Clase 6

Problemas abiertos en DL y ML

Enzo Ferrante

eferrante@sinc.unl.edu.ar

@enzoferrante

¿Cómo aprender nuevas representaciones de los datos?

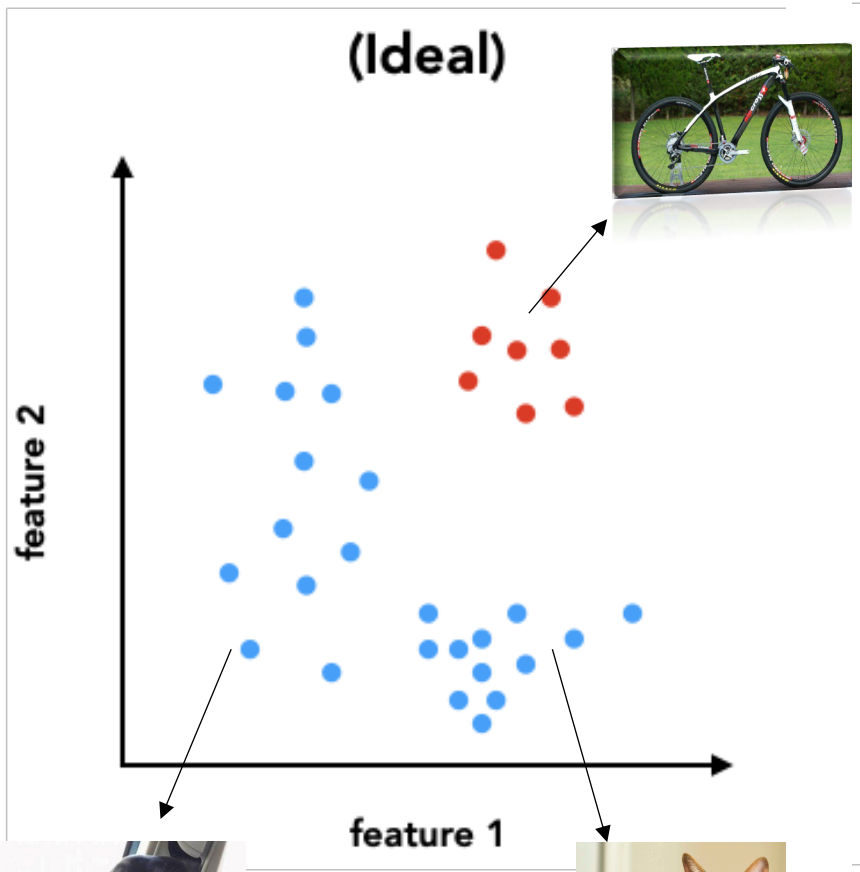
¿Cómo generalizar a múltiples dominios de datos?

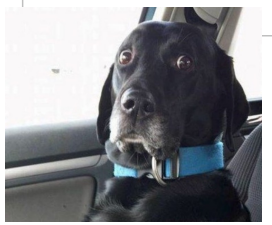
¿Cómo interpretar los modelos entrenados?

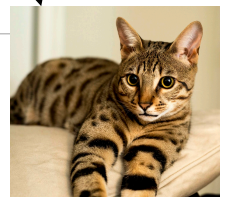
¿Cómo evitar el sesgo en los modelos entrenados?

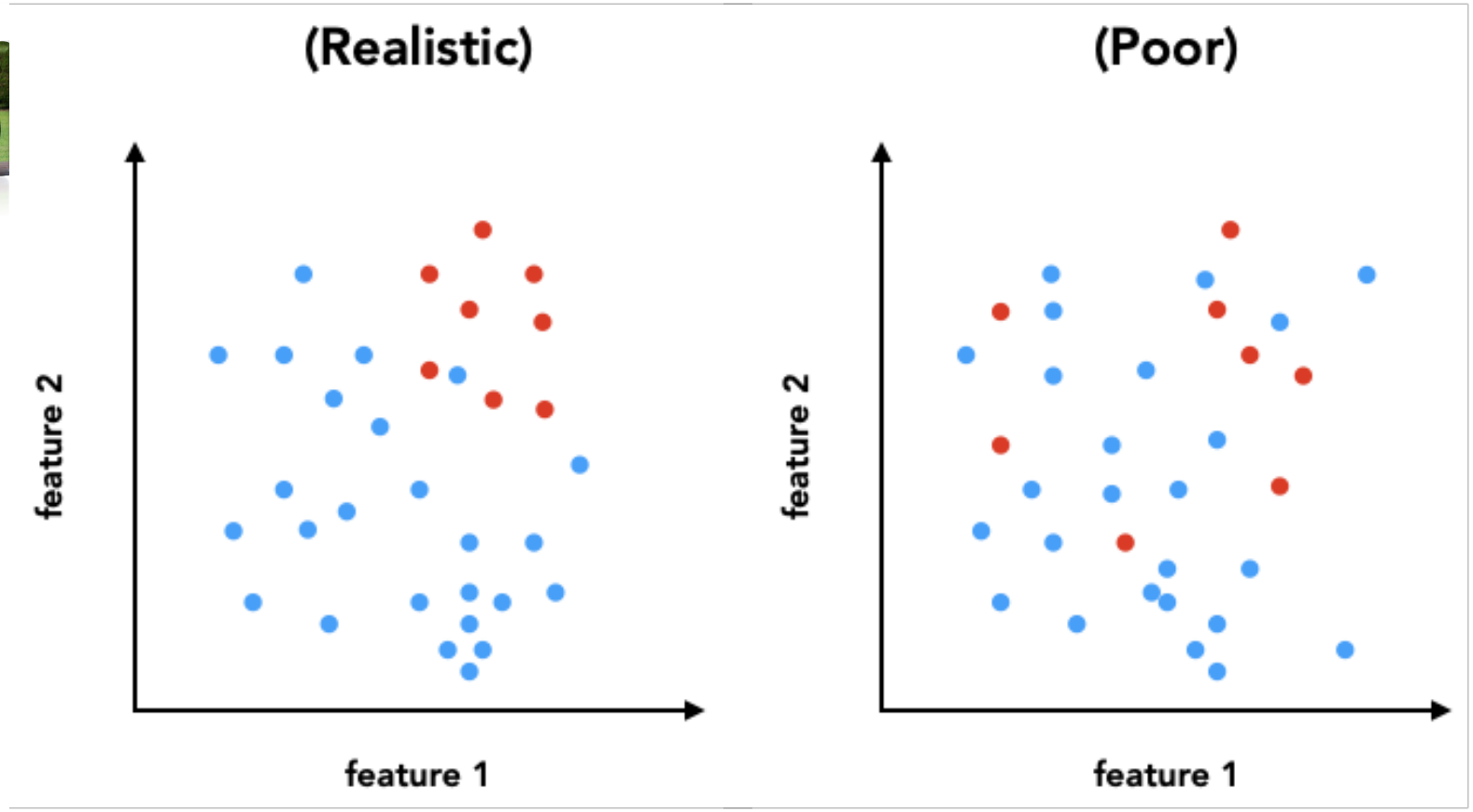
¿Cómo aprender nuevas representaciones de los datos?

Nuevas representaciones









Extraído de: https://towardsdatascience.com/concept-learning-and-feature-spaces-45cee19e49db

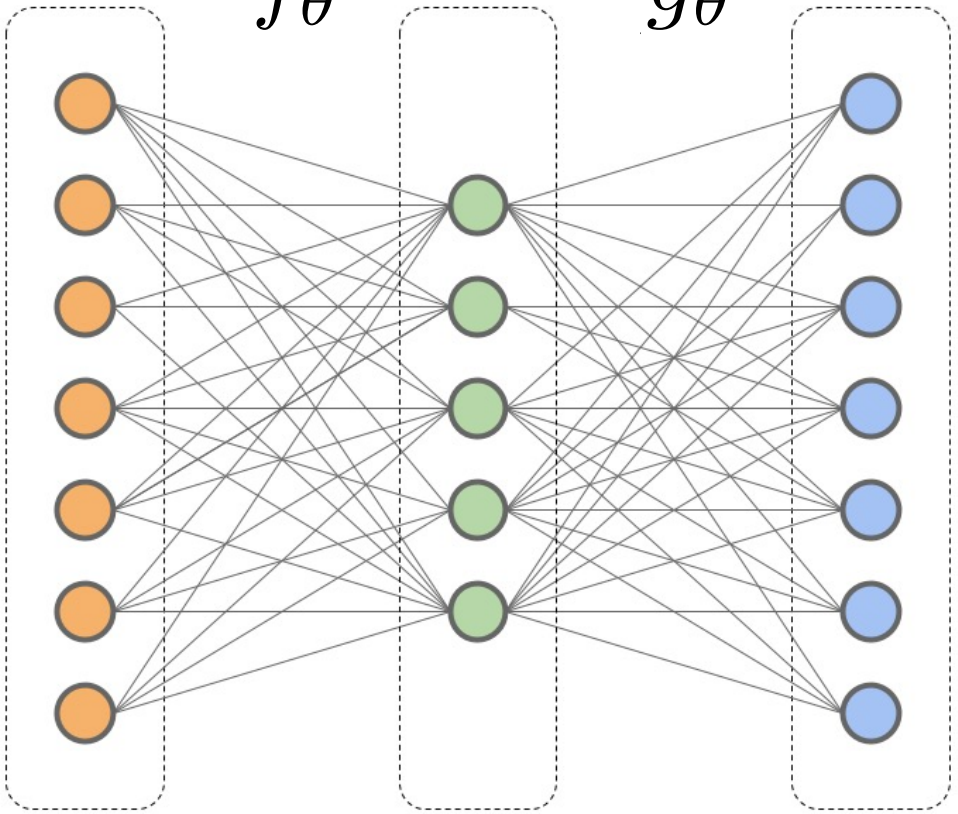
Aprendizaje de representaciones no supervisado

Autocodificadores

Encoder f_θ

Decoder g_θ

x



$h = f_\theta(x)$ Espacio latente

Función de pérdida

$$L(x;\theta) = (g_\theta(f_\theta(x)) - x)^2$$

Decoder Encoder Imagen original

Autocodificadores convolucionales

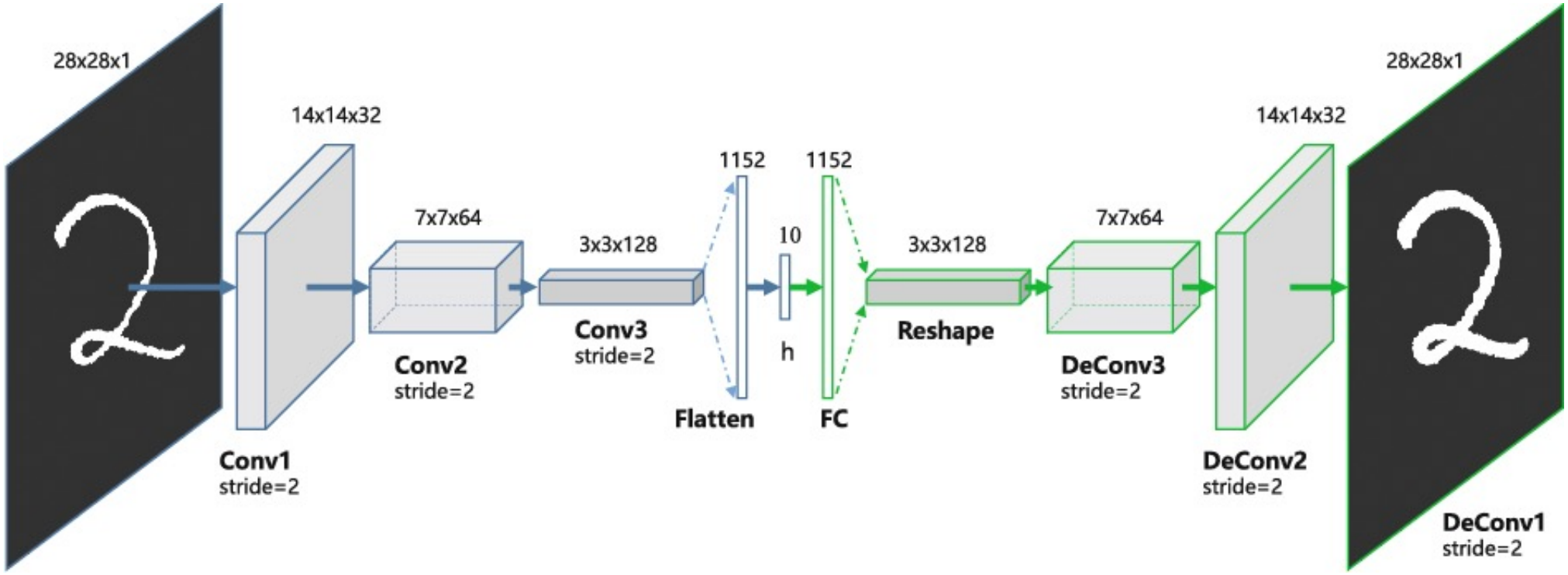


Imagen Extraída de Vincent et al, 2010 (JMLR)

Denoising autoencoders

Journal of Machine Learning Research 11 (2010) 3371-3408

Submitted 5/10; Published 12/10

Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion

Pascal Vincent PASCAL.VINCENT@UMONTREAL.CA
Département d'informatique et de recherche opérationnelle
Université de Montréal
2920, chemin de la Tour
Montréal, Québec, H3T 1J8, Canada

Hugo Larochelle LAROCHEH@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, Ontario, M5S 3G4, Canada

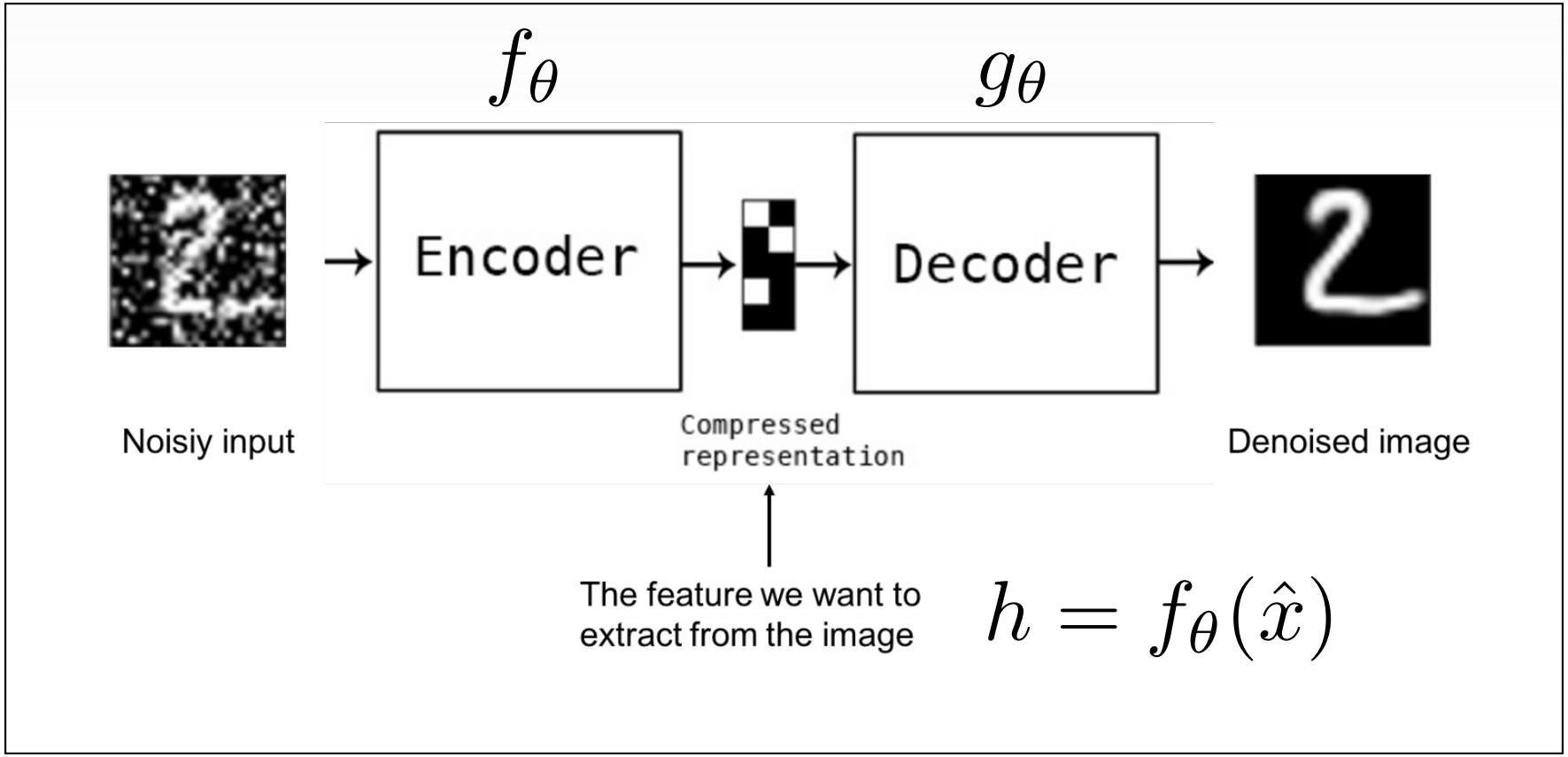
Isabelle Lajoie ISABELLE.LAJOIE.1@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Pierre-Antoine Manzagol PIERRE-ANTOINE.MANZAGOL@UMONTREAL.CA
Département d'informatique et de recherche opérationnelle
Université de Montréal
2920, chemin de la Tour
Montréal, Québec, H3T 1J8, Canada

Editor: Léon Bottou

Abstract

We explore an

Denoising autoencoders



Loss function
Reconstruction MSE

$$L(x;\theta) = (g_\theta(f_\theta(\hat{x})) - x)^2$$

Imagen con ruido

Decoder Encoder

Imagen original

Imagen Extraída de https://blog.sicara.com/keras-tutorial-content-based-image-retrieval-convolutional-denoising-autoencoder-dc91450cc511

Denoising autoencoders

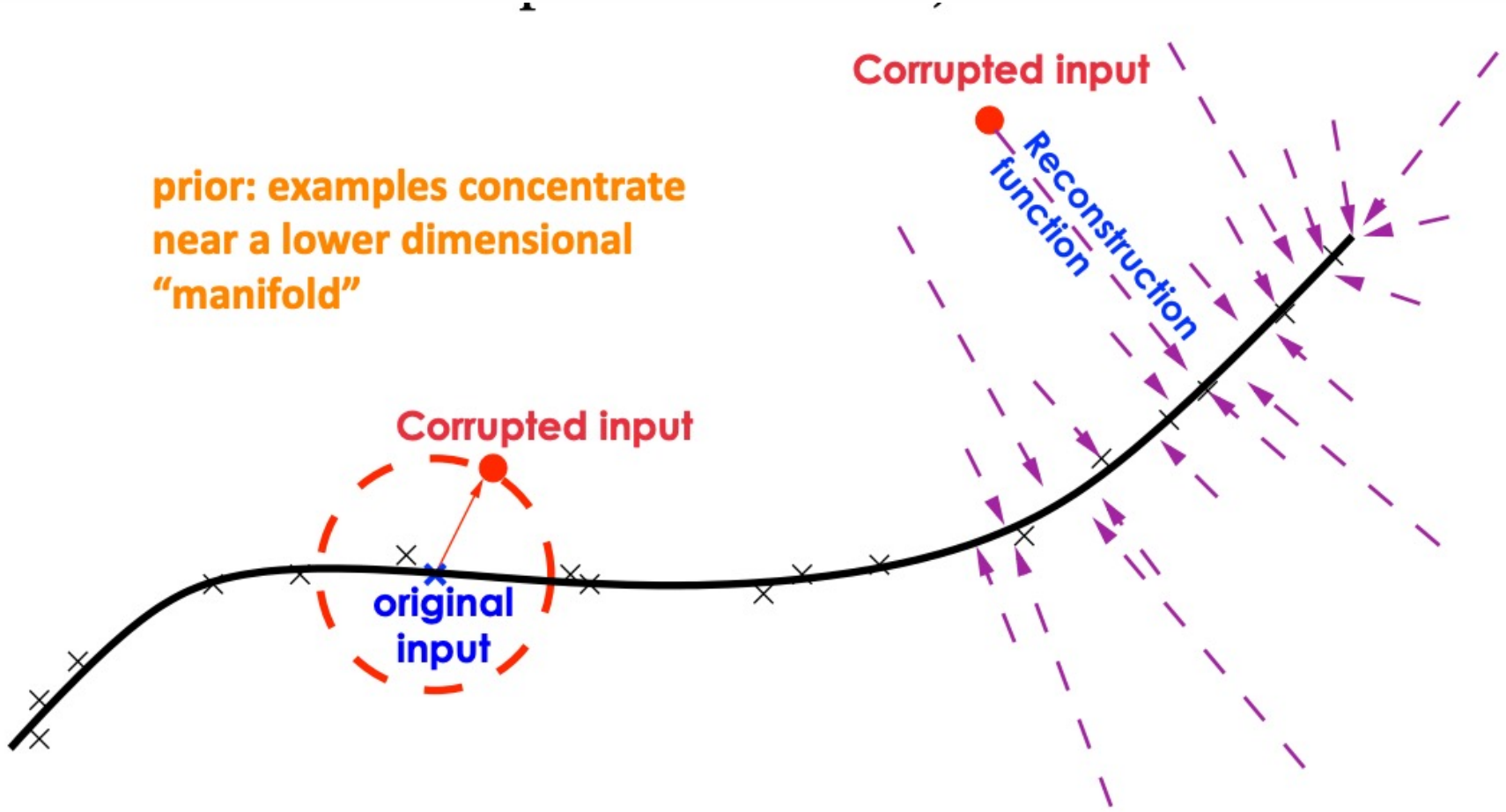
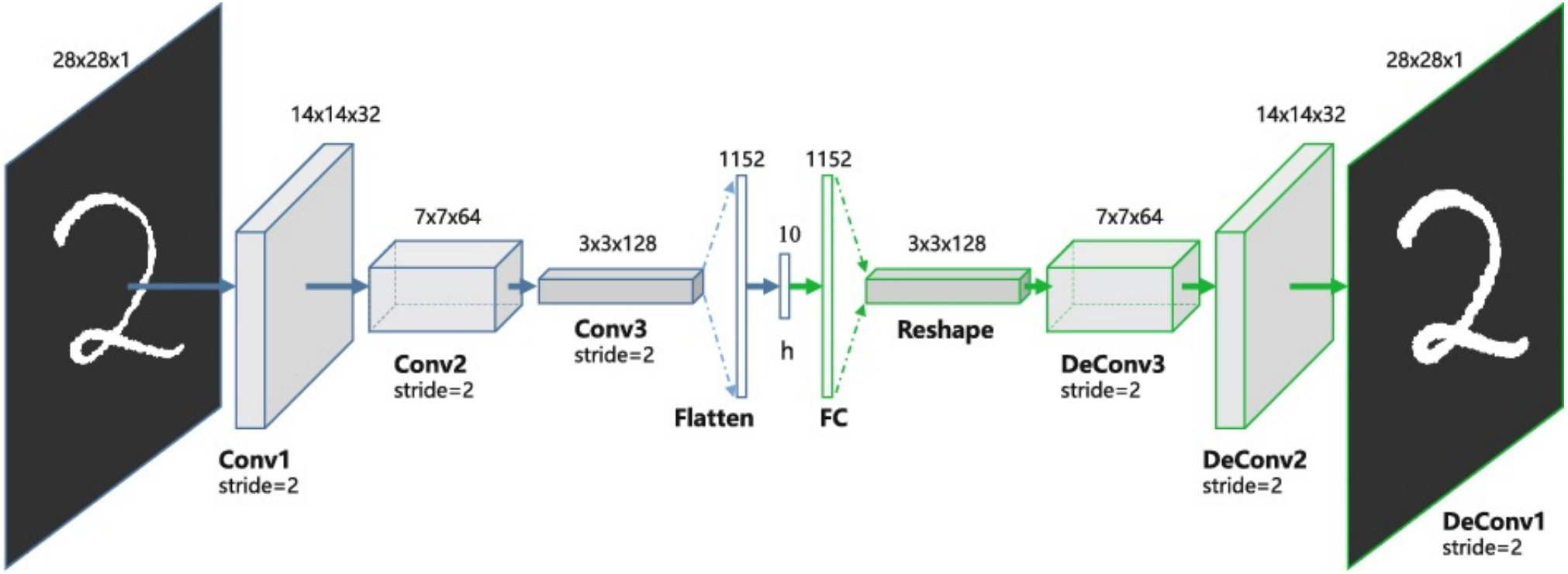
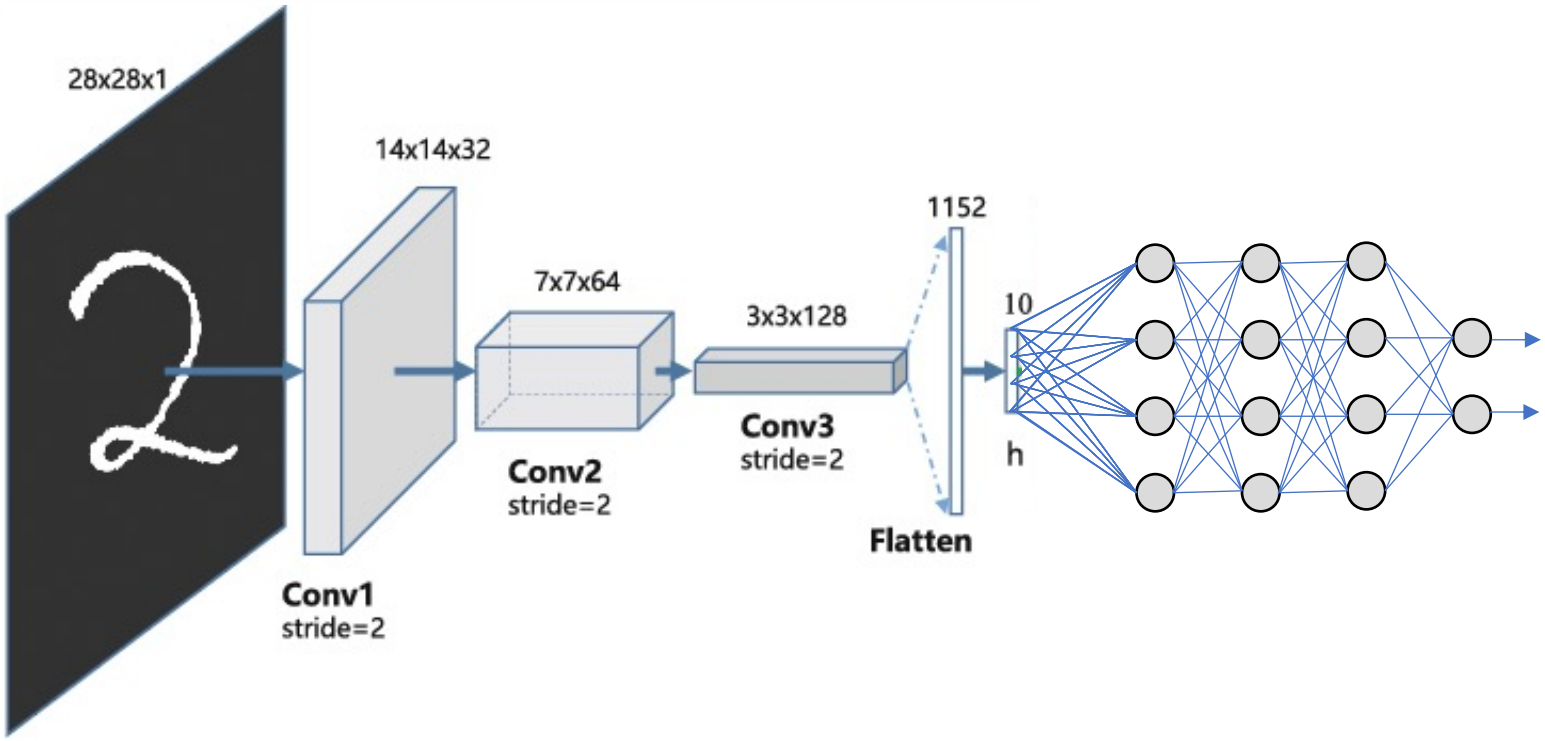


Imagen Extraída de Bengio et al, 2014 ” Representation Learning: A Review and New Perspectives”

Aplicación 1: Inicialización de modelos



Aplicación 1: Inicialización de modelos



- El modelo se inicializa con los pesos del autoencoder, y se continua entrenando para la tarea de clasificación
- Permite usar datos sin etiquetas para el pre-entrenamiento

Aplicación 2: Detección de anomalías

Pregunta: Cómo usar un autocodificador para realizar detección de anomalías?

Paper: https://arxiv.org/pdf/1807.07356.pdf

Aplicación 2: Detección de anomalías

- **Entreno** un autocodificador con casos 'normales'

- Dado un dataset de prueba, lo **reconstruyo** utilizando el autocodificador

- **Computo el MSE** entre la reconstrucción y los datos originales

- Defino un **threshold** en el MSE para determinar cuando un dato debe ser considerado outlier

Aplicación 2: Detección de anomalías

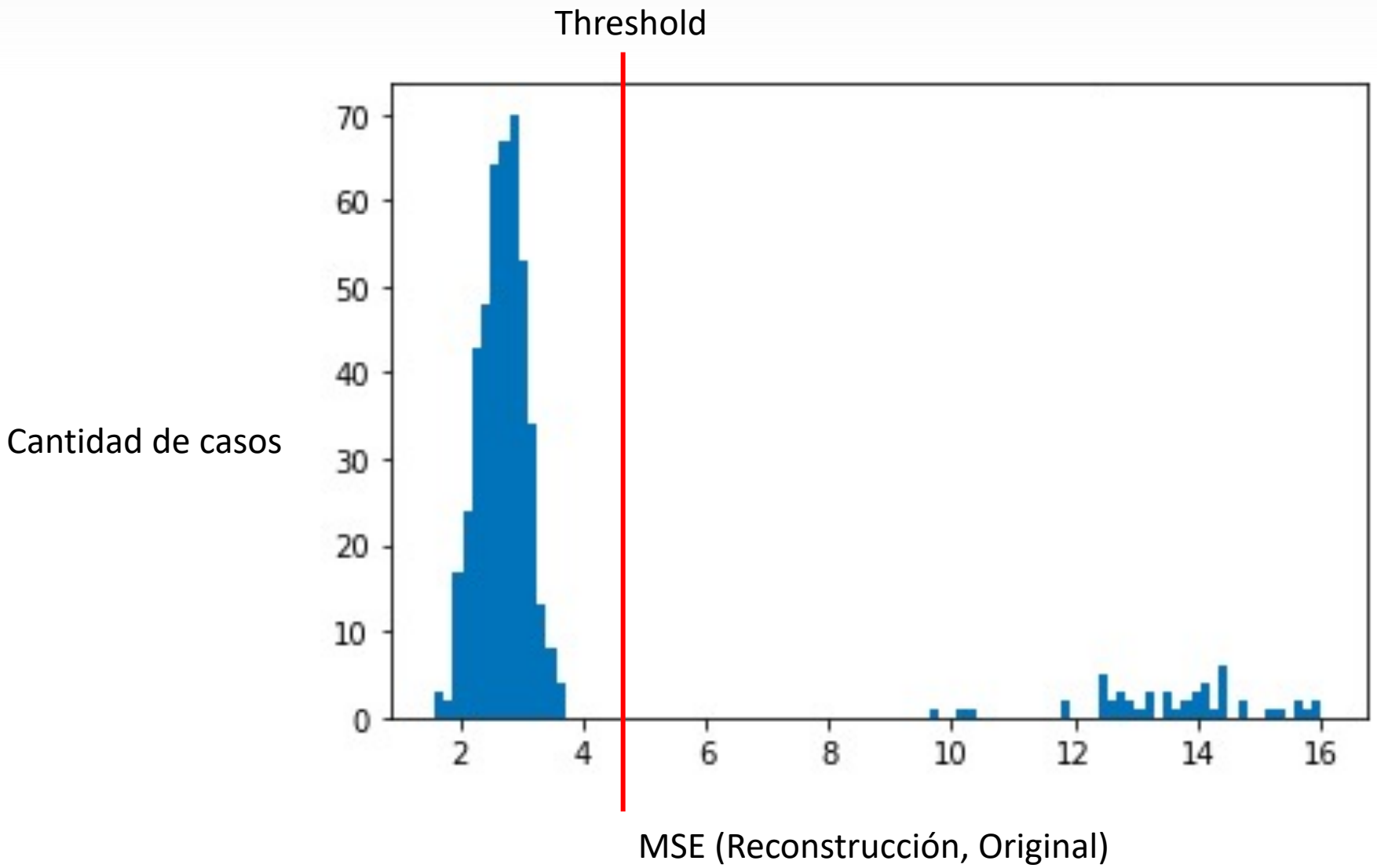
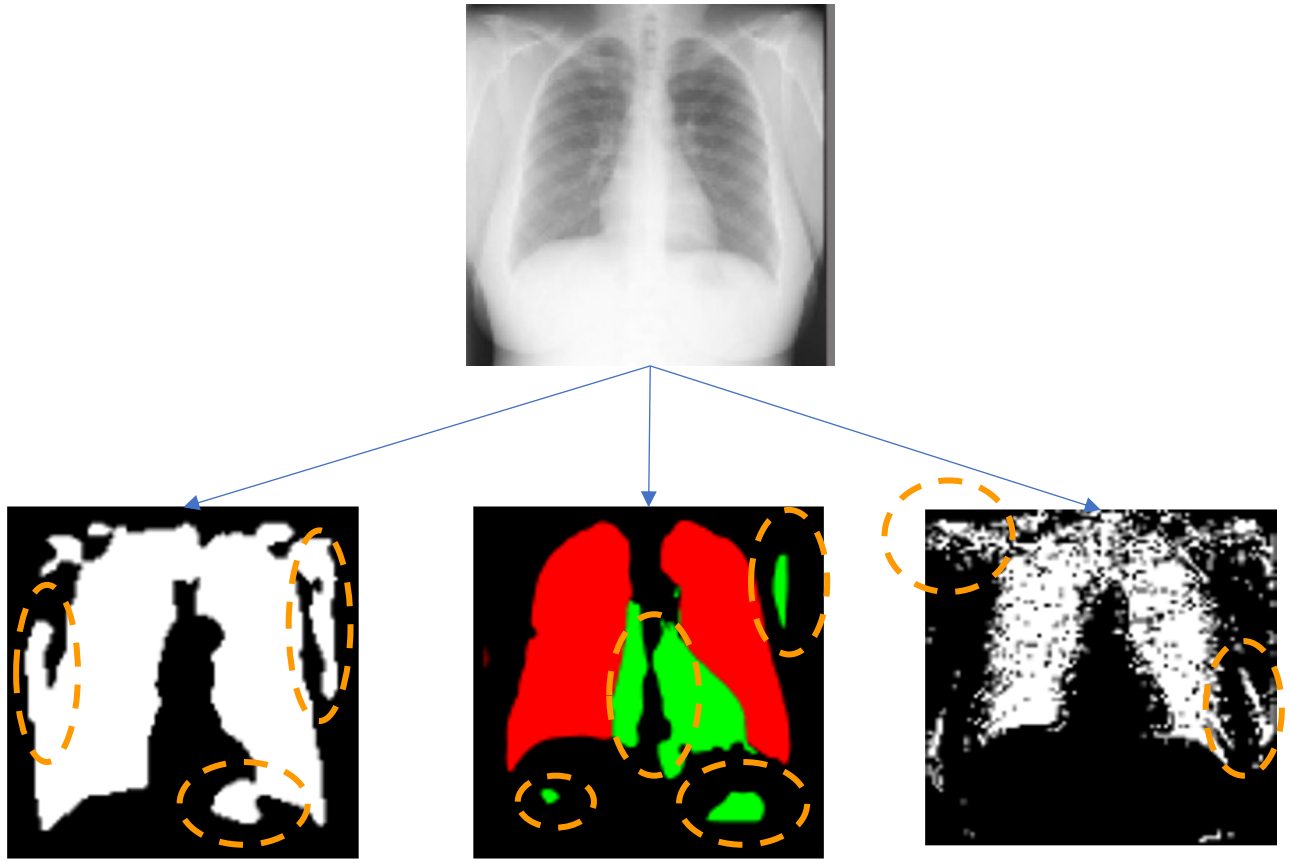


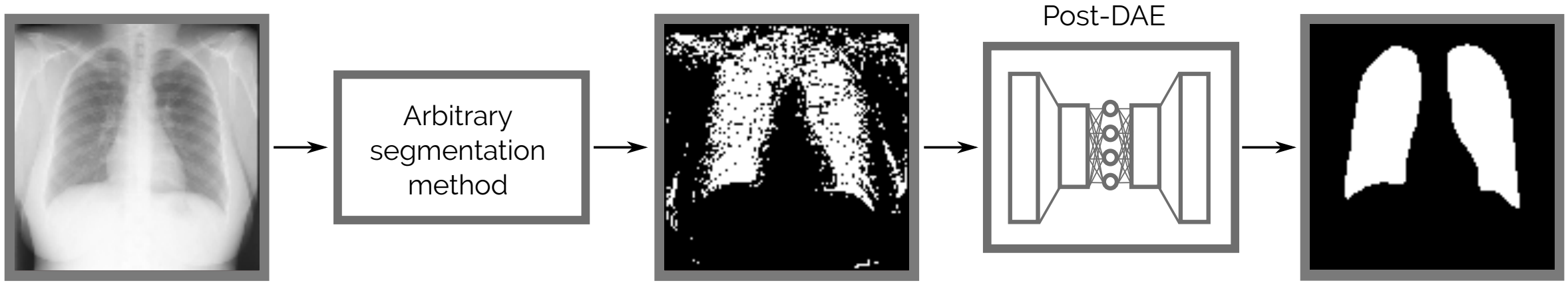
Imagen extraída de https://towardsdatascience.com/anomaly-detection-with-autoencoder-b4cdce4866a6

Aplicación 3: Autoencoders como etapa de post-procesamiento



Segmentación de estructuras anatómicas

Aplicación 3: Autoencoders como etapa de post-procesamiento



Larrazabal, Martinez & Ferrante (MICCAI 2019)

Autoencoders como etapa de post-procesamiento

Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders

Agostina J. Larrazabal, Cesar Martinez, Enzo Ferrante

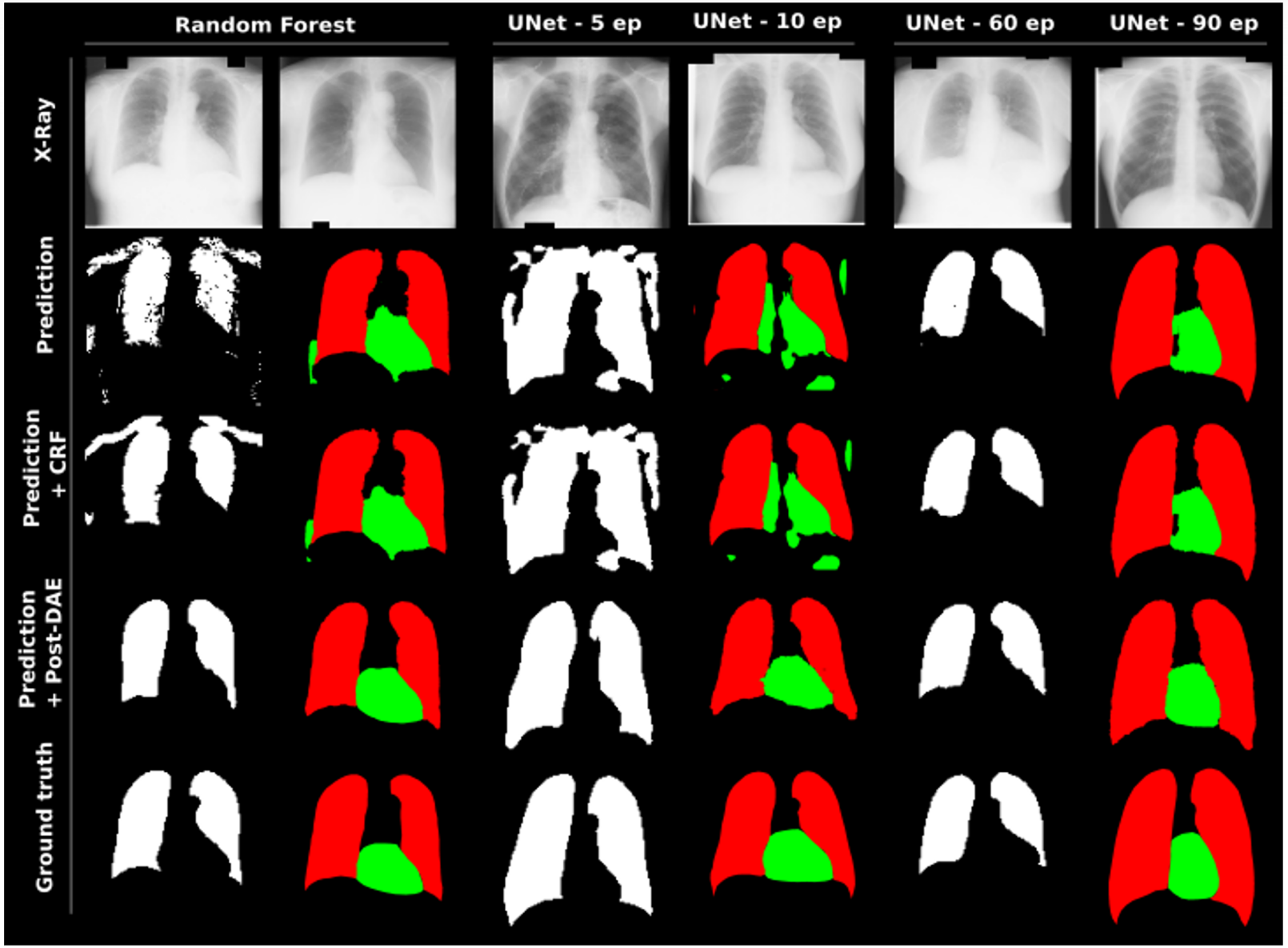
Research institute for signals, systems and computational intelligence, sinc(i),
FICH-UNL / CONICET, Santa Fe, Argentina

Abstract. Deep convolutional neural networks (CNN) proved to be highly accurate to perform anatomical segmentation of medical images. However, some of the most popular CNN architectures for image segmentation still rely on post-processing strategies (e.g. Conditional Random Fields) to incorporate connectivity constraints into the resulting masks. These post-processing steps are based on the assumption that objects are usually continuous and therefore nearby pixels should be assigned the same object label. Even if it is a valid assumption in general, these methods do not offer a straightforward way to incorporate more complex priors like convexity or arbitrary shape restrictions.

In this work we propose Post-DAE, a post-processing method based on denoising autoencoders (DAE) trained using only segmentation masks. ... learn a low-dimensional space of anatomically plausible segmenta- ... processing step to impose shape constraints ... tation methods.

MICCAI 2019

Autoencoders como etapa de post-procesamiento

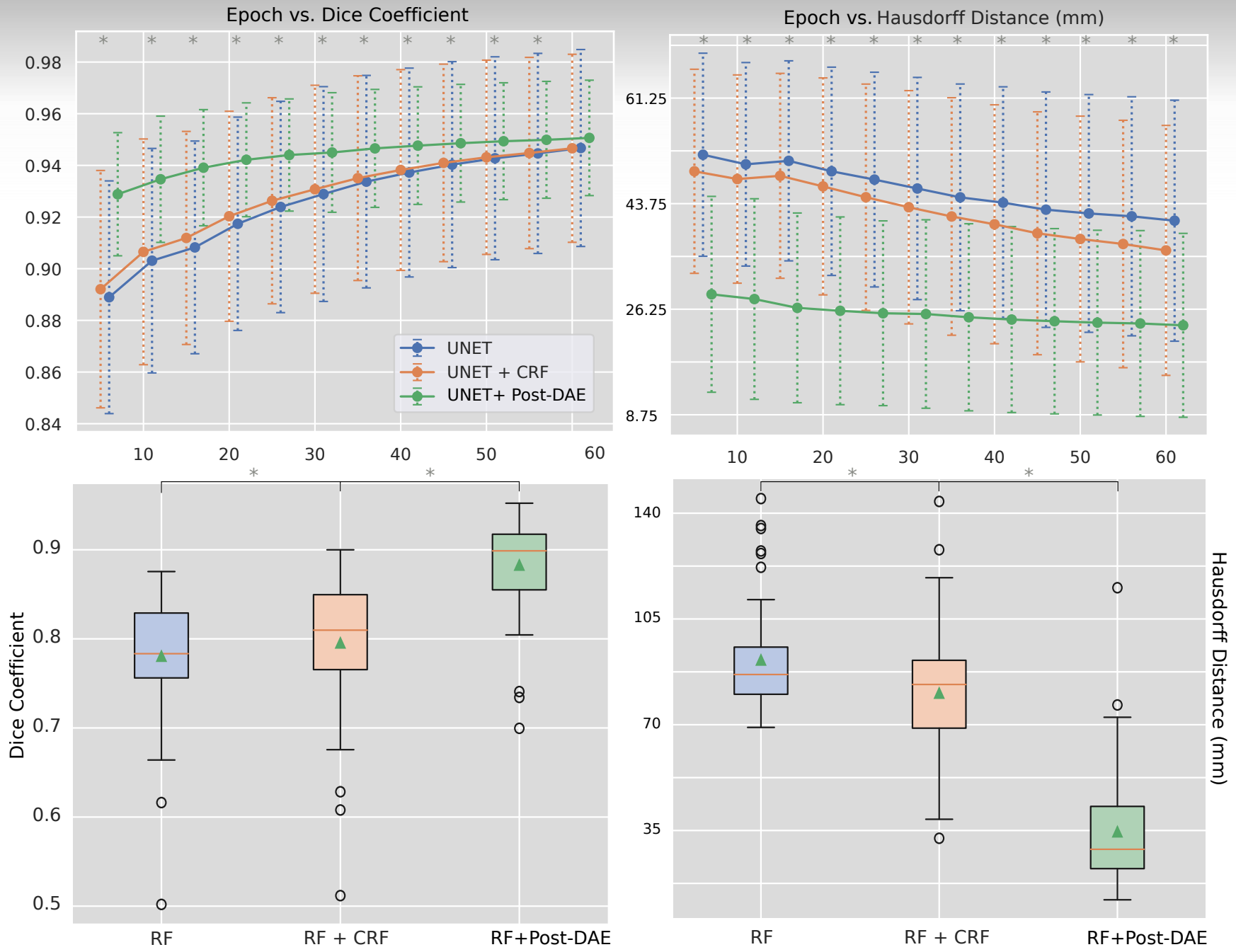


Larrazabal, Martinez & Ferrante (MICCAI 2019)

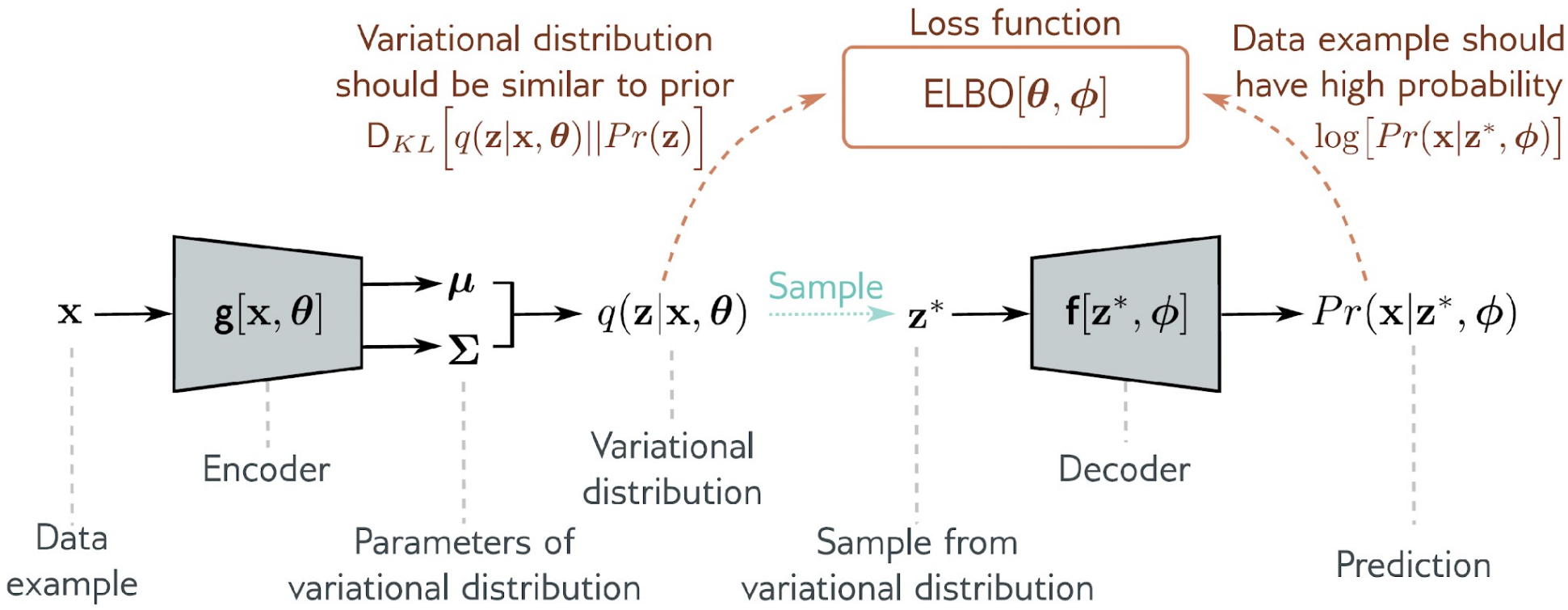
Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders

MICCAI 2019

Visualization of segmentation masks before and after post-processing using Post-DAE


Epoch vs. Dice Coefficient
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
10
20
30
40
50
60
UNET
UNET + CRF
UNET+ Post-DAE
Epoch vs. Hausdorff Distance (mm)
61.25
43.75
26.25
8.75
10
20
30
40
50
60
Dice Coefficient
0.9
0.8
0.7
0.6
0.5
RF
RF + CRF
RF+Post-DAE
140
105
70
35
Hausdorff Distance (mm)
RF
RF + CRF
RF+Post-DAE

Variational Autoencoders



Variational Autoencoders

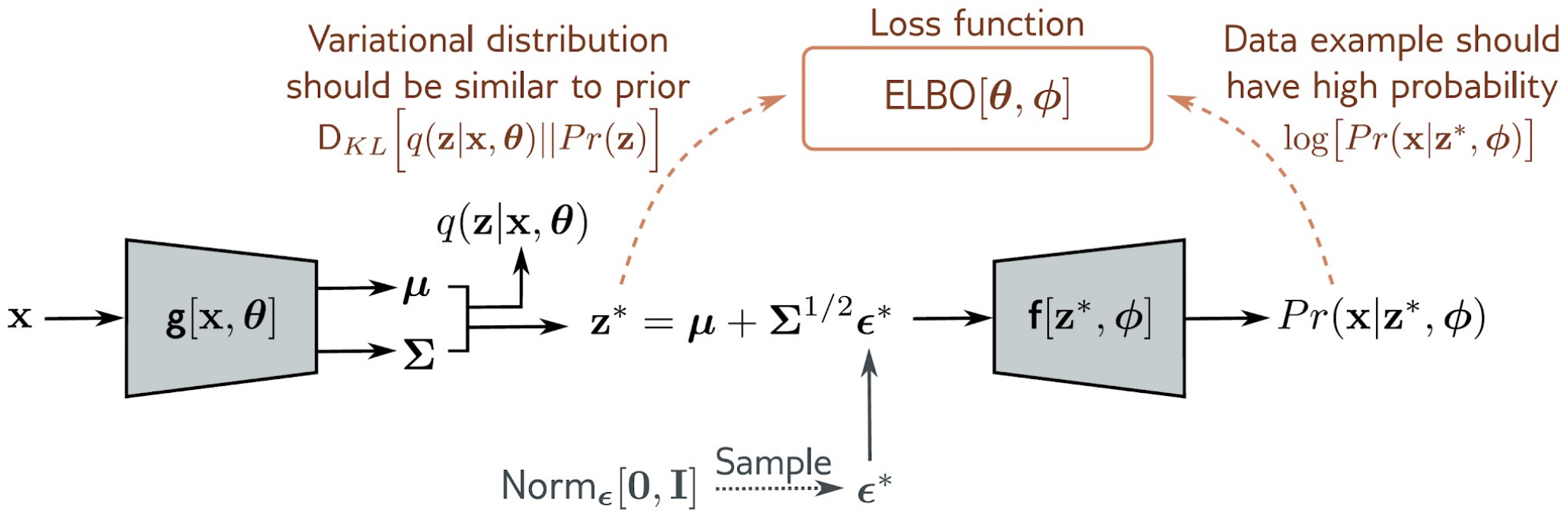
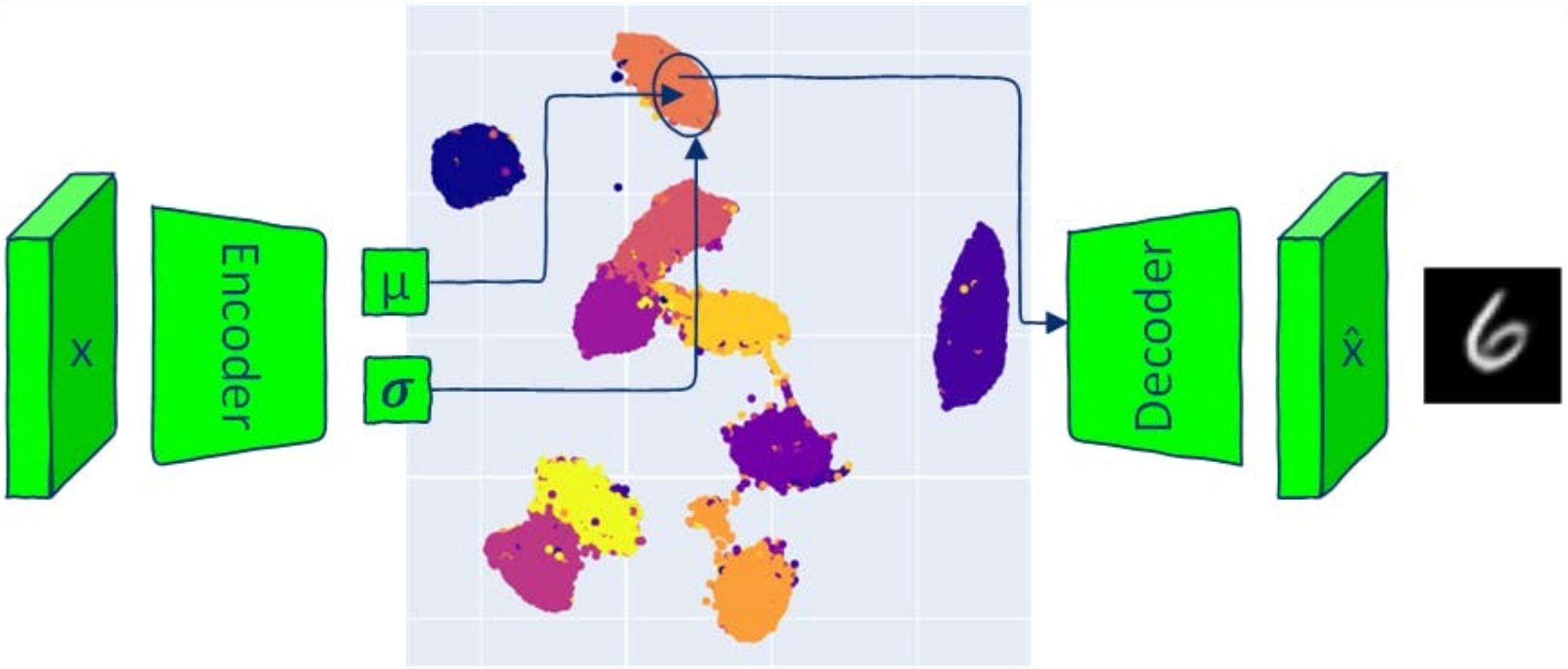


Figure 17.11 Reparameterization trick. With the original architecture (figure 17.9), we cannot easily backpropagate through the sampling step. The reparameterization trick removes the sampling step from the main pipeline; we draw from a standard normal and combine this with the predicted mean and covariance to get a sample from the variational distribution.

Sampling from a Variational Autoencoders



¿Cómo generalizar a múltiples dominios de datos?

Adaptación de dominio

Se produce un cambio de dominio en los datos de entrada,
pero la tarea que queremos resolver es la misma en ambos dominios

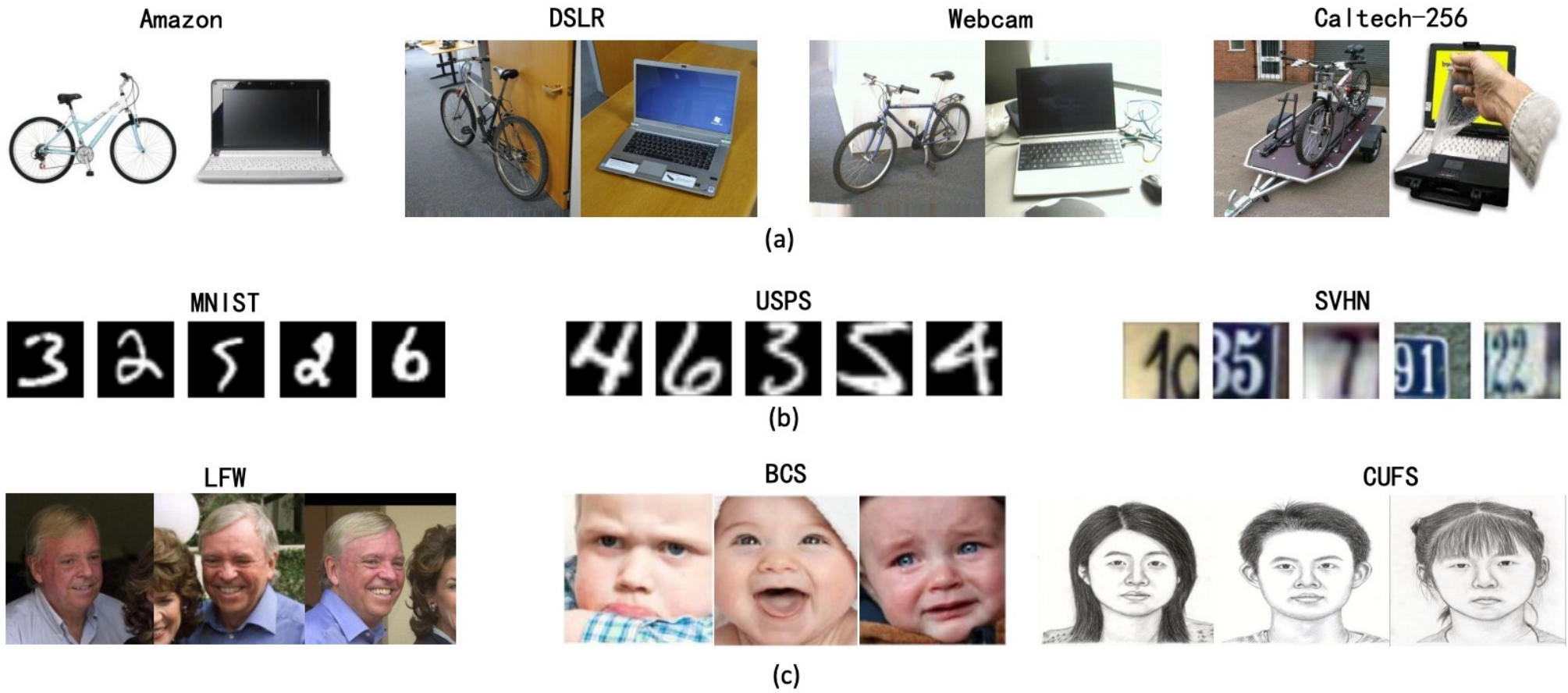


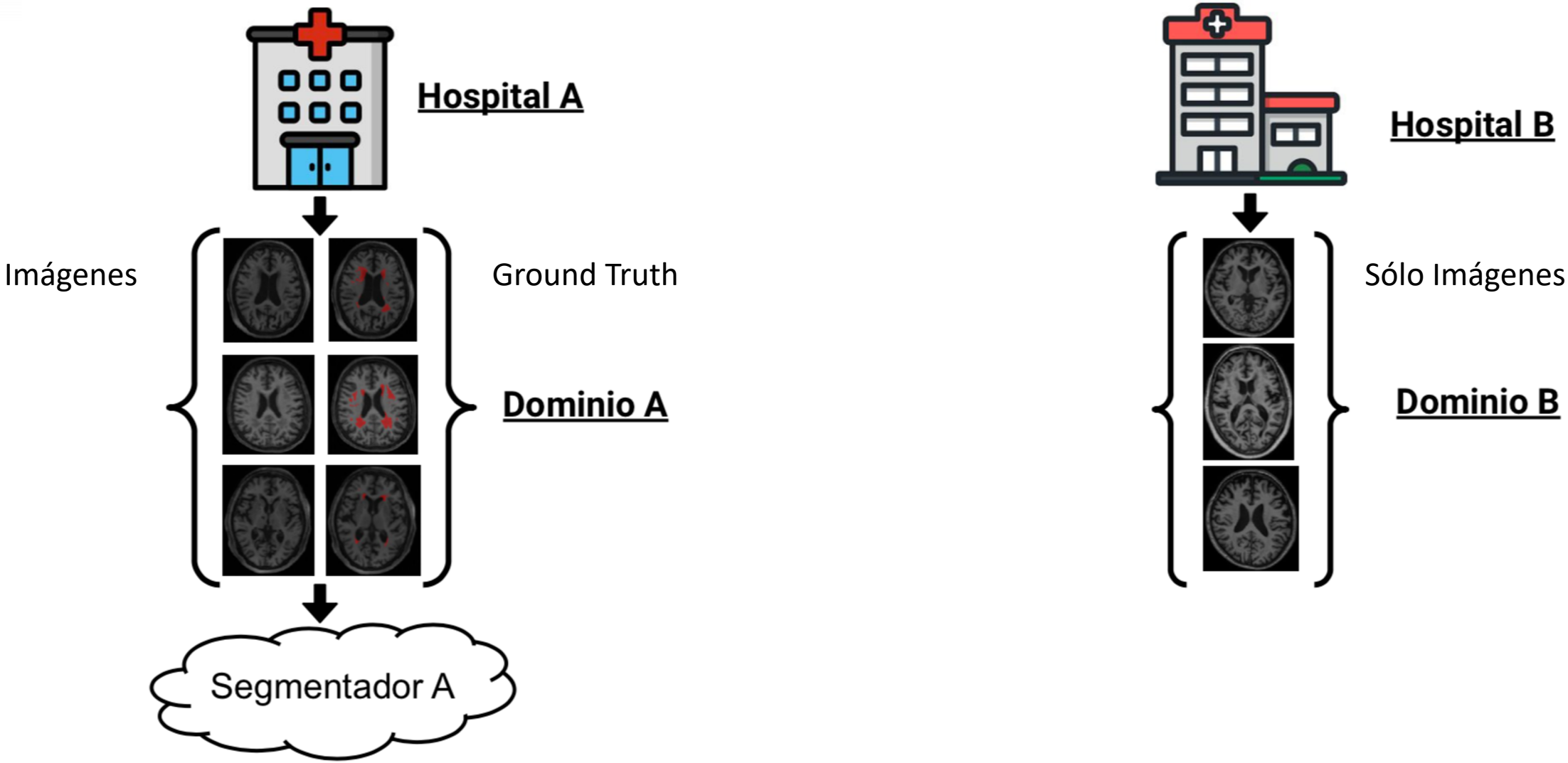
Imagen extraída de (Wang & Deng, 2018, Neurocomputing)

Adaptación de dominio

Problema multisitio

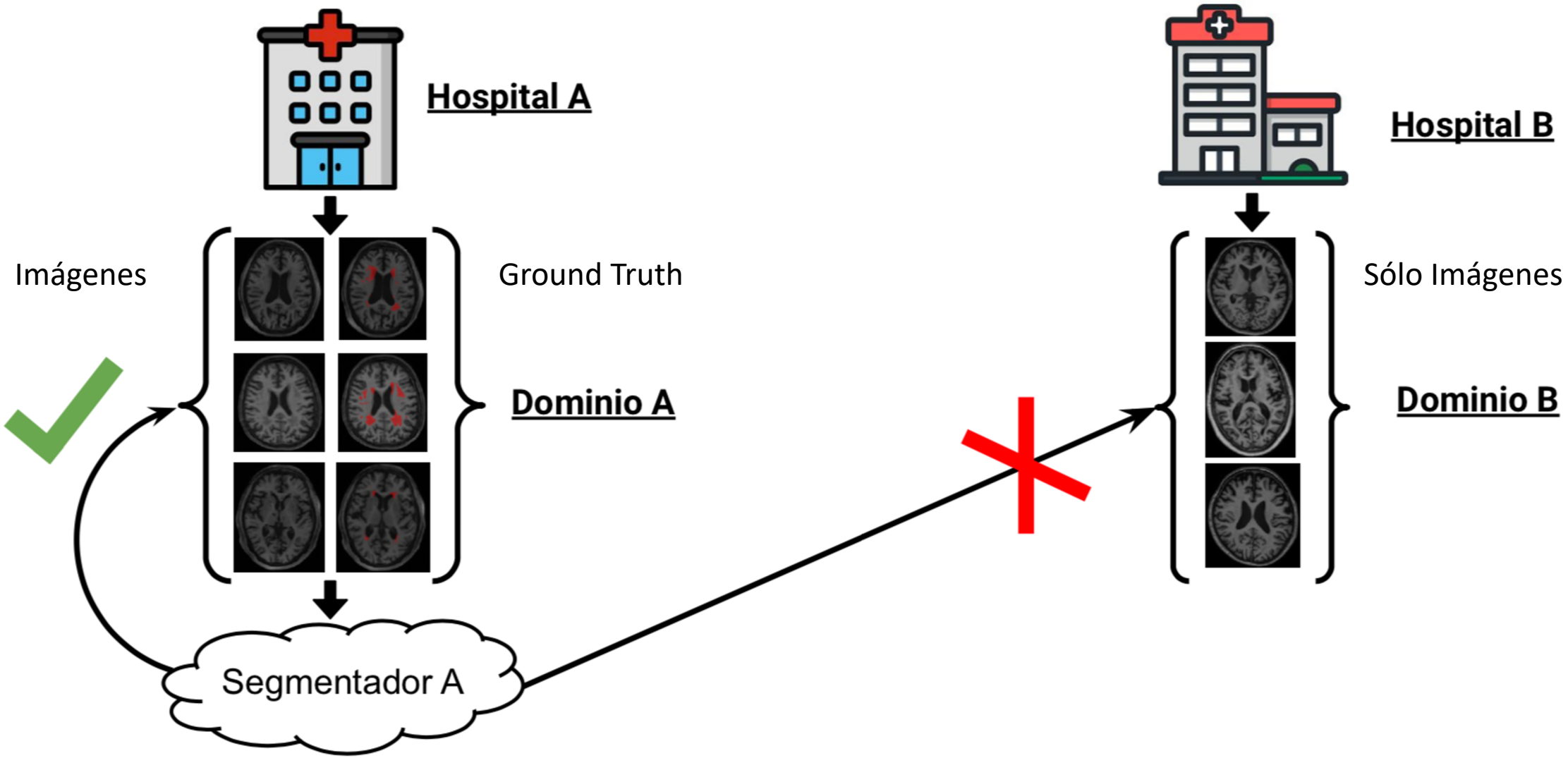
Adaptación de dominio

Problema multi-sitio



Adaptación de dominio

Problema multi-sitio



Si tengo etiquetas en todos los dominios

Solución 1: Entrenar con datos de todos los sitios (en caso de tener anotaciones para todos ellos)

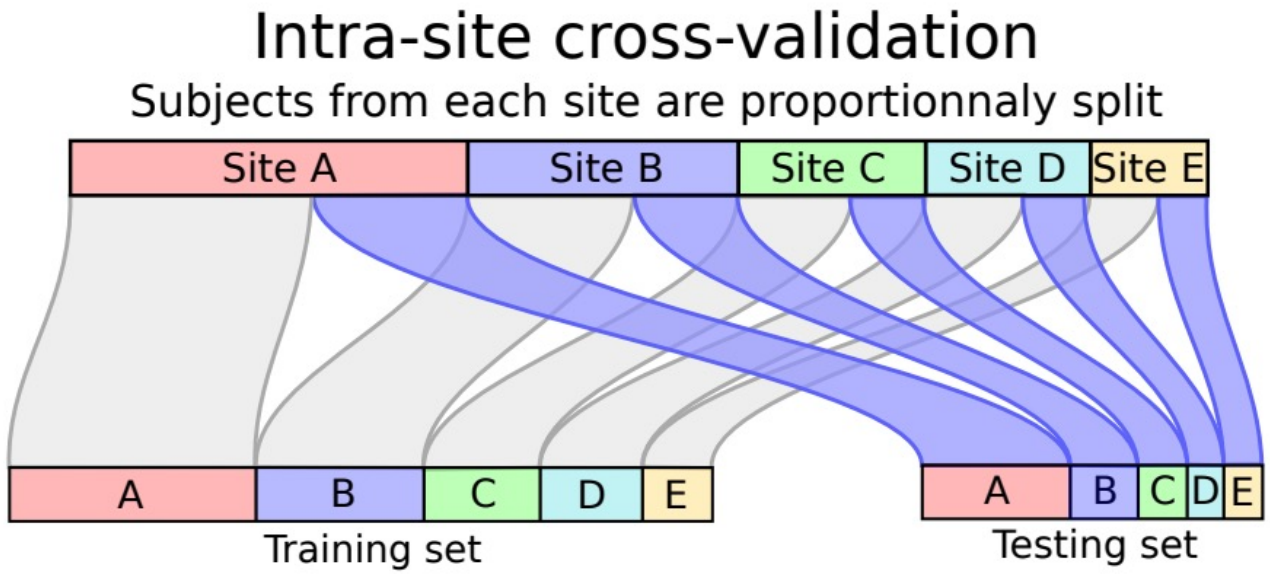


Imagen extraída de (Abraham et al, 2016, Neuroimage)

Si tengo etiquetas en todos los dominios

Solución 2: Utilizo transferencia de aprendizaje por medio de fine-tuning

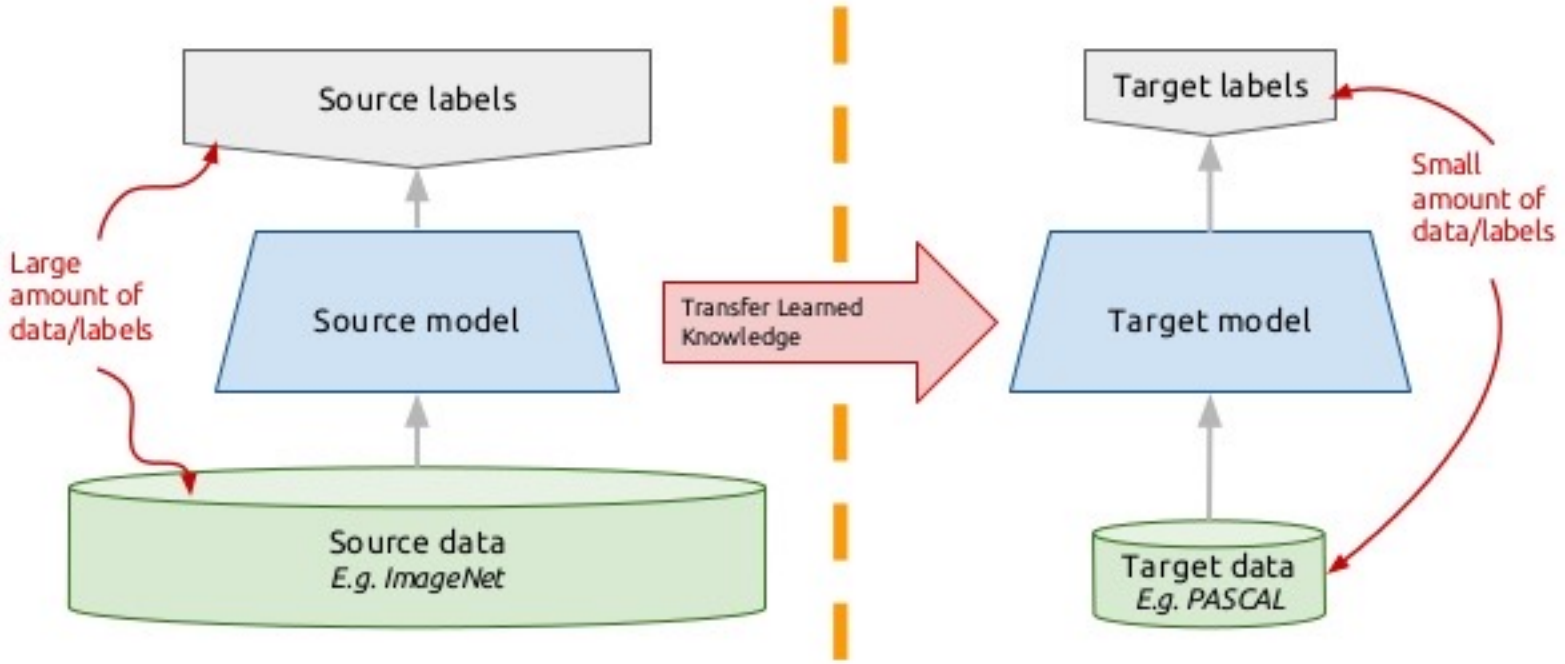


Imagen extraída de https://medium.com/the-official-integrate-ai-blog/transfer-learning-explained-7d275c1e34e2

Si NO tengo etiquetas en el dominio target

Adaptación de dominio NO supervisada

Adaptación de dominio no supervisada: ideas para resolverlo

1. Mapear la distribución de los datos del Sitio 1 al Sitio 2, y re-entrenar nuestro modelo

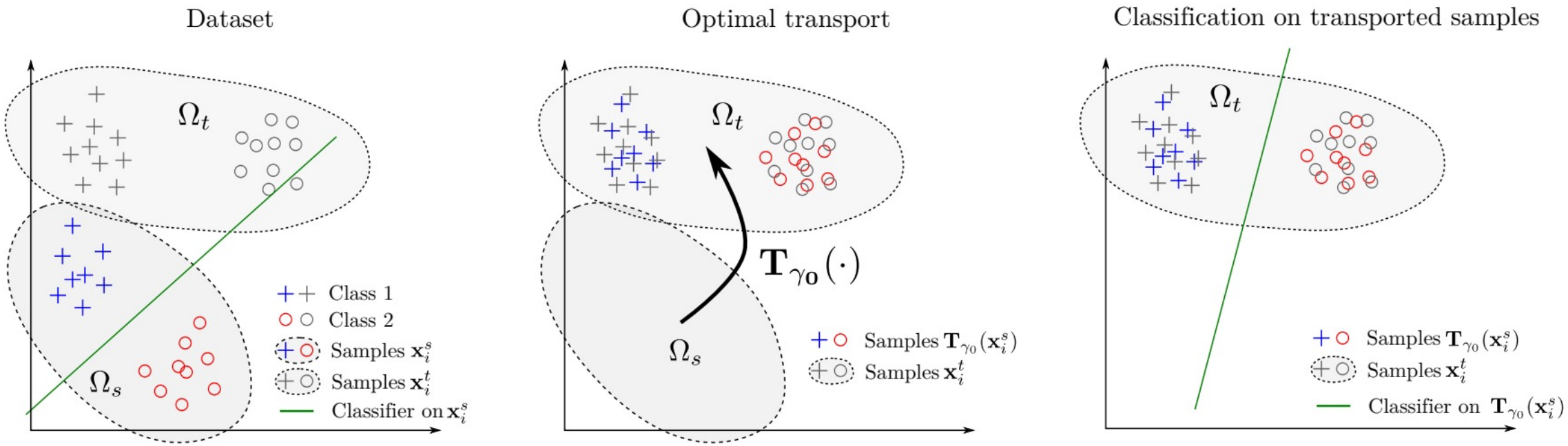


Fig. 1: Illustration of the proposed approach for domain adaptation. (left) dataset for training, *i.e.* source domain, and testing, *i.e.* target domain. Note that a classifier estimated on the training examples clearly does not fit the target data. (middle) a data dependent transportation map $\mathbf{T}_{\gamma 0}$ is estimated and used to transport the training samples onto the target domain. Note that this transformation is usually not linear. (right) the transported labeled samples are used for estimating a classifier in the target domain.

Courty et al, IEEE T-PAMI, 2016

Adaptación de dominio no supervisada: ideas para resolverlo

2. Aprendizaje de features invariantes al dominio

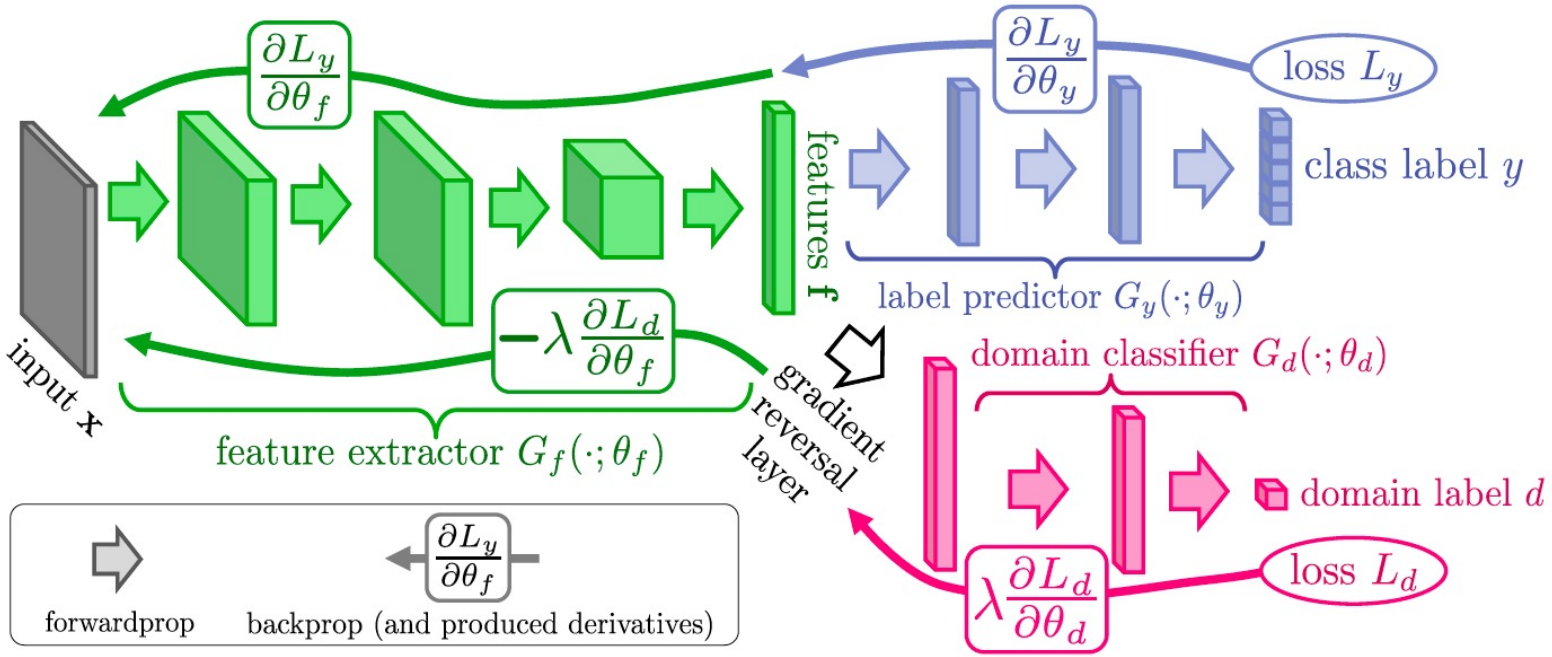


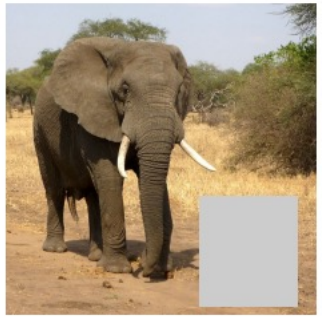
Figure 1: The **proposed architecture** includes a deep *feature extractor* (green) and a deep *label predictor* (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a *domain classifier* (red) connected to the feature extractor via a *gradient reversal layer* that multiplies the gradient by a certain negative constant during the backpropagation-based training. Otherwise, the training proceeds standardly and minimizes the label prediction loss (for source examples) and the domain classification loss (for all

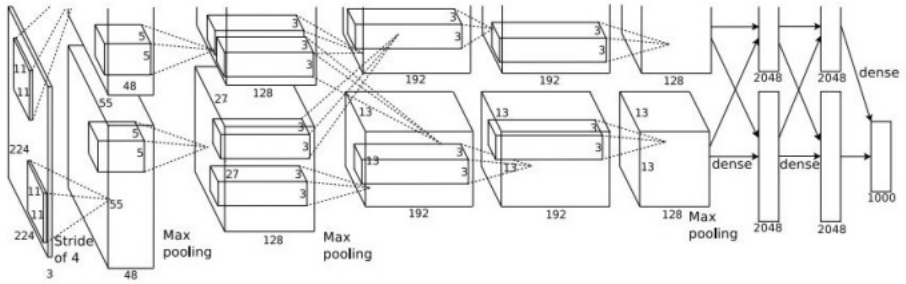
Ganin et al, JMLR, 2016

¿Cómo interpretar los modelos entrenados?

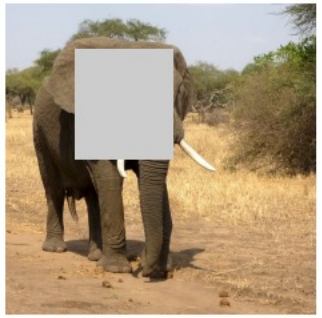
Mapas de saliencia por medio de oclusión

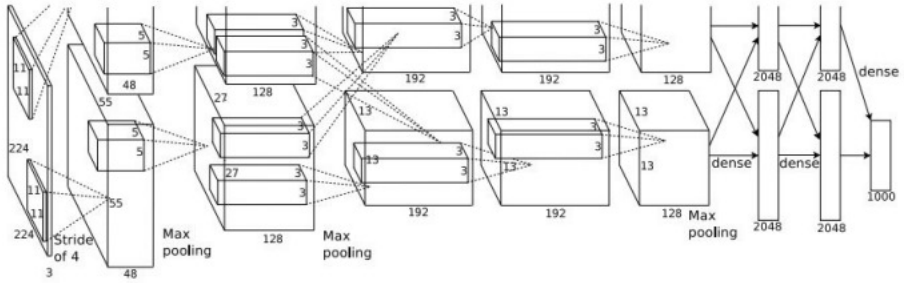
Mapas de saliencia por medio de oclusión





P(elephant) = 0.95



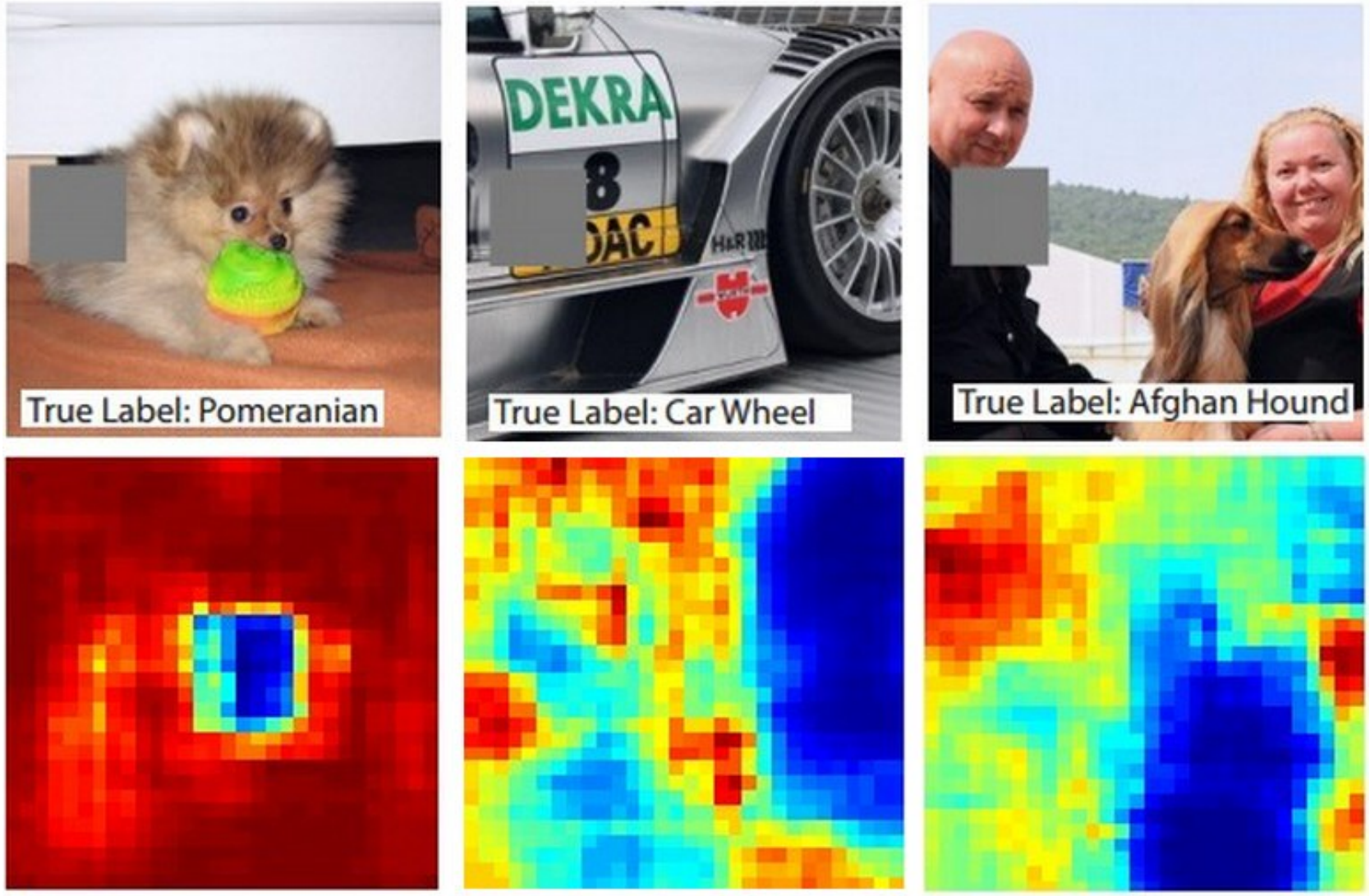


P(elephant) = 0.75

Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Basado en: http://cs231n.stanford.edu/

Mapas de saliencia por medio de oclusión



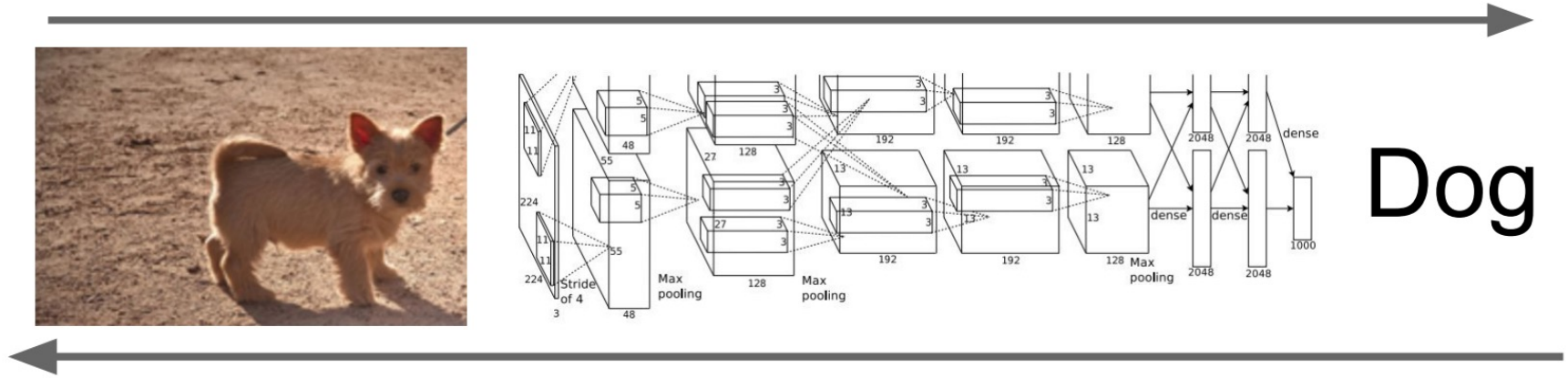
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

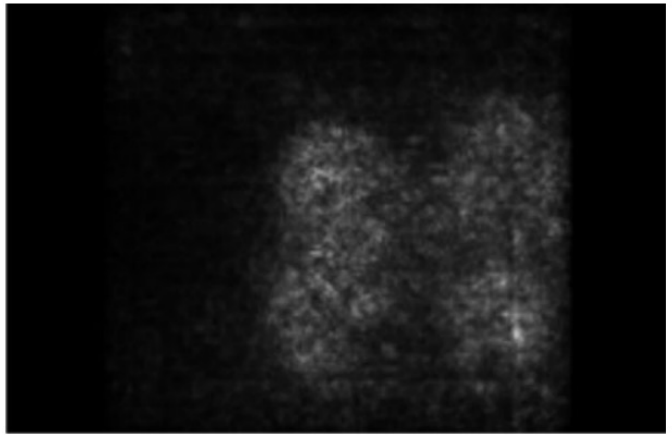
Basado en: http://cs231n.stanford.edu/

Mapas de saliencia por retropropagación

Mapas de saliencia por retropropagación

Computamos el gradiente del score de la clase de interés respecto a los píxeles de entrada





Tomamos su valor absoluto
y el máximo por los canales RGB

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).

Basado en: http://cs231n.stanford.edu/

Class activation maps

Class activation maps

Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

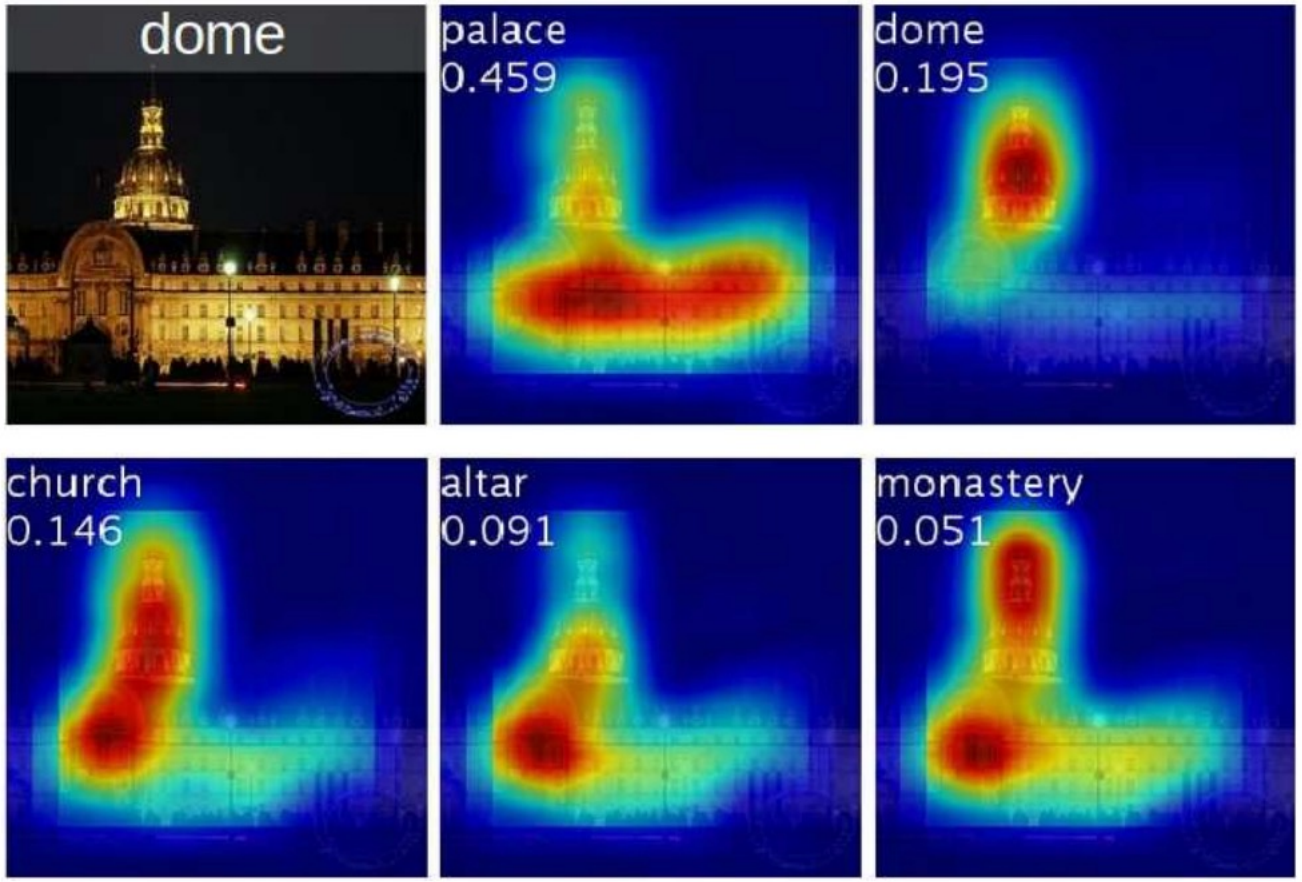
Abstract

In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network (CNN) to have remarkable localization ability despite being trained on image-level labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that exposes the implicit attention of CNNs on an image. Despite the apparent simplicity of global average pooling, we are able to achieve 37.1% top-5 error for object localization on ILSVRC 2014 without training on any bounding box annotation.We demonstrate in a variety of experiments that our network is able to localize the discriminative image regions despite just being trained for solving classification task[1].

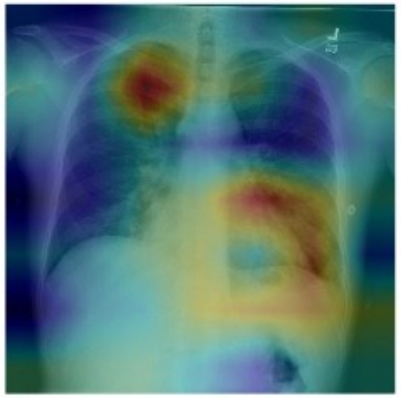
Figure 1. A simple modification of the global average pooling layer combined with our class activation mapping (CAM) technique allows the classification-trained CNN to both classify the image and localize class-specific image regions in a single forward-pass e.g., the toothbrush for *brushing teeth* and the chainsaw for *cutting trees*.

for a wide variety of tasks, even those that the network

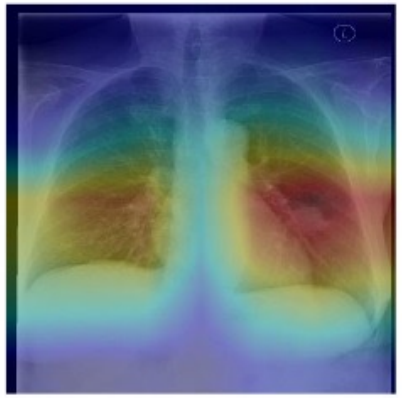
Class activation maps



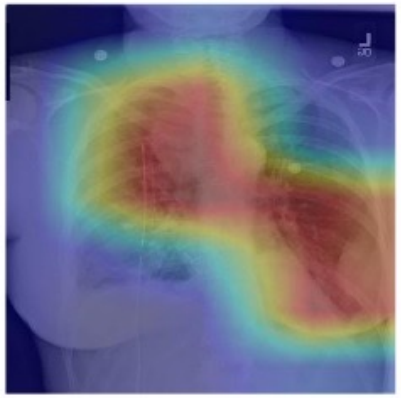
Class activation maps



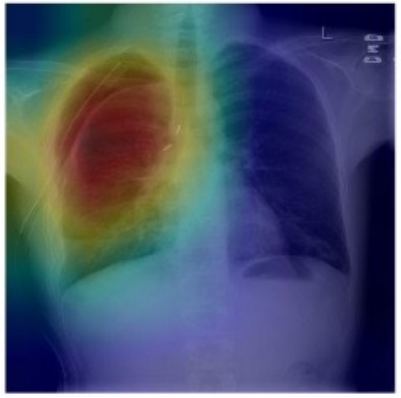
(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.



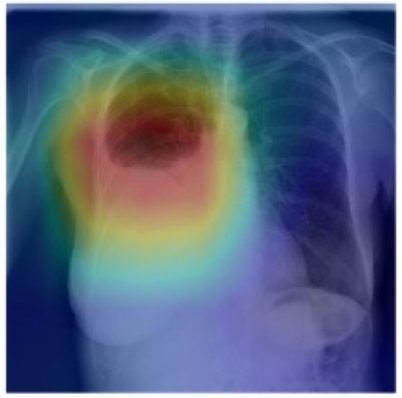
(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.



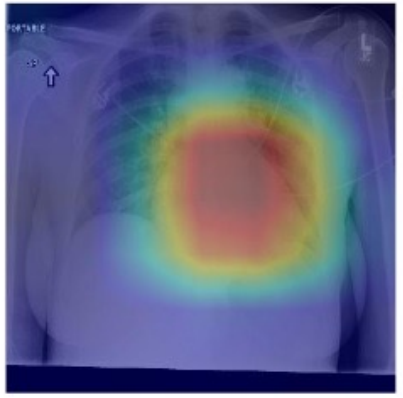
(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.



(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).

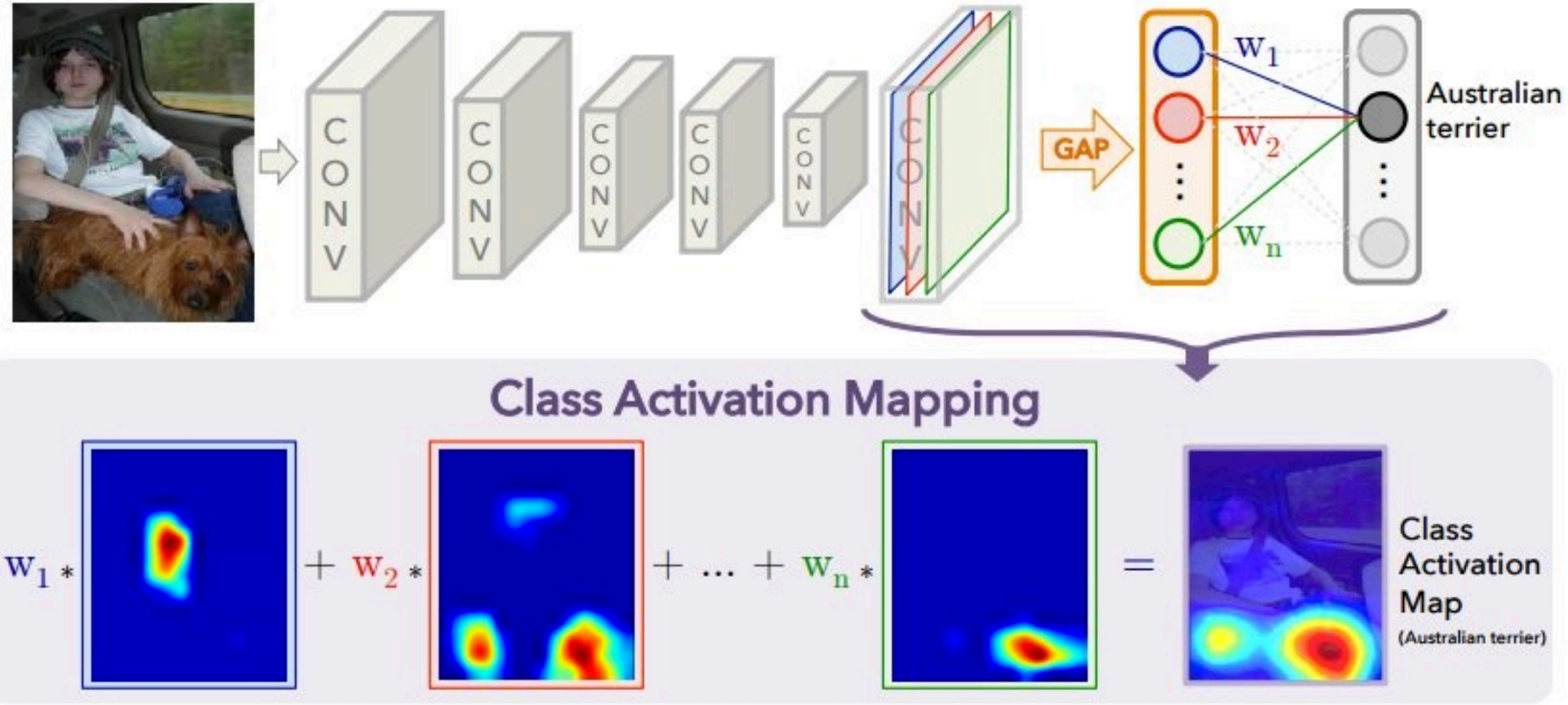


(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.



(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Class activation maps

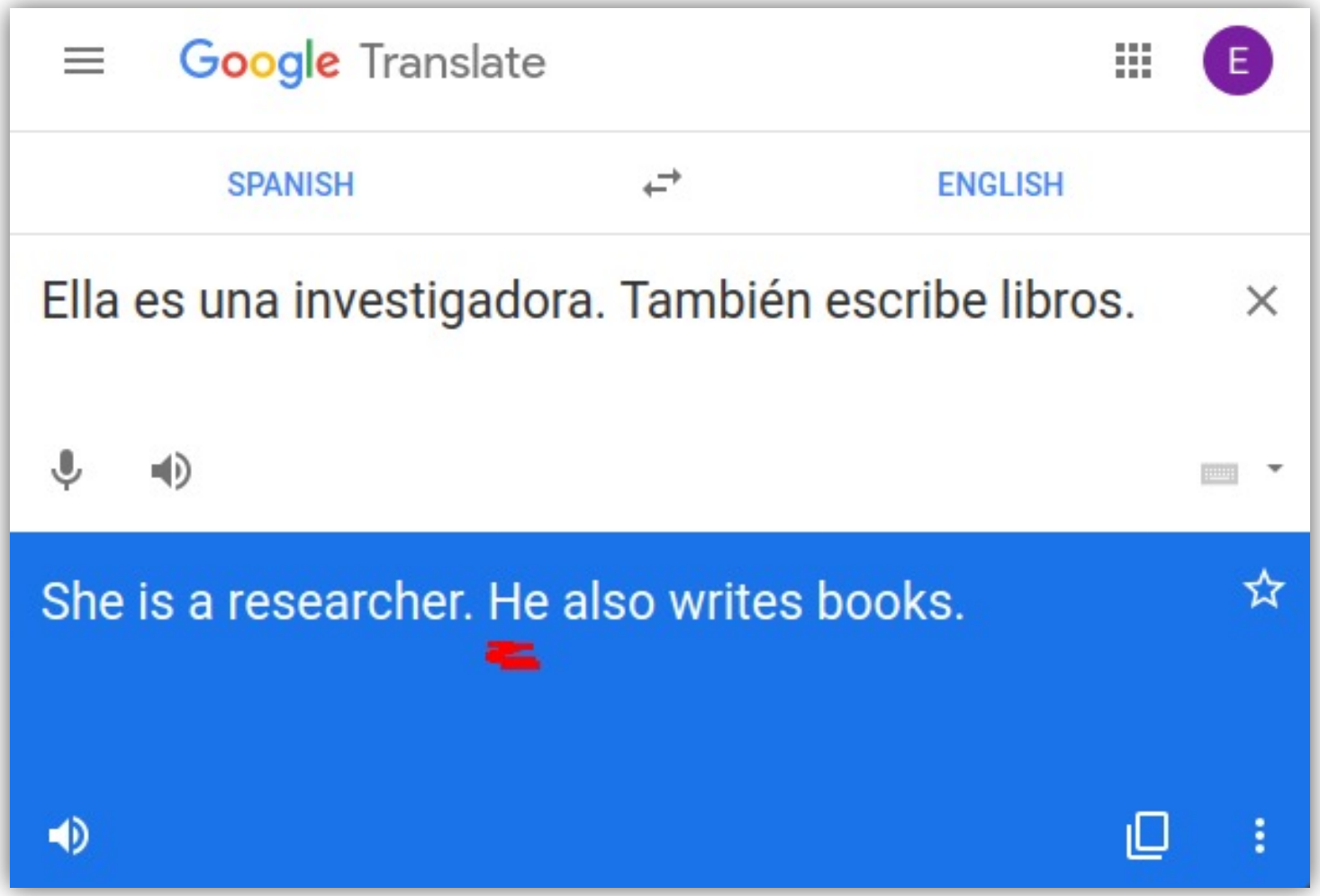


¿Cómo evitar el sesgo en los modelos de deep learning?

Sesgo en los modelos de IA

Sesgo en los modelos de IA

Sesgo de género en la traducción de textos



Sesgo en los modelos de IA

Novia o disfraz?

Sesgo en los modelos de IA

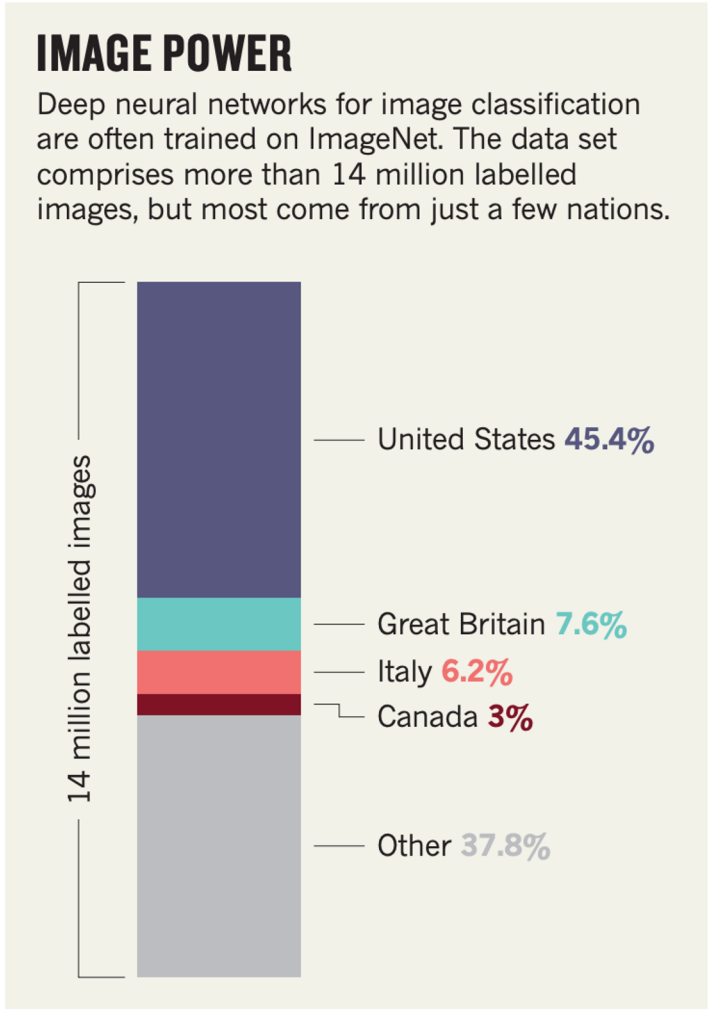
Sesgo en el reconocimiento de rostros

Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively are found to suffer from achieving much lower accuracy on females with darker skin color(see Fig4, Buolamwini and Gebru, 2018).

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Sesgo en los modelos de IA

Datos sesgados



“Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations.”

Fuente: Zou, James, and Londa Schiebinger. "AI can be sexist and racist—it’s time to make it fair." Nature, (2018): 324.

Sesgo en los modelos de IA

Datos sesgados

COMMENT · 18 JULY 2018

AI can be sexist and racist — it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

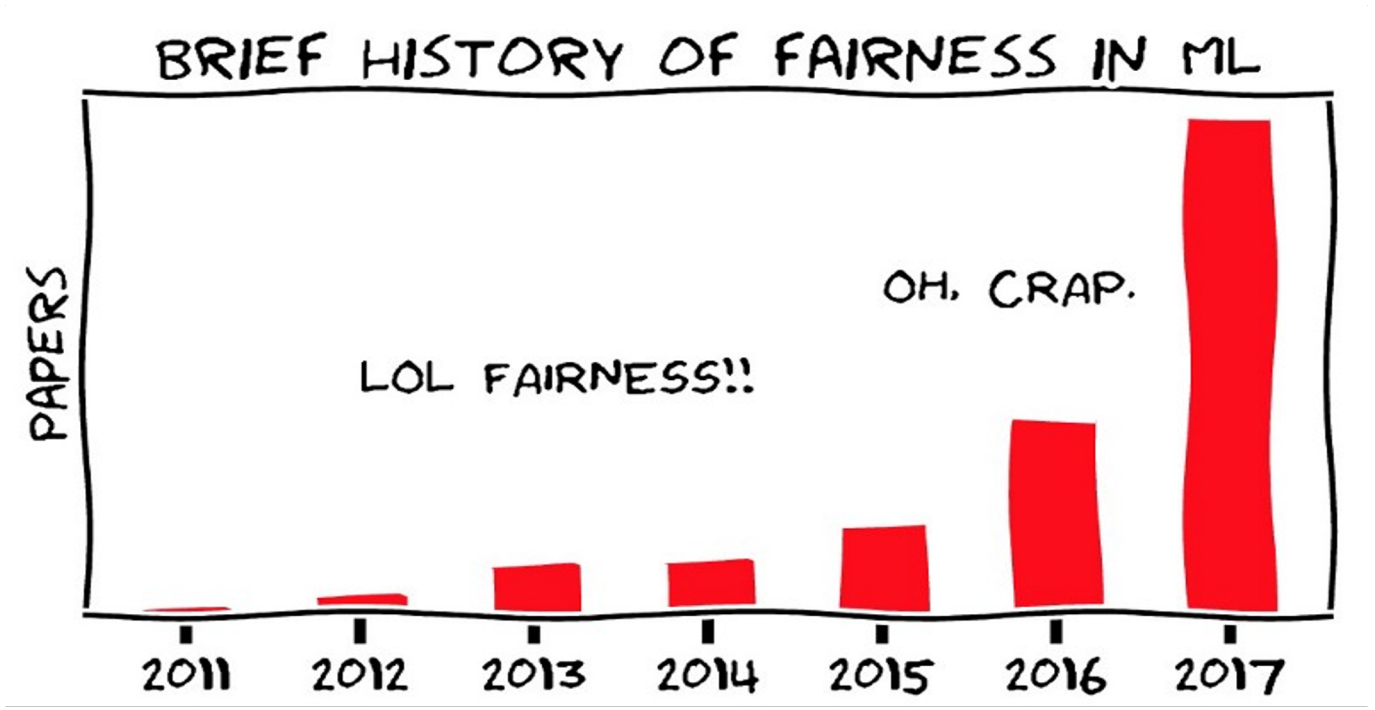
James Zou & Londa Schiebinger

Atributos protegidos

Son variables contra las cuales queremos proteger el sesgo algorítmico como género, etnia, país de procedencia, edad, etc.

Ej: En un algoritmo que recomienda si deberíamos o no otorgarle un crédito a una persona, el género podría ser una variable protegida.

Fairness en AI



Definición formal de ”fairness” en IA

Paridad demográfica

- La clasificación tiene que ser independiente del atributo protegido.

- A = Atributo protegido
- $\hat{Y}$ = Predicción
- Y = Ground Truth

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$$

- **Ej:** La probablidad de darle un crédito a una persona de género femenino es la misma que a una persona de género masculino

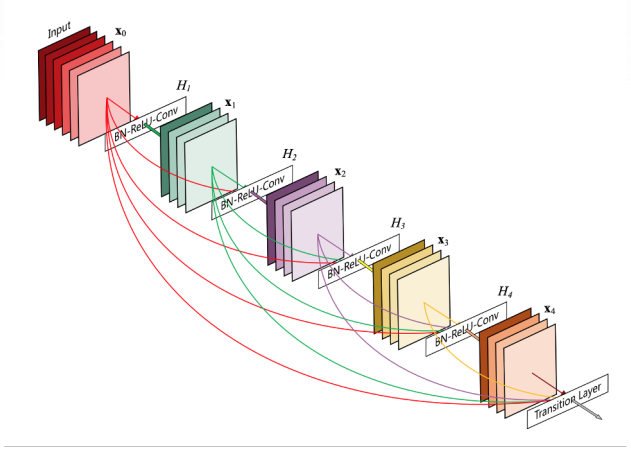
Para profundizar: ”Equality of Opportunity in Supervised Learning” Hardt et al, 2016. NIPS

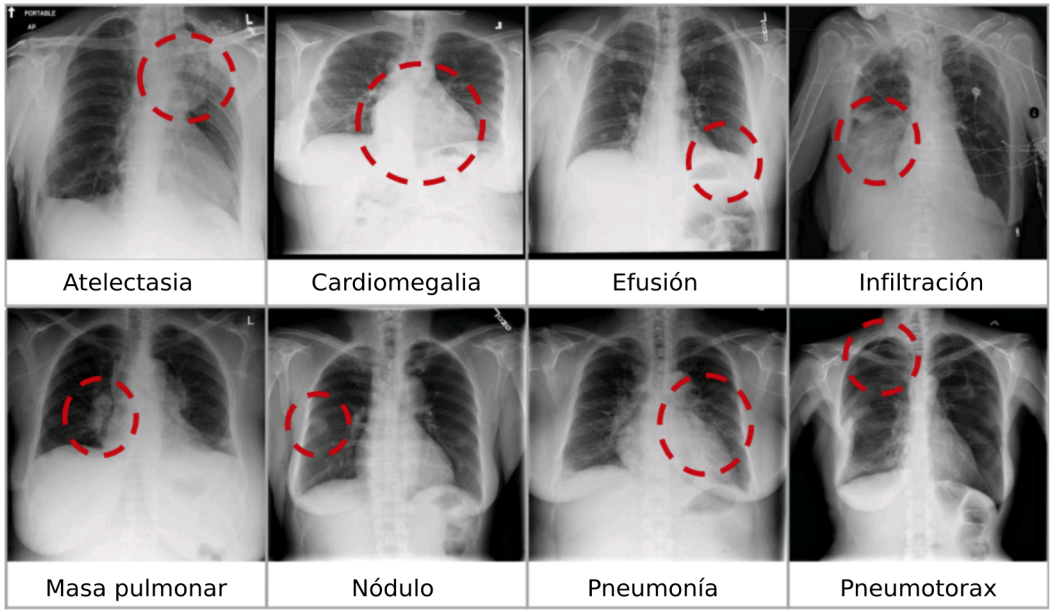
Qué pasa con el balance de datos?

Experimento

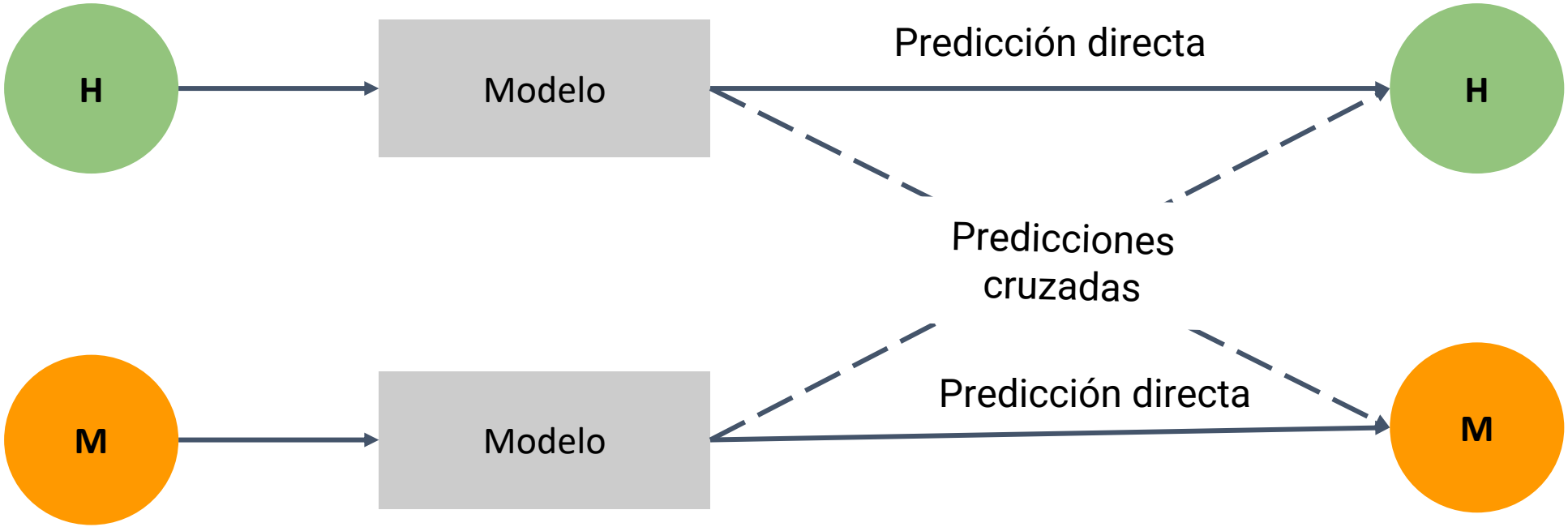
- Utilizamos imágenes de una **base datos pública** de 112.000 radiografías torácicas [Wang 2017] con diagnósticos e información demográfica disponible

- Generamos **particiones de entrenamiento independientes** para género masculino y femenino, con igual cantidad de pacientes por enfermedad y género (48568).

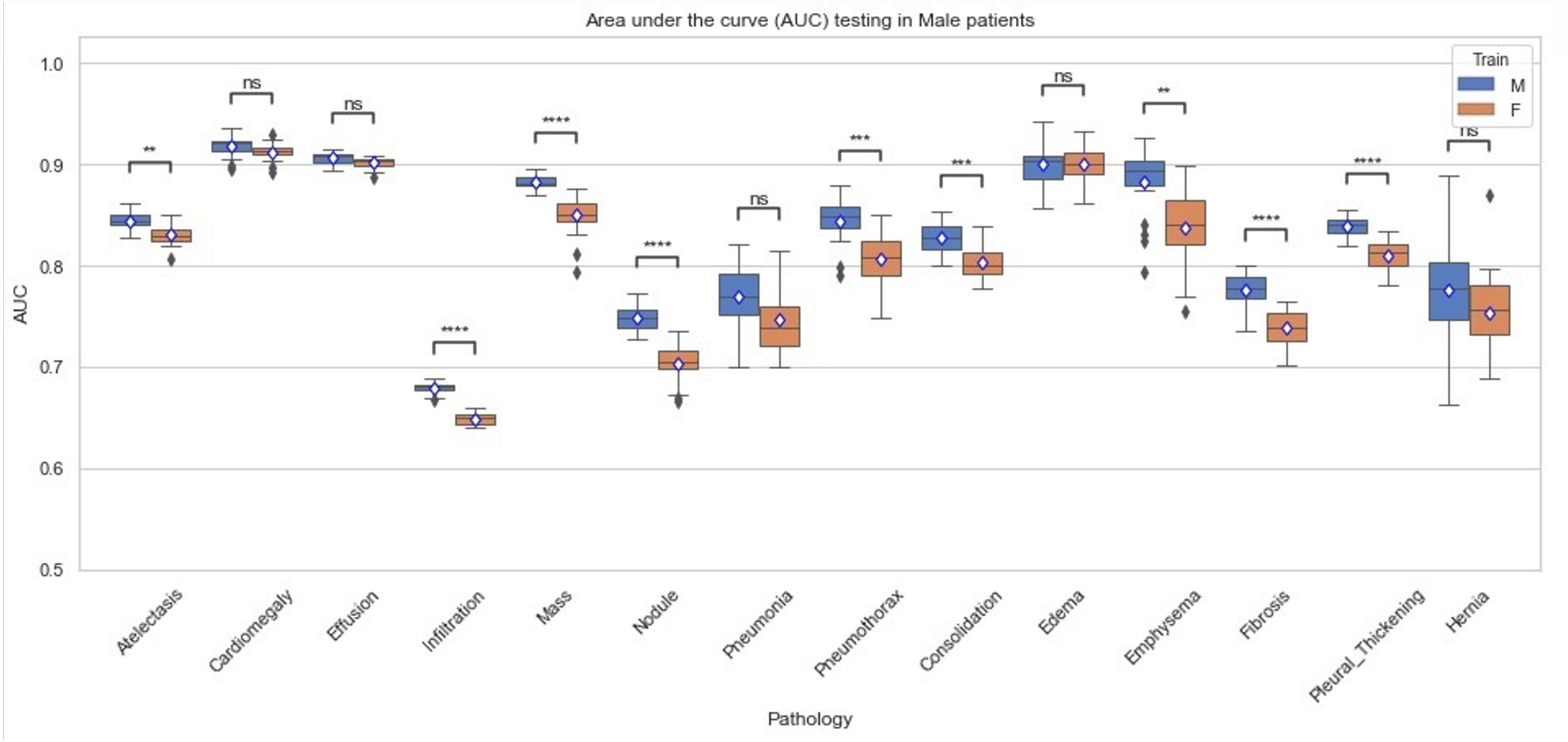




Experimento con desbalances extremos



Experimento con desbalances extremos



PNAS

Home Articles

NEW RESEARCH IN Phy

BRIEF REPORT

Gender imba
datasets pro
computer-a

Agostina J. Larrazab
Enzo Ferrante

PNAS first published May 26, 2020 https://d

Edited by David L. Donoho, Stanford Univers
30, 2019)

SCIENTIFIC AMERICAN®

Subscribe

POLICY | OPINION

Health Care AI Systems Are Biased

We need more diverse data to avoid perpetuating inequality in medicine

By Amit Kaushal, Russ Altman, Curt Langlotz on November 17, 2020

El problema no son sólo los datos

The Guardian

Artificial intelligence (AI)

'Disastrous' lack of diversity in AI industry perpetuates bias, study finds

Report says an overwhelmingly white and male field has reached 'a moment of reckoning' over discriminatory systems

Kari Paul *in San Francisco*

Wed 17 Apr 2019 01.47 BST

▲ Biased AI systems can be largely attributed to the lack of diversity among those who design and build them, the report said. Photograph: Jens Schlüter/EPA

Lack of diversity in the artificial intelligence field has reached "a moment of reckoning", according to new findings published by a New York University research center. A "diversity disaster" has contributed to flawed systems that perpetuate gender and racial biases found the survey, published by the AI Now Institute, of more than 150 studies and reports.

Last week, AI Now, a research group at New York University, released a study about A.I.'s diversity crisis. The report said that a lack of diversity among the people who create artificial intelligence and in the data they use to train it has created huge shortcomings in the technology.

For example, 80% of university professors who specialize in A.I. are men, the report said. Meanwhile, at leading A.I. companies like Facebook, women comprise only 15% of the A.I. research staff while at Google, women account for only 10%.

Furthermore, Timnit Gebru, who is an A.I. researcher at Google, is cited in the report as saying "she was one of six black people—out of 8,500 attendees" at a leading A.I. conference in 2016.

TECH • ARTIFICIAL INTELLIGENCE

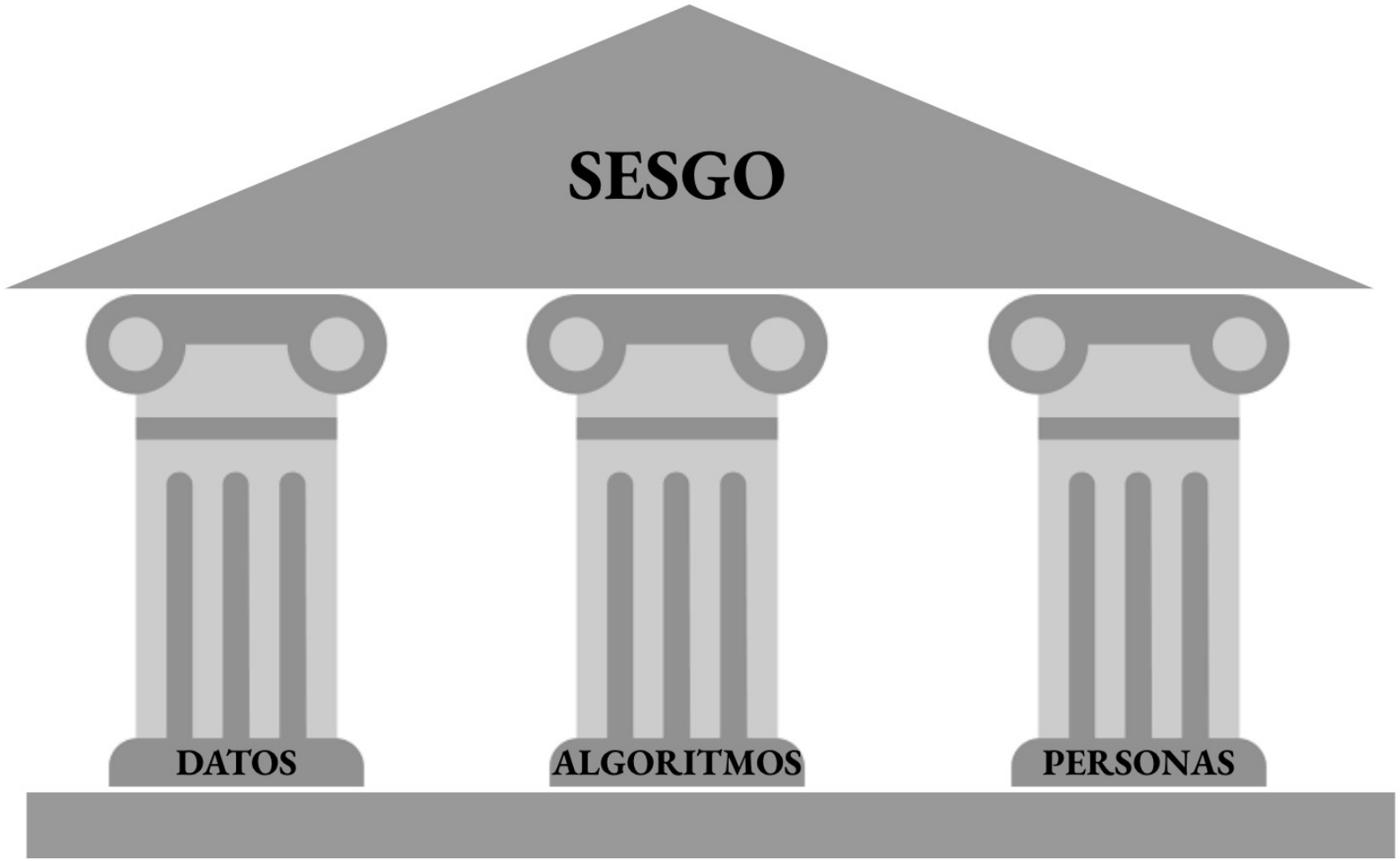
Eye on A.I.— How to Fix Artificial Intelligence's Diversity Crisis

By Jonathan Vanian April 23, 2019

Documental recomendado

Coded Bias (Prejuicio Cifrado, 2020)

Link → https://www.netflix.com/title/81328723


SESGO
DATOS
ALGORITMOS
PERSONAS

Clase 6

Problemas abiertos en DL y ML

Enzo Ferrante

eferrante@sinc.unl.edu.ar

@enzoferrante