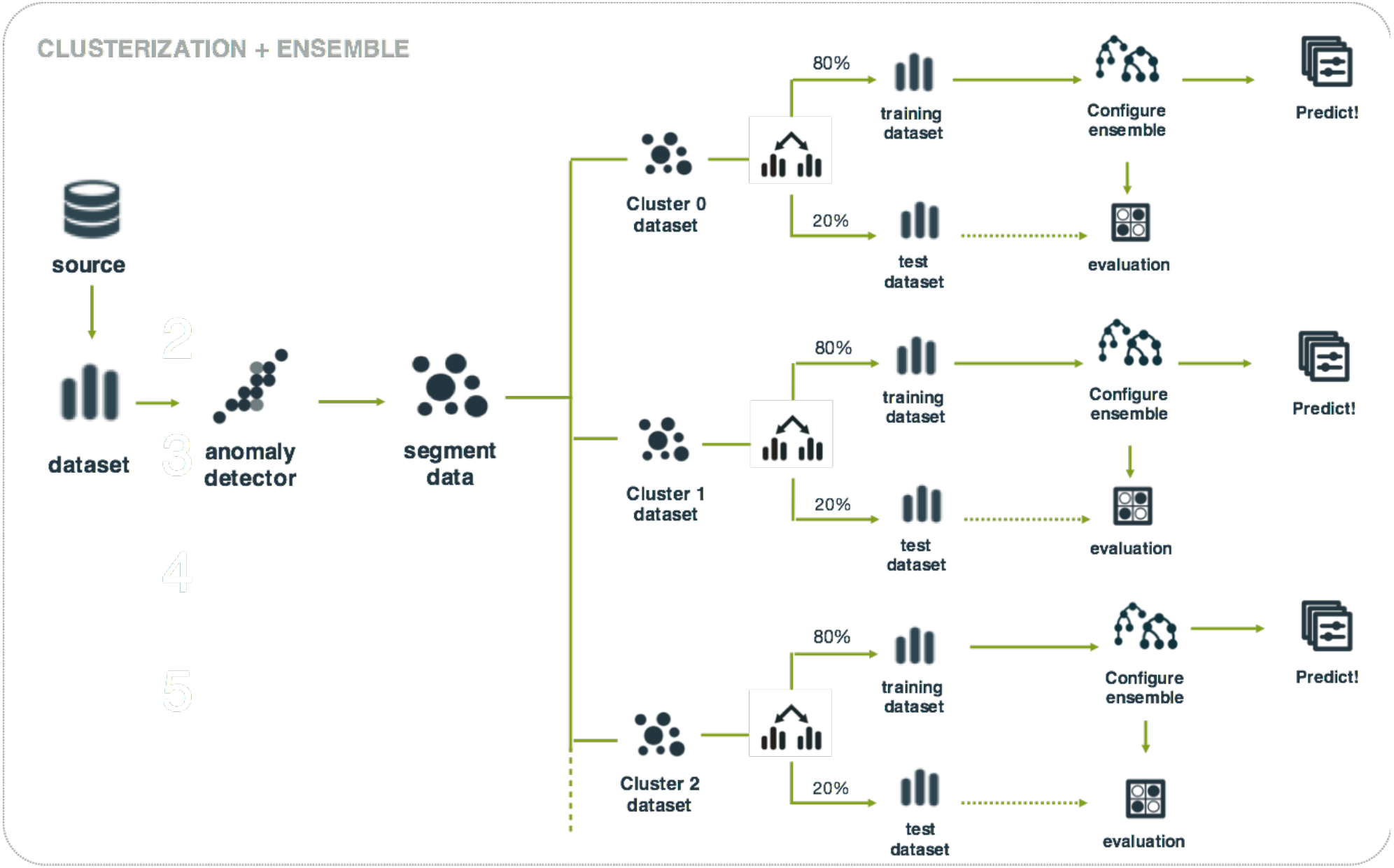
Ensemble, sin Customers, con features adicionales

ENSEMBLE

MEAN ABSOLUTE ERROR

816.59

Proceso más Depurado



Resultados (III)

Modelo simple sin Customers

MODEL

MEAN ABSOLUTE ERROR

1,581.51

Modelo simple, sin Customers, con features adicionales

MODEL

MEAN ABSOLUTE ERROR

865.36

Ensemble, sin Customers, con features adicionales

ENSEMBLE

MEAN ABSOLUTE ERROR

816.59

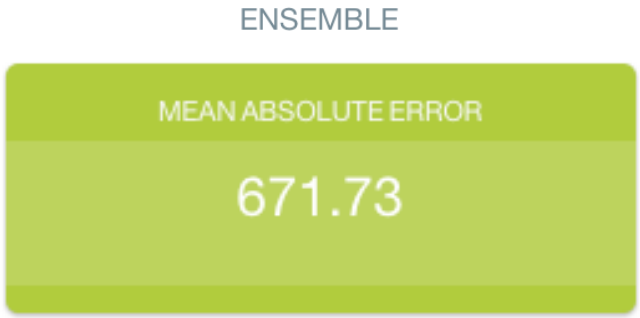
Ensemble sobre Segmento de Tiendas

ENSEMBLE

MEAN ABSOLUTE ERROR

671.73

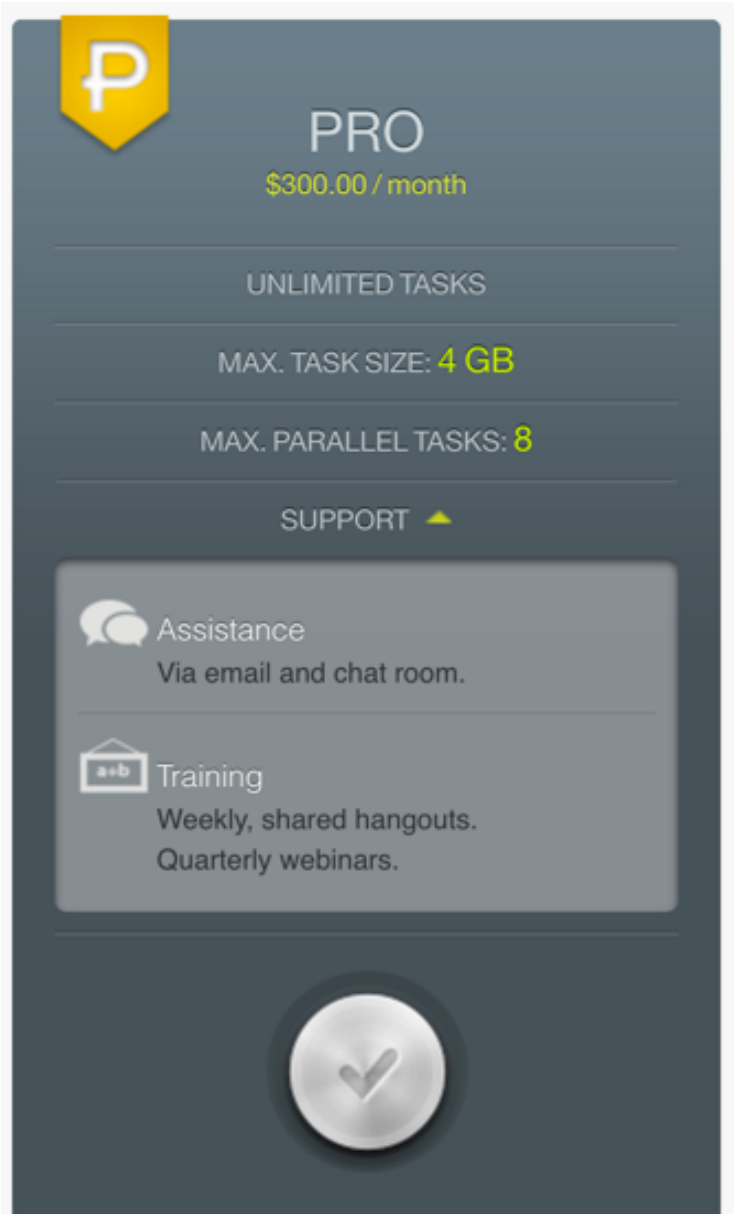
Conclusiones

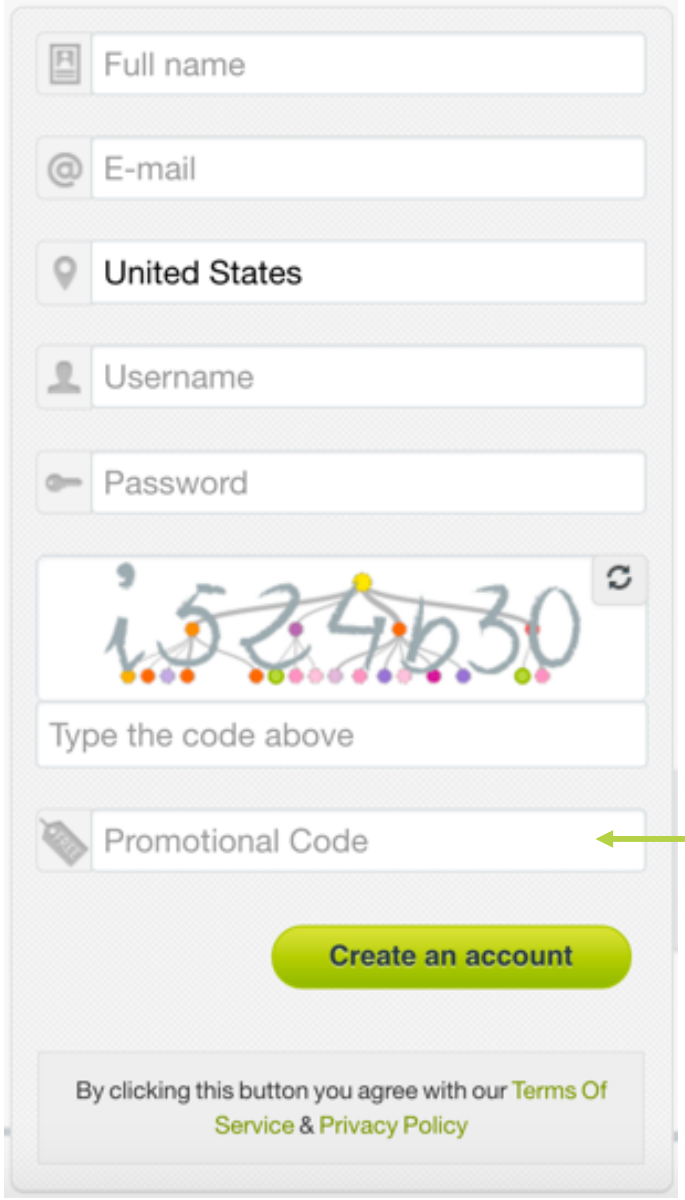


% desviación diaria media = 672/ 6,480 ≈ **10%**

- Ajustar modelo por función de costes – Mejor pasarme que quedarme corto?
- Añadir nuevas variables: Datos de producto y categorías, Cross-selling, Climatología, Fechas especiales, Partidos de Fútbol, Calendario de Congresos…
- Modelizar por producto-tienda / día
- Segmentar días por vacaciones, o vísperas de vacaciones o promos. O tratar directamente la estacionalidad.
- Segmentar mejor tiendas, por espacio, por mejores/peores ventas…
- Derivar mejores métricas para relacionar variables
- Iterar hasta alcanzar el mejor mix (*automatizar este proceso*)

Código para los asistentes





ctpml

Teresa Alvarez Soler

Product Manager

alvarez@bigml.com