# Machine Learning
# Introduction

Alvaro Soto

Computer Science Department (DCC), PUC

# Machine Learning

- Machine Learning is part of a broader field known as Artificial Intelligence.
- What is Artificial Intelligence?
  - Easy part: Artificial.
  - Hard part: Intelligence.

# Biological World




# Artificial World


## Artificial Intelligence

Study of computational models that allows machines to perceive, reason, and act with great flexibility.

## Artificial Intelligence

Study of computational models that allows machines to acquire certain level of understanding of its world.

# Some History

- 60's : Era of optimism or naiveness.
  - Deductive learning.
- 70's - 80's: Era of pessimism.
- 90's: Era of reborn.
- 2000's: Era of Machine Learning.
  - Inductive learning.

# Inductive learning

This bird can fly

This bird can fly

This bird can fly

This bird can fly

Can this bird fly ?

# Inductive learning

This bird can fly

This bird can fly

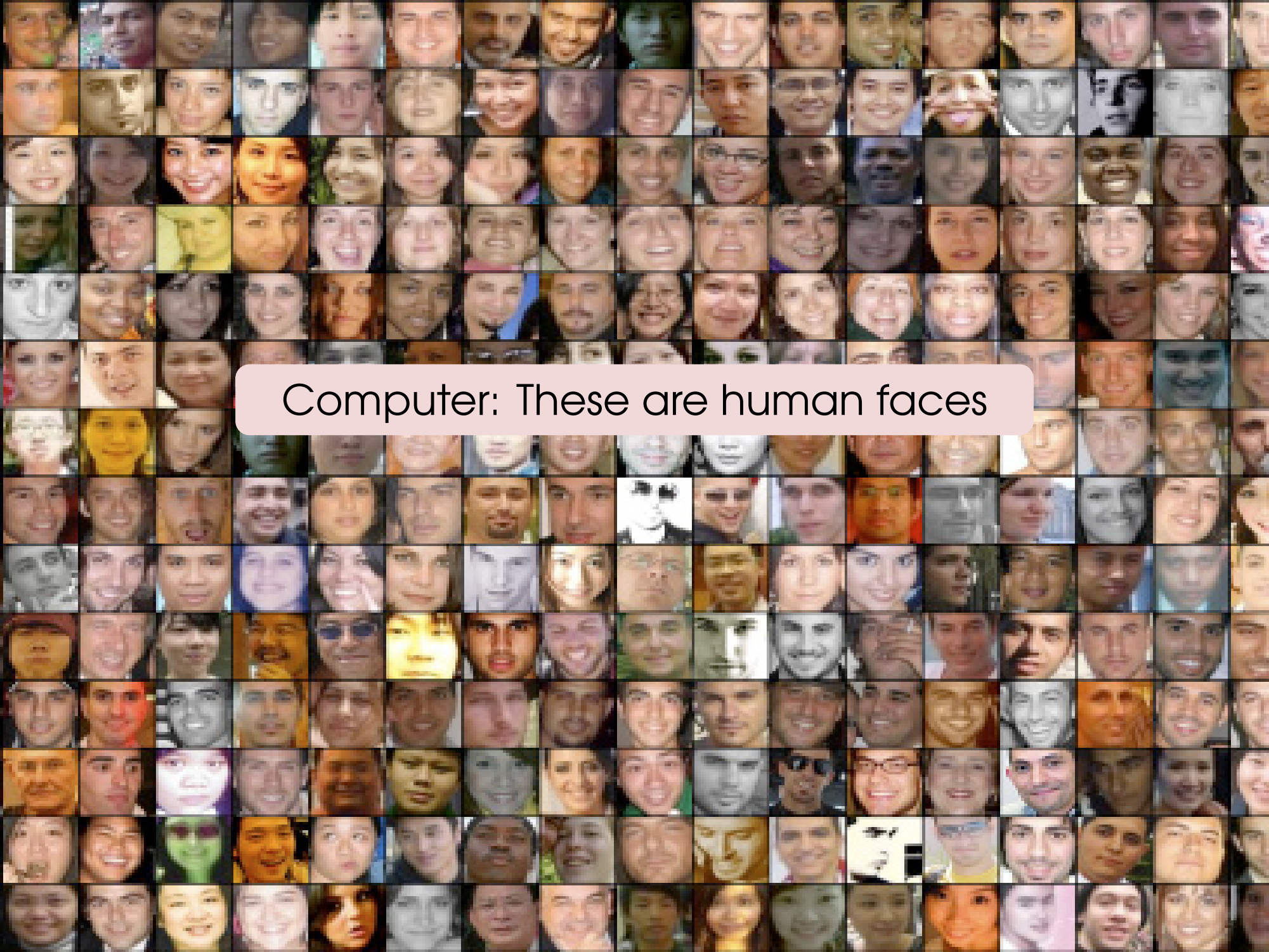
This bird can fly

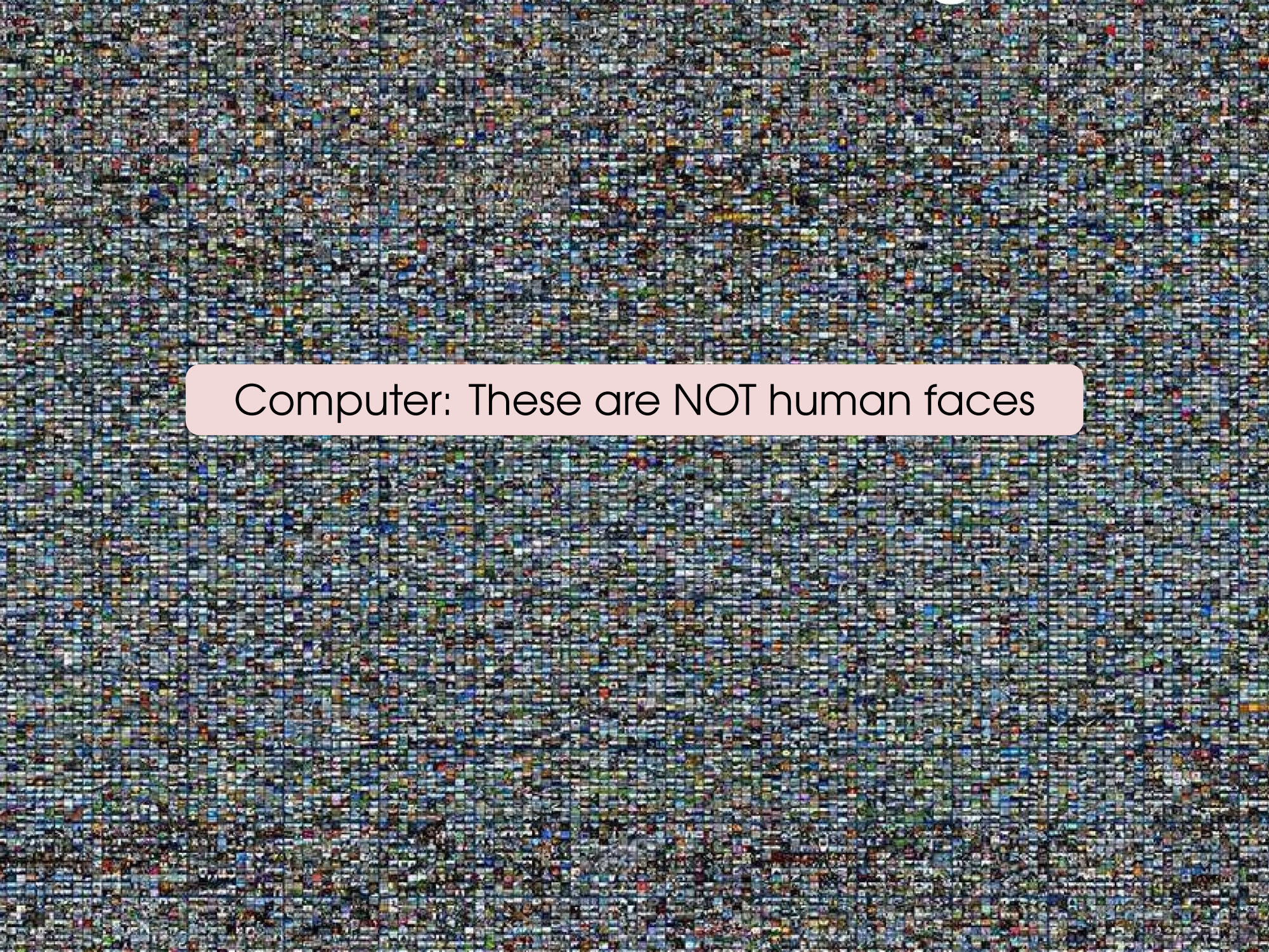
This bird can fly

Can this bird fly ?

# Machine Learning

Computational programs (algorithms) that learn from experience, i.e., data.


Computer: These are human faces


Computer: These are NOT human faces

Computer: Any human face?

# Inductive learning in real world

This is a chair

This is a chair

Most times a difficult problem

This is a chair

What are these?

## AI: Modern View (Rusell and Norvig)

An intelligent agent (or intelligent machine) perceives, reasons, and acts with great flexibility.

## Machine learning (ML)

An intelligent agent learns from experience

## AI: This course perspective

An intelligent agent is able to acquire certain level of understanding of its world.

## ML: This course perspective

- Learning from experience to build useful representations to understand the world.
- Then use these representation to make prediction, take decision, etc.

# Experience

- Machine learning techniques operate over multidimensional data.
- Every data example or instance is given by a set of relevant measurements or attributes.

Examples: Datasets from UCI repository (`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 |

# Experience: Training set

The dataset used to learn a model using a machine learning technique is known as: 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 |

# Experience: Training set

Attributes, dimensions, variables, or features.

Class or Label

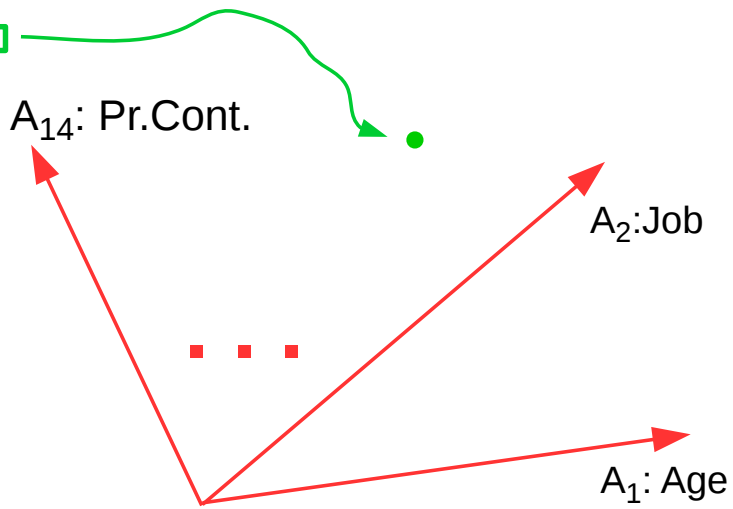
Examples, registers, or instances.

| 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 |

# Experience: Feature space

- Every training example can be considered as a vector that lies in a 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 |



# Generalization: The ultimate goal

## Training set used during the learning process.

| 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 |

## New instances unknown to the model.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | |

# Can we measure generalization capabilities?

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 |

Test Set: instances with known label, but not included during training. They are used to evaluate generalization capabilities of the resulting model (why?).

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | |

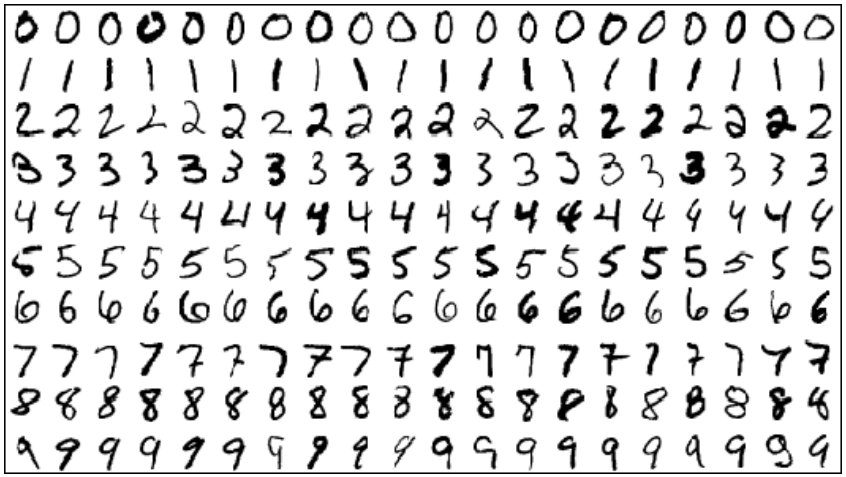
# Traing, test, and validation sets

- One needs data to train a ML algorithm: training set.
- Also, in most cases, one needs to measure the performance that can be expected from a ML model: test set.
- Rule of thumb: from the available data, reserve 70-80% for training and 30-20% for test.

## Validation set

- Machine learning techniques usually need to adjust structural parameters. Ex.: number of iterations, model structure, etc.
- To adjust these structural parameters, one needs data that is independent from the training and test sets.
- This data is known as: validation set.
- Rule of thumb: from the available data, reserve 60-80% for training, 10-20% for test, and 10-20% for validation.

# Case example: MNIST dataset

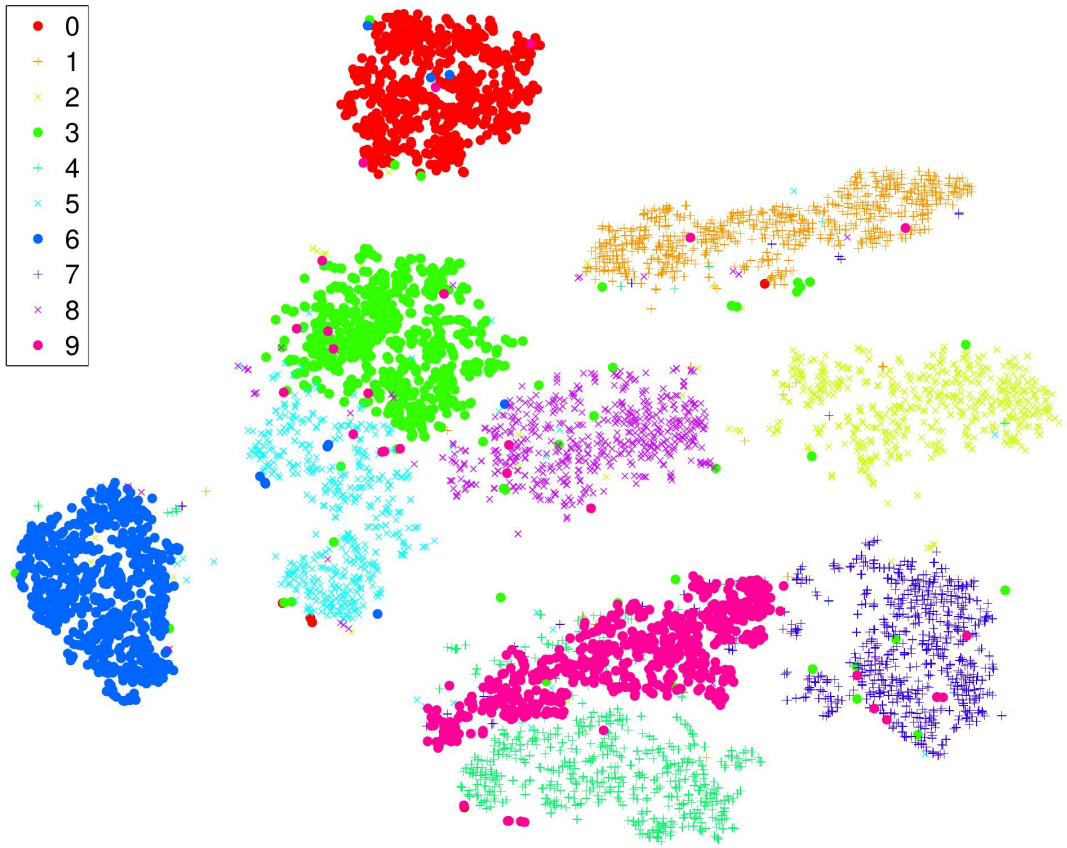


- MNIST is a dataset of images corresponding to handwritten digits.
- Each image displays a single digit from 0 to 9. Images are binary with a resolution of 28x28 pixels.
- Goal: Build a classifier that can recognize the digit in each image.
- Dataset consists of 60.000 training examples and 10.000 test cases.

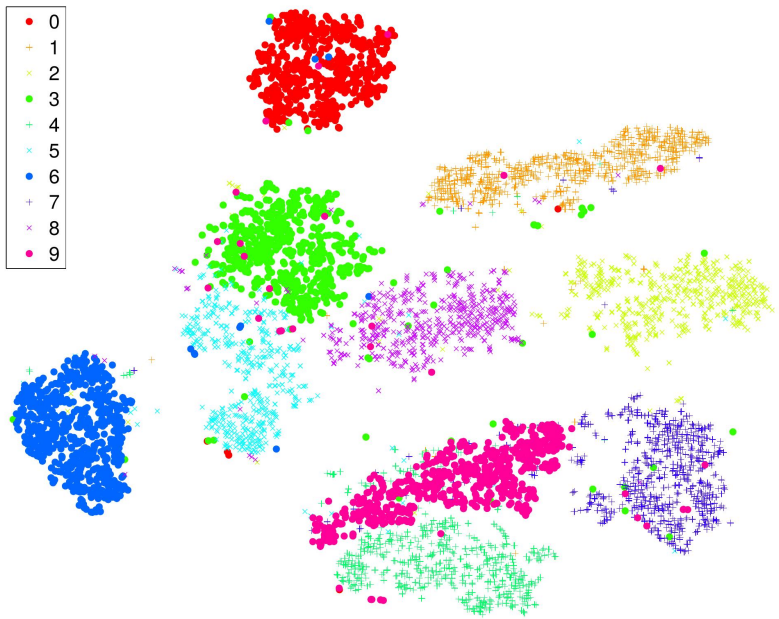
What size is the feature space of the MNIST dataset?

# Ej. MNIST dataset

- It is not possible to provide a direct representation of the MNIST feature space using a 2D visualization.
- However, we can use dimensionality reduction techniques such as t-SNE (`http://lvdmaaten.github.io/tsne`) to obtain a suitable approximation:



# Ej. MNIST dataset




- This visualization illustrates the complexity of the target classifier.

Resulting confusion matrix after applying a K-nearest neighbors classifier over the MNIST test set.

Prediction

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 |

These results are based on K=1 using Euclidean distance in the space of 784 dimensions (why 784?).

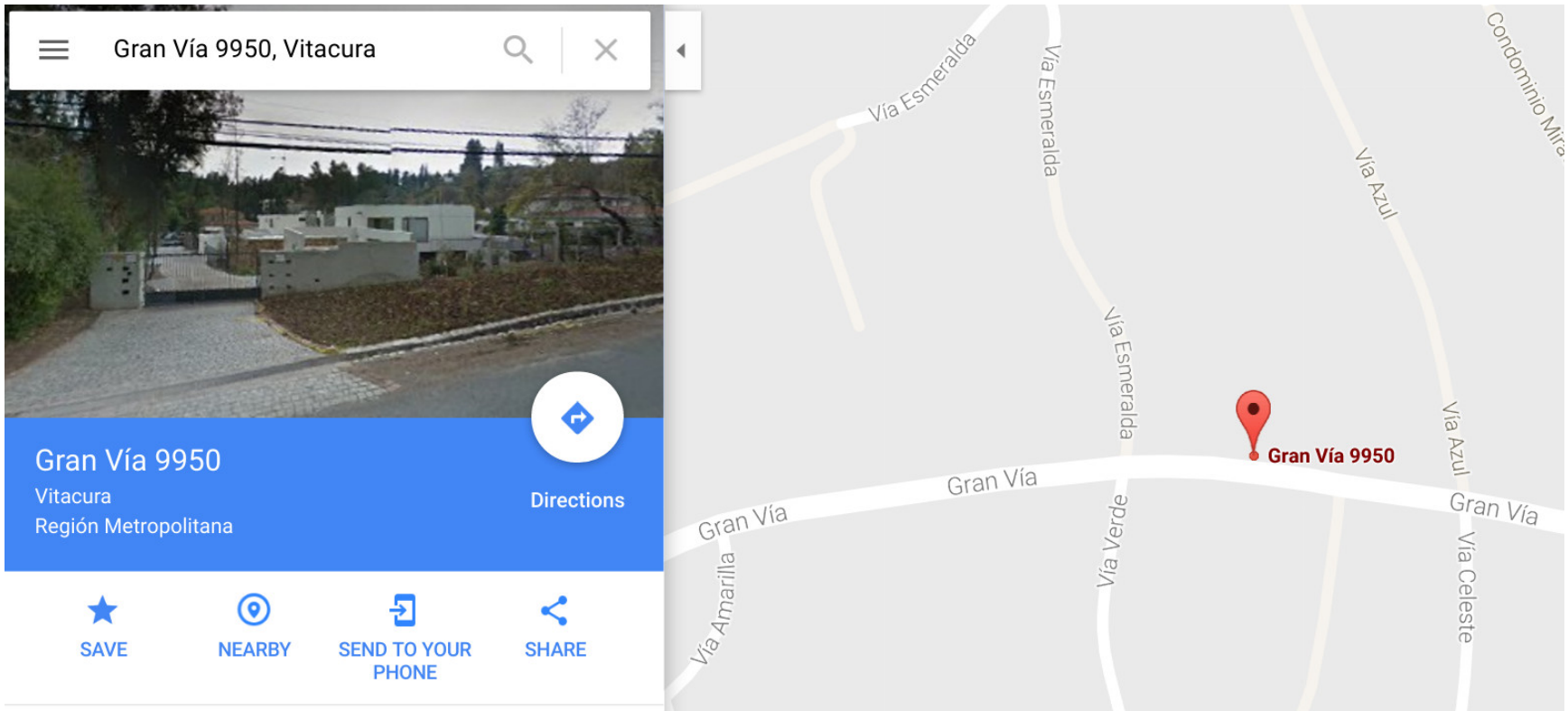
# Machine Learning and Big Data

- As we discussed, Machine Learning is about learning from data.
- Currently, Big Data is it's great ally.
- By providing large sources of data, Big Data is increasing the accuracy of machine learning techniques.
- The synergistic combination between machine learning and big data is a key element behind the sucess of deep learning.

Good experience, i.e. data, is key to learn.

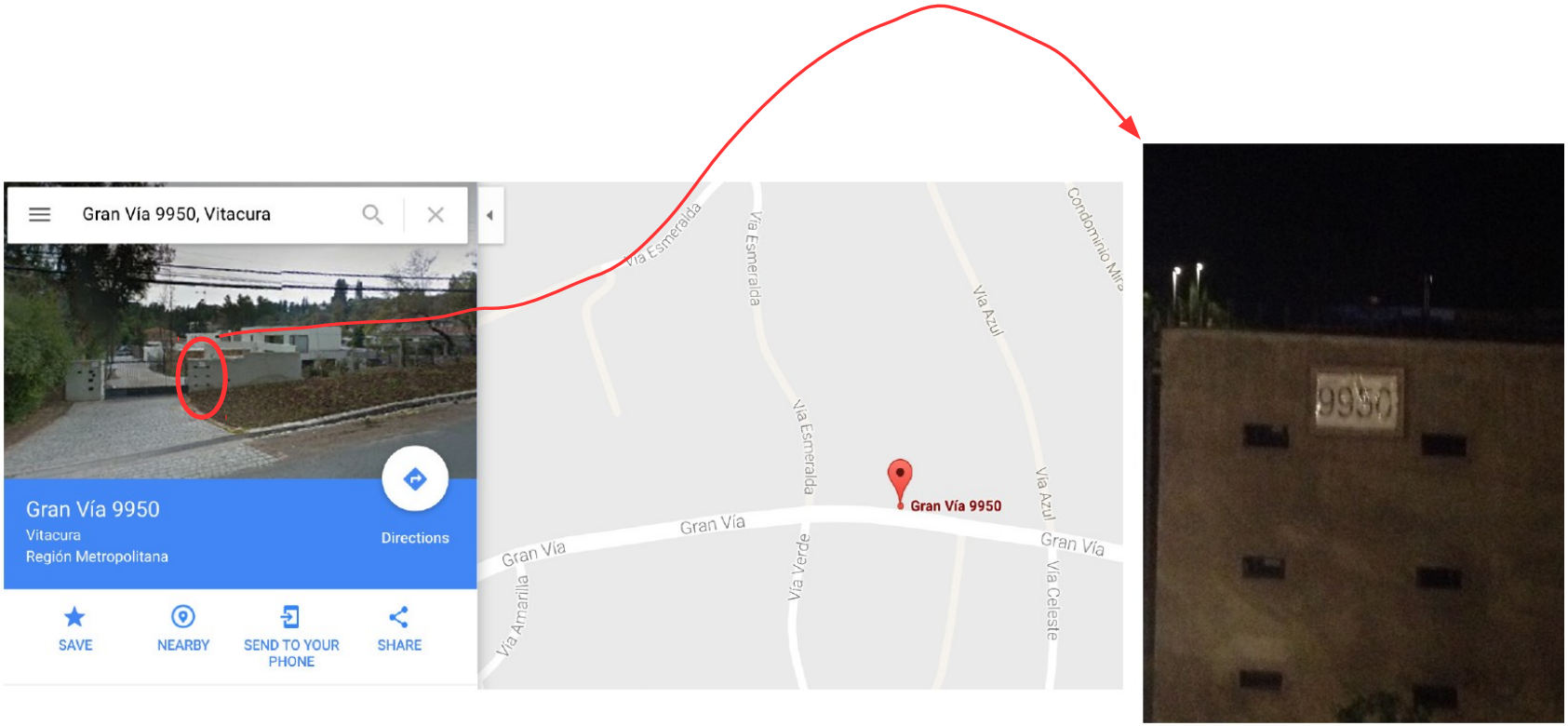
# Experience is key

## Example: Google Street View.



# Experience is key

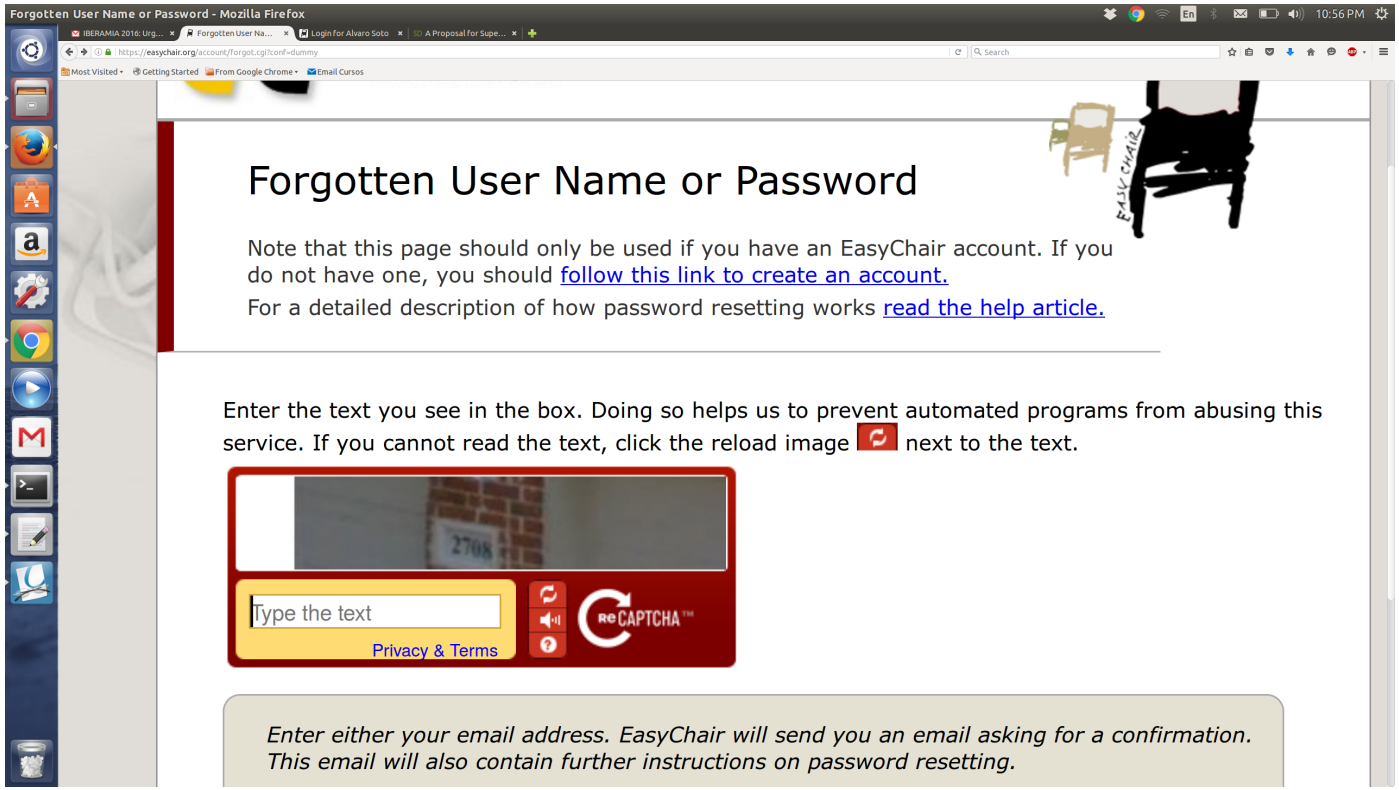
Example: Google Street View.



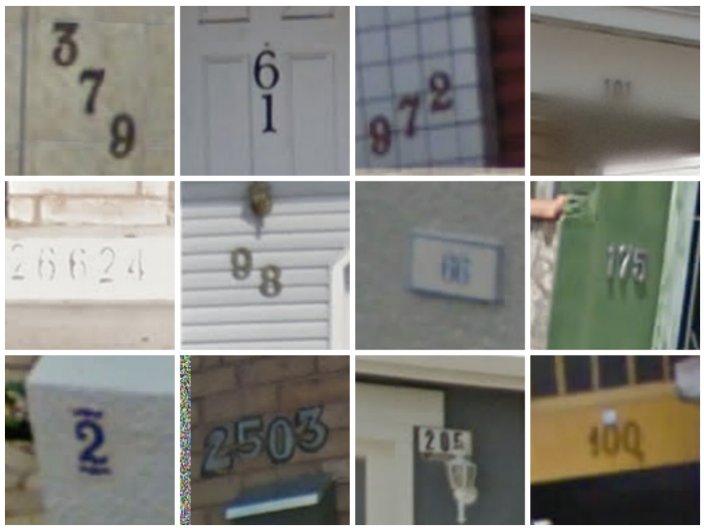

How can they do this?

# Experience is key

## Example: Google Street View.

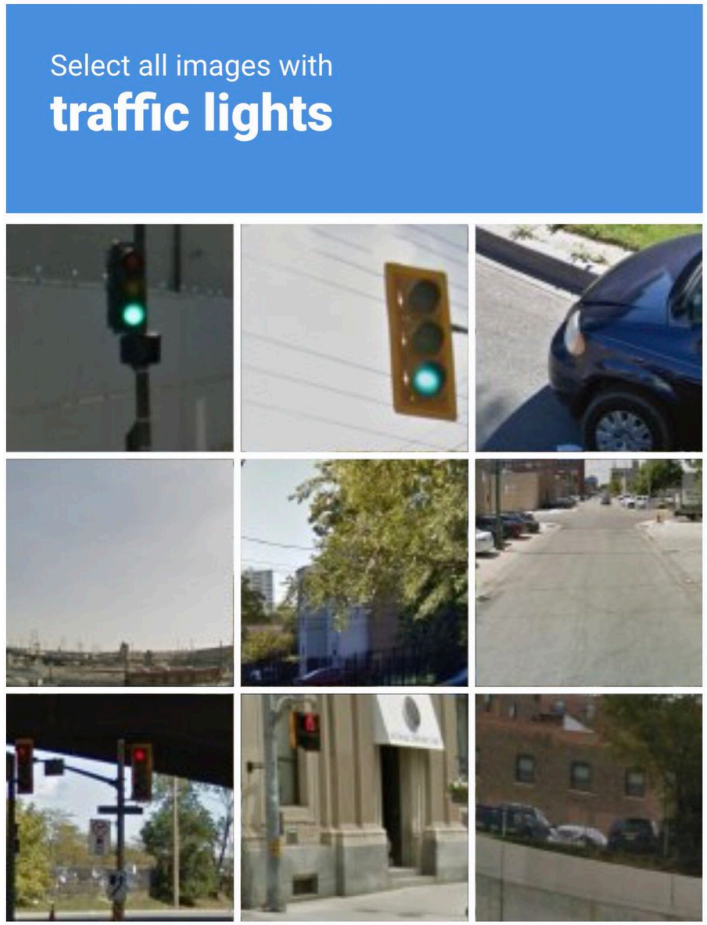


# Experience is key: Training Data

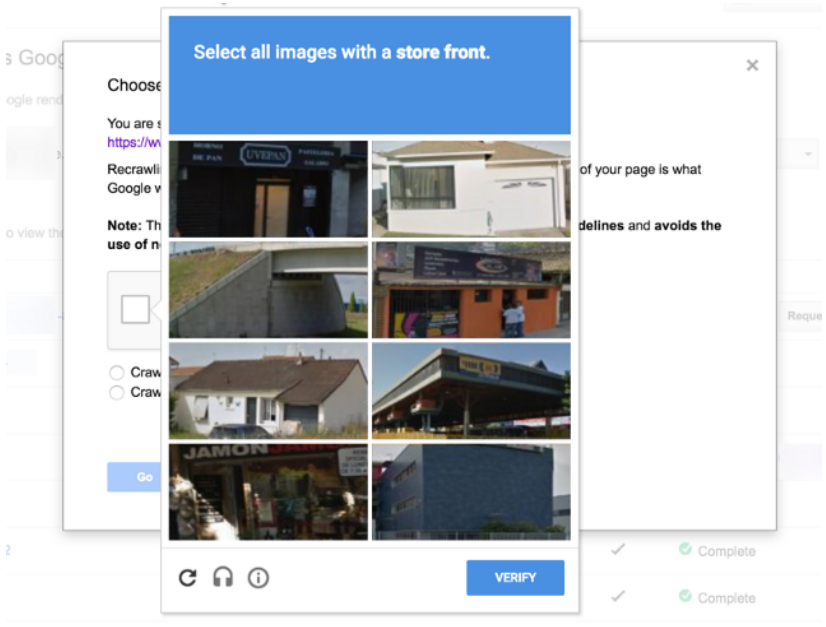


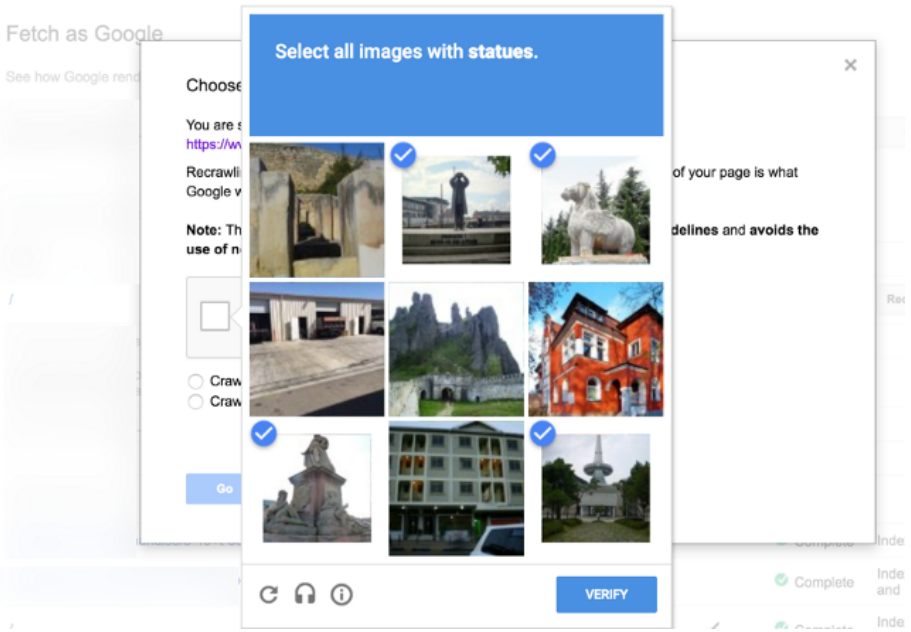
http://research.google.com/pubs/pub42241.html

# Experience is key: Training Data

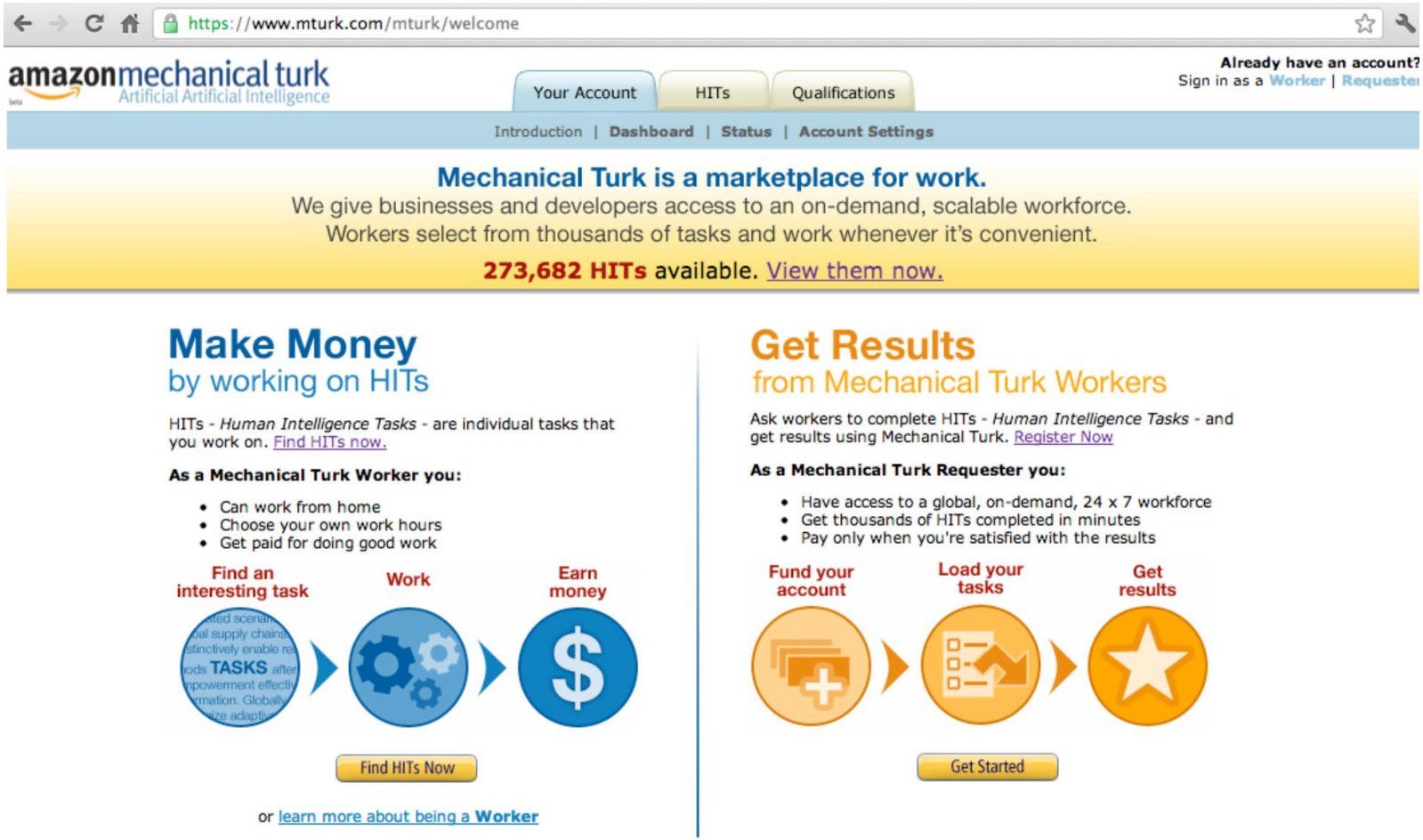


# Experience is key: Training Data

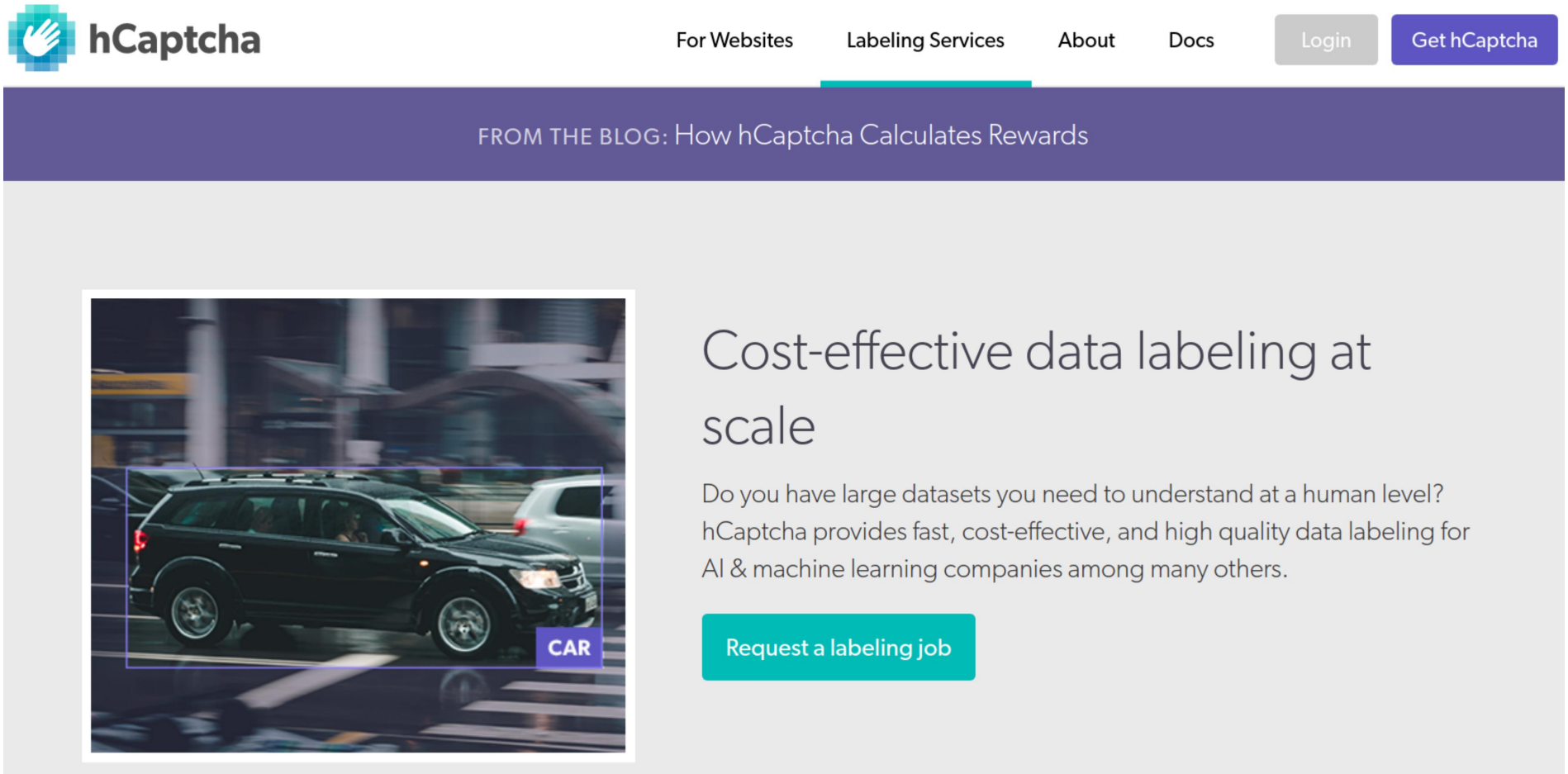


# Experience is key: Crowdsourcing and Human Computation



# Experience is key: Crowdsourcing and Human Computation



# Experience is key: Crowdsourcing and Human Computation

**Appen acquires Figure Eight for up to $300M, bringing two data annotation companies together**

Anthony Ha @anthonyha / 4 hours ago

Comment

10/03/2019

# Experience

So far we have been focusing on a learning paradigm known as: Supervised learning. However, depending of the type of training data available and the goals of the learning process, we can find several popular learning paradigms:

- Supervised learning.
- Unsupervised learning.
- Reinforcement learning.
- Semi-supervised learning.
- Active learning.
- Structural learning.
- Supervised clustering.
- Instance based learning.
- ...

As we will discuss in this course, all these learning frameworks are just different ways to provide semantic to the learning process (grounding).

## Machine Learning

Any computer program that improves its performance at some task through experience.

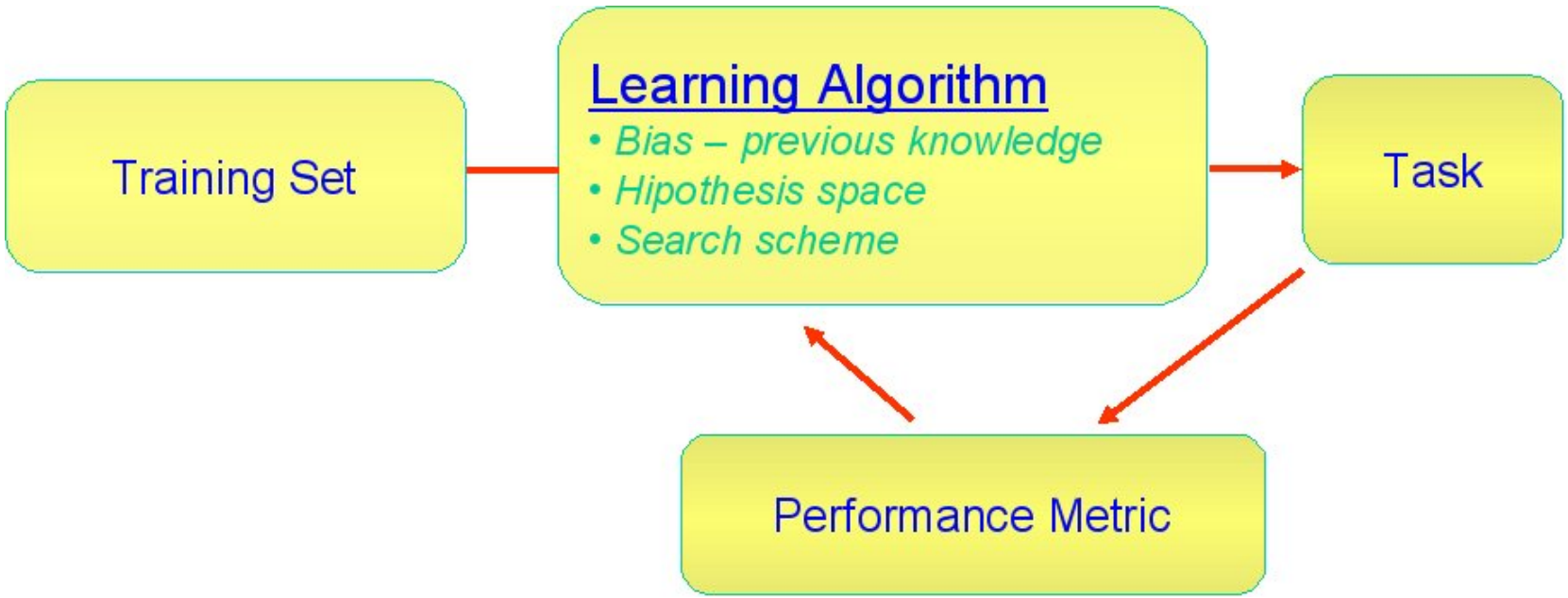
A computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance for tasks in T, as measured by P, improves with experience E.

**Key assumption of Inductive Machine Learning**

Any hypothesis that approximates a target function well over a sufficiently large set of training examples will also approximate this target function well over unobserved examples.

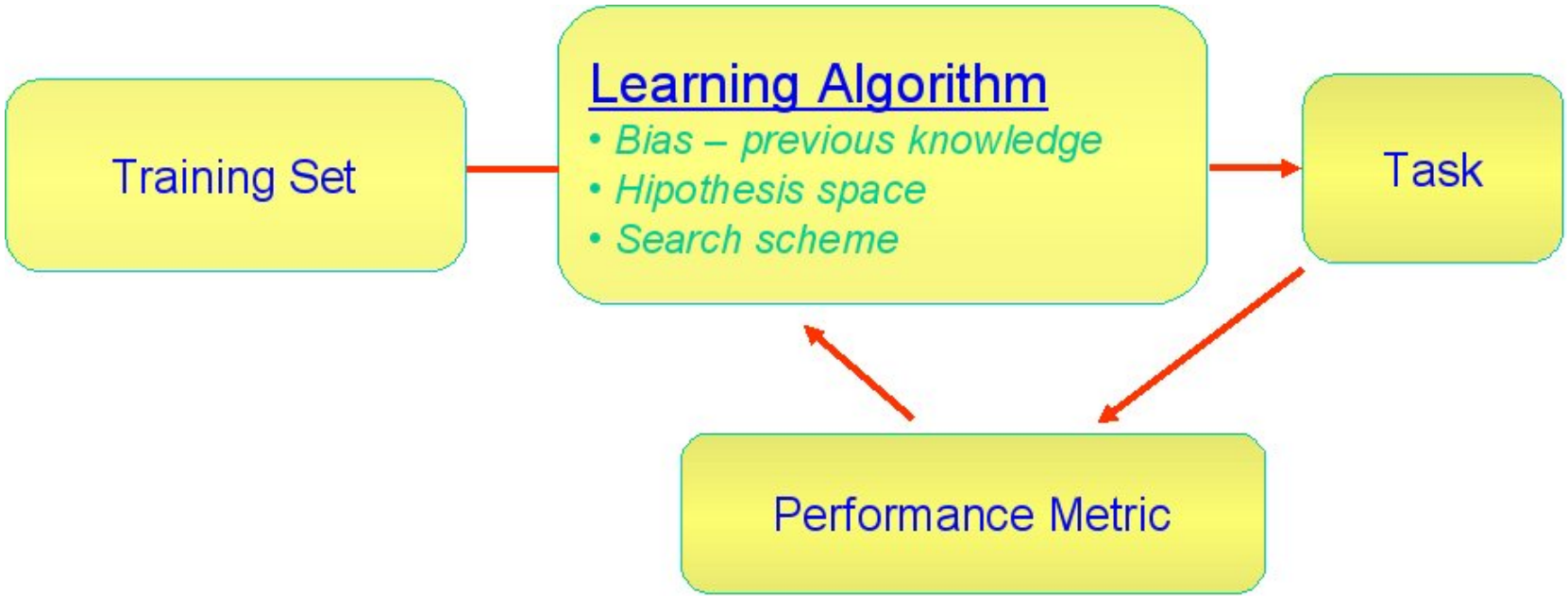
- OJO The previous assumption requires that the training data must be representative of the target concept or function.

Learning involves searching through a space of possible hypotheses to find a hypothesis that provides a suitable (best) fit to the available training data and prior constraints (previous knowledge).




OJO There is a new element: previous knowledge…Any comment?

Machine learning, 3 main ingredients:
Representation + Performance + Optimization.



Machine learning, 3 main ingredients:
Representation + Performance + Optimization.

## Machine learning problem

$$f^* = \arg\min_{f \in \mathcal{H}} \mathcal{L}(f(x)) = \arg\min_{f \in \mathcal{H}} \int_{x_i \in T} \mathcal{L}(f(x_i))\, d_T$$

We usually approximate this using a training set Tr:

$$f^* \approx f^*_{Tr} = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{x_i \in Tr}^{N} \mathcal{L}(f(x_i))$$

$\mathcal{H}$: hypothesis space.
$\mathcal{L}$: loss function
$f^*$: optimal hypothesis in $\mathcal{H}$ under $\mathcal{L}$.

## Generic machine learning loss

$$f^* \approx f^*_{Tr} = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{x_i \in Tr}^{N} \mathcal{L}(f(x_i))$$

## Supervised learning

$$f^*_{Tr} = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{x_i, y_i \in Tr}^{N} \mathcal{L}(f(x_i), y_i)$$

# Hypothesis space

- Models have limits (suitable representation?).
- Search in complex spaces (non-lineal or convex) is usually a hard and exhausting process (right optimization tool and performance metric?).

Risks,

- If the hypothesis space is too flexible, there is a higher risk to converge to misleading hypothesis (suffer overfitting problems).
- If the search tool is too limited, there is a higher risk to being unable to find a good hypothesis (converge to local optimals).

# Overfitting?

- When one applies a ML technique, one is interested in the generalization capabilities of the resulting model, why?.
- Generalization? $\rightarrow$ Inductive learning.
- In other words, one is interested in obtaining good predictions for new instances of the target domain.

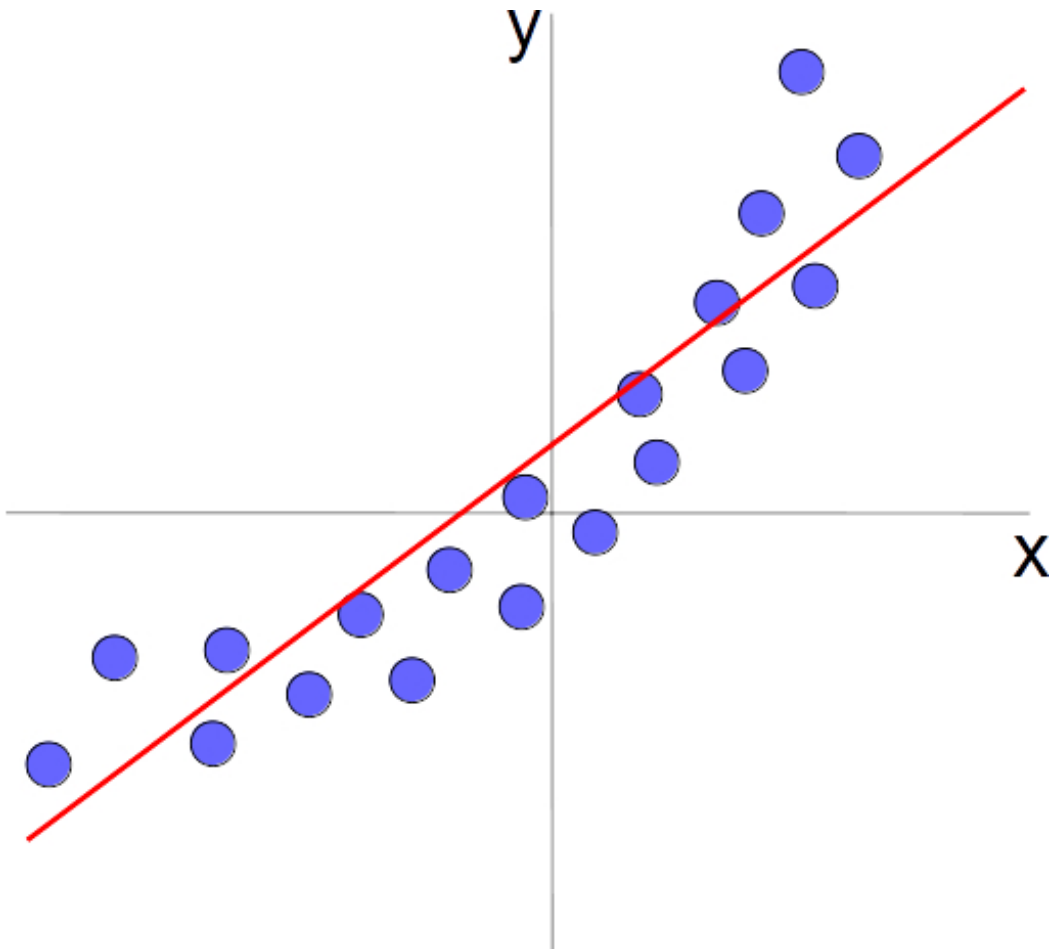
Overfitting: It is a situation when a ML model starts to memorize (overfit) the traininig data, decreasing its capability to correctly predict new instances.

**Jorge Luis Borges: ``Funes el memoriozo".**

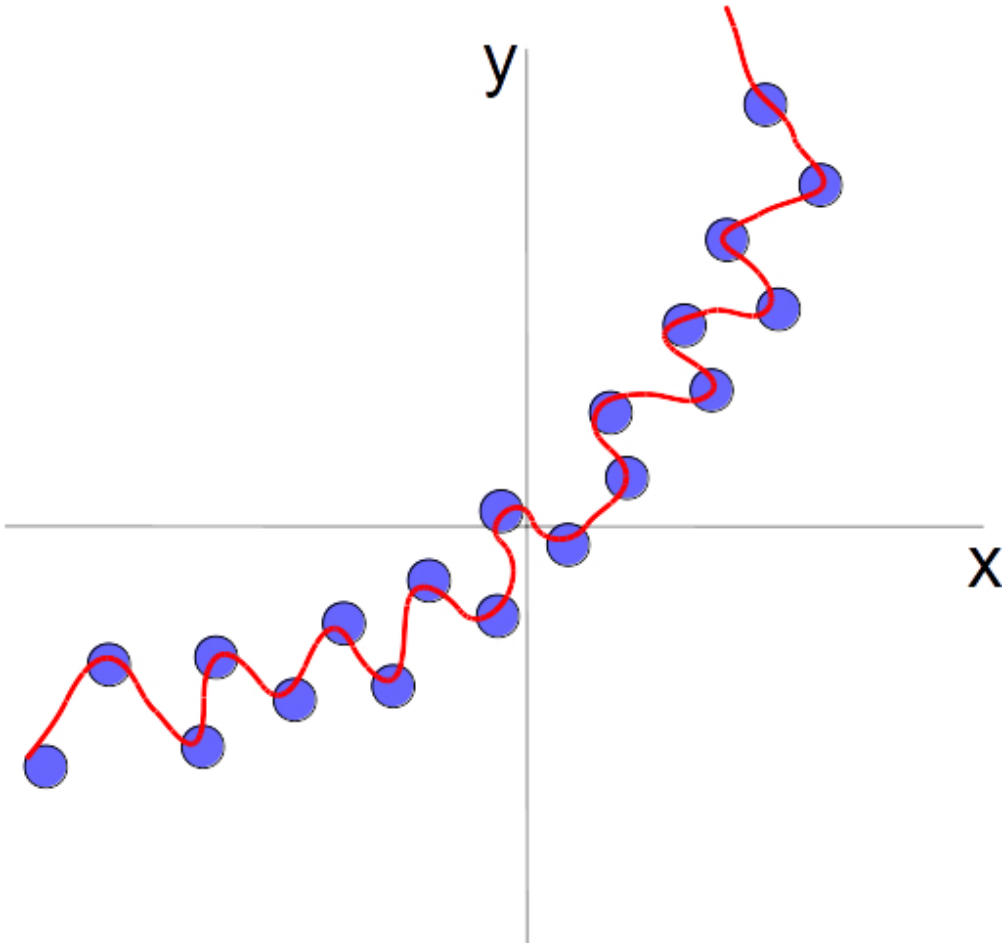
``...Había aprendido sin esfuerzo el inglés, el francés, el portugués, el latín. Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es **generalizar, abstraer**. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos."

Jorge Luis Borges: "Funes el memoriozo".

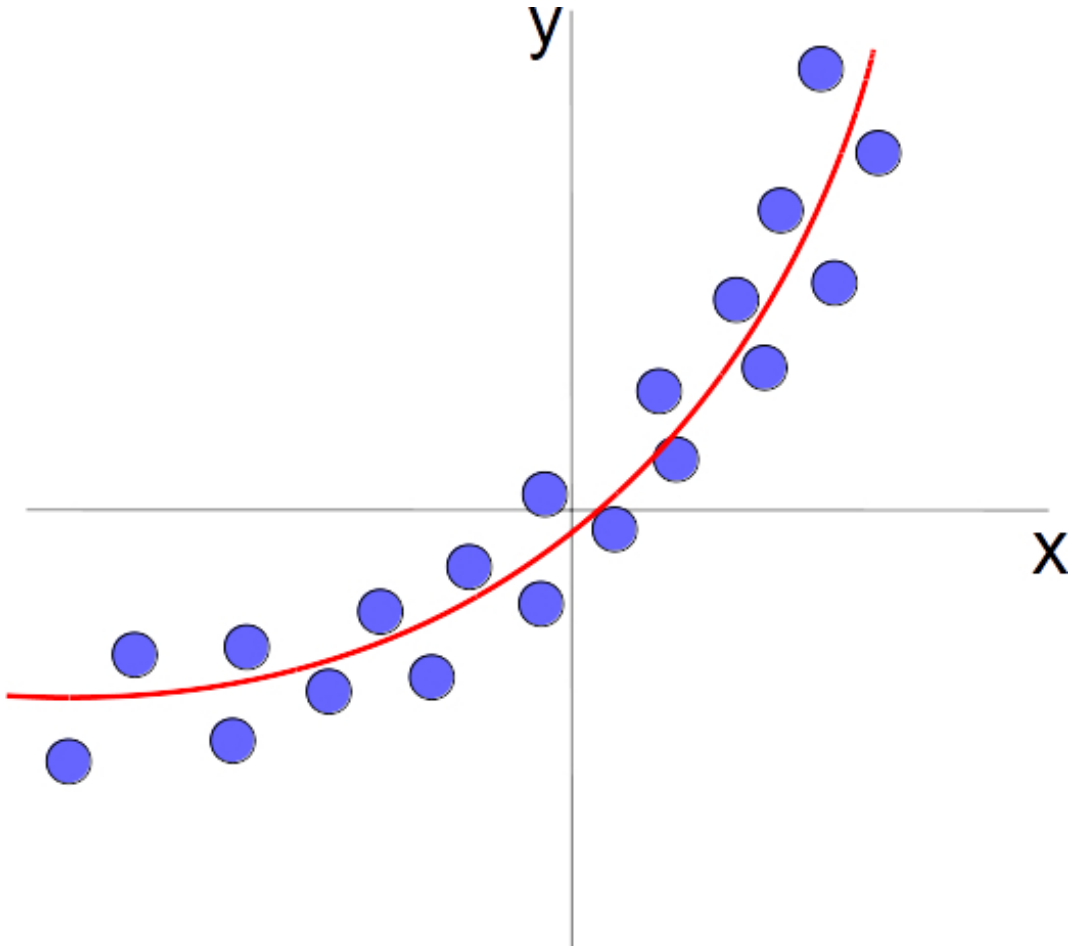
# Underfitting, Overfitting, Good fit



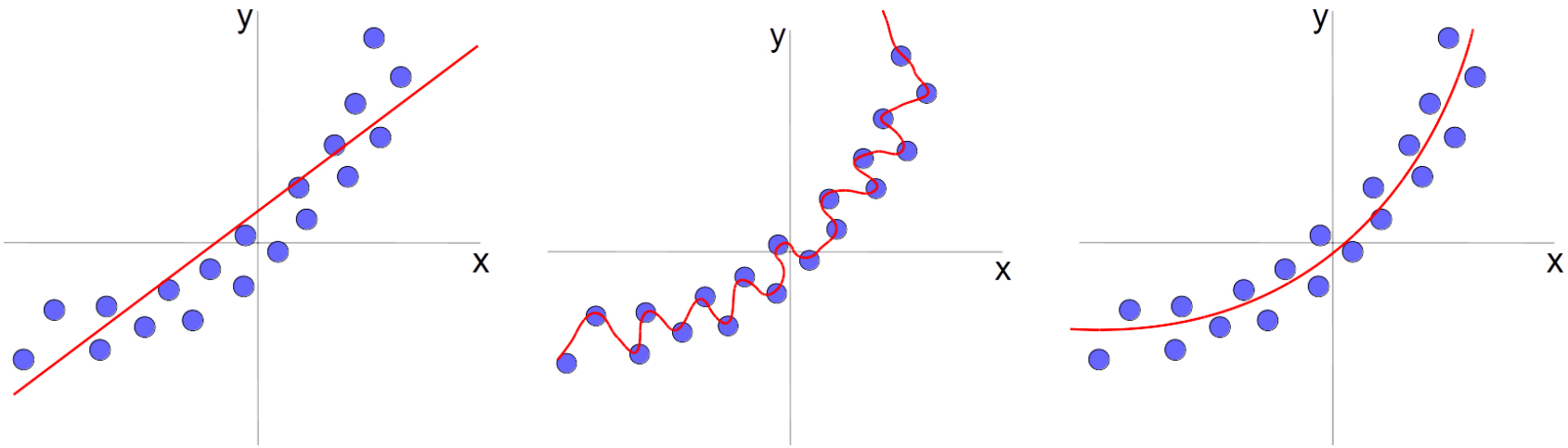
# Underfitting, Overfitting, Good fit



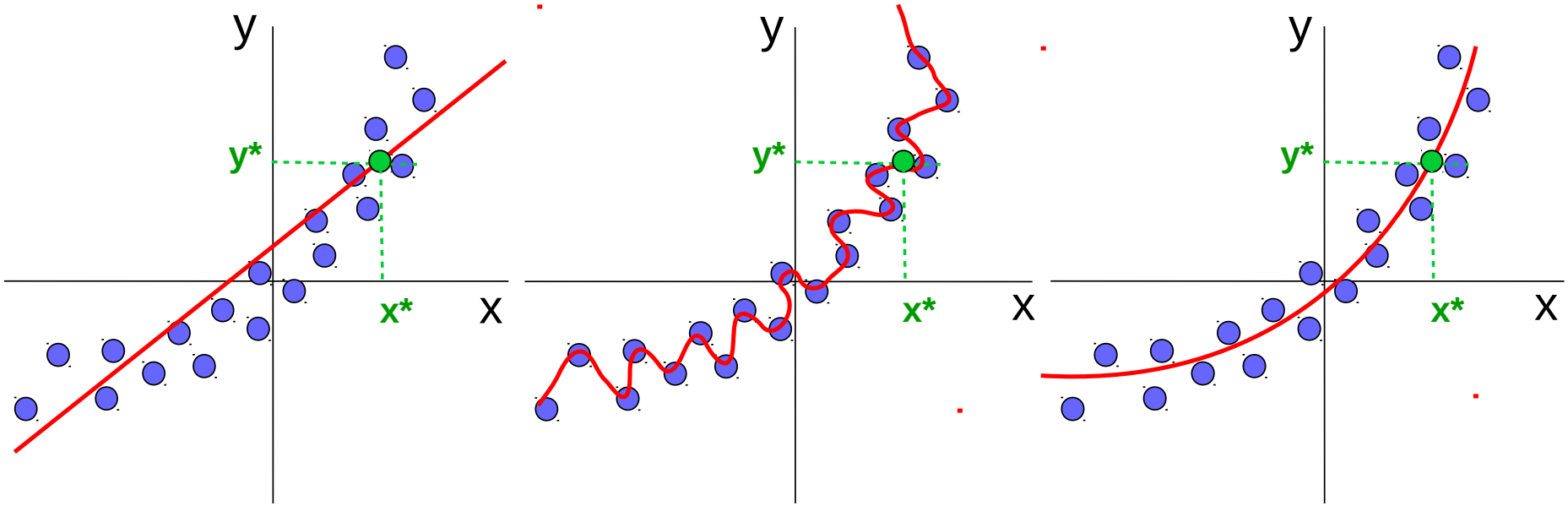
# Underfitting, Overfitting, Good fit



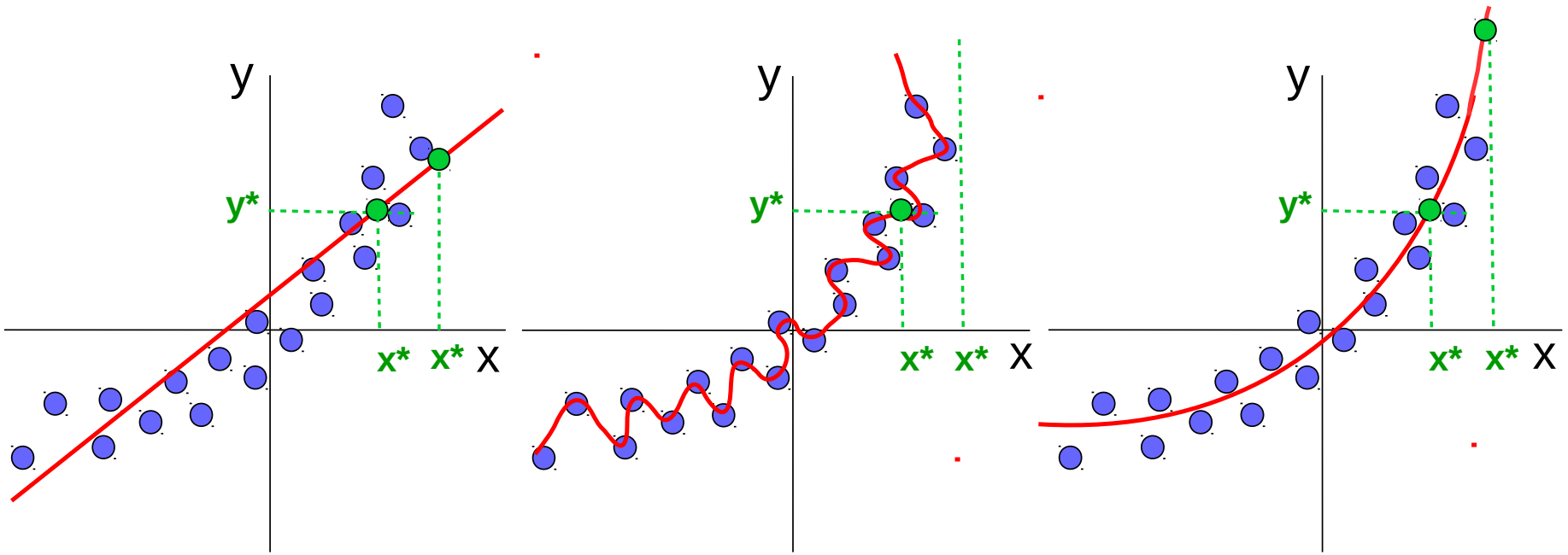
# Underfitting, Overfitting, Good fit



# Generalization



# Generalization



Concepts that you should know about ML.

- Classical AI and ML views.
- Generic view of ML.
- Main elements of a ML technique: Representation + Performance + Optimization.
- Supervised and unsupervised ML.
- Overfitting.
- Training, test, validation sets.

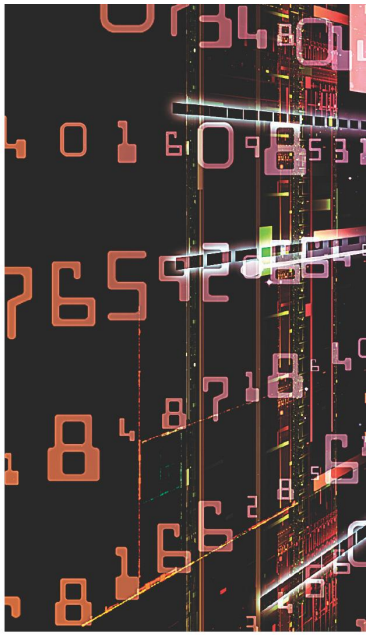
Recommended concepts that you should review. We will discuss some of them in future lectures:

- All the ideas in this week lecture: "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.**