Aprendizaje de Máquina
Introducción

Alvaro Soto

Departamento de Ciencia de la Computación (DCC), PUC

Aprendizaje de Máquina

- Aprendizaje de Máquina (machine learning) es una subárea de la Inteligencia Artificial.

- Inteligencia Artificial?
  - Parte sencilla: Artificial.
  - Parte compleja: Inteligencia.

Mundo Biológico



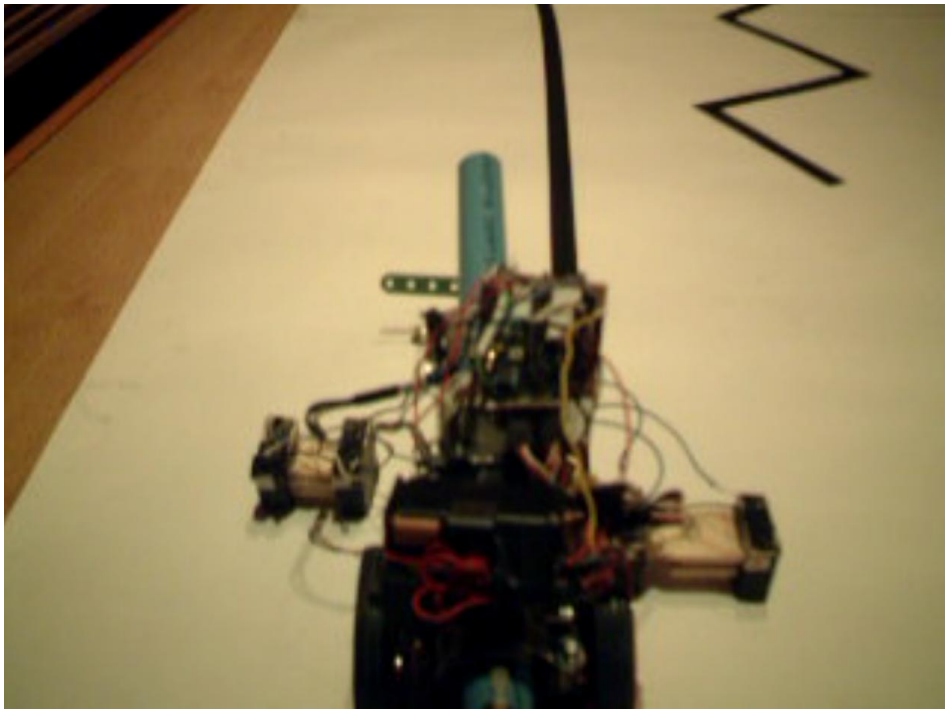

Mundo Artificial

Inteligencia Artificial

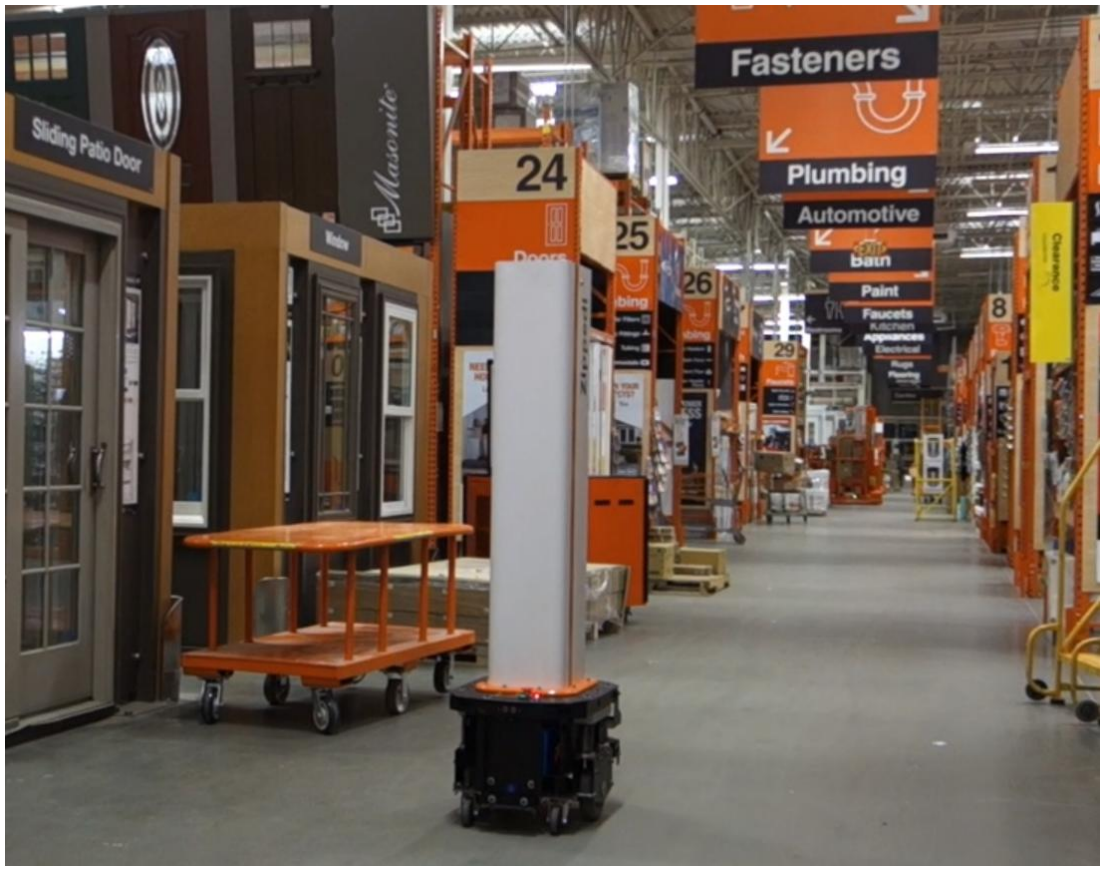
Estudio de modelos computacionales (algoritmos) que permitan construir máquinas capaces de percibir, razonar y actuar para lograr sus objetivos.

Desarrollo de una máquina cognitiva capaz de entender el estado del ambiente y razonar para lograr sus objetivos.

Understanding: Year 2001



Understanding: Year 2019



Algo de Historia

- 60's : Era del optimismo o ingenuidad.
  - Aprendizaje deductivo.
- 70's - 80's: Era del pesimismo.
- 90's: Era del renacimiento.
- 2000's: Era del Aprendizaje de Máquina.
  - Aprendizaje inductivo.

Aprendizaje Inductivo

This bird can fly

This bird can fly

This bird can fly

This bird can fly

Can this bird fly ?

Aprendizaje Inductivo

This bird can fly

This bird can fly

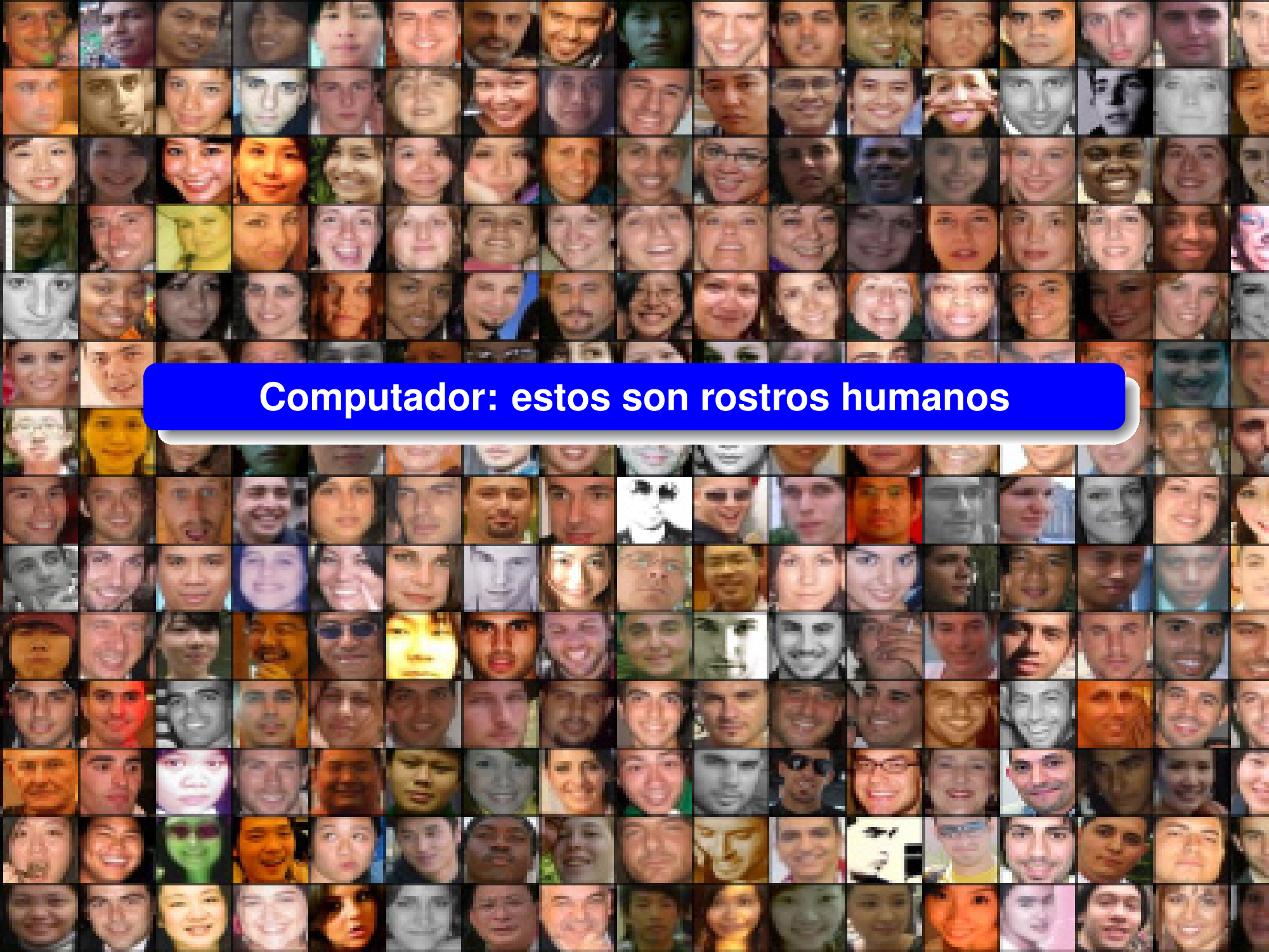
This bird can fly

This bird can fly

Can this bird fly ?

Aprendizaje de Máquina

Programas computacionales (algoritmos) que aprenden de la experiencia, i.e., datos.


Computador: estos son rostros humanos


Computador: estos NO son rostros humanos

Computador: algún rostro humano?

Esto es una silla

Esto es una silla

A veces un problema muy difícil

Esto es una silla

¿Qué es esto ?

Inteligencia Artificial: Una Visión Moderna (Rusell and Norvig)

Un agente inteligente es capaz de percibir, razonar y actuar con alta flexibilidad.

Perspectiva del Aprendizaje de Máquina

Un agente inteligente es capaz de aprender de la experiencia.

Aprender de la experiencia

Adquirir conocimiento en forma inductiva con objeto de:

- Entender el estado de un sistema (percibir).
- Realizar predicciones.
- Tomar decisiones.
- . . . , y en general actuar en forma *inteligente*.

Experiencia adecuada (datos) es clave

Experiencia: Espacio de características (feature space)

- Los algoritmos de aprendizaje de máquina operan sobre datos multidimensionales.
- Cada dato o registro está caracterizado por una serie de mediciones o atributos relevantes.

Ejemplos tomados del repositorio UCI (`http://archive.ics.uci.edu/ml`):

Wine Data Set

Download: Data Folder, Data Set Description

Abstract: Using chemical analysis determine the origin of wines

Data Set Characteristics:	Multivariate	**Number of Instances:**	178	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	13	**Date Donated**	1991-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	573523

Bank Marketing Data Set

Download: Data Folder, Data Set Description

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classif

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	248768

Set de entrenamiento

Los datos usados para ajustar un modelo de aprendizaje de máquina son denominados: set de entrenamiento (training set).

Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no

Set de entrenamiento

Attributes, dimensions, variables, or features.

Class or Label

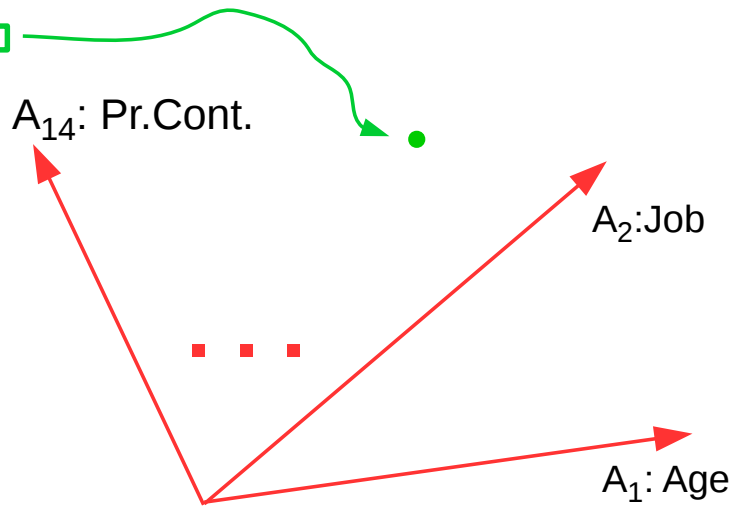
Examples, registers, or instances.

Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no

Espacio de características: Feature space

- Cada registro o tupla del set de entrenamiento puede ser considerada como un vector en el espacio de características o feature space.

Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no



Generalización el GRAN objetivo

Set de entrenamiento usado por el modelo en su aprendizaje.

Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no

Nuevas instancias no vistas previamente por el modelo.

20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	?
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	

Generalización el GRAN objetivo

Set de entrenamiento.

Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no

Generalización: capacidades de predecir correctamente nuevas instancias, no incluidas en set de entrenamiento.

20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	?
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	

Sets de: i) Entrenamiento, ii) Test y iii) Validación

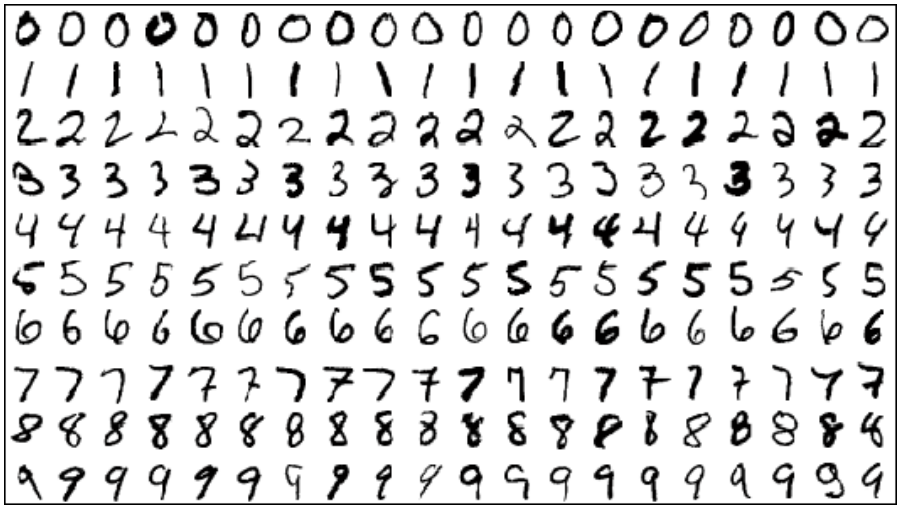
(i) Set de entrenamiento: datos usados para entrenar modelo.

(ii) Set de test: instancias con rótulos conocidos, pero no usadas durante el entrenamiento. Útil para evaluar capacidades de generalización.

(iii) Set de validación: instancias con rótulos conocidos usadas durante el entrenamiento para ajustar hiper parámetros estructurales del algoritmo de aprendizaje de máquina.

Regla práctica: reservar 70-80% datos para entrenamiento, 10-20% para test y 10-20% para validación.

Ej. MNIST dataset

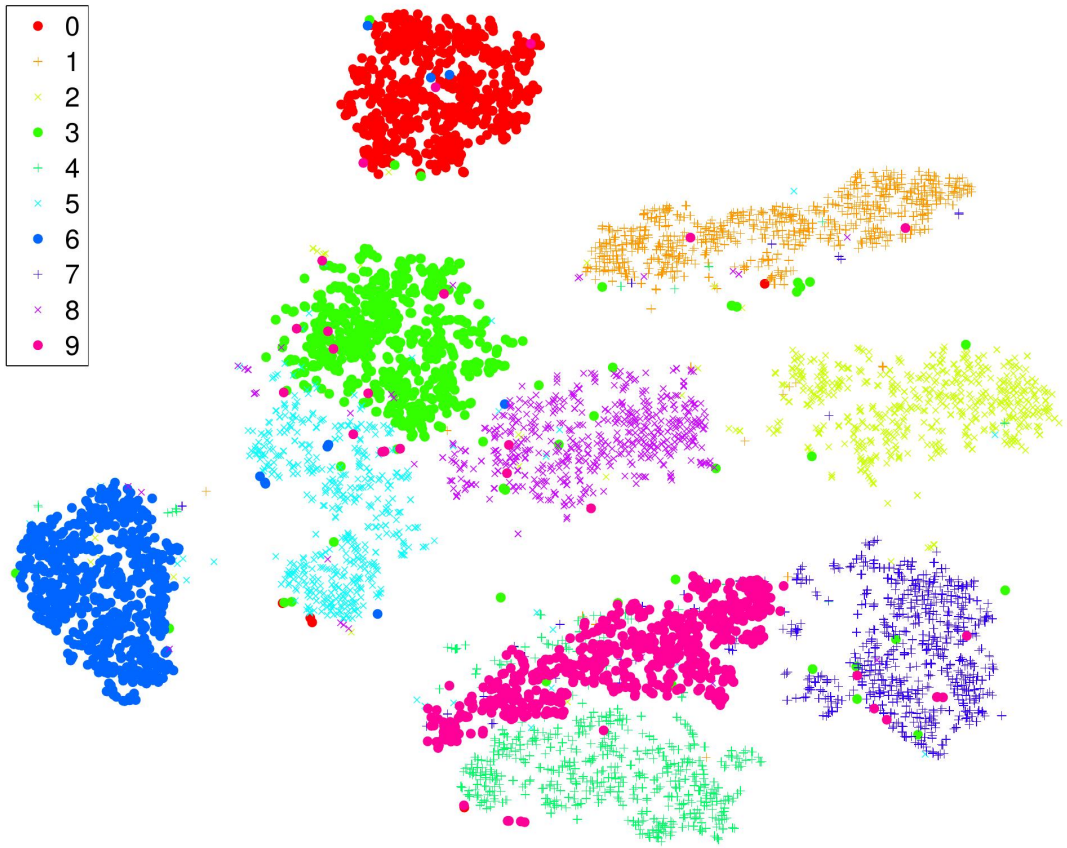


- MNIST es un dataset de imágenes sobre dígitos escritos a mano.
- Cada imagen contiene un dígito entre 0 y 9. Las imágenes son binarias con una resolución de 28x28 pixeles.
- El objetivo es construir un clasificador que permita reconocer el dígito en cada imagen.
- El dataset consiste de 60.000 ejemplos de entrenamiento y 10.000 de test.

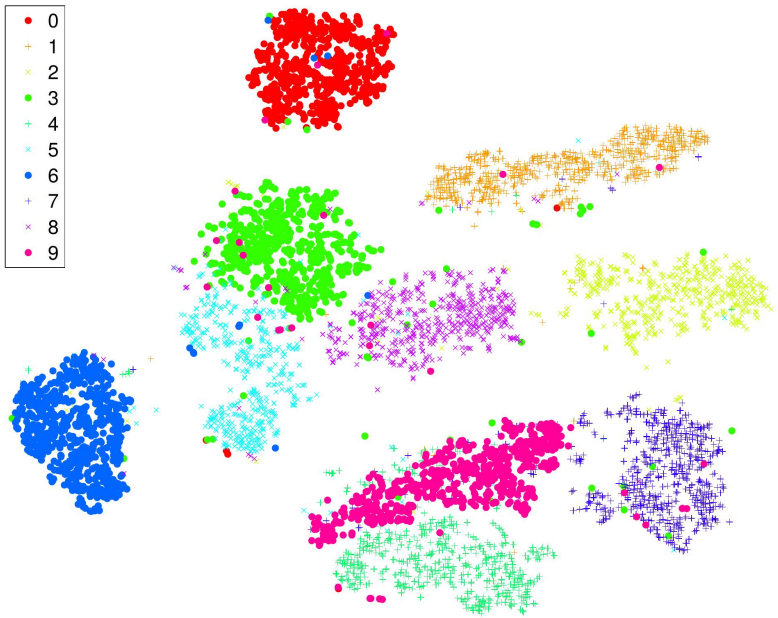
¿Cuál es el tamaño del feature space del set MNIST?

Ej. MNIST dataset

- No es posible una representación directa de MNIST en una visualización 2D.
- Sin embargo, utilizando la alquimia de técnicas de reducción de dimensionalidad, tal como t-SNE (`http://lvdmaaten.github.io/tsne`), es posible lograr la siguiente aproximación:



Ej. MNIST dataset



- La visualización 2D ilustra la tarea que debe cumplir un clasificador
- ¿Cuál es esta tarea?

Un primer clasificador

- Matrix de confusión al aplicar un clasificador del tipo K-vecinos cercanos sobre el set de test de MNIST.
- En este caso el clasificador usa un valor de K=1 y distancia euclideana en el espacio de 784 dimensiones (¿por qué 784?).

Predicción

Real

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	972	1	1	0	0	1	3	1	0	0
1	0	1129	3	0	1	1	1	0	0	0
2	7	6	992	5	1	0	2	16	3	0
3	0	1	2	970	1	19	0	7	7	3
4	0	7	0	0	944	0	3	5	1	22
5	1	1	0	12	2	860	5	1	6	4
6	4	2	0	0	3	5	944	0	0	0
7	0	14	6	2	4	0	0	992	0	10
8	6	1	3	14	5	13	3	4	920	5
9	2	5	1	6	10	5	1	11	1	967

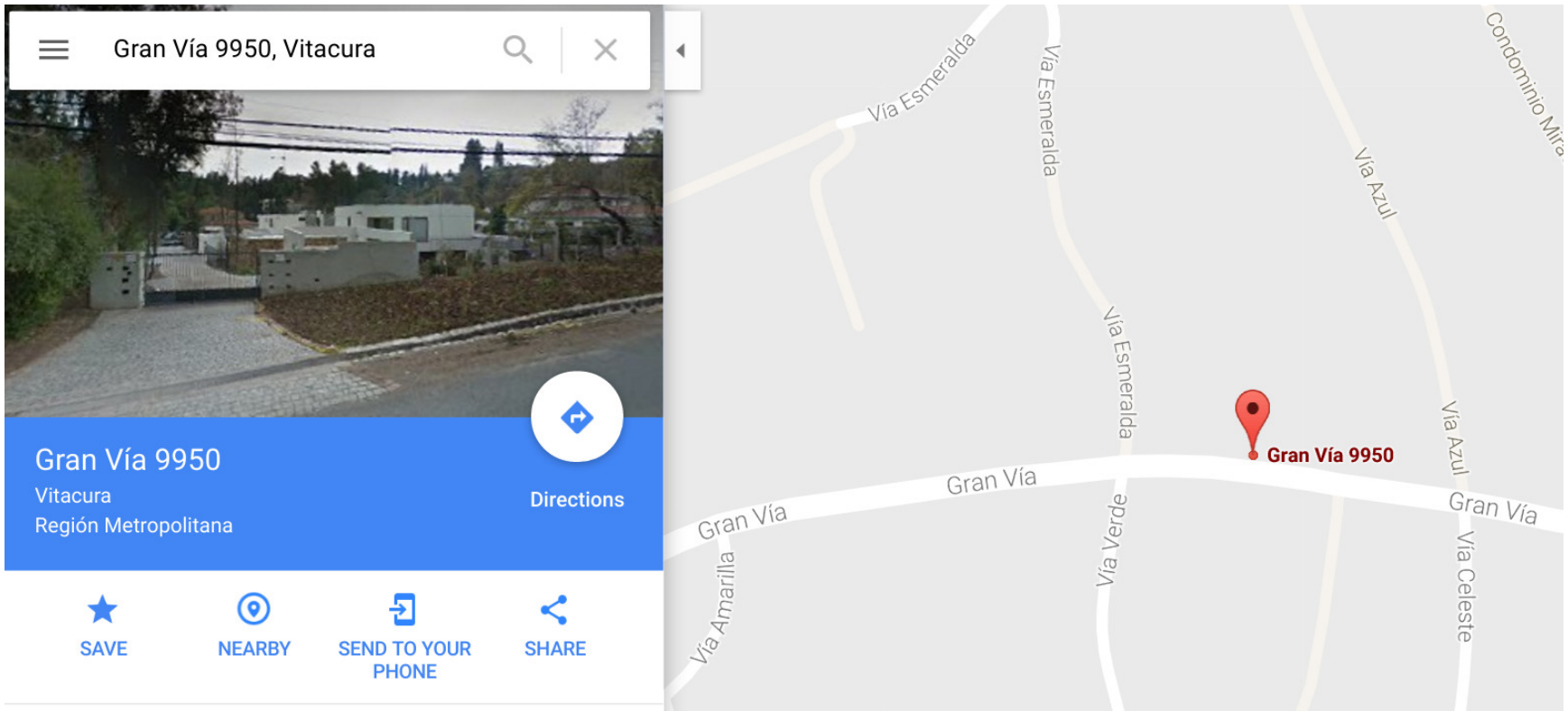
Aprendizaje de máquina y set de entrenamiento

- Como mencionamos, el foco principal del aprendizaje de máquina es aprender desde fuentes de datos.
- De hecho, hoy el paradigma denominado Big Data es el gran aliado del aprendizaje de máquina.
- Como veremos en el curso:

> Un buen set de entrenamiento es clave para poder lograr un buen aprendizaje.

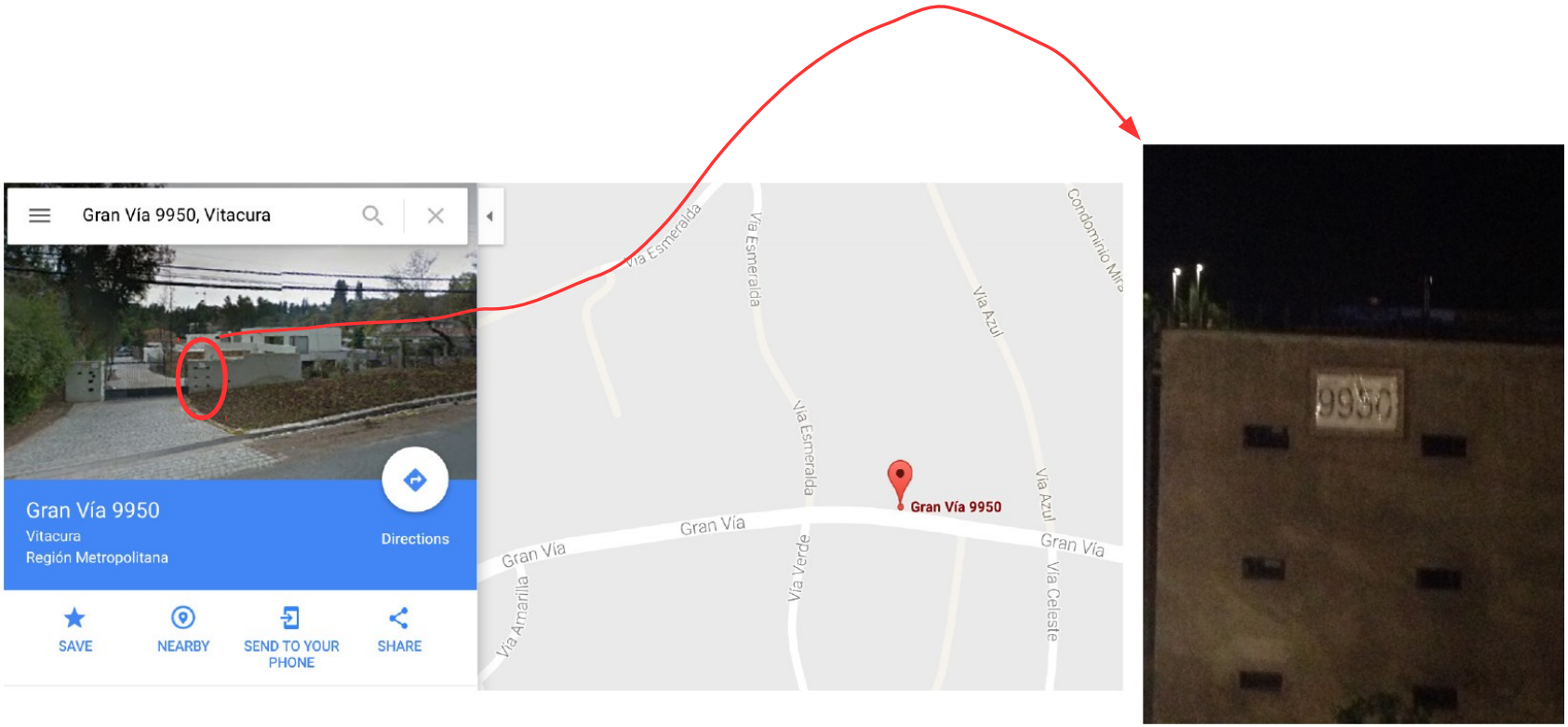
Un buen set de entrenamiento es clave

Ejemplo: Google Street View.



Un buen set de entrenamiento es clave

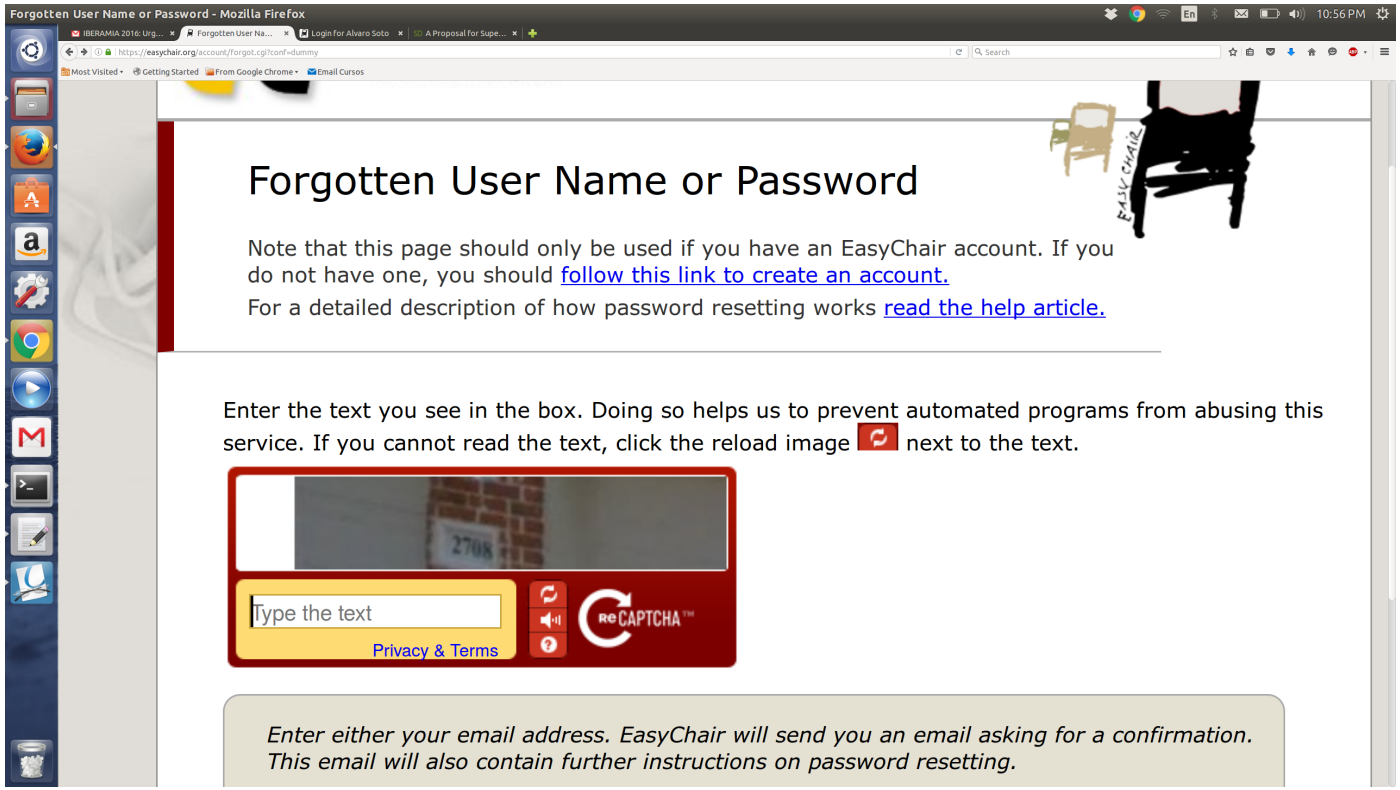
Ejemplo: Google Street View.



¿Cómo hacen esto?

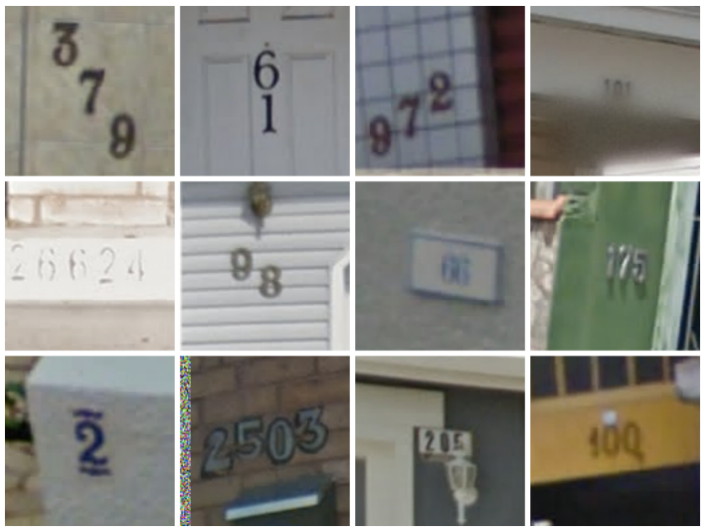
Un buen set de entrenamiento es clave

Ejemplo: Google Street View.



Un buen set de entrenamiento es clave

Ejemplo: Google Street View.



http://research.google.com/pubs/pub42241.html

Un buen set de entrenamiento es clave

Ejemplo: Reconocimiento de poses en consola Xbox

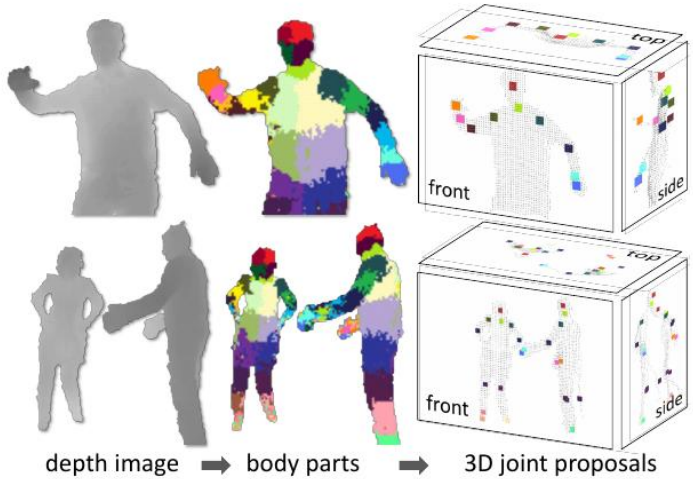
Real-Time Human Pose Recognition in Parts from Single Depth Images

Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio
Richard Moore Alex Kipman Andrew Blake
Microsoft Research Cambridge & Xbox Incubation

Abstract

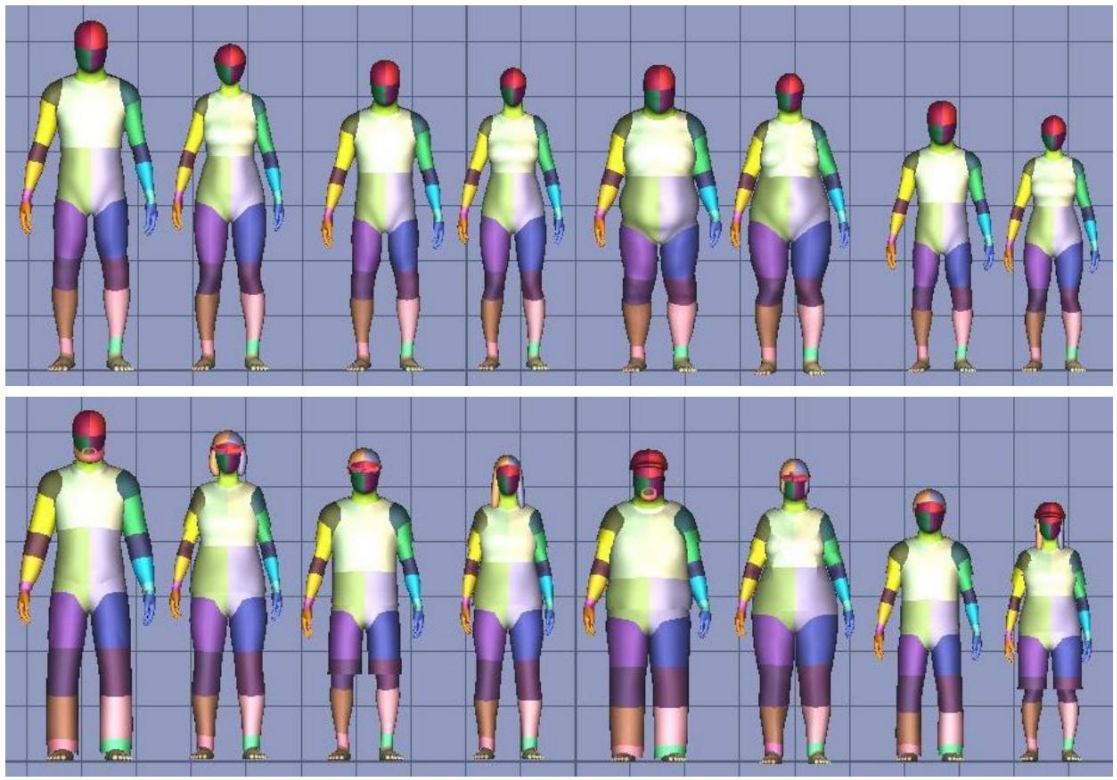
We propose a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. We take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Our large and highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. *Finally we generate confidence-scored 3D proposals of several body joints by reprojecting the classification result and finding local modes.*

The system runs at 200 frames per second on consumer



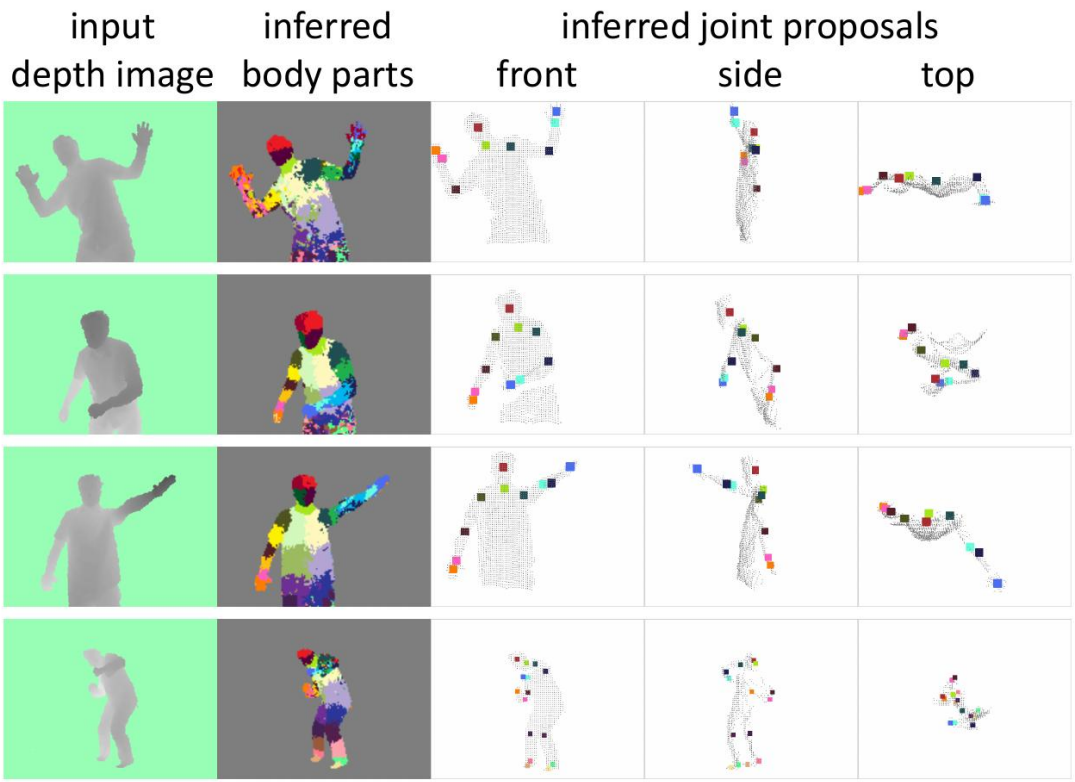
Un buen set de entrenamiento es clave

Ejemplo: Reconocimiento de poses en consola Xbox



Un buen set de entrenamiento es clave

Ejemplo: Reconocimiento de poses en consola Xbox



Sobreajuste?

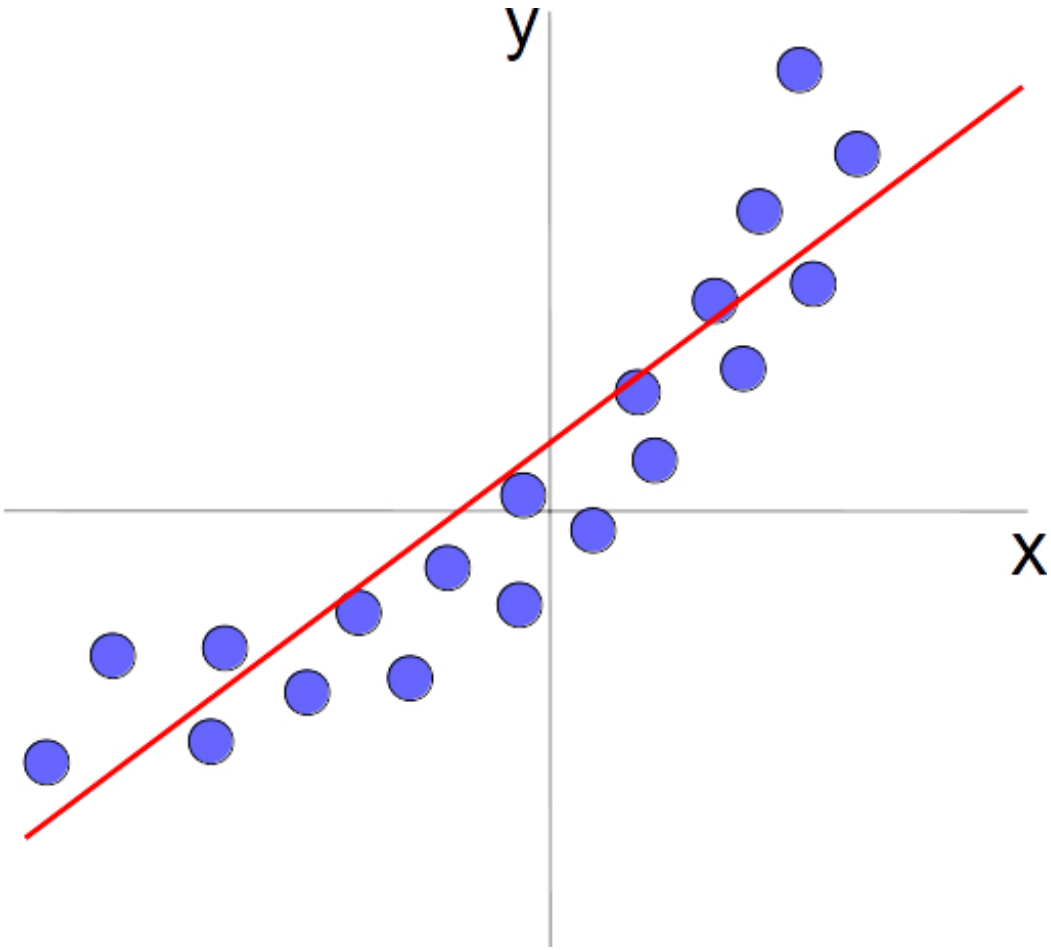
- Como mencionamos, generalización es lo fundamental para una técnica de aprendizaje de máquina, ¿por qué?.
- Generalización $\rightarrow$ Aprendizaje inductivo.
- En otras palabras, lo importante es lograr una buena predicción en instancias nuevas del dominio de interés.

Sobreajuste (Overfitting): ocurre cuando un modelo de aprendizaje de máquina comienza a memorizar los datos de entrenamiento perdiendo sus capacidades de generalización.

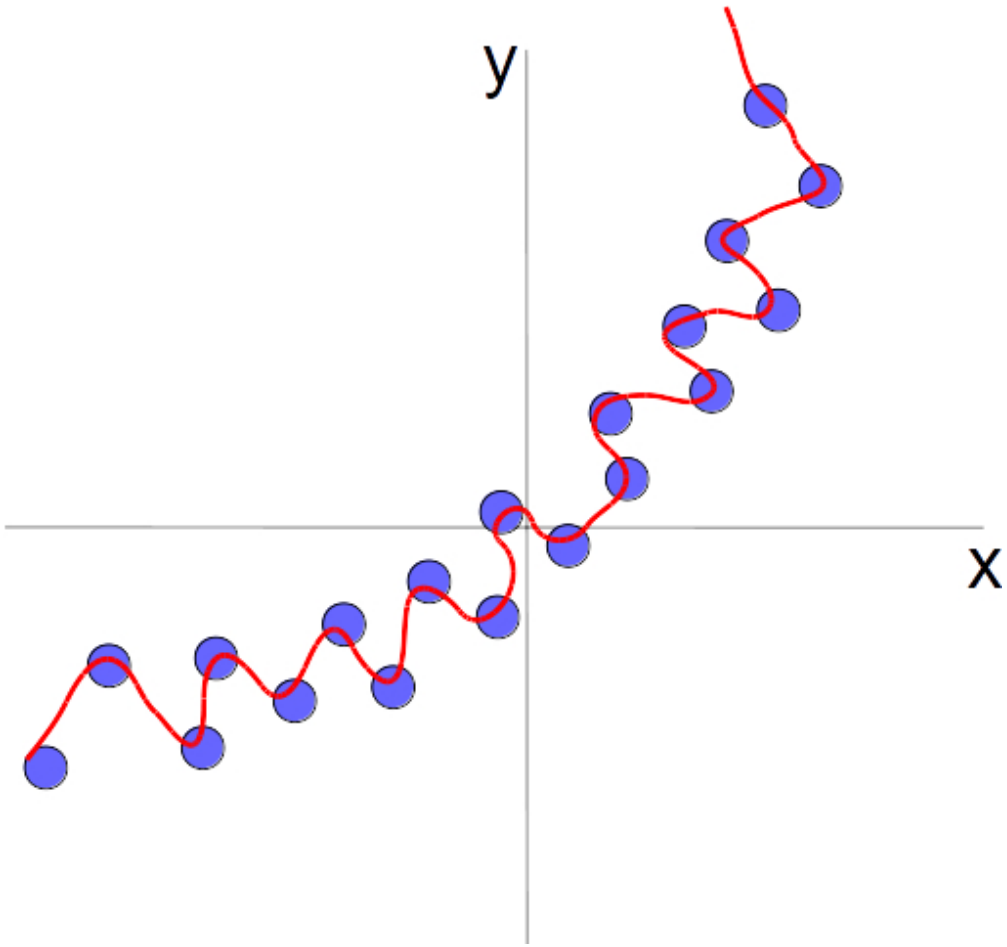
Jorge Luis Borges: ``Funes el memoriozo".

``...Había aprendido sin esfuerzo el inglés, el francés, el portugués, el latín. Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es **generalizar, abstraer**. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos."

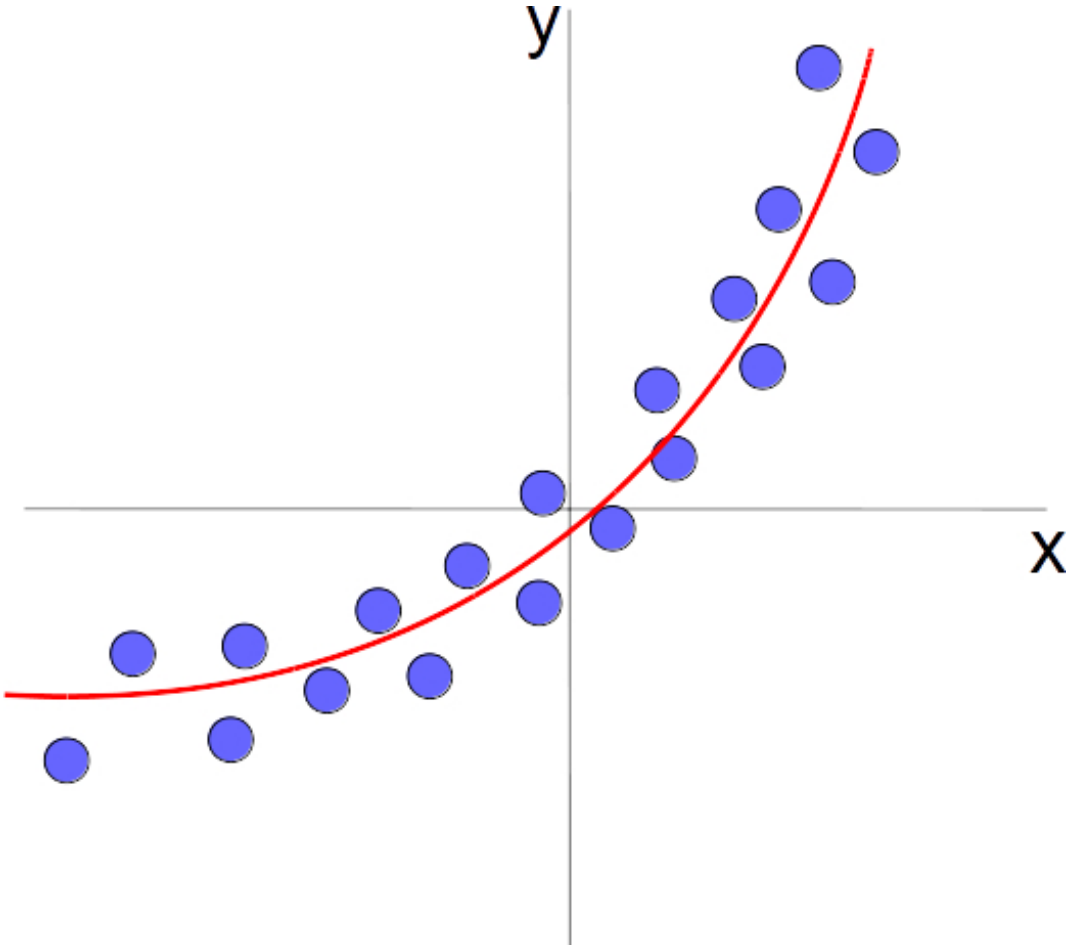
Underfitting, Overfitting, Good fit

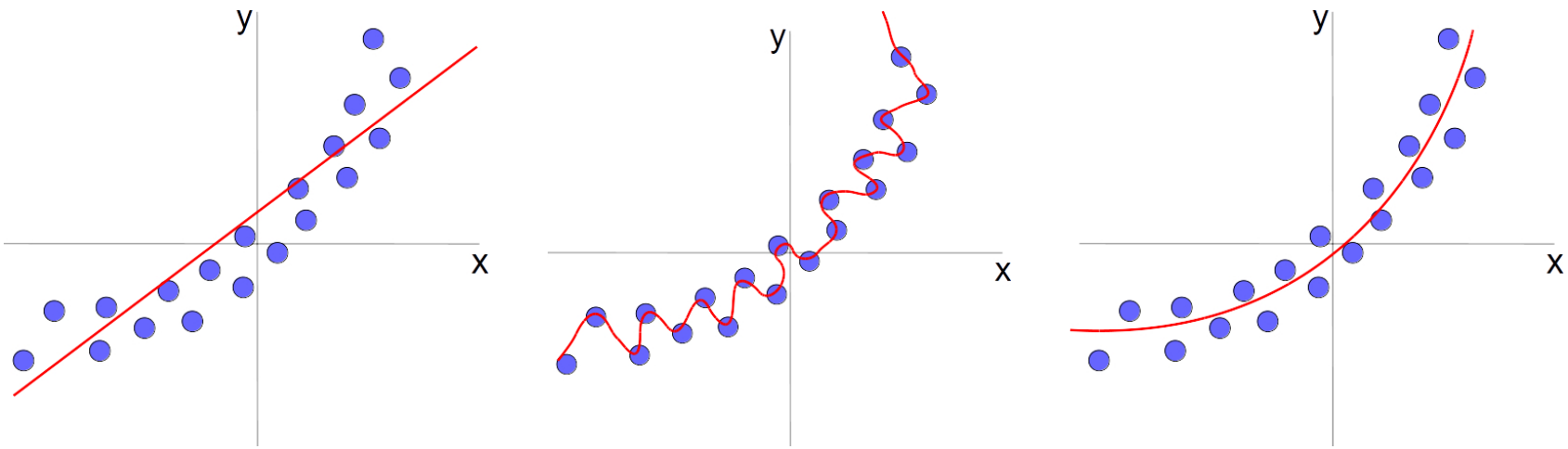


Underfitting, Overfitting, Good fit



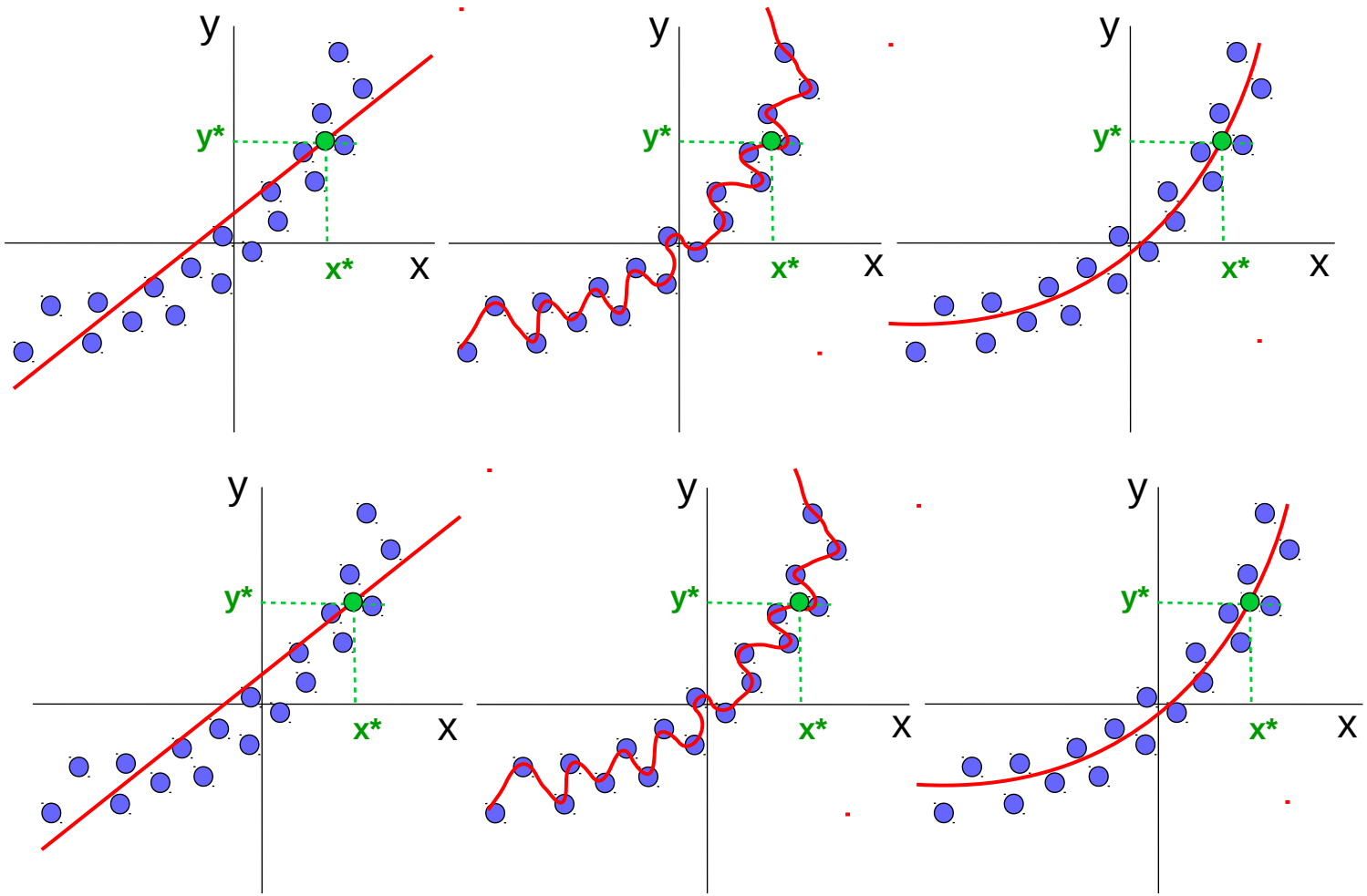
Underfitting, Overfitting, Good fit





Un buen modelo captura patrones (tendencias) relevantes en los datos. Una condición importante es lograr un buen nivel de robustez ante ruido en los datos de entrenamiento.

Generalización



Resumen

Conceptos importantes:

- Inteligencia Artificial vs Aprendizaje de Máquina (ML).
- Visión genérica de un algoritmo de ML.
- Elementos principales de un algoritmo de ML: Representación + Rendimiento + Optimización.
- Relevancia de datos de entrenamiento.
- Espacio de características.
- Aprendizaje supervisado, no supervisado, reforzado.
- Sobreajuste.
- Set de entrenamiento, validación y test.

Lectura recomendada:

- “A Few Useful Things to Know About Machine Learning” by Pedro Domingos.

Tapping into the "folk knowledge" needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

MACHINE LEARNING SYSTEMS automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation.[15] Several fine textbooks are available to interested practitioners and researchers (for example, Mitchell[16] and Witten et al.[24]). However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less-than-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This is the purpose of this article.

» key insights

- **Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled.**
- **Machine learning is widely used in computer science and other fields. However, developing successful machine learning applications requires a substantial amount of "black art" that is difficult to find in textbooks.**
- **This article summarizes 12 key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions.**