

# Machine Learning Práctico

Pablo Zivic

@ideasrapidas



# Agenda

- Objetivo del curso
- Dinámica
  - Pre-requisitos
  - Classroom
  - Grupo de facebook
  - Teóricas / Prácticas
- Teórica
  - Definiciones
  - Metodología
- Práctica
  - Presentación de la problemática que vamos a trabajar en el curso
  - Coding slots

Link a slides



# Objetivo del curso

Utilizar un caso de negocio ficticio como excusa  
para abordar algunas técnicas y metodologías útiles  
para crear valor con Machine Learning

[illegible]

# Pre-requisitos: la herramienta para *enfocar* el curso

- **Gradiente:**

- Significado y utilización en optimización de funciones. Regla de la cadena

- **Probabilidad:**

- Noción de esperanza, esperanza condicional y test de hipótesis

- **Programación:**

- Funciones, módulos, algunas técnicas de visualización, paquetes estadísticos.

# Dinámica

- Clase práctica
  - Clase de consultas
  - Revisitamos algún contenido
  - Repasamos algún ejercicio de la guía
  - Revisamos alguna pregunta del cuestionario de la clase anterior
- Clase teórica
  - Clase Teórica
  - Clase Práctica
  - In class tasks
- Guías de ejercicios por clase
- Multiple choice por clase para entregar por clase

# Anuncios y material


- Todo el material va a estar en <http://bit.ly/ml-practico>
  - Lo vamos a subir a medida que vamos dando los contenidos
- Los anuncios y demás ocurrirán en [el grupo de facebook:](http://bit.ly/ml-practico-fb)  
<http://bit.ly/ml-practico-fb>

Si no te sumaste aun, usá el link para solicitar sumarte















# Preguntas y respuestas


**Dave Taylor** created a poll.  
19 hrs


What's your vote for best science fiction movie of all time?

<input type="checkbox"/>	Star Wars IV: A New Hope <i>Added by you</i>			+39
<input type="checkbox"/>	The Day the Earth Stood Still (1951) <i>Added by Matt Hart</i>			+26
<input checked="" type="checkbox"/>	2001: A Space Odyssey <i>Added by you</i>			+18
<input type="checkbox"/>	Alien <i>Added by Lee Reeves Allen</i>			+8
<input type="checkbox"/>	Blade Runner <i>Added by you</i>			+5
6 More Options...				

 10

22 Comments

 Like

 Comment

**Empecemos!**

# Algunas definiciones

- Artificial Intelligence
  - The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. (Oxford)

# Algunas definiciones

- Artificial Intelligence

- The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. (Oxford)

- Machine Learning

- “Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience [2].” An algorithm can be thought of as a set of rules/instructions that a computer programmer specifies, which a computer is able to process. (Mitchell)

# Algunas definiciones

- Artificial Intelligence

- The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. (Oxford)

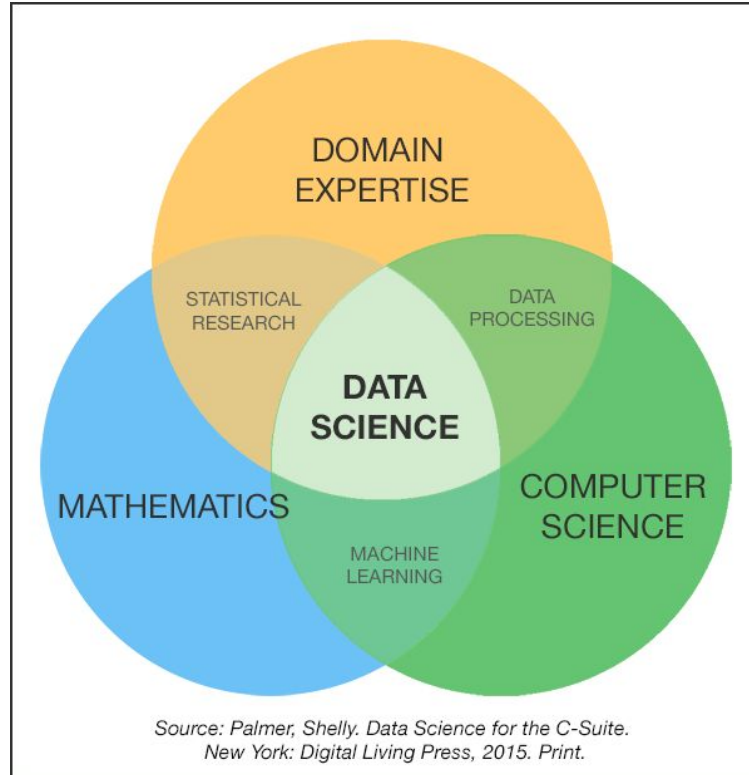
- Machine Learning

- “Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience [2].” An algorithm can be thought of as a set of rules/instructions that a computer programmer specifies, which a computer is able to process. (Mitchell)

- Data Science

- 

# Data Science



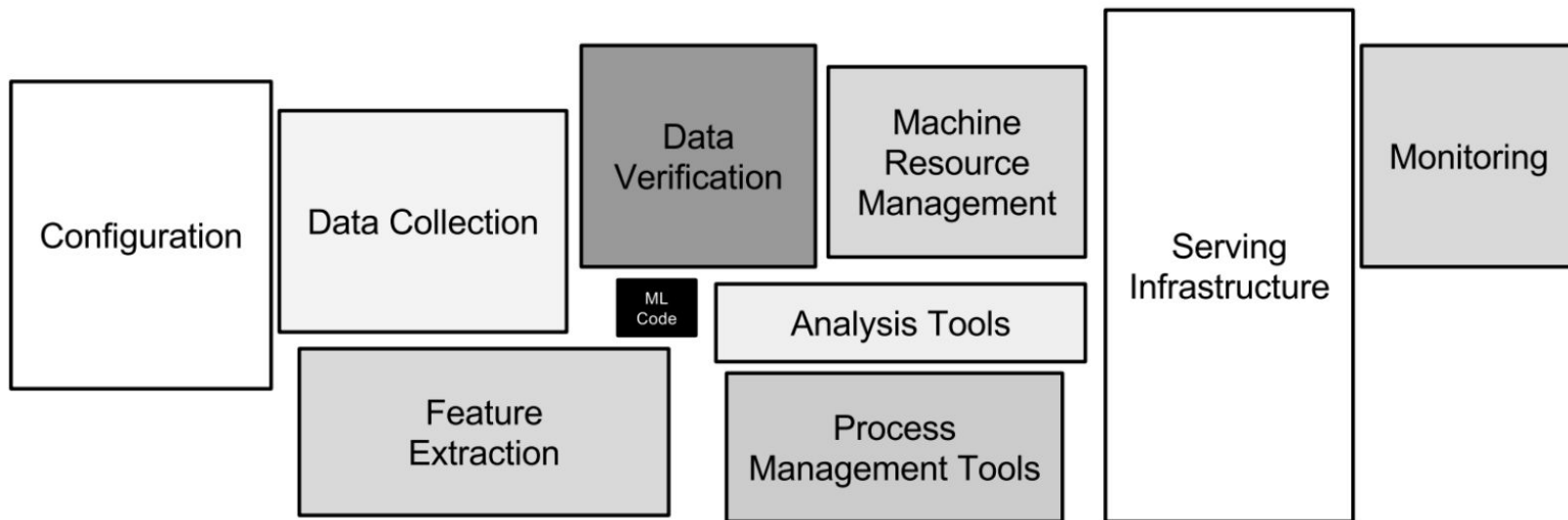
# Equipos que construyen data products

- Backend engineer
- Machine Learning engineer
- Data Engineer
- Data Scientist
- Frontend engineer
- UX
- Manager

Machine Learning en el eter



# Todas las complejidades con las que no vamos a lidiar



# Data products

- Productos. Value proposition
- Productos de datos
- Key Performance Indicators (KPIs)
- Metricas offline
- Machine Learning es una particita

# Business case

- Qué quiero lograr?
- Cómo me voy a dar cuenta que lo logré?
- Condiciones de frontera
  - Time to market
  - Real time vs batch
  - Datos disponibles
  - etc

Espacio de soluciones  
posibles

# Business case: Ejemplo de recomendaciones en e-commerce



Pasa el mouse encima de la imagen para aplicar zoom

## Google Pixel - 64GB - Desbloqueado (Renovado) Negro liso

[Visita la tienda de Amazon Renewed](#)

★★★★☆ 108 calificaciones | 24 preguntas respondidas

Precio nuevo: US\$ 699.00

Precio: **US\$ 400.00 & Envío GRATIS**

Ahorras: US\$ 299.00 (55%)

[Get \\$60 off instantly: Pay \\$340.00 upon approval of the Amazon.com Store Card.](#)

Disponible a un precio menor de [otros vendedores](#) que podrían no ofrecer envío Prime gratis.

Color: **Negro liso**



Tamaño: **64GB**



Estilo: **Pixel 4 XL**




El producto funciona y se ve como nuevo. Producto respaldado por la Garantía de Amazon Renewed de 90 días.

- Este producto seminuevo ha sido inspeccionado, probado y limpiado profesionalmente por proveedores calificados de Amazon.

- No habrá imperfecciones estéticas visibles cuando se sostenga a un brazo de distancia.

- Los productos con baterías superarán el 80% de capacidad con respecto a la nueva.

# Business case: Ejemplo de recomendaciones en e-commerce



Google Pixel - 64GB - Desbloqueado  
(Renovado) Negro liso  
[Visita la tienda de Amazon Renewed](#)  
★★★★☆ 108 calificaciones | 24 preguntas respondidas

Precio nuevo: US\$ 699.00  
Precio: **US\$ 400.00 & Envío GRATIS**  
Ahorrar: US\$ 299.00 (55%)

Get \$60 off instantly: Pay \$340.00 upon approval of the Amazon.com Store Card.

Disponible a un precio menor de otros vendedores que podrían no ofrecer envío Prime gratis.

Color: **Negro liso**

Tamaño: **64GB**

Estilo: **Pixel 4 XL**

El producto funciona y se ve como nuevo. Producto respaldado por la Garantía de Amazon Renewed de 90 días.

- Este producto seminuevo ha sido inspeccionado, probado y limpiado profesionalmente por proveedores calificados de Amazon.
- No habrá imperfecciones estéticas visibles cuando se sostenga a un brazo de distancia.
- Los productos con baterías superarán el 80% de capacidad con respecto a la nueva.

Pasa el mouse encima de la imagen para aplicar zoom

## Productos populares inspirados por este artículo

Página 1 de 7



Teléfono celular  
Samsung Galaxy A10s  
(32 GB, 2 GB de RAM),  
pantalla HD+ Infinity-V...  
★★★★☆ 3,412  
US\$128.00



Samsung Galaxy A10e  
32GB A102U GSM -  
Teléfono desbloqueado,  
color negro  
★★★★☆ 301  
US\$116.45



Samsung Galaxy A70  
A705M 128GB DUOS  
GSM Teléfono Android  
desbloqueado con...  
★★★★☆ 2,843  
5 ofertas desde  
US\$289.99



ZTE Maven 3 Z835 |  
(8GB, 1GB RAM) | 5.0"  
Full HD Display | 5MP  
Cámara Trasera |...  
★★★★☆ 608  
US\$59.99



Apple iPhone 7 negro  
mate 32 GB Verizon  
desbloqueado  
(certificado...  
★★★★☆ 706  
US\$178.95



Samsung Galaxy A20s  
(32GB 2GB RAM) 6.5"  
HD+ Triple Cámara SM-  
A207F/DS 4G LTE...  
★★★★☆ 236  
US\$163.98



LG Phoenix 2 AT&T  
Prepagado  
★★★★☆ 556  
US\$63.87



# Business case: Ejemplo de recomendaciones en e-commerce

- Qué quiero lograr?
  - Mejorar la experiencia del sitio a través de recomendaciones
    - Aumentando el engagement
    - Facilitando que el usuario pueda encontrar lo que busca

# Business case: Ejemplo de recomendaciones en e-commerce

- Qué quiero lograr?
  - Mejorar la experiencia del sitio a través de recomendaciones
    - Aumentando el engagement
    - Facilitando que el usuario pueda encontrar lo que busca
- Cómo me voy a dar cuenta que lo logré?
  - Ventas atribuidas, Click Through Rate

# Business case: Ejemplo de recomendaciones en e-commerce

- Qué quiero lograr?
  - Mejorar la experiencia del sitio a través de recomendaciones
    - Aumentando el engagement
    - Facilitando que el usuario pueda encontrar lo que busca
- Cómo me voy a dar cuenta que lo logré?
  - Ventas atribuidas, Click Through Rate
- Condiciones de frontera
  - Time to market: Primer solución en 3 meses
  - Real time vs batch: Se puede precalcular, simplifica la complejidad de infraestructura
  - Datos disponibles: Historia de navegación y compras de los usuarios



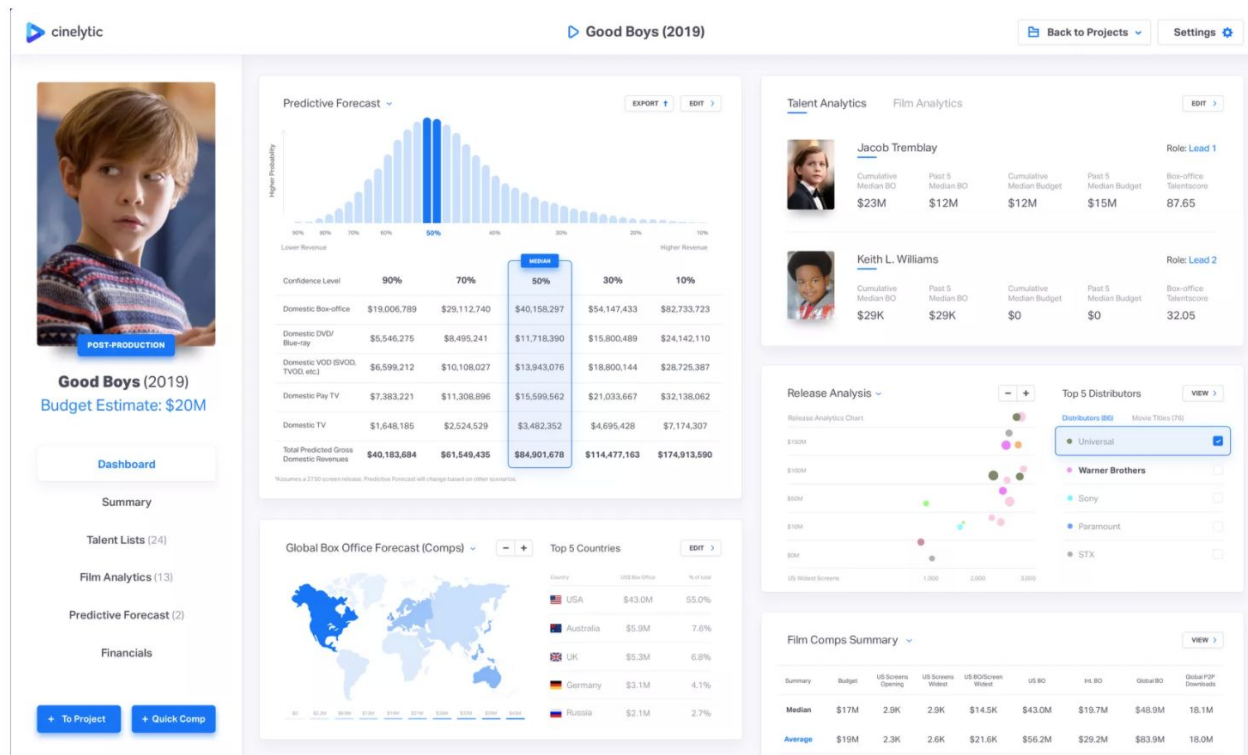
# Problema con el que vamos a trabajar

Una importante productora quiere asistir a sus guionistas proporcionándoles información que les permita anticiparse a características clave que determinen el éxito de una película

# Disclaimer

- Este business case es “inventado”
- No conozco sobre la industria del cine y que tipo de análisis se suelen hacer
  - Lecturas: [sofy.tv](https://sofy.tv), [theverge.com](https://theverge.com), [neilpatel](https://neilpatel.com)
- Los datos necesarios para hacer este ejercicio de forma realista “no están” disponibles de forma pública

“You can compare them separately, compare them in the package. Model out both scenarios with Emma Watson and Jennifer Lawrence, and see, for this particular film ... which has better implications for different territories,” Queisser tells *The Verge*.



An example of Cinelytic's software.

Before green-lighting House of Cards, Netflix knew:

- A lot of users watched the David Fincher directed movie The Social Network from beginning to end.
- The British version of “House of Cards” has been well watched.
- Those who watched the British version “House of Cards” also watched Kevin Spacey films and/or films directed by David Fincher.

Each of these 3 synergistic factors had to contain a certain volume of users.

Otherwise, House of Cards might belong to a different network right now. Netflix had a lot of users in all 3 factors.

Business case: con el que vamos a trabajar

# Business case: con el que vamos a trabajar

- Qué quiero lograr?
  - Aumentar la cantidad de pitches que reciben inversión
  - Aumentar el retorno de inversión de las películas que se filman
- Cómo me voy a dar cuenta que lo logré?
  - Hay un incremento de un 40% en el porcentaje de pitches que reciben inversión
  - Un 50% de las películas producidas tienen que generar 5 veces su inversión
- Condiciones de frontera
  - Time to market: 3 meses
  - Real time vs batch: batch
  - **Datos disponibles**
    - **Datos de IMDB**

# Aprendizaje supervisado

# Idea

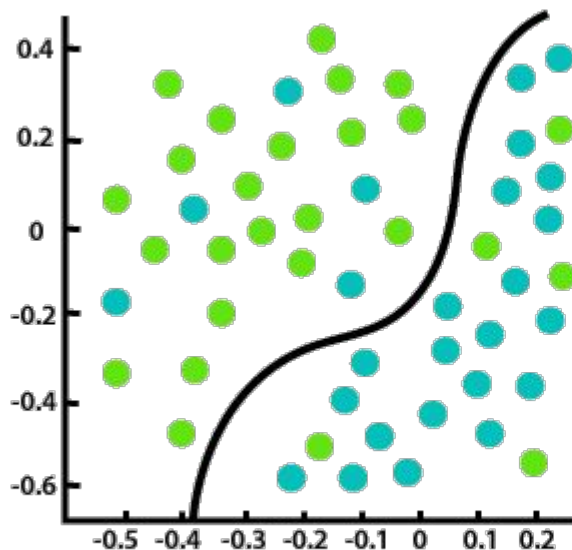
- Tenemos  $N$  observaciones sobre una **entidad**
- La entidad la podemos describir por **características** (features en ingles)
- Queremos predecir algo de la entidad (target variable en ingles)
- Matematicamente:
  - $x_i$  = vector de características de la  $i$ -ésima entidad
  - $y_i$  = target variable asociada a la  $i$ -ésima entidad



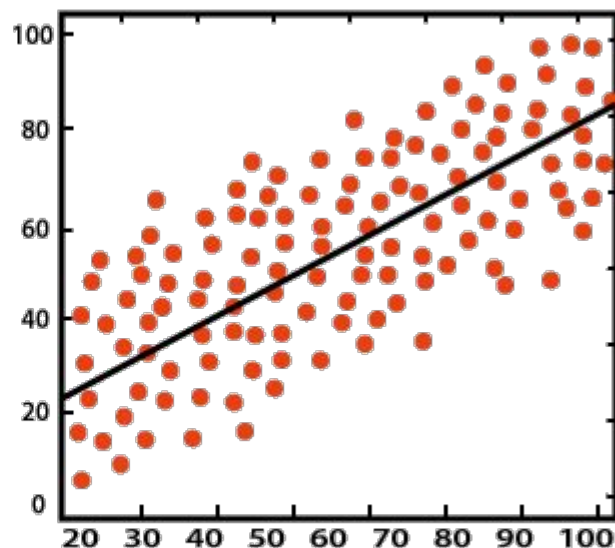
# Ejemplos de aprendizaje supervisado

- Predecir si una imagen contiene a un gato
  - $x_i$  = La imagen
  - $y_i$  = 1 si hay un gato, 0 si no
- Detectar spam
  - $x_i$  = El mail
  - $y_i$  = 1 si el mail es spam, 0 si no
- Predecir el precio de las casas
  - $x_i$  = <barrio, antigüedad, ambientes, superficie, ABL, etc>
  - $y_i$  = precio de venta de la propiedad

## Dos tipos de aprendizaje supervisado: Clasificación y Regresión



Classification



Regression

# Abriendo la caja: Regresión lineal

Dependent Variable  $\rightarrow Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Population Y intercept  $\rightarrow \beta_0$

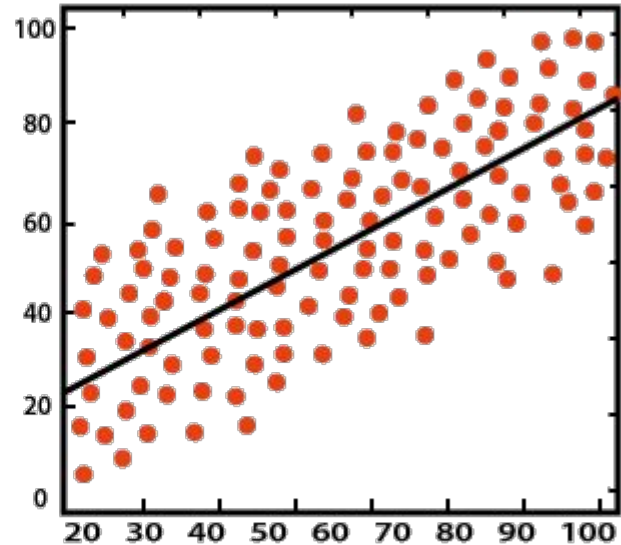
Population Slope Coefficient  $\rightarrow \beta_1$

Independent Variable  $\rightarrow X_i$

Random Error term  $\rightarrow \varepsilon_i$

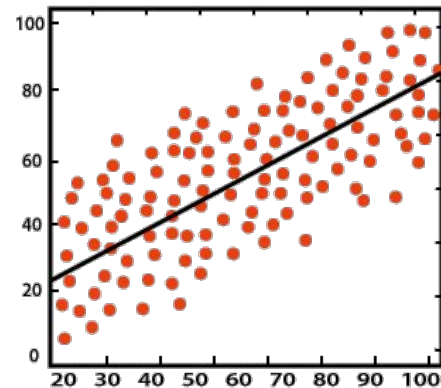
Linear component  $\underbrace{\beta_0 + \beta_1 X_i}$

Random Error component  $\underbrace{\varepsilon_i}$



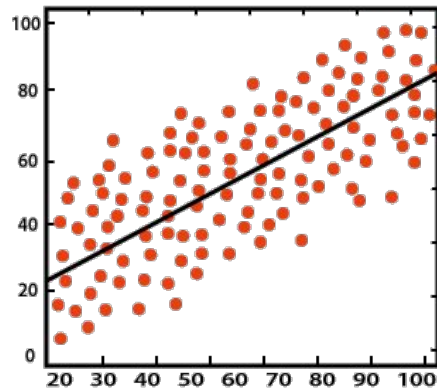
# Y cómo se qué valores usar para el vector $\beta$ ?

- Para eso tenemos los datos
- La idea es que **ajuste a los datos**



# Y cómo se qué valores usar para el vector $\beta$ ?

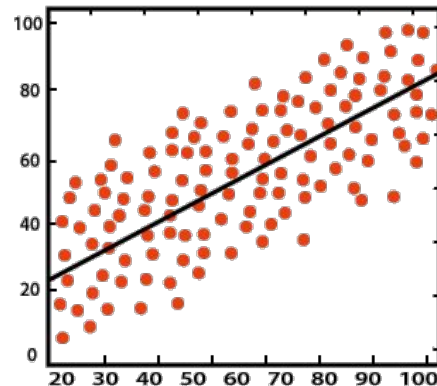
- Para eso tenemos los datos
- La idea es que **ajuste a los datos**
- Para eso tenemos que poder **cuantificar cuán buenos** son valores para  $\beta$ .
  - La función que cuantifica **cuán mala** es una solución se llama *loss function*
- Hay muchas formas de cuantificar



# Y cómo se qué valores usar para el vector $\beta$ ?

- La forma más común es **minimizando el error cuadrático**

$$L(\beta) = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$



Y como se minimiza? Descenso por el gradiente

$$L(\beta) = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\beta_0 \leftarrow \beta_0 - \alpha \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \alpha \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1}$$

# Y como se minimiza? Descenso por el gradiente

10/06/2019 14:00:00

$$L(\beta) = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\beta_0 \leftarrow \beta_0 - \alpha \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \alpha \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1}$$



# Abstrayendo

- Estamos **ajustando una función paramétrica**  $f_{\beta}(x_i)$ 
  - A esta función se le llama modelo
  - En el caso de regresión lineal en una sola dimensión  $f_{\beta}(x_i) = \beta_0 + \beta_1 x_i$
- Ajustar el modelo corresponde a encontrar los **parámetros  $\beta$  que minimizan la loss function**
  - En todos los modelos hay mínimo global? No!

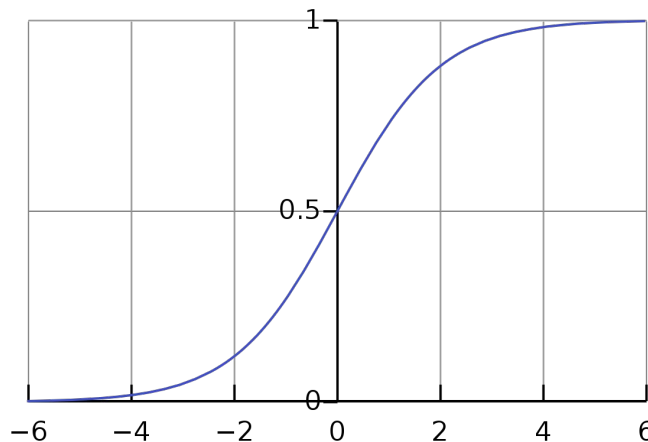
# Y qué pasa cuando quiero hacer clasificación?

- Basta con cambiar
  - El **modelo** para que el conjunto imagen sean los valores 0 y 1 recordar que en clasificación binaria,  $y_i = 0$ , o  $y_i = 1$ 
    - Se suele lograr teniendo una función que mapee al intervalo (0, 1) y determinando un umbral de corte
  - La **loss function** por una que represente mejor el problema

# Y qué pasa cuando quiero hacer clasificación?

- Basta con cambiar
  - El **modelo** para que el conjunto imagen sea el intervalo  $(0, 1)$  recordar que en clasificación binaria,  $y_i = 0$ , o  $y_i = 1$
  - La **loss function** por una que represente mejor el problema

- Modelo  $f_{\beta}(x_i) = \text{sigmoid}(\beta_0 + \beta_1 x_i)$   
 $\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$



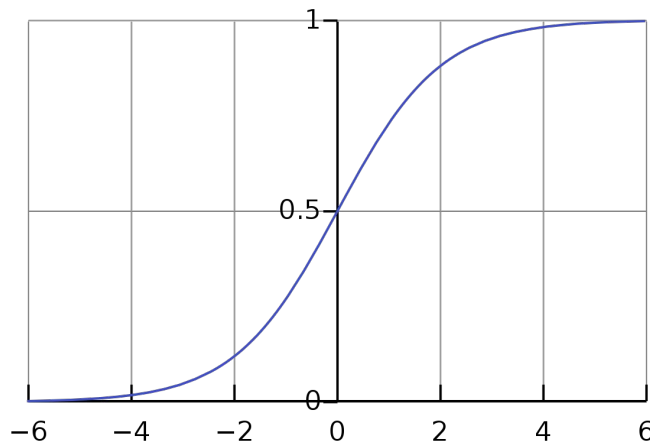
# Y qué pasa cuando quiero hacer clasificación?

- Basta con cambiar
  - El **modelo** para que el conjunto imagen sea el intervalo  $(0, 1)$  recordar que en clasificación binaria,  $y_i = 0$ , o  $y_i = 1$
  - La **loss function** por una que represente mejor el problema

- Modelo
$$f_{\beta}(x_i) = \text{sigmoid}(\beta_0 + \beta_1 x_i)$$
$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

- Loss function

$$\sum_i y_i \log(f_{\text{beta}}(x_i)) + (1 - y_i) \log(1 - f_{\text{beta}}(x_i))$$



# Qué otro tipo de modelos hay?

- Árboles
- Support vector machines
- Redes neuronales
- Ensembles de árboles
- Ensembles de cualquiera de los de arriba
- Hay más!!

# Árboles de decisión

- Ejemplo: queremos decidir si vamos a salir a andar en bici
- Tenemos el pronóstico y la humedad de muchos días ( $x_i$ )
- Tenemos anotado si la pasamos bien o no ( $y_i$ )

# Árboles de decisión

Decision Tree Diagram



$$InformationGain(Y, a) = H(Y) - H(Y|a)$$

# Árboles de decisión

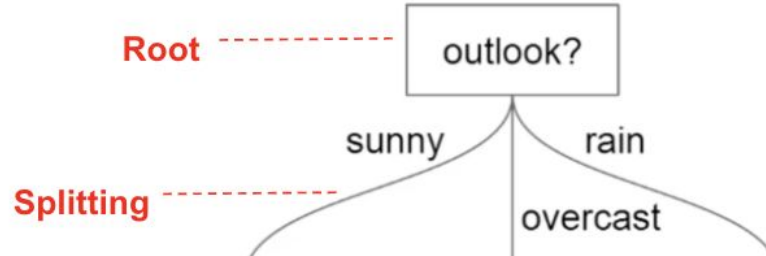
## Decision Tree Diagram





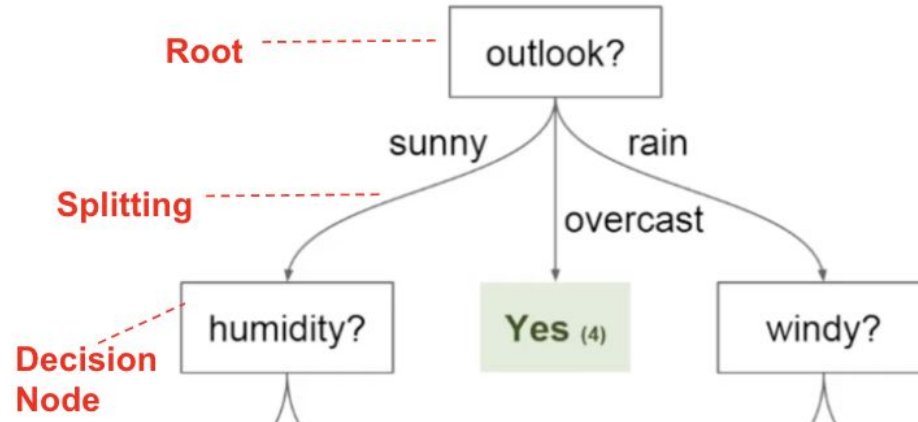
# Árboles de decisión

Decision Tree Diagram



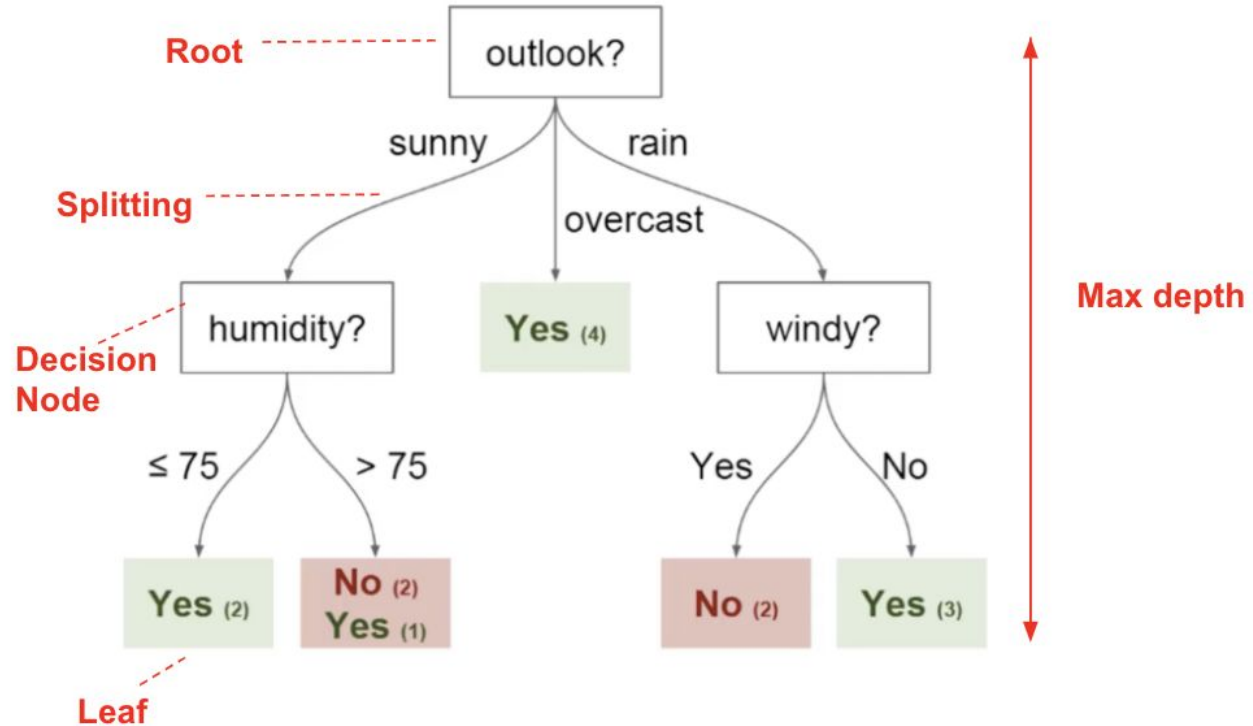
# Árboles de decisión

Decision Tree Diagram



# Árboles de decisión

Decision Tree Diagram



Métricas para evaluar a los modelos

# Lema

- No siempre la métrica con la que evaluamos a nuestros modelos es la misma que optimizamos
- Principalmente porque la métrica puede ser difícil de optimizar
  - Hay métricas no diferenciables
  - Métricas costosas computacionalmente
  - Optimizarlas directamente nos limita las familias de modelos que podemos utilizar
  - Queremos evaluar en 3 métricas a nuestro modelo y solo podemos optimizar una

# Métricas para evaluar modelos de regresión

$$\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Mean squared error

$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$R^2$

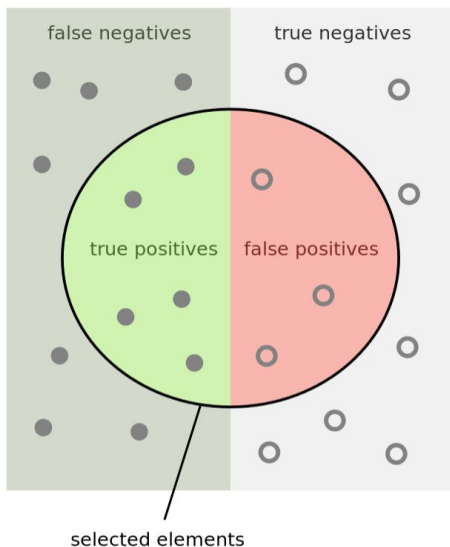
$$\frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Mean absolute percentage error

$$\frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Mean absolute error

# Métricas para evaluar modelos de clasificación



$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

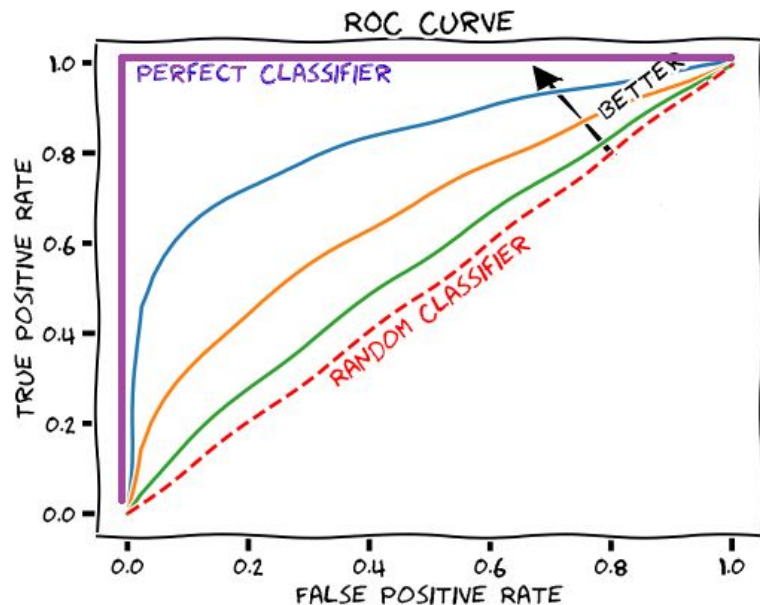
How many selected items are relevant?

Precision =  $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

How many relevant items are selected?

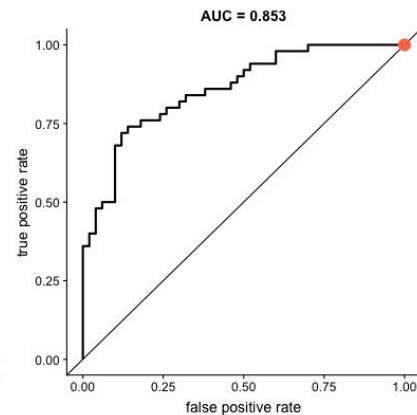
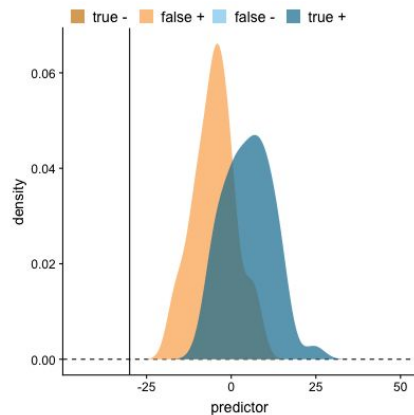
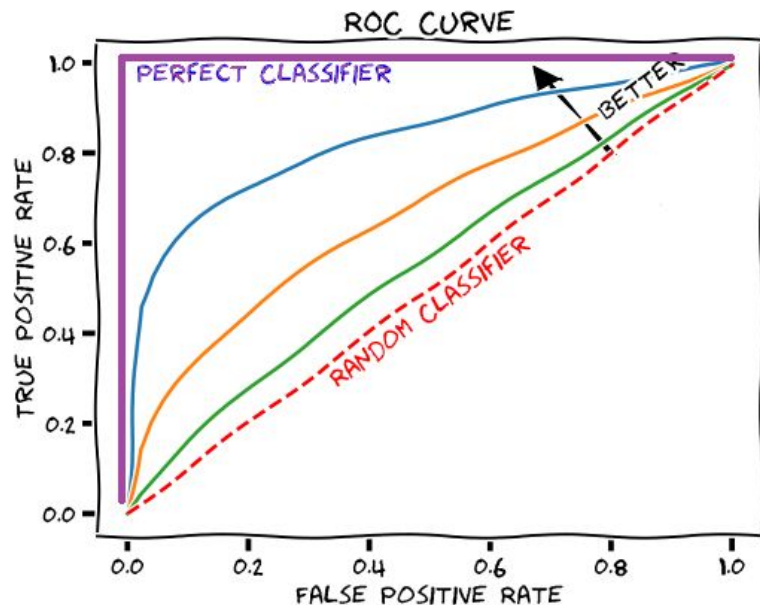
Recall =  $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

# Métricas para evaluar modelos de clasificación





# Métricas para evaluar modelos de clasificación



Fuente de la [animación](#): [este](#) post

# Metodología

trying not to drift and find value

# Market pull vs technology push

Qué necesito para resolver este **problema**

vs

Quiero usar esta **técnica** re linda

# Be Lean!



Are the people who solve them superhuman data scientists who can come up with better ideas in five minutes than most people can in a lifetime? Are they magicians of applied math who can cobble together millions of lines of code for high-performance machine learning in a few hours? **No**. Many of them are incredibly smart, but meeting big problems head-on usually isn't the winning approach. There's a method to solving data problems that avoids the big, heavyweight solution, and instead, concentrates building something quickly and iterating. Smart data scientists don't just solve big, hard problems; they also have an instinct for making big problems small.

Lectura recomendada: [data.juuitsu](https://data.juuitsu)

DJ Patil

# Criterio pareto: lo excelente es enemigo de lo bueno

Muchas veces con cubrir el 80% del problema alcanza

# Machine Learning yearning

- El libro es oro puro
  - Este índice es una parte →
- Cosas que quiero rescatar
  - Crear un end2end es tu primera misión
  - Error analysis is your friend
  - Training set distribution
  - Cómo debería elegir mi métrica offline?
- Hay mucho mas!!!  
pero ahora no lo voy a mencionar :P

[1 Why Machine Learning Strategy](#)

[2 How to use this book to help your team](#)

[3 Prerequisites and Notation](#)

[4 Scale drives machine learning progress](#)

[5 Your development and test sets](#)

[6 Your dev and test sets should come from the same distribution](#)

[7 How large do the dev/test sets need to be?](#)

[8 Establish a single-number evaluation metric for your team to optimize](#)

[9 Optimizing and satisficing metrics](#)

[10 Having a dev set and metric speeds up iterations](#)

[11 When to change dev/test sets and metrics](#)

[12 Takeaways: Setting up development and test sets](#)

[13 Build your first system quickly, then iterate](#)

[14 Error analysis: Look at dev set examples to evaluate ideas](#)

[15 Evaluating multiple ideas in parallel during error analysis](#)

[16 Cleaning up mislabeled dev and test set examples](#)

[17 If you have a large dev set, split it into two subsets, only one of which you look at](#)

[18 How big should the Eyeball and Blackbox dev sets be?](#)

[19 Takeaways: Basic error analysis](#)

[20 Bias and Variance: The two big sources of error](#)

[21 Examples of Bias and Variance](#)

[22 Comparing to the optimal error rate](#)

[23 Addressing Bias and Variance](#)

[24 Bias vs. Variance tradeoff](#)

[25 Techniques for reducing avoidable bias](#)

# Crear un end2end (solución completa) es tu primera misión

- Comenzar sencillo, luego complejizar

# Crear un end2end (solución completa) es tu primera misión

- Comenzar sencillo, luego complejizar
- No perderse en un detalle que podría ser irrelevante



# Crear un end2end (solución completa) es tu primera misión

- Comenzar sencillo, luego complejizar
- No perderse en un detalle que podría ser irrelevante
- Asegurarse que todo tiene sentido
  - ETL (Extract Transform and Load, a.k.a. data extraction and processing)
  - training
  - evaluation
  - serving

# Crear un end2end (solución completa) es tu primera misión

- Comenzar sencillo, luego complejizar
- No perderse en un detalle que podría ser irrelevante
- Asegurarse que todo tiene sentido
  - ETL (Extract Transform and Load, a.k.a. data extraction and processing)
  - training
  - evaluation
  - serving
- Nos permite invertir nuestro tiempo en donde más duele en futuras iteraciones

# Crear un end2end (solución completa) es tu primera misión

- Comenzar sencillo, luego complejizar
- No perderse en un detalle que podría ser irrelevante
- Asegurarse que todo tiene sentido
  - ETL (Extract Transform and Load, a.k.a. data extraction and processing)
  - training
  - evaluation
  - serving
- Nos permite invertir nuestro tiempo en donde más duele en futuras iteraciones
- Asegurar cierta reproducibilidad

# Métrica offline vs KPI

- KPI
  - Mide el efecto de nuestro modelo en la realidad
  - Generalmente no se puede calcular en el laboratorio
- Métrica offline
  - Es una métrica standard (e.g. accuracy, f1, roc auc)
  - Nos permite benchmark contra papers, blogs, etc
  - Debería ser un proxy de la calidad del modelo

# Train and validation set distribution

Queremos que represente lo más fielmente posible el escenario donde será utilizado el modelo

## Alternativas y trade offs

- Random splitting
  - e.g. cats and dogs prediction
- Stratified
  - e.g. imbalanced classification
- By timestamp
  - e.g. spam filtering, time series

# Error analysis is your friend

- Sirve para iterar el modelo
- Tomar N ejemplos donde el modelo falla en la predicción
- Revisarlos a mano: qué puede estar pasando?
- Elaborar posibles problemas y contar porotos
- Abordar el problema que más aparece
  - es el que más impacto va a tener en la metrica

# Error analysis: ejemplo con predictor de gatos

People with no idea  
about AI, telling me my  
AI will destroy the world

Me wondering why my  
neural network is  
classifying a cat as a dog..



# Error analysis: ejemplo con predictor de gatos

Image from Dev Set	Dogs	Great Cats (Tiger, Panthers)	Improve on blurry images	Insta filter error
1	✓			✓
2		✓		
3			✓	
4	✓			✓
5			✓	
.				
..		✓		
...			✓	
100				✓
% incorrect classification	8%	40%	60%	20%



# Marco para tomar decisiones

- Hipótesis
- Experimentos
- Decisiones
- Costos del experimento
- Riesgo de no hacer nada (default action, hipótesis nula)

# Marco para tomar decisiones

- Contexto: La métrica (accuracy, f1, roc auc) me da muy baja en test

# Marco para tomar decisiones

- Contexto: La métrica (accuracy, f1, roc auc) me da muy baja en test
- Hipotesis: La distribución del train y test set difieren

# Marco para tomar decisiones

- **Contexto:** La métrica (accuracy, f1, roc auc) me da muy baja en test
- **Hipotesis:** La distribución del train y test set difieren
- **Experimento:** Entreno un modelo para distinguir entre ejemplos de training y test set

# Marco para tomar decisiones

- **Contexto:** La métrica (accuracy, f1, roc auc) me da muy baja en test
- **Hipótesis:** La distribución del train y test set difieren
- **Experimento:** Entreno un modelo para distinguir entre ejemplos de training y test set
- **Decisión:** Hacer el train / test splitting diferente
- **Costo:** Bajo
- **Riesgo:** Estamos a ciegas respecto a la performance del modelo en producción

# Marco para tomar decisiones

- **Contexto:** Nos piden hacer un forecasting para optimizar la compra de insumos en un negocio

# Marco para tomar decisiones

- **Contexto:** Nos piden hacer un forecasting para optimizar la compra de insumos en un negocio
- **Hipótesis:** Se puede hacer una reducción de al menos 15% en las pérdidas con un forecasting
- **Experimento:** Usamos los datos de oráculo como forecasting perfecto para simular cuanto hubiéramos comprado
- **Decisión:** Hacer o no forecasting
- **Costo:** Bajo
- **Riesgo:** Podríamos terminar haciendo un sistema que no tenga el efecto esperado

# Técnicas para evaluar una hipótesis

## Foco: Rigurosidad vs Costo vs Riesgo

- Test de hipótesis estadístico

- Hay que asegurarse de usarlo bien. No hacer p-hacking! Significancia estadística vs Diferencia relevante

- Visualización

- Usar tu GPU nativa (👁👁) para ver cambios relevantes. Boxplots, error bars and bootstrapped confidence intervals allowed ;)

- A/B test

- The cheapest control double blinded experiment. Possibly expensive

- Human in the middle

- Human performance on a task as a proxy



# Statistical inference in one sentence



Cassie Kozyrkov

[Follow](#)

Jan 11, 2019 · 9 min read



“Does the evidence that we collected make our null hypothesis look ridiculous?”

[full post here](#)

# Statistical inference in one sentence



Cassie Kozyrkov

[Follow](#)

Jan 11, 2019 · 9 min read



- Step 1: What's the default action?
- Step 2: What's the alternative action?
- Step 3: What's the null hypothesis?
- Step 4: What's the alternative hypothesis?
- Collect data
  - How you frame your decision-making is important. Not all decisions lend themselves to the approach taught in STAT101.
  - You should get into the habit of **learning nothing** more often, because if you **insist on learning something** beyond the data every time you test hypotheses, you will learn something stupid. p-hacking

# Ejemplo: personalización en un motor de búsqueda

- **Default action:** implementar personalización con preferencias del usuario en los resultado
- **Alternative action:** no implementar personalización  
y trabajar en otra iniciativa en ese tiempo. Costo de oportunidad
- **Null hypothesis:** Los usuarios exhiben patrones de búsqueda distintivos que pueden ser explotados para mostrar resultados más relevantes **medible en un 10% menor de bounce rate**
- **Alternative hypothesis:**

# ¿Contra qué compite?

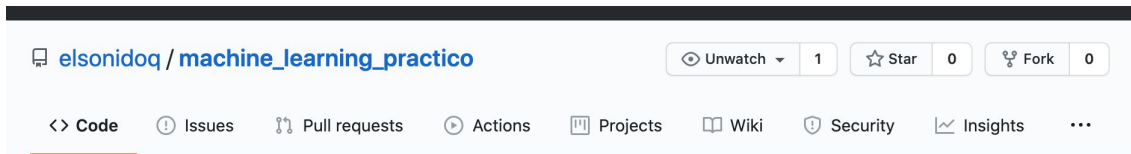
- Costo de oportunidad
- Cuando hacemos X no hacemos **otra cosa**
- Se deriva de estar con mindset **market pull** (y no technology push)

**Hands on!**

# Forkeá el repo

[https://github.com/elsonidoq/machine\\_learning\\_practico](https://github.com/elsonidoq/machine_learning_practico)

Fork me on GitHub



# A word on languages

- Todo el material de notebooks esta en Python porque no se R
- Las guías de ejercicios son agnósticas al lenguaje
- **Es posible correr código R en colab**

[https://github.com/elsonidoq/machine\\_learning\\_practico](https://github.com/elsonidoq/machine_learning_practico)



Repository:

[elsonidoq/machine\\_learning\\_practico](#) ▼

Branch:

[master](#) ▼

Path



[notebooks/clase-1/01\\_get\\_the\\_data.ipynb](#)



[notebooks/clase-1/02\\_running\\_R\\_on\\_colab.ipynb](#)



# A word on languages

- Todo el material de notebooks esta en Python porque no sé R
- Las guías de ejercicios son agnósticas al lenguaje
- **Es posible correr código R en colab**



02\_running\_R\_on\_colab.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text  Copy to Drive



```
# Load in the r magic
%reload_ext rpy2.ipython
%config IPCompleter.greedy=True
%config InlineBackend.figure_format = 'retina'
```

```
[ ] %%R
```

```
install.packages("mlbench")
library(mlbench)
```



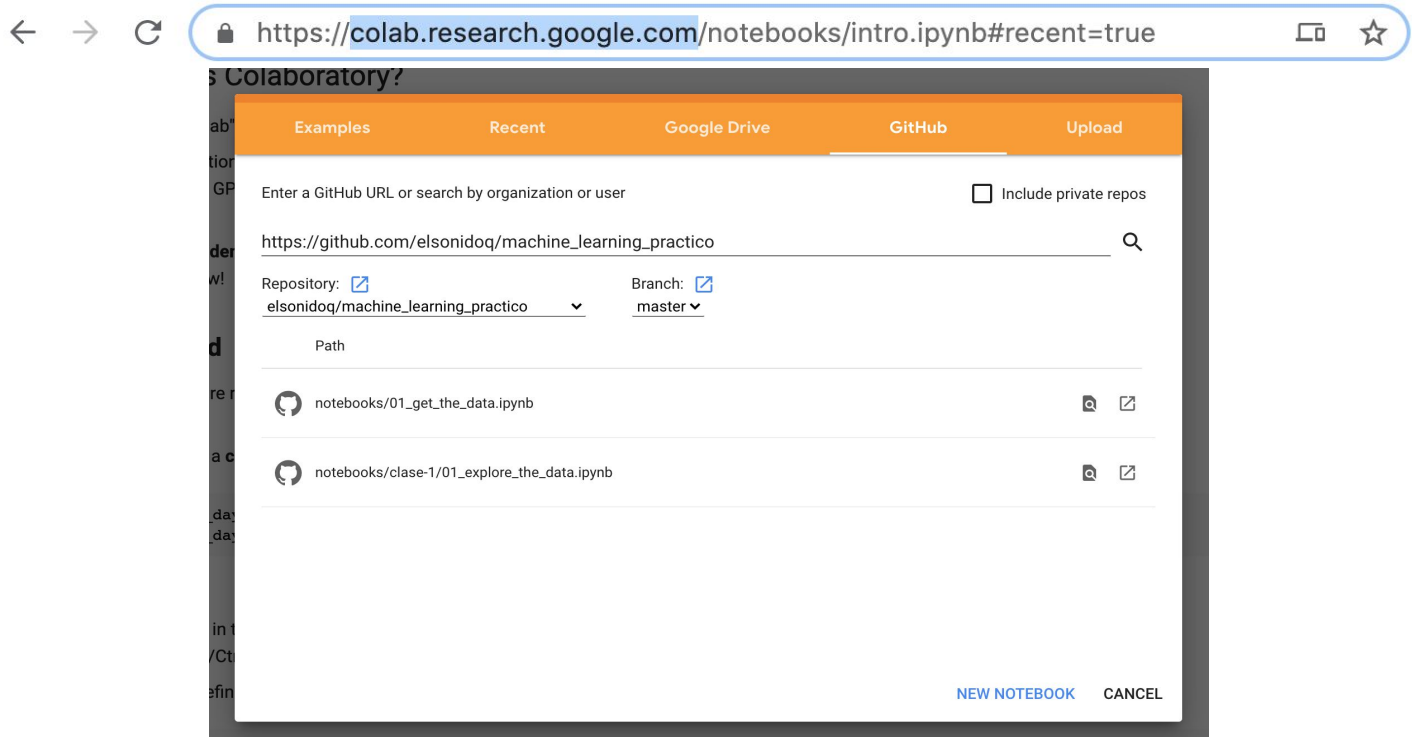
# Open Source

Si hiciste algo que te parece bueno,  
mandame un un pull request

# Si lo vas a usar local: Git ABC

- git clone [https://github.com/elsonidoq/machine\\_learning\\_practico.git](https://github.com/elsonidoq/machine_learning_practico.git)
  - O con tu repo forkeado!
- git add <archivo que quiero subir a github>
- git commit -m “avance en XYZ”
- git push origin HEAD

# Getting the data into your google drive



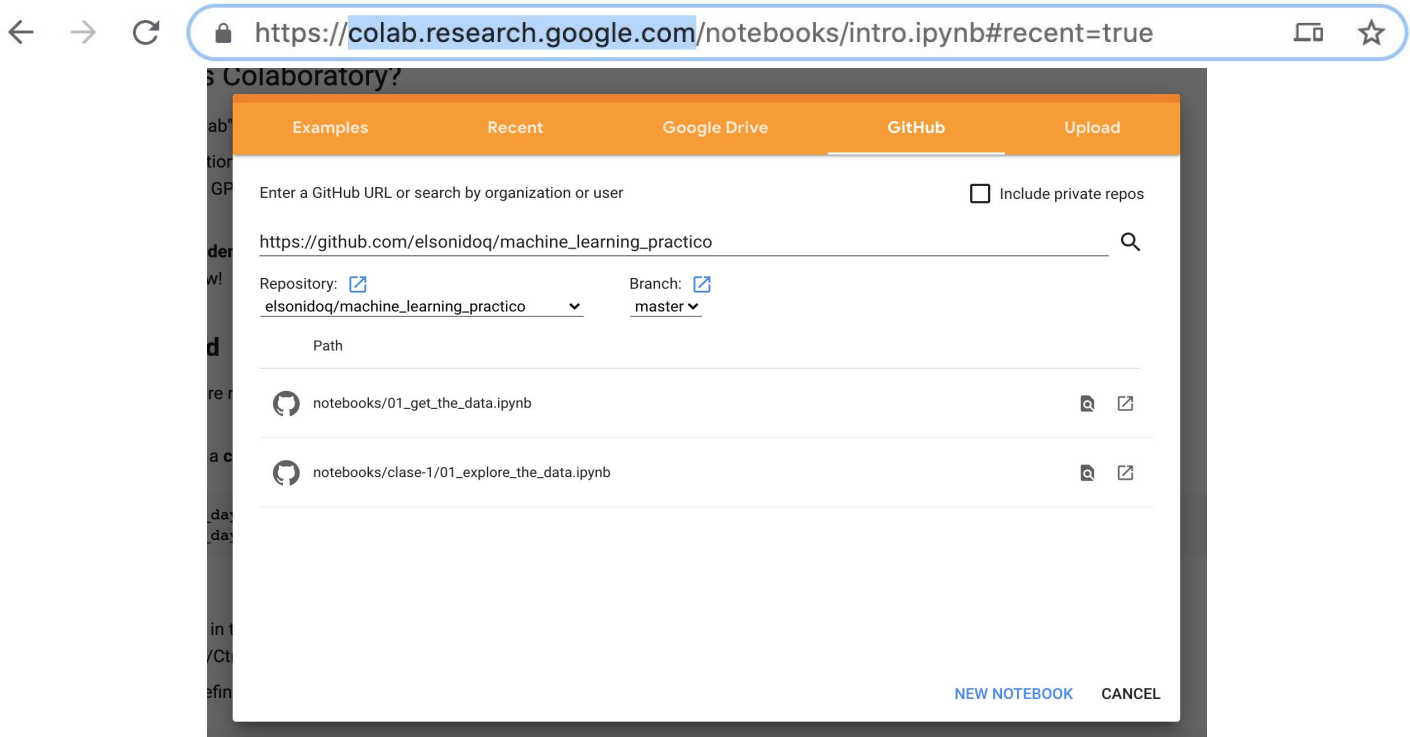
# Ahora es tu turno

Lleva los datos a tu google drive (15 minutos)

# Guia de ejercicios



# Exploring the data



# Ahora es tu turno

Se te ocurre algun otro check para hacer??

- Qué pasa con esas peliculas que duran menos de ~45 minutos?
- Qué pasa con esas que estan entre 45 y 60 minutos? Debemos descartarlas?
- Qué pasa con los años de las peliculas? Queremos quedarnos con todos? Algunos? Por qué?
- Similar con los ratings. Queremos quedarnos con todas? Solo las que tuvieron al menos X puntaje? Al menos X votos?
- Nos pueden llegar a servir aquellas peliculas que no tenemos informacion de gross?
- Aparece algo más que llame la atencion si miramos a la base de datos de actores?

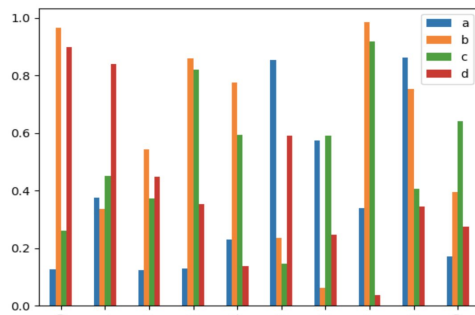


# Como seguir...

← → ↺ [pandas.pydata.org/docs/user\\_guide/visualization.html](https://pandas.pydata.org/docs/user_guide/visualization.html)

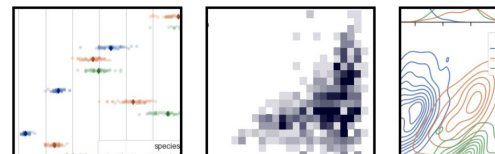
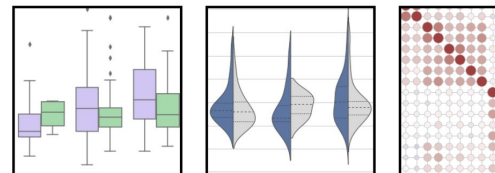
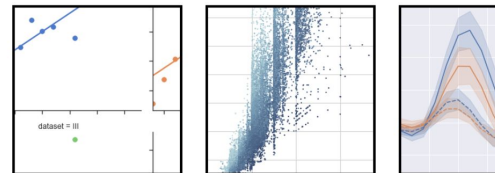
```
In [20]: df2 = pd.DataFrame(np.random.rand(10, 4), columns=['a', 'b', 'c', 'd'])
```

```
In [21]: df2.plot.bar();
```



🔒 [seaborn.pydata.org/examples/index.html](https://seaborn.pydata.org/examples/index.html)

## Example gallery



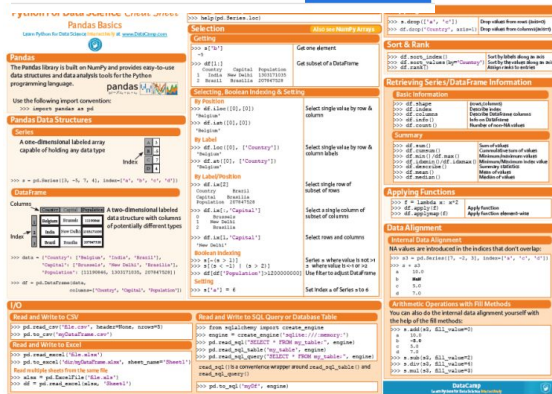
← → ↺ 🔒 matplotlib.org/gallery.html



## pandas cheat sheet

Q Todos  Imágenes

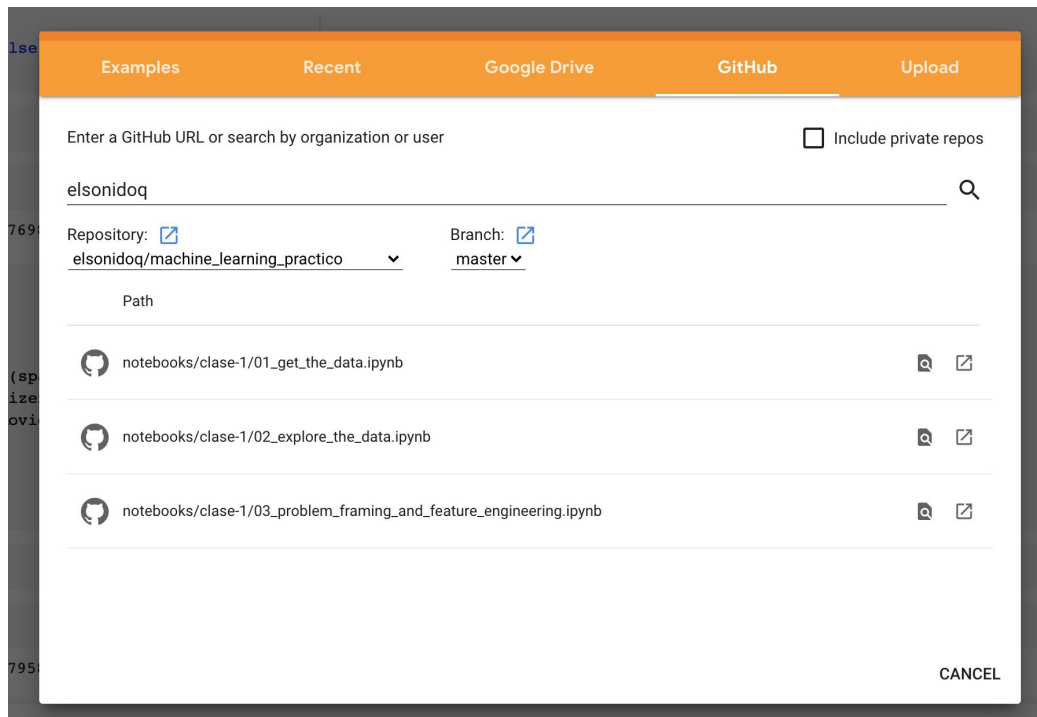
 Imágenes



# Problem framing

- Regresión? Predecir la cantidad dólares de de gross?
- Clasificación? de las películas que generaron al menos X dólares?
- Otro?
- Cómo se va a usar el modelo?
- Los ratings son un buen proxy?
- Qué features tiene sentido incluir?
- Qué métrica de evaluación vamos a usar?
- Cómo vamos a partir el training set del test set?

# Problem framing



# The rules of machine learning

[Terminology](#)

[Overview](#)

[Before Machine Learning](#)

[Rule #1: Don't be afraid to launch a product without machine learning.](#)

[Rule #2: Make metrics design and implementation a priority.](#)

[Rule #3: Choose machine learning over a complex heuristic.](#)

[ML Phase I: Your First Pipeline](#)

[Rule #4: Keep the first model simple and get the infrastructure right.](#)

[Rule #5: Test the infrastructure independently from the machine learning.](#)

[Rule #6: Be careful about dropped data when copying pipelines.](#)

[Rule #7: Turn heuristics into features, or handle them externally.](#)

[Monitoring](#)

[Rule #8: Know the freshness requirements of your system.](#)

[Rule #9: Detect problems before exporting models.](#)

[Rule #10: Watch for silent failures.](#)

[Rule #11: Give feature sets owners and documentation.](#)

[Your First Objective](#)

[Rule #12: Don't overthink which objective you choose to directly optimize.](#)

[Rule #13: Choose a simple, observable and attributable metric for your first objective.](#)

[Rule #14: Starting with an interpretable model makes debugging easier.](#)

[Rule #15: Separate Spam Filtering and Quality Ranking in a Policy Layer.](#)

[ML Phase II: Feature Engineering](#)

[Rule #16: Plan to launch and iterate.](#)

[Rule #17: Start with directly observed and reported features as opposed to learned features.](#)

Para más detalles: [PDF](#), [Pagina](#)

mostrar esto?

<http://karpathy.github.io/2019/04/25/recipe/>