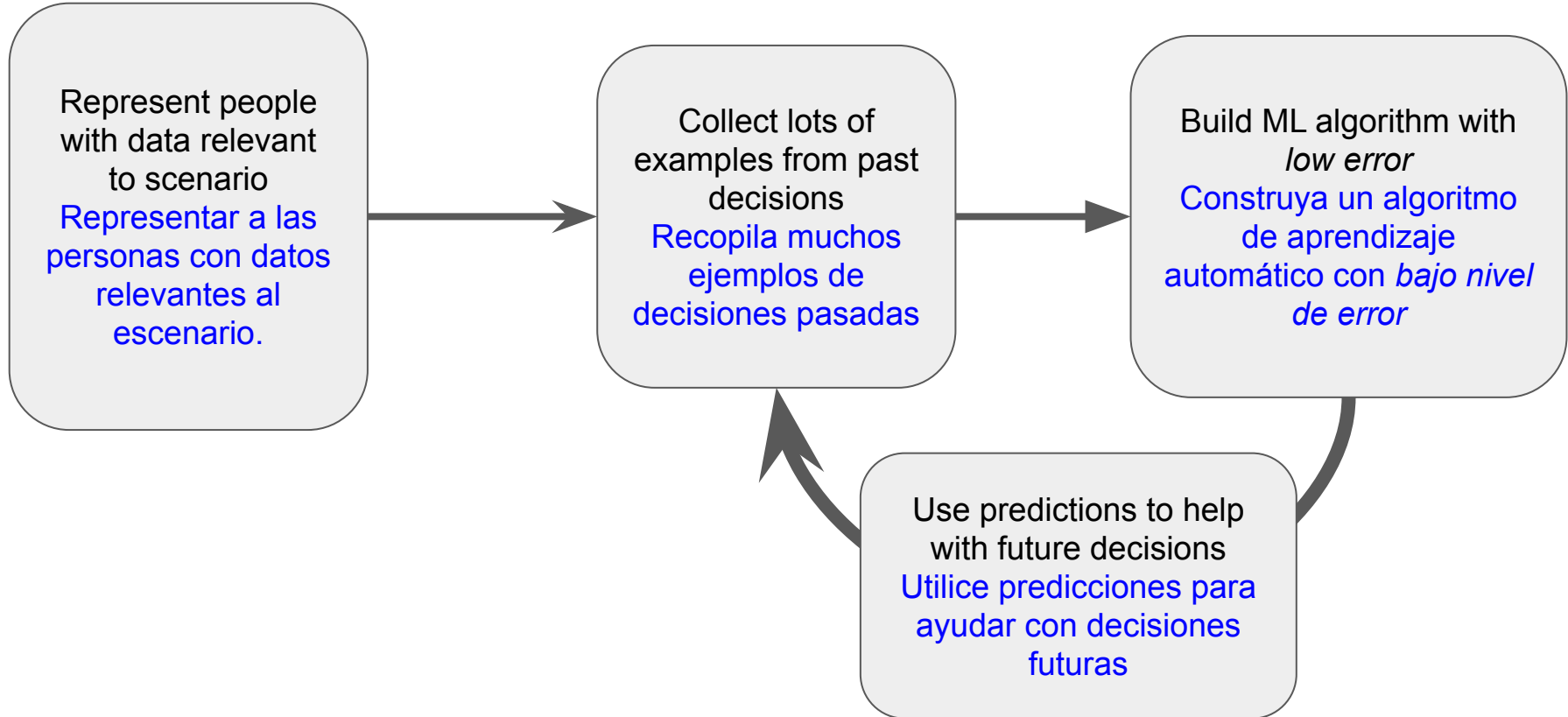


Evaluating the success of AI  
algorithms

*Evaluación del éxito de los  
algoritmos de IA*

Prof. Jessie Finocchiaro, Boston College

# AI pipeline *Canalización de IA*



# Representing people with data Representando personas con datos

Maria is applying for a loan to buy a new car

- Works in customer service in San Luis Potosi
- A mother of 3 young children
- Recently left a partner who incurred a lot of debt

María está solicitando un préstamo para comprar un auto nuevo.

- Trabaja en atención al cliente en San Luis Potosí.
- Es madre de tres niños pequeños.
- Recientemente dejó a su pareja, quien acumuló muchas deudas.



# Representing people with data Representando personas con datos

In a simple scenario, use her **income** and **credit score** to decide whether or not Maria will repay her loan.

En un escenario simple, use sus **ingresos** y su **puntaje crediticio** para decidir si María pagará o no su préstamo.

Income: 70000, Credit score: 675

Ingresos: 70000, Puntaje crediticio: 675



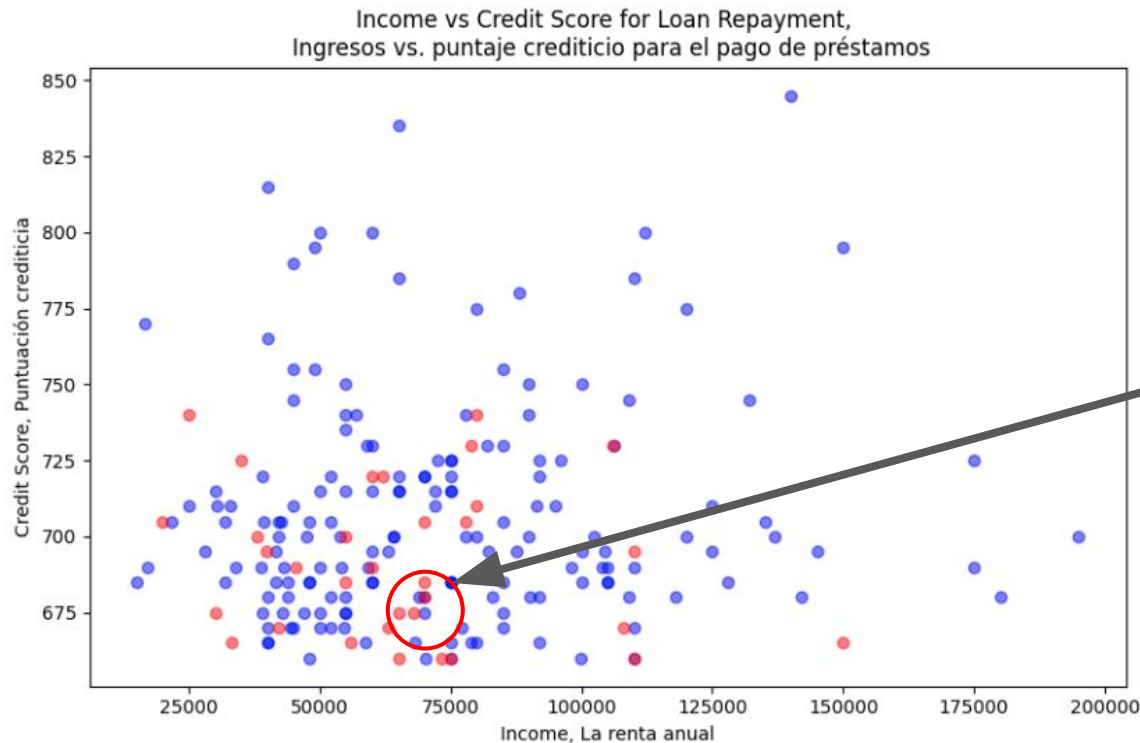
# Using past examples to predict for Maria

## Usando ejemplos pasados para predecir el futuro de María



Data/datos:  
[LendingClub](#)

Based on historical data, what do we expect from Maria?  
Basándonos en datos históricos, ¿qué esperamos de María?



# Comparing a few different predictions for Maria

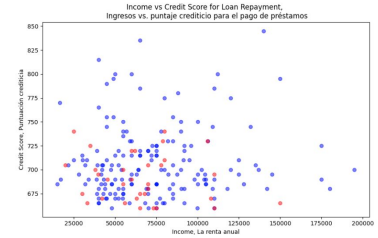
Choose



with **small error** comparing



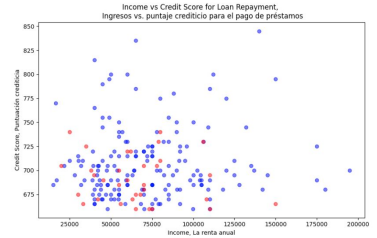
vs



Elija con un **pequeño error** comparando



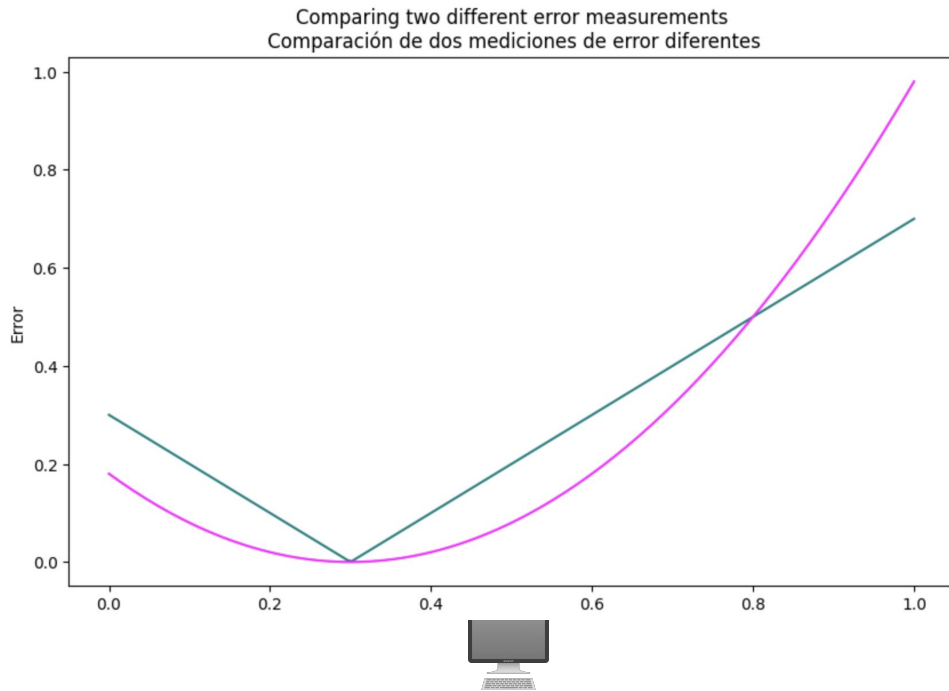
con



# “Small error” “Pequeño error”

$$\text{error}(\text{monitor}) = \begin{cases} \text{monitor} & \text{if default} \\ (1 - \text{monitor}) & \text{if repay loan} \end{cases}$$

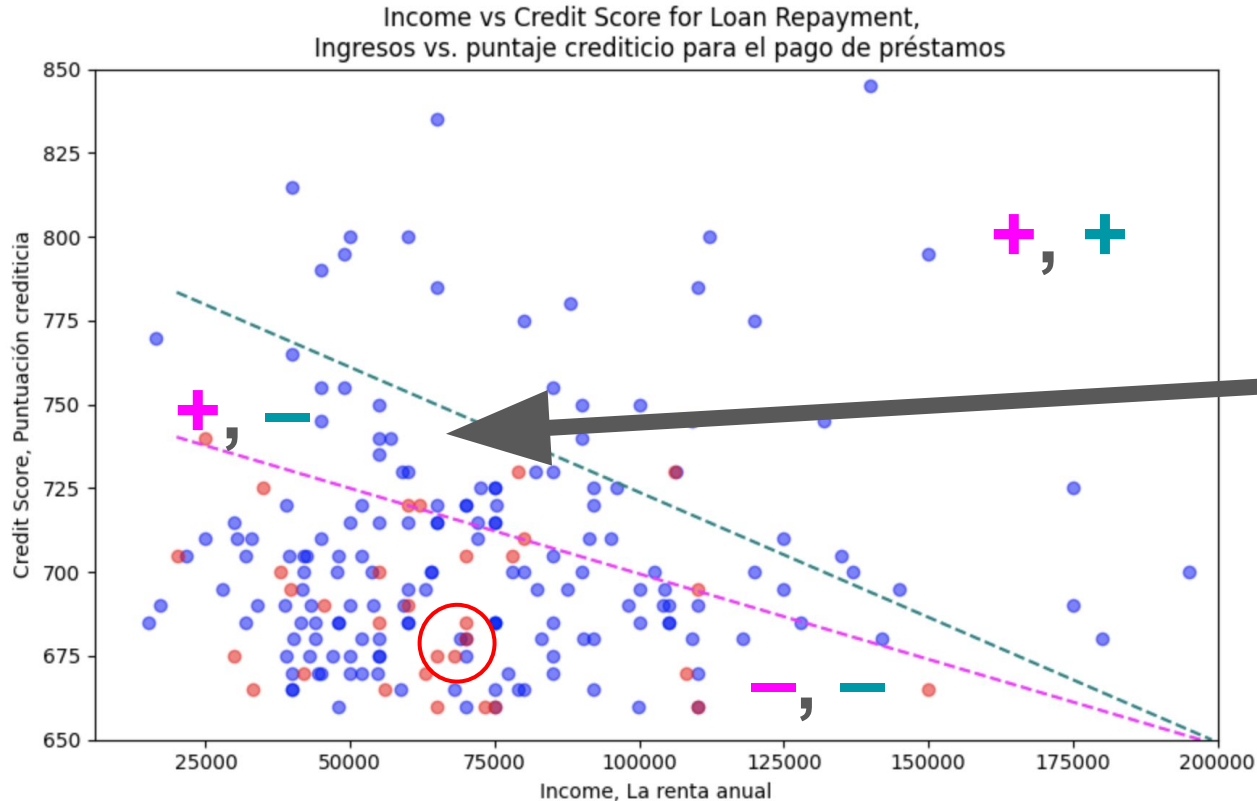
$$\text{error}(\text{monitor}) = \begin{cases} \text{monitor}^2 & \text{if default} \\ (1 - \text{monitor})^2 & \text{if repay loan} \end{cases}$$



For the rest of this talk, we compare two AI algorithms: each using the teal/pink error measurement  
Durante el resto de esta charla, comparamos dos algoritmos de IA: cada uno utiliza la medición de error verde azulado/rosa.



# Comparing two AI algorithms Comparación de dos algoritmos de IA

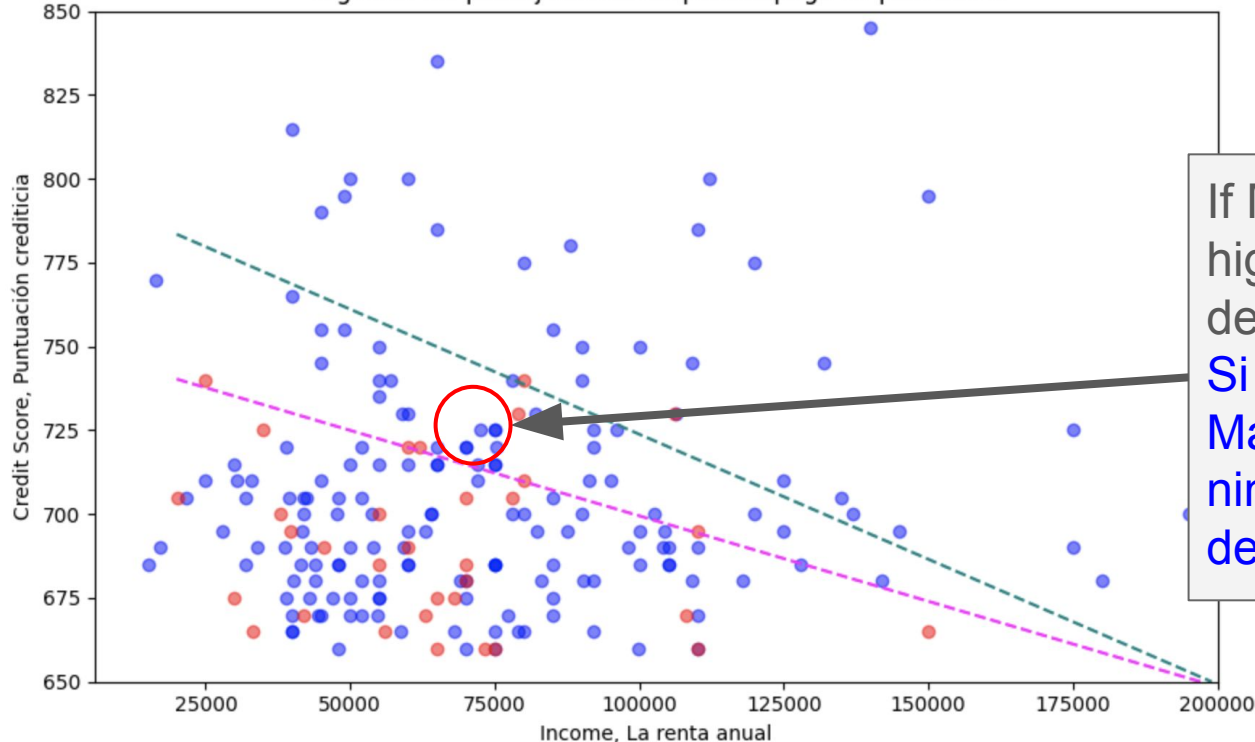


People with medium credit score and lower income are treated differently by the two algorithms

Las personas con puntuación crediticia media e ingresos más bajos reciben un trato diferente por parte de los dos algoritmos

Remember Maria's former partner incurred a lot of debt  
Recuerda que la expareja de María contrajo muchas deudas

Income vs Credit Score for Loan Repayment,  
Ingresos vs. puntaje crediticio para el pago de préstamos



If Maria's credit score was a bit higher, no similar applicants default on their loan

Si la puntuación crediticia de María fuese un poco más alta, ningún solicitante similar dejaría de pagar su préstamo.

Data/Datos: [LendingClub](https://lendingclub.com)

# Errors are inevitable



In a world of messy data, “perfect prediction” is typically impossible

En un mundo de datos desordenados, la “predicción perfecta” suele ser imposible



How we measure errors changes the algorithms we obtain, and how we evaluate these algorithms changes our perception of their quality

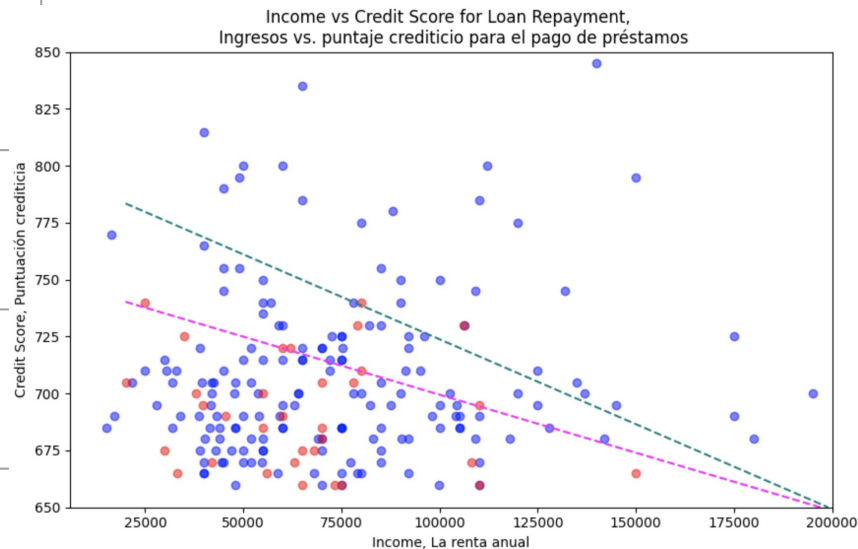
La forma en que medimos los errores cambia los algoritmos que obtenemos, y la forma en que evaluamos estos algoritmos cambia nuestra percepción de su calidad.

# Measuring quality of decisions

Accuracy exactitud	Average proportion of predictions where  is correct Proporción promedio de predicciones en las que la  acierta
Precision precisión	When you predict someone will repay, how often they repay Cuando predice que alguien pagará, con qué frecuencia lo hará.
Recall recuperación	Of all the people who would repay, how often are they predicted to repay? De todas las personas que pagarían, ¿con qué frecuencia se prevé que lo hagan?
Calibration calibración	When the algorithm predicts “70% change of repaying,” people repay 70% of the time Cuando el algoritmo predice “70% de cambio en el reembolso”, la gente reembolsa el 70% de las veces.
	... fairness, weighted welfare, etc. ...equidad, bienestar ponderado, etc.

# Let's compare **Vamos a comparar**

		
Accuracy ↑	<b>.435</b>	<b>.300</b>
Precision ↑	<b>.8857</b>	<b>.935</b>
Recall ↑	<b>.3173</b>	<b>.1736</b>
Calibration error ↓	<b>.1378</b>	<b>.1375</b>



# Mechanism design challenges for AI

Strategic behavior **Comportamiento estratégico**

Limited resources **Recursos limitados**

Fairness concerns **Preocupaciones de equidad**

# Summary Sumario

- AI algorithms are often trained and evaluated using some error metric
- Los algoritmos de IA suelen entrenarse y evaluarse utilizando alguna métrica de error.
- The choice of error metric changes how we view our algorithms
- La elección de la métrica de error cambia nuestra perspectiva sobre nuestros algoritmos.
- These evaluation metrics must be carefully chosen in partnership with stakeholders
- Estas métricas de evaluación deben elegirse cuidadosamente en colaboración con las partes interesadas.



Thank you Muchas gracias

Email: [finocch@bc.edu](mailto:finocch@bc.edu)