Aprendizaje Automático

2-2021 UTFSM

Objetivos

- Introducir brevemente el curso.
- Presentar el programa y calendario.

Preguntas

- ¿Qué es el aprendizaje automático (machine learning)?
- ¿Para qué sirve o cuándo es útil?
- ¿Cuáles son las dificultades más relevantes?

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed"

Arthur Samuel (1959)

Entrenar versus Programar

El aprendizaje automático es un conjunto de herramientas, modelos y métodos que permiten a un programa/máquina/sistema aprender una determinada tarea sin definir explícitamente el algoritmo necesario para obtener el resultado que deseamos.

En vez de programar explícitamente las reglas para producir el resultado, queremos entrenar el programa, es decir mostrarle ejemplos de lo que esperamos que haga. Decimos que el programa aprende cuando logra reproducir el comportamiento que queremos, es decir, cuando logra inferir de los ejemplos el algoritmo que produce las repuestas deseadas.

Entrenar versus Programar

Ejemplo: Filtrado de correo spam. Queremos un programa que lea nuestros correos y filtre automáticamente aquellos maliciosos o aquellos que simplemente no deseamos leer.

Entrenar versus Programar

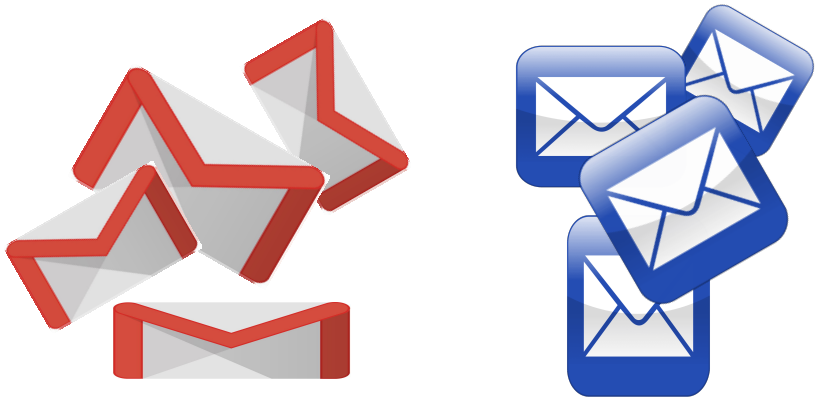
Asumamos que un “correo” se encuentra representado como una estructura de 3 campos (dirección de origen, título y cuerpo), donde el primero es una cadena de la forma name@domain y los últimos dos consisten en listas de palabras.

Por *“programar explícitamente una máquina”* para esta tarea, nos referimos a definir manualmente un conjunto de reglas que, operando sobre la representación disponible, permitan determinar si el correo debe ser filtrado o no. Por ejemplo:

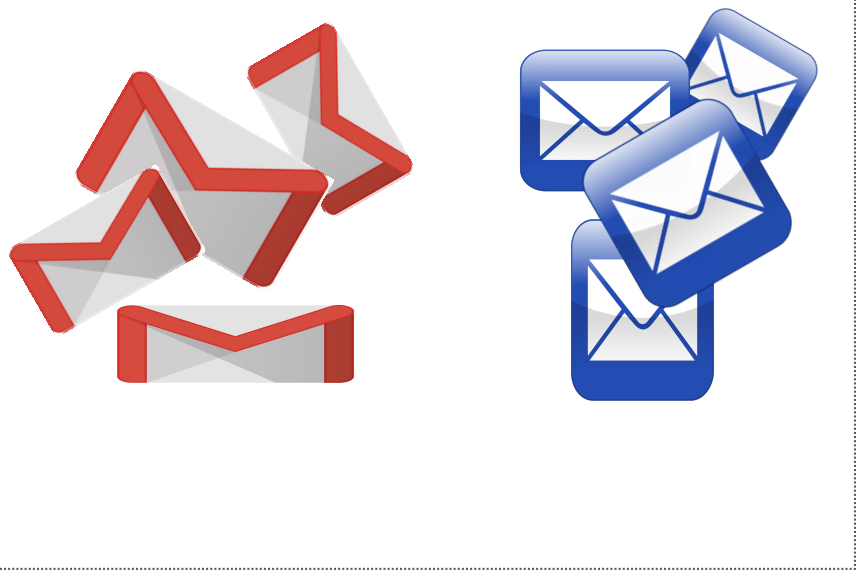
IF email.source.domain == ‘utfsm’ THEN type=spam
IF (‘tarea’ IN email.header) THEN type=spam

Entrenar versus Programar

Por “entrenar un programa” para la tarea, nos referimos a dar a un programa ejemplos de correos SPAM (que hemos filtrado en el pasado) y posiblemente ejemplos de correos no SPAM (=HAM), de modo que él mismo infiera las reglas que permiten distinguir futuros correos entrantes.



Ejemplos (datos)



Máquina (programa)

Programa deseado

IF email.source.domain == ‘utfsm’ **THEN** type=spam
IF (‘tarea’ IN email.header) **THEN** type=spam

Hacia una Definición

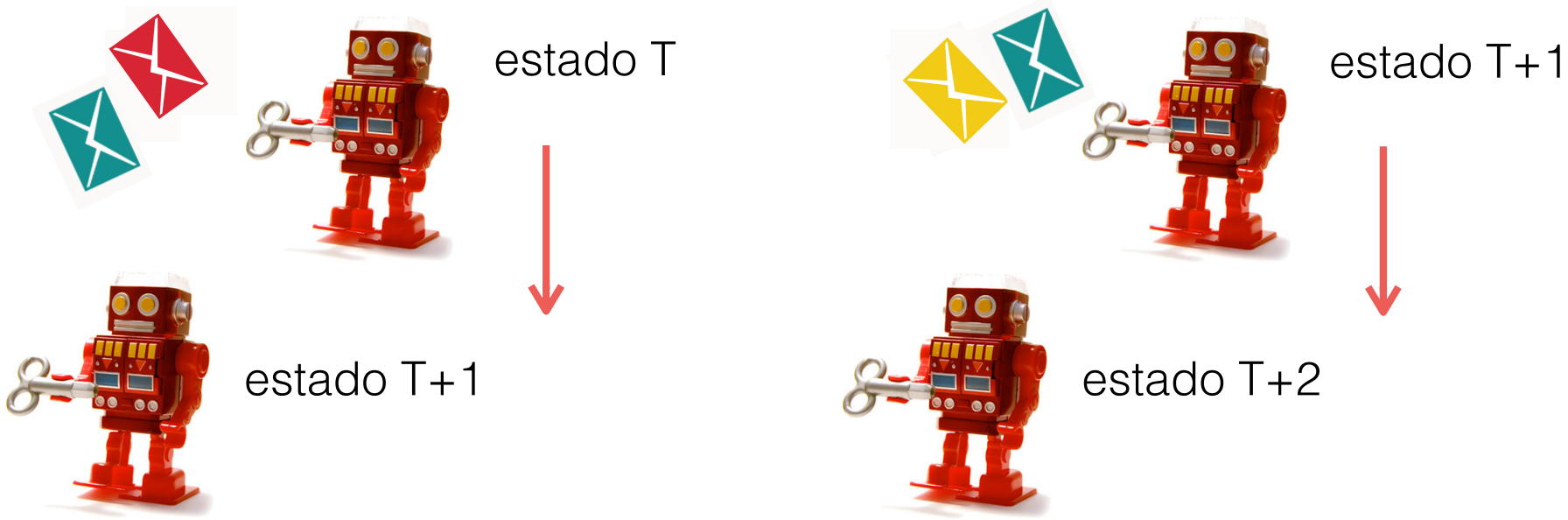
El punto de vista anterior es un poco "estático" y corresponde a un escenario que se denomina "batch" (por lotes). Consideremos una definición alternativa (de un libro clásico del área):

"A program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**"

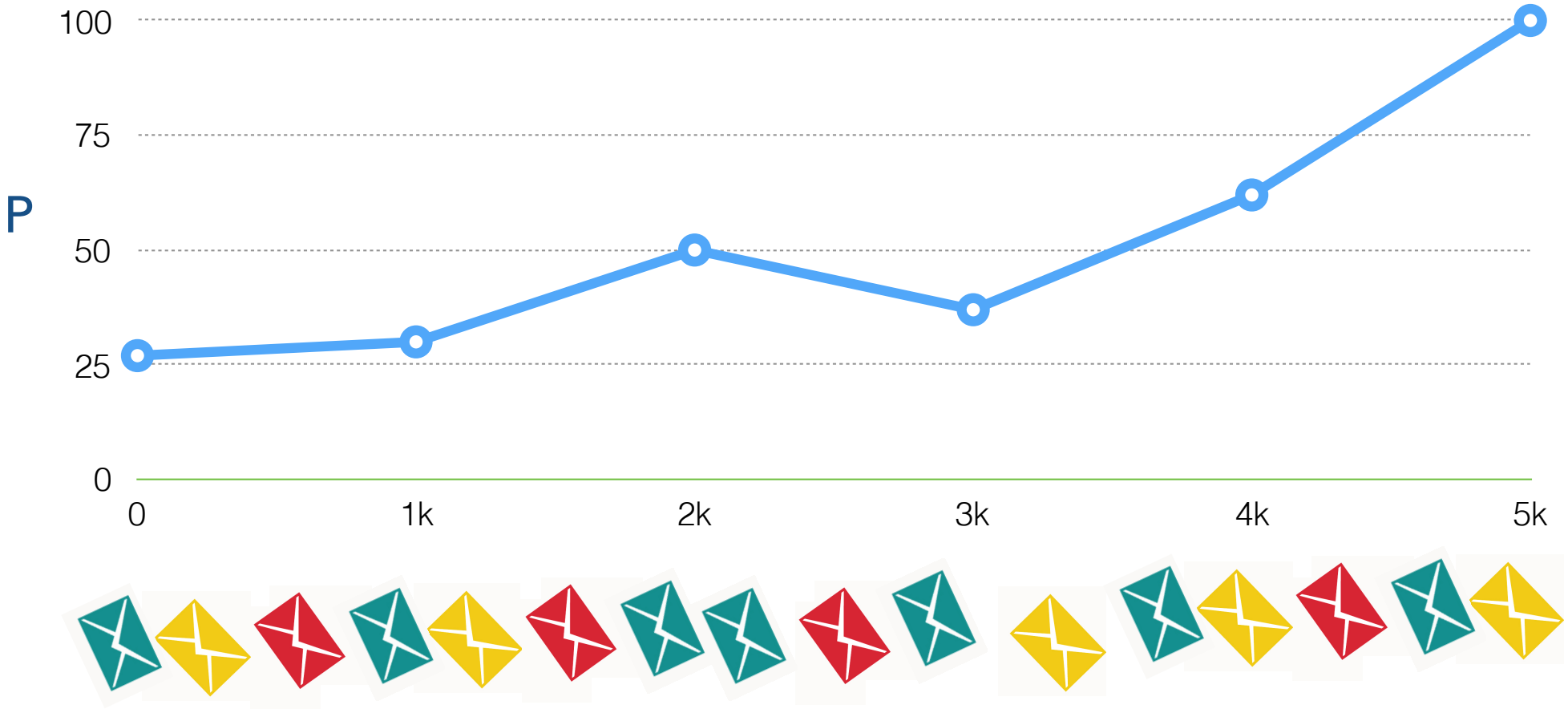
Tom Mitchell, *Machine Learning*, 1997.

Hacia una Definición

Este punto de vista es más consistente con lo que se denomina *aprendizaje online* (online learning), escenario en que no se asume que la máquina dispone de <u>todos los ejemplos desde un comienzo</u> sino que los observa ejemplos de manera continua y aprende de manera continua.



Desempeño (P) vs Experiencia (E)



E = DATOS

¿Cuándo es esto útil?

El aprendizaje automático es útil para resolver problemas en que resulta difícil especificar un algoritmo que produzca el comportamiento deseado, y es mucho más fácil recolectar ejemplos que reflejen el comportamiento deseado.

El aprendizaje automático es también útil para problemas en que se requiere una adaptación constante al cambios en el ambiente (por ejemplo, nuevos ataques spam).

El aprendizaje automático es también útil para problemas en que se requiere un alto grado de personalización de la solución (por ejemplo, filtrado de noticias depende del gusto del usuario).

¿Es esto posible?

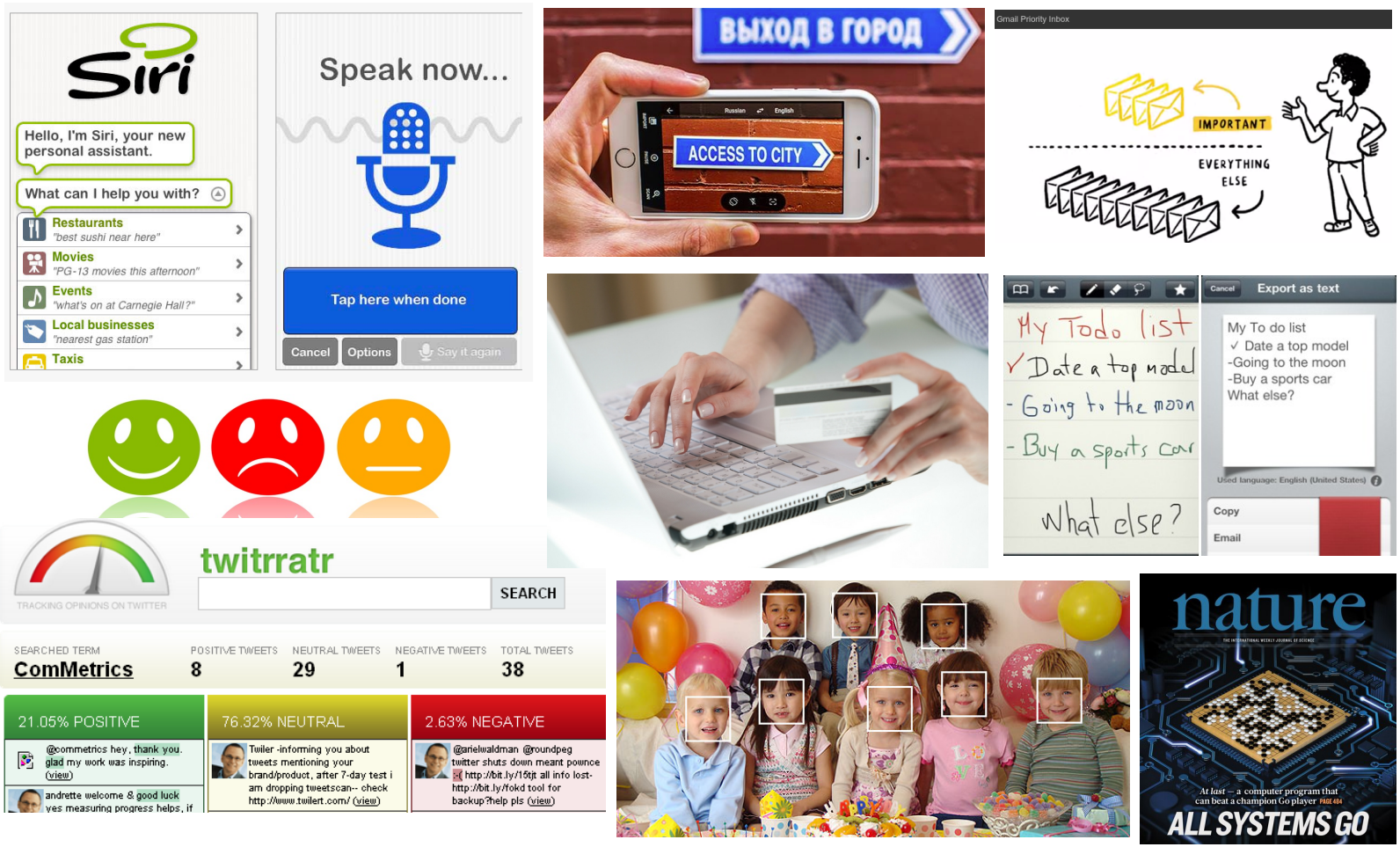
Sí, si tenemos un conjunto de datos suficientemente grande y un modelo suficientemente flexible como para aprovecharlos.

Sí, si podemos acotar suficientemente bien el problema (a uno del tipo entrada-salida) y lo conocemos suficientemente bien como para identificar los *atributos* que contienen la información necesaria para que la máquina lo resuelva.

Sí, si conocemos los problemas que aparecen y estamos disponibles a hacer mucha experimentación (prueba y error) para resolverlos.

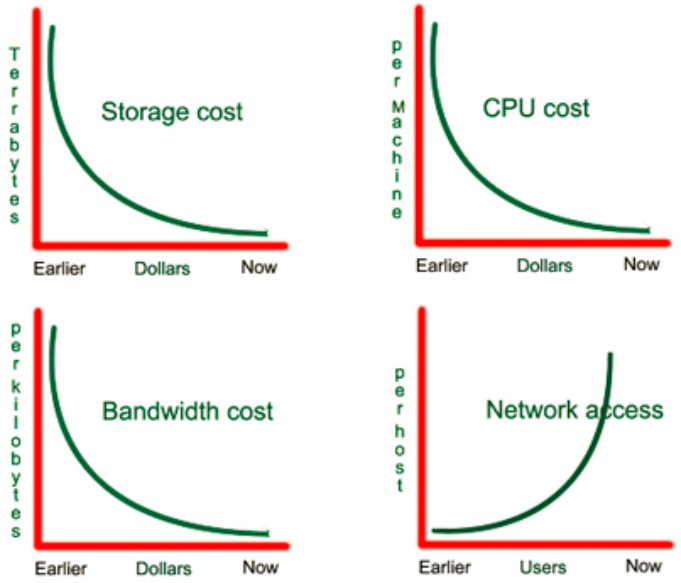
Motivación

Los avances de los últimos años han hecho que el aprendizaje automático se convierta en tecnología muy utilizada en la industria.



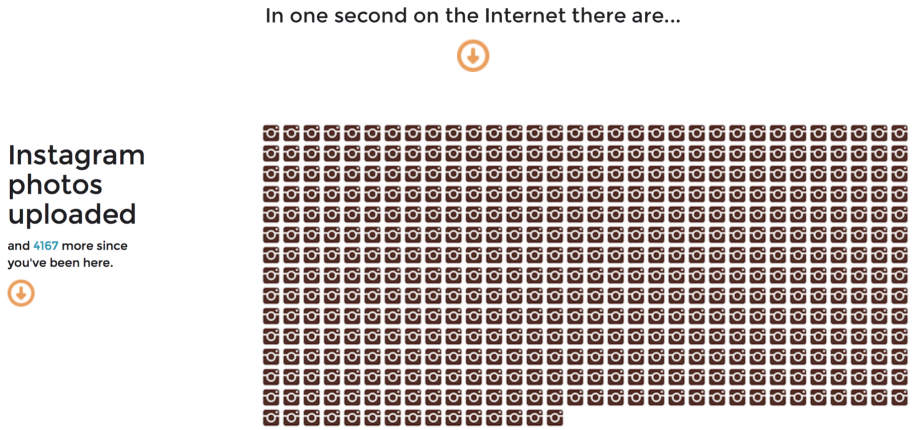
Motivación

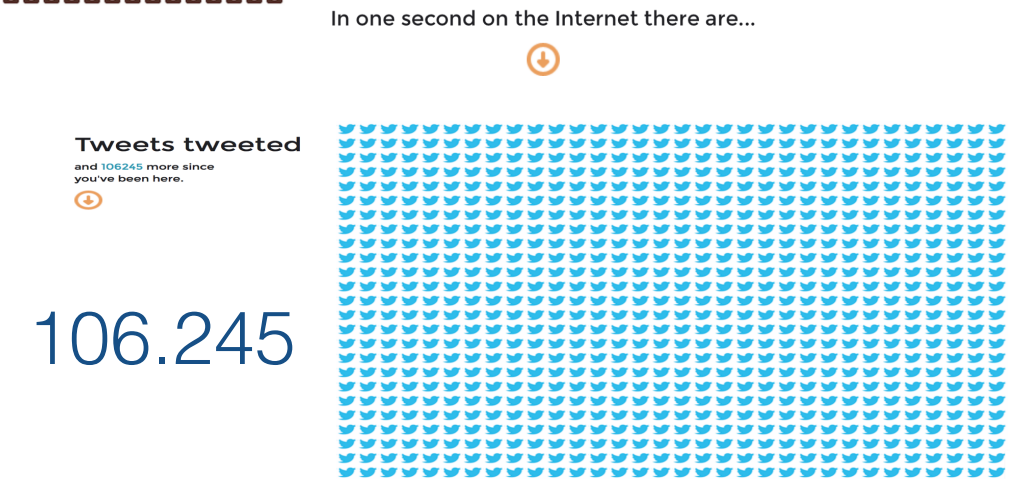
Los avances han sido posibles gracias al progreso en el hardware disponible (paralelismo, GPU's, etc), en el software disponible (librerías especializadas) y sobre todo en la cantidad de datos disponibles.



Motivación

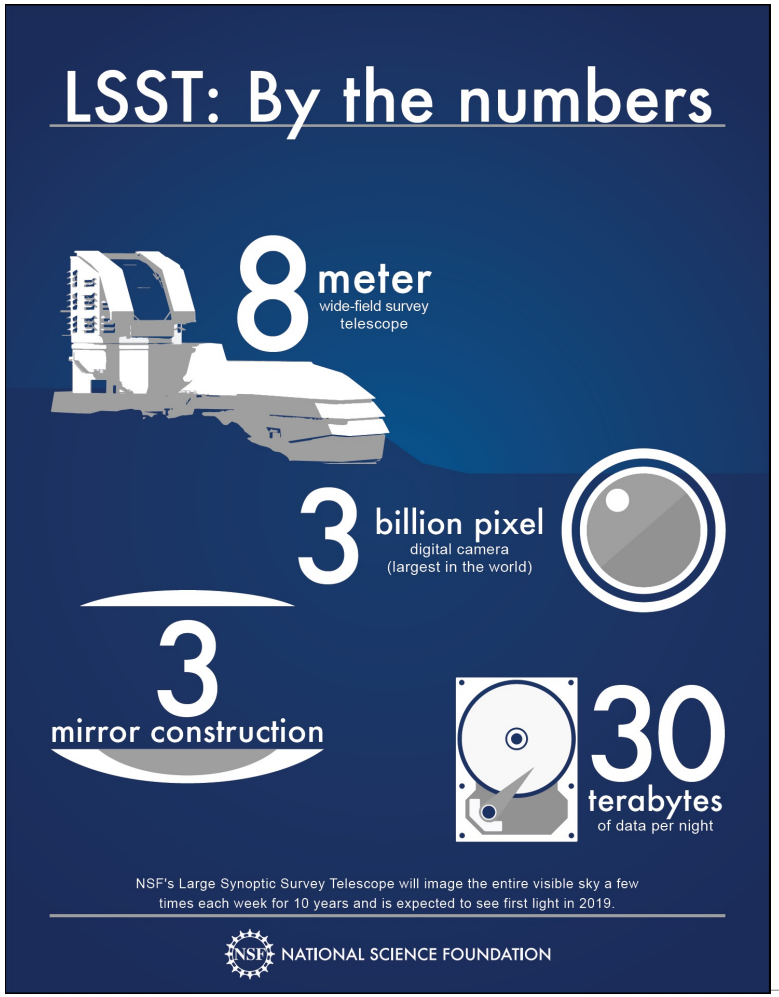
La sobre-abundancia de datos es lo que hace **posibles** los enormes progresos que se están produciendo en el área …





Motivación

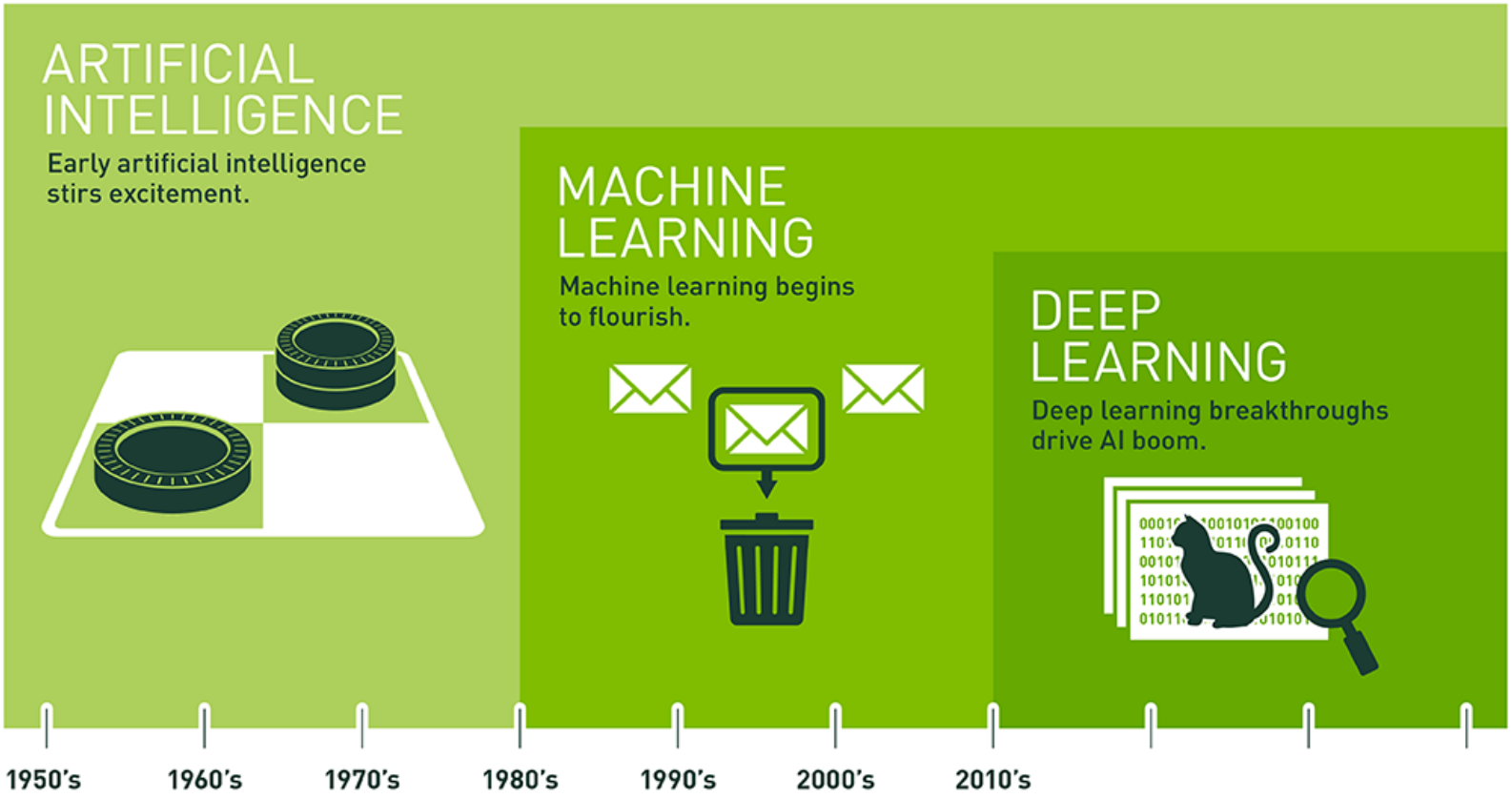
y es también lo que los hace **necesarios** …



LSST (Large Synoptic Survey Telescope)

Motivación

Los avances actuales han provocado un aumento de las **expectativas** que se tienen con respecto a lo que la inteligencia artificial puede lograr en los próximos años.



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine Learning en la Prensa

WIRED — Facebook Taps 'Deep Learning' Giant for New AI Lab

BUSINESS | CULTURE | GEAR | IDEAS | SCIENCE

CADE METZ BUSINESS 12.09.13 03:14 PM

FACEBOOK TAPS 'DEEP LEARNING' GIANT FOR NEW AI LAB

Twitter pays up to $150M for Magic Pony Technology, which uses neural networks to improve images

THE NEW YORKER

NEWS DESK

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

FOOD TECH

Artificial Intelligence Chef Wants To Disrupt Factory Farming With Innovative Vegan Products

A Chilean startup is using a machine-learning algorithm that cooks plant-based products, in a bid to disrupt the factory-farm food system.

WIRED

ROBERT MCMILLAN BUSINESS 03.13.13 06:30 AM

GOOGLE HIRES BRAINS THAT HELPED SUPERCHARGE MACHINE LEARNING

Google DeepMind pairs with NHS to use machine learning to fight blindness

'Deep learning' research company will use 1m anonymised eye scans to train a neural network to identify early signs of degenerative eye conditions

BIG DATA

Apple acquires machine learning startup Turi, formerly known as GraphLab and Dato

ARTIFICIAL INTELLIGENCE | INTEL | BIG DATA

Intel is paying more than $400 million to buy deep-learning startup Nervana Systems

The chip giant is betting that machine learning is going to be a big deal in the data center.

"Prensa" Científica:

SPECIAL SECTION

RISE OF THE MACHINES

By **Jelena Stajic, Richard Stone, Gilbert Chin,** *and* **Brad Wible**

Although most would agree that the average person is smarter than the average cat, comparing humans and machines is not as straightforward. A computer may not excel at abstract reasoning, but it can process vast amounts of data in the blink of an eye. In recent years, researchers in artificial intelligence (AI) have used this computational firepower on the scads of data accumulating online, in academic research, in financial records, and in virtually all walks of life. The algorithms they develop help machines learn from data and apply that knowledge in new situations, much like humans do. The ability of computers to extract personal information from seemingly innocuous data raises privacy concerns. Yet many AI systems indisputably improve our lives; for example, by making communication easier through machine translation, by helping diagnose illness, and by providing modern comforts, such as your smartphone acting as your personal assistant. This special issue presents a survey of the remarkable progress made in AI and outlines challenges lying ahead.

Many AI systems are designed for narrow applications, such as playing chess, flying a jet, or trading stocks. AI researchers also have a grander aspiration: to create a well-rounded and thus more humanlike intelligent agent. Scaling that research peak is daunting. But triumphs in the field of AI are bringing to the fore questions that, until recently, seemed better left to science fiction than to science: How will we ensure that the rise of the machines is entirely under human control? And what will the world be like if truly intelligent computers come to coexist with humankind?

The editors gratefully acknowledge the advice of Eric Horvitz (Microsoft) on the Reviews in this special issue.

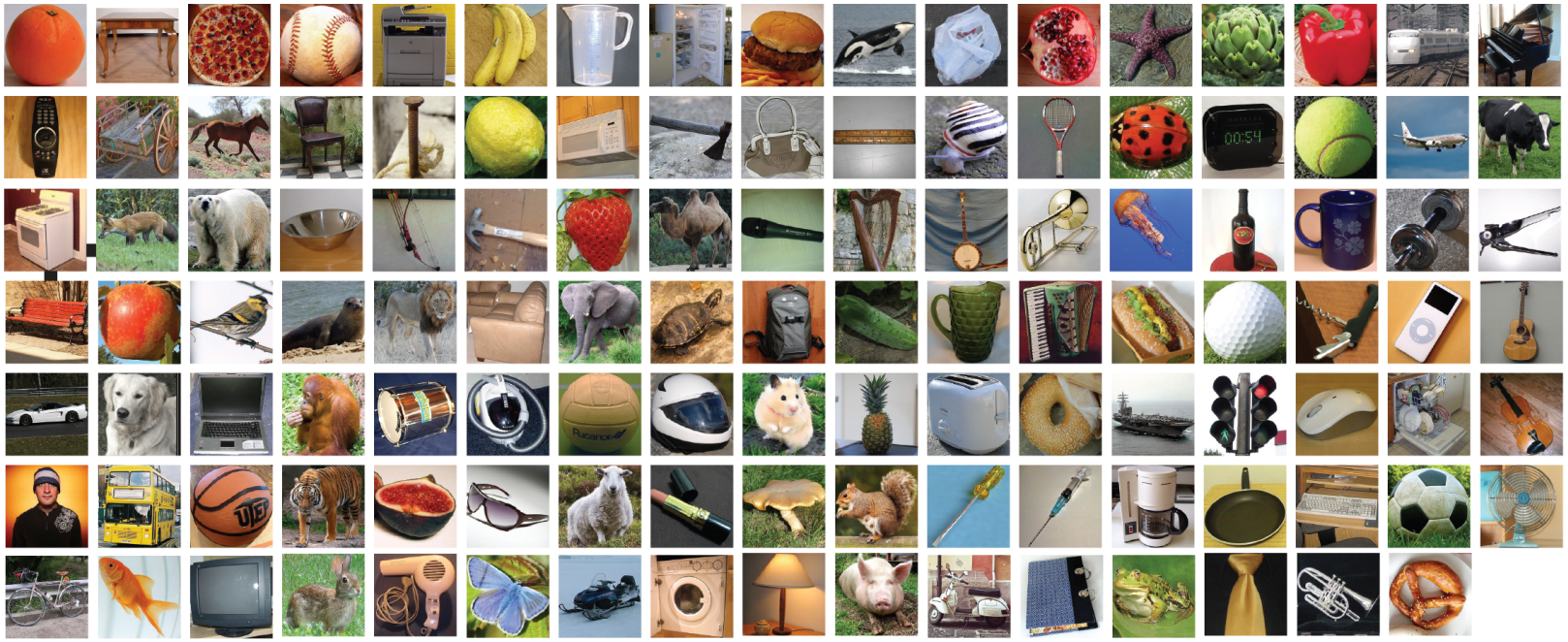
INSIDE

Science. Julio 2015.

Nature. Mayo 2017.

Avances Recientes

Desafío *Imagenet* (reconocer entre 22.000 tipos de objetos)



Accuracy humana: 94.5%

Estado del arte al 2010: 75%

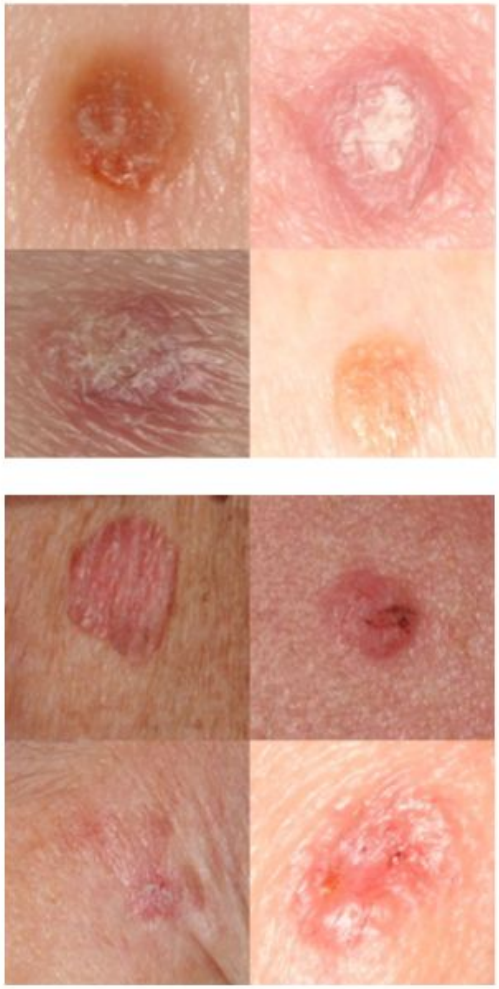
Algoritmo Resnet (Microsoft) al 2015: 96.43%*

* Ensamblado de redes neuronales profundas

Avances Recientes

Benign

Malignant



Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.5% *
CNN - PA	**72.0%** **

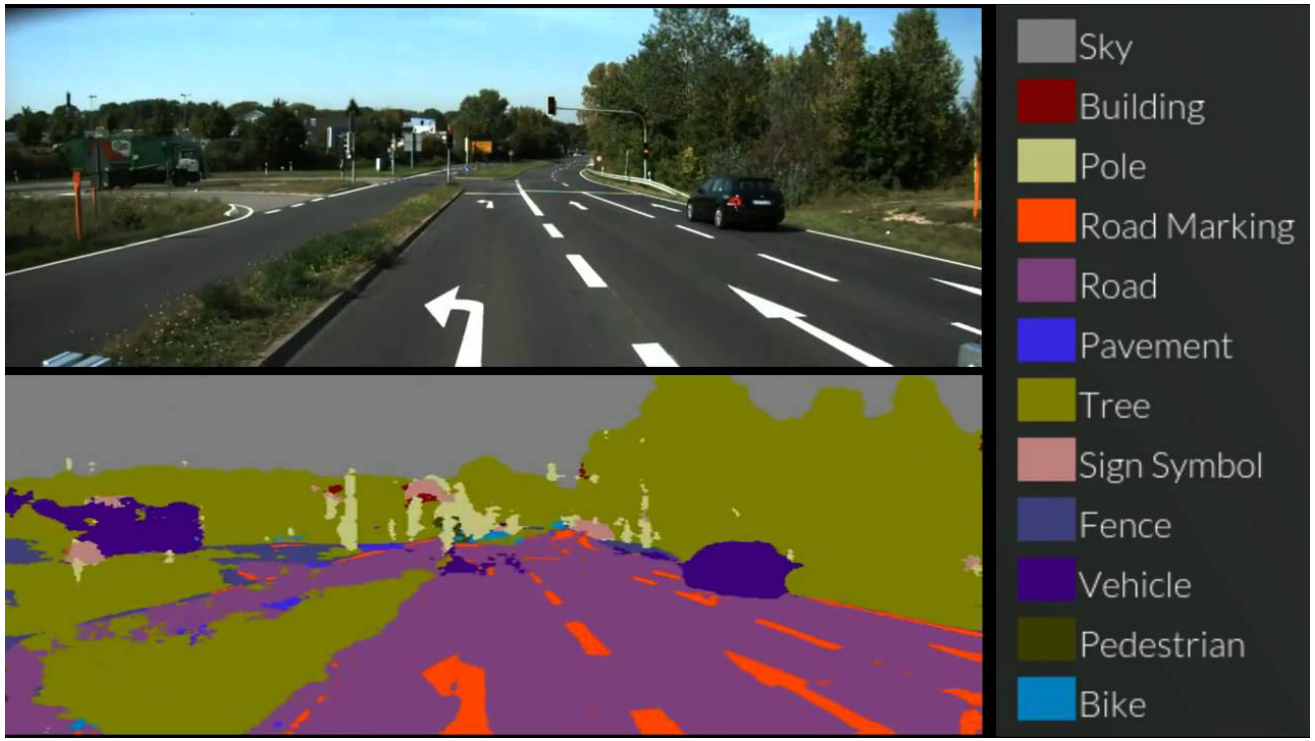
Disease classes:
three-way classification

0. Benign single lesions
1. Malignant single lesions
2. Non-neoplastic lesions

* Red neuronal profunda entrenada sobre 3 clases.

** Red neuronal profunda entrenada sobre 757 clases.

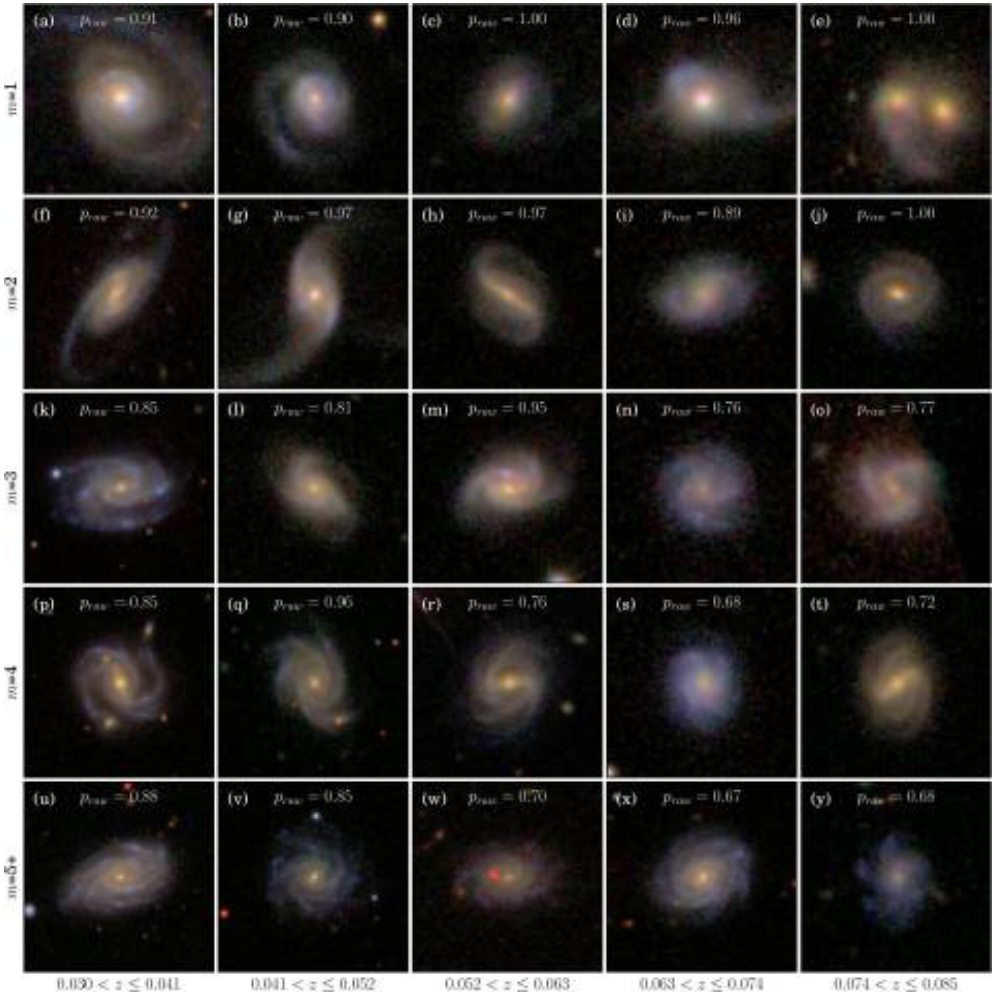
Avances Recientes



*

* Deep net
  + Reinforcement learning

Avances Recientes



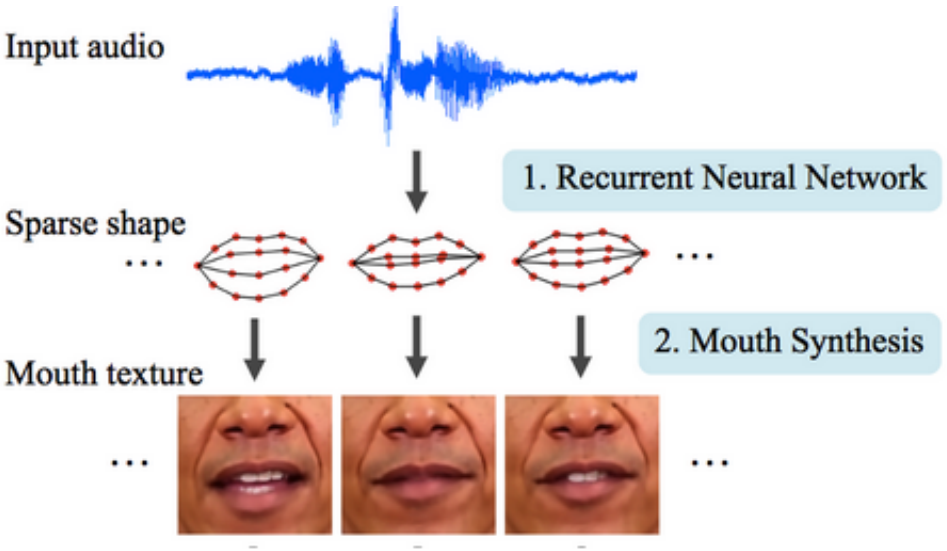
Accuracy SVM 88%
Accuracy Red Neuronal: 98%

Avances Recientes

El texto que describe esta imagen fue generado por una red neuronal artificial.

A woman is throwing a **frisbee** in a park.

Avances Recientes



Avances Recientes

Deep learning + Reinforcement learning

Advertencia

- Este no es un curso de programación, ingeniería de software o negocios. En este curso no veremos los métodos como cajas negras o como funciones de una librería. Nos interesan los fundamentos teóricos y conceptuales detrás de las técnicas.

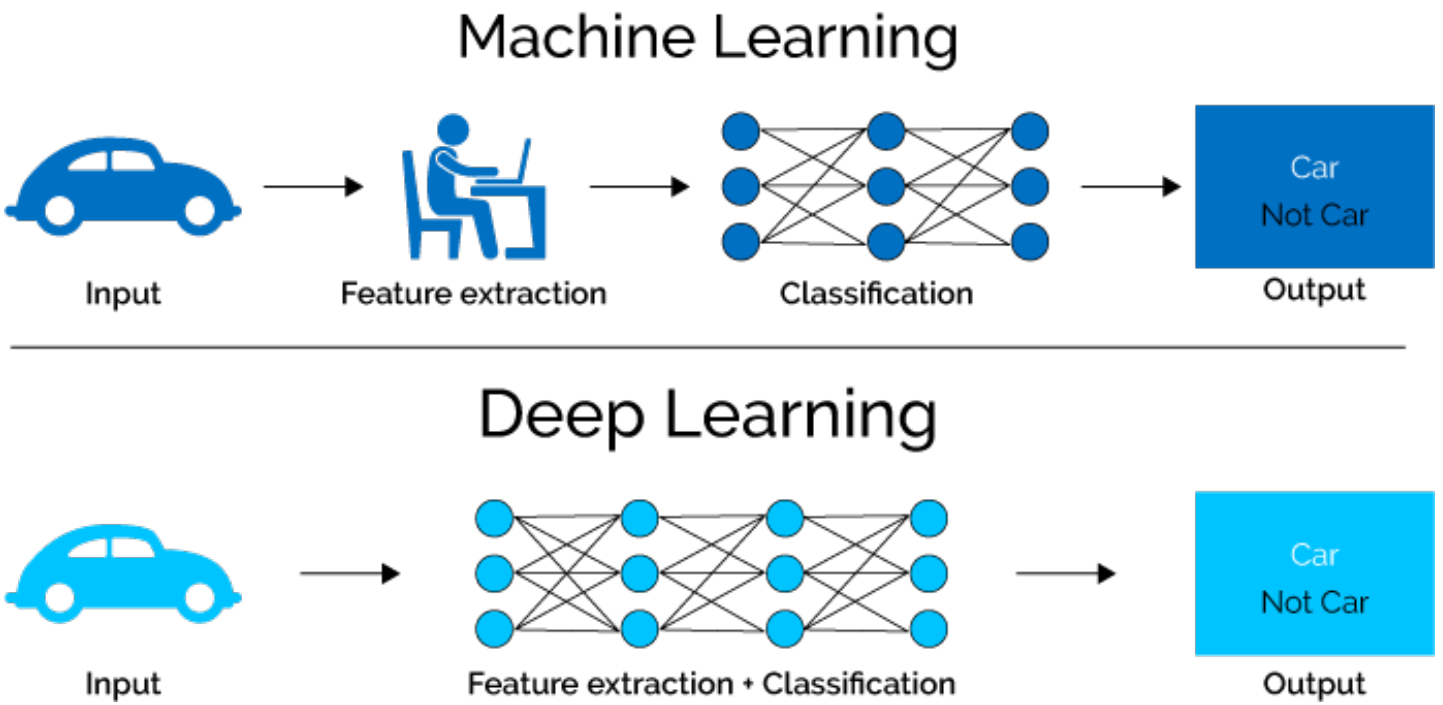
Advertencia

- Sin desmedro de lo anterior, en este curso nos interesa que aprendan a resolver problemas de análisis de datos usando las técnicas que estudiaremos. Para ello complementaremos las discusiones teóricas con ejemplos prácticos y talleres que representan como mínimo el 60% de la nota final (el 40% restante también incluye pequeños problemas prácticos).

Advertencia

- Este es un curso introductivo al área. Veremos los métodos más clásicos y una breve introducción a temas más avanzados. En particular, este no es un curso de deep learning. Un curso más especializado sobre ese tema es INF395 o INF477 (redes neuronales artificiales) que se dicta los primeros semestres de cada año.

Deep learning vs Classic ML



Deep learning vs Classic ML

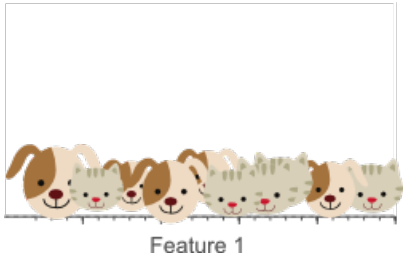
A Few Useful Things to Know about Machine Learning

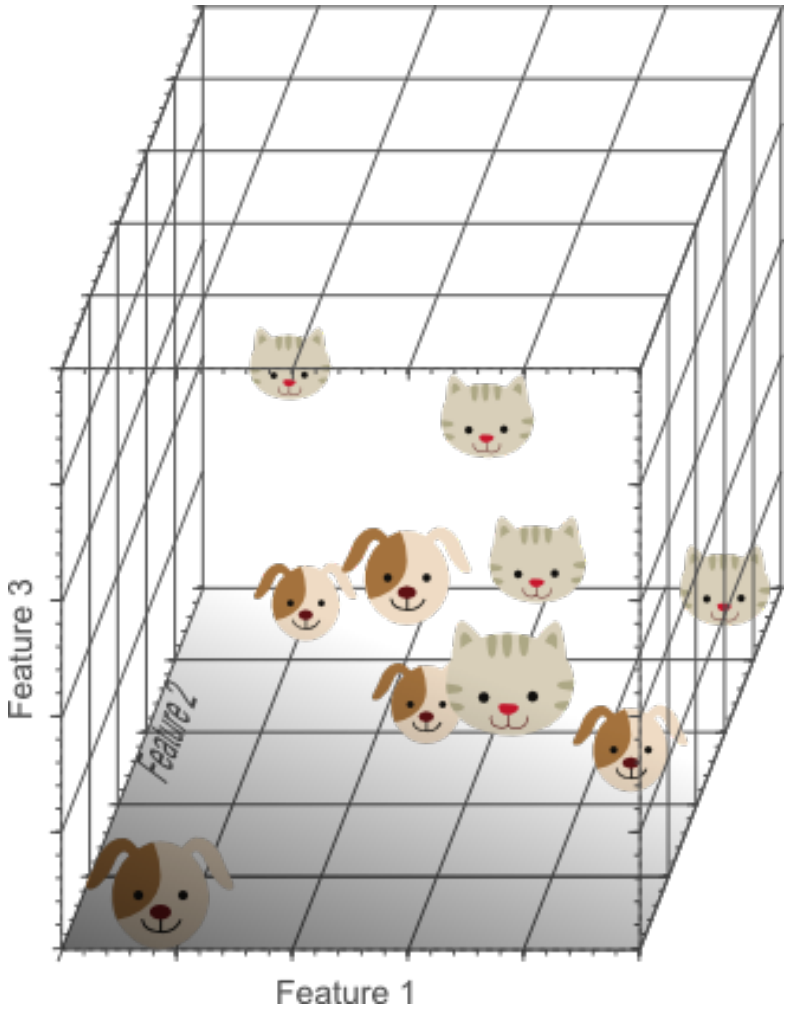
Pedro Domingos
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350, U.S.A.
pedrod@cs.washington.edu

8. FEATURE ENGINEERING IS THE KEY

At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. If you have many independent features that each correlate well with the class, learning is easy. On the other hand, if the class is a very complex function of the features, you may not be able to learn it. Often, the raw data is not in a form that is amenable to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes. It is often also one of the most interesting parts, where intuition, creativity and "black art" are as important as the technical stuff.

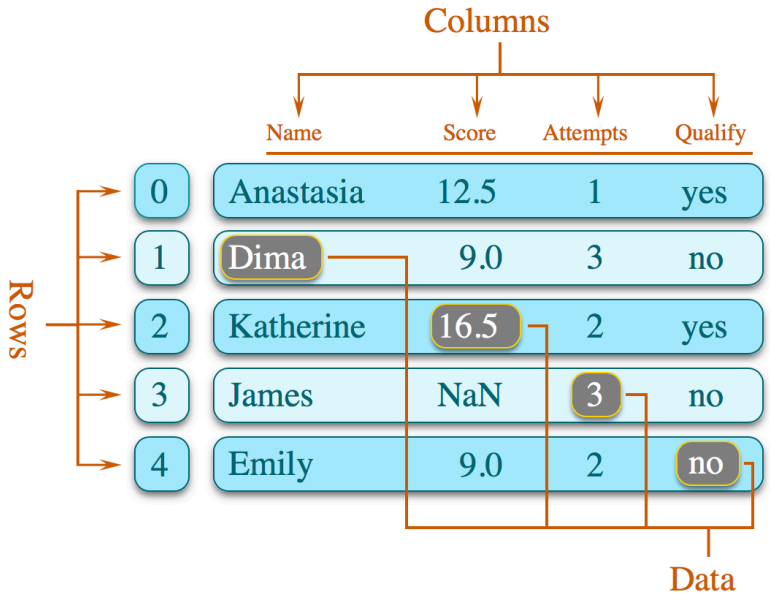
Deep learning vs Classic ML





Temas de Talleres

- Dataframes



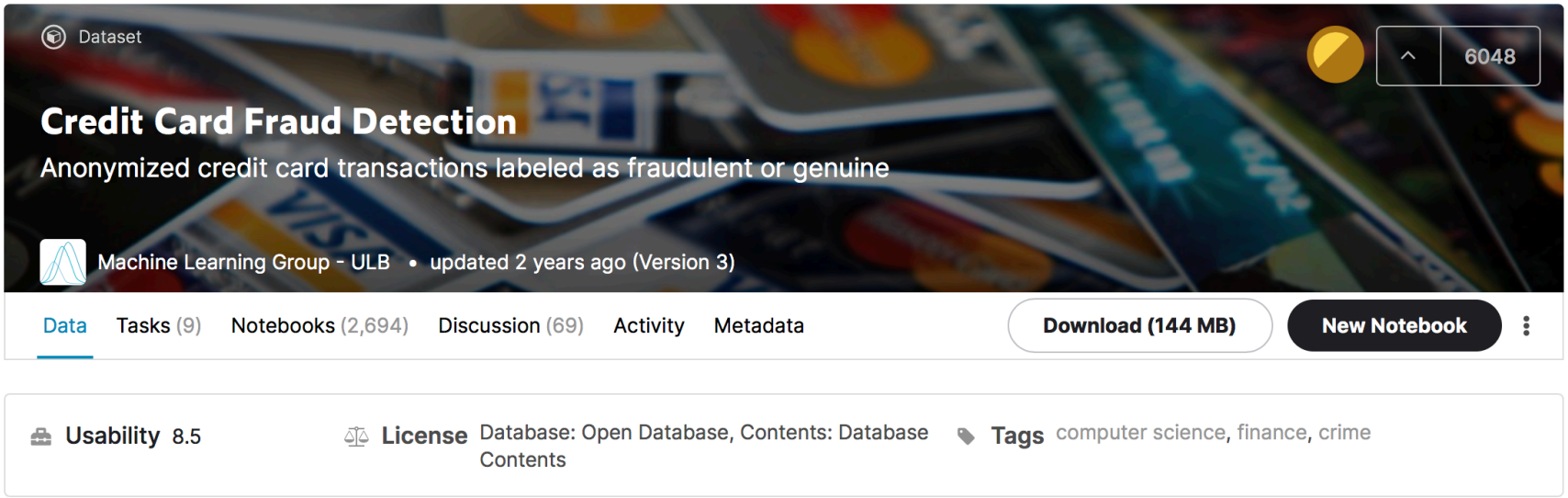

```
In [305]: ventas = pd.DataFrame({
              "Producto":["A", "B", "C", "B", "A", "A"],
              "Ventas": [6, 2, 1, 4, 5, 2]
          })
          ventas
```

Out[305]:

	Producto	Ventas
0	A	6
1	B	2
2	C	1
3	B	4
4	A	5
5	A	2

Temas de Talleres

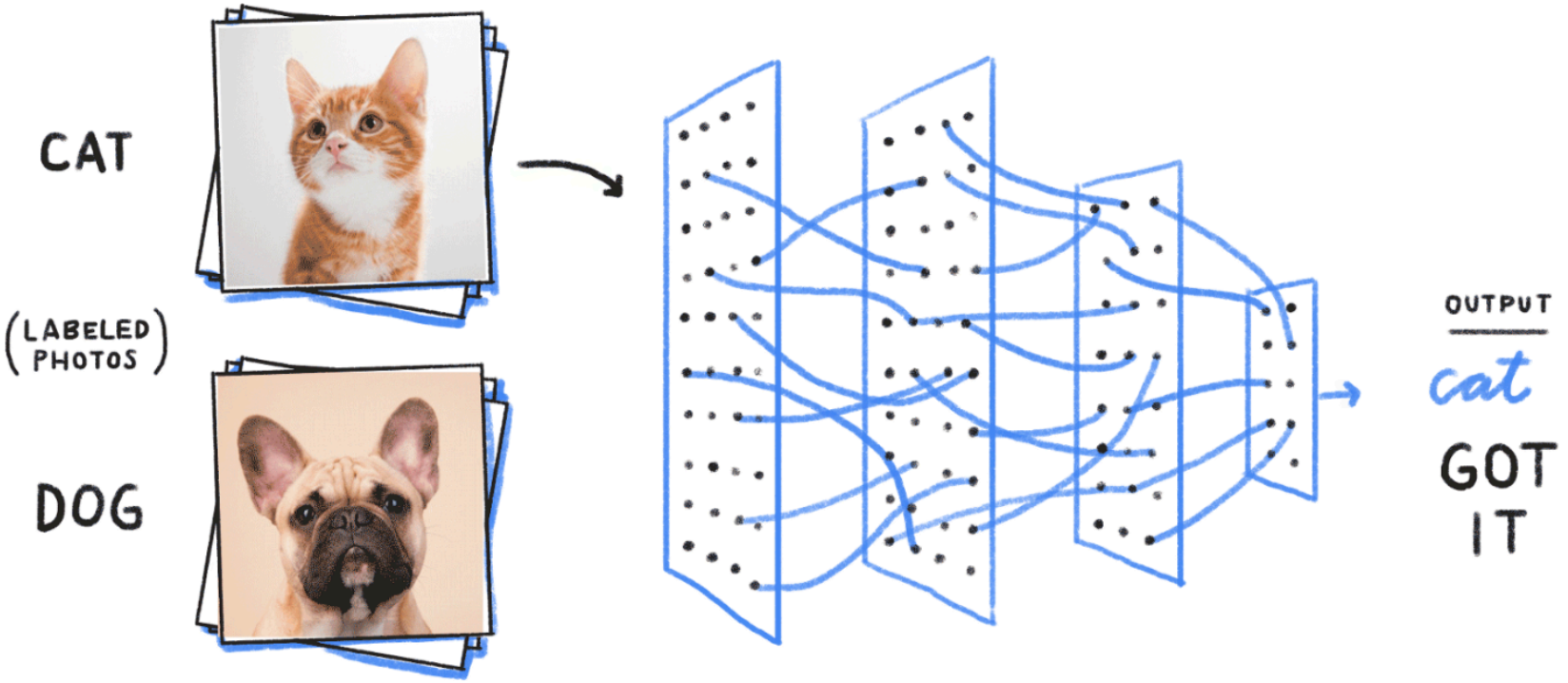
- Contextos interesantes / desafiantes



https://www.kaggle.com/

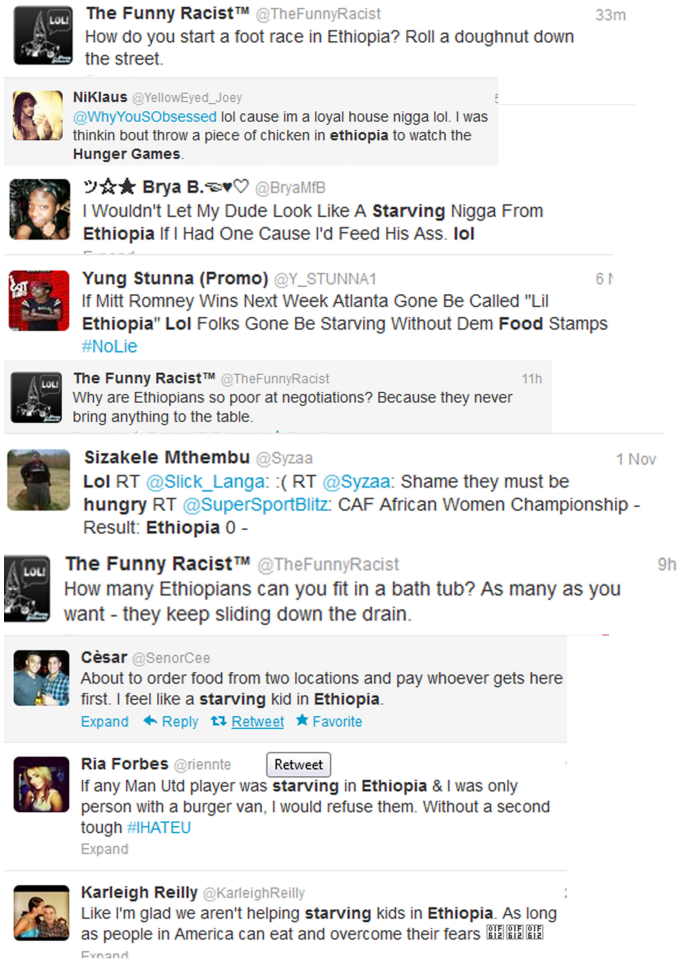
Temas de Talleres

- Manipulación de imágenes.

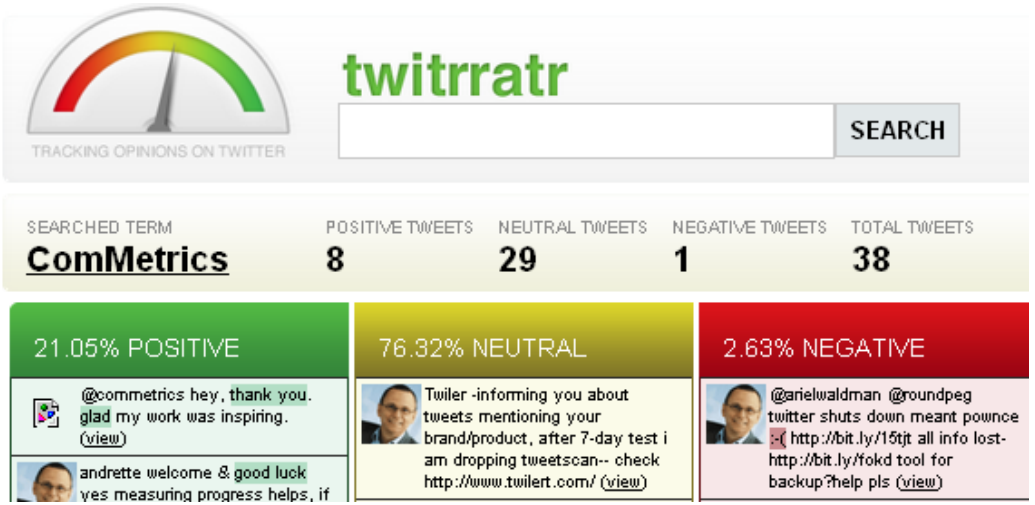


Temas de Talleres

- Manipulación de Texto.

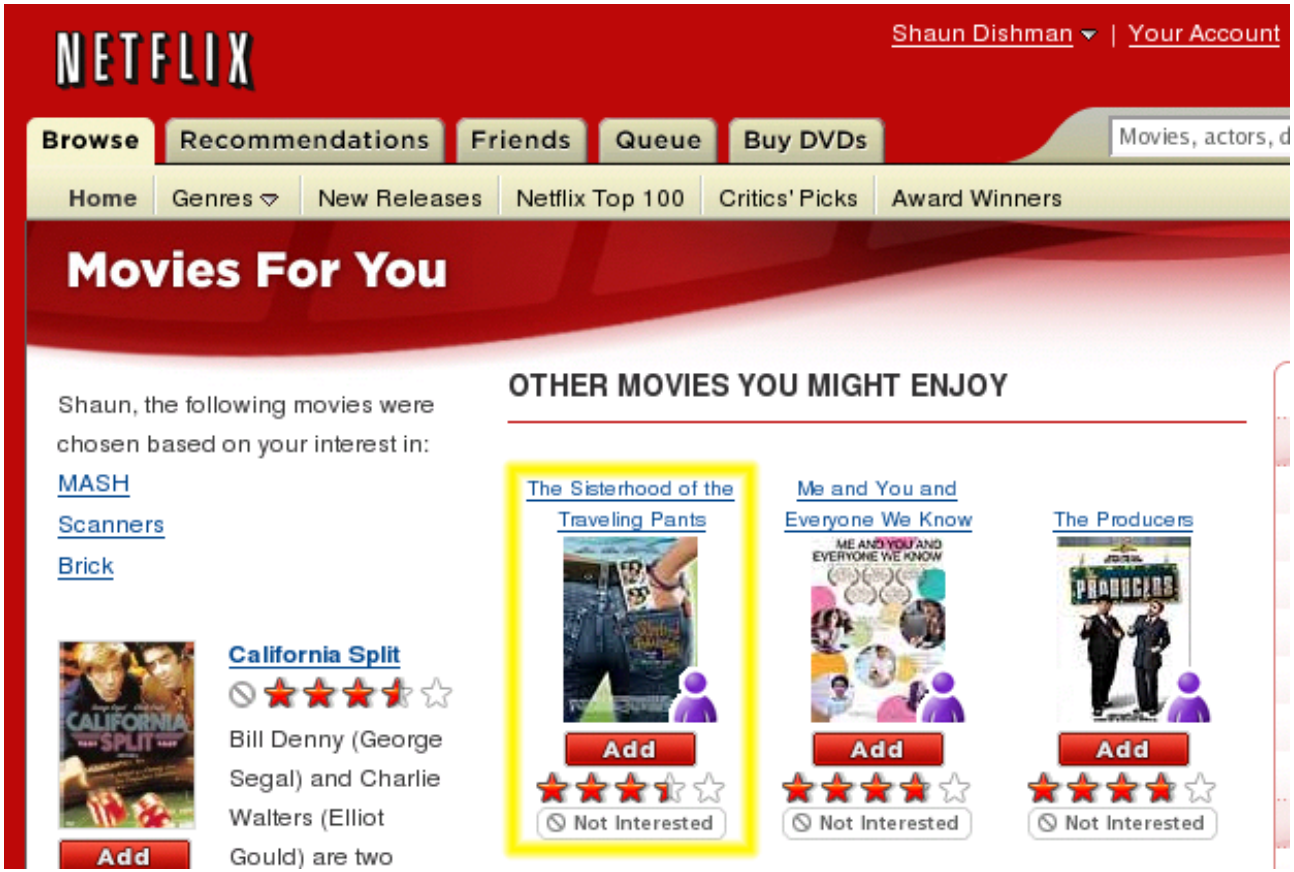


Sentiment Analysis



Temas de Talleres

- Filtrado colaborativo

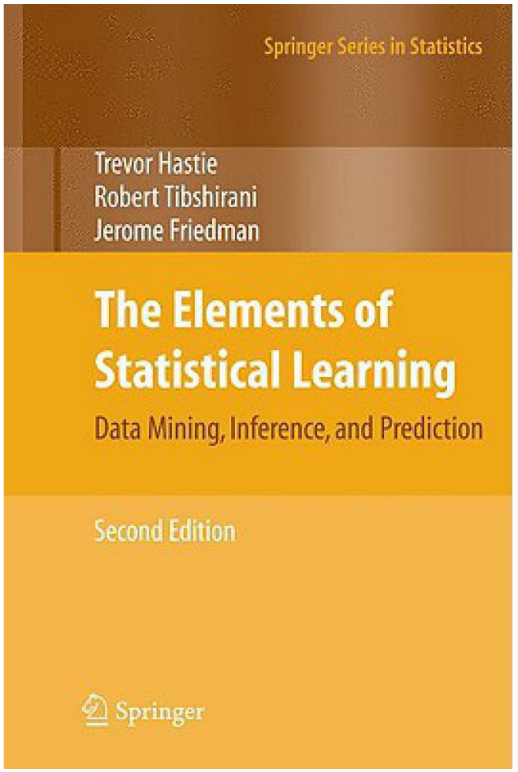


Syllabus 2020

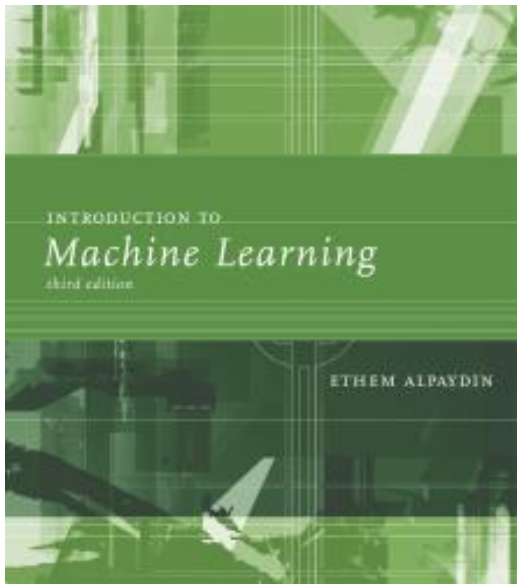
- Revisar lista de temas y calendario en syllabus.
- Evaluado mediante tareas y quices.
- Quices conceptuales y/o pequeño ejercicio con código.
- Talleres: tareas en equipos en que se aborda una pregunta de investigación o un problema práctico específico. En ambos casos, se propone una metodología de solución, se experimenta sobre datos, se definen métricas y se elaboran conclusiones.
- Defensas: presentaciones orales exponen soluciones y resultados. Todas las entregas incluyen un jupyter notebook y un video explicativo de <=25 minutos.
- Evaluaciones cruzadas de pares.

Referencias

- T. Hastie et al. *The elements of statistical learning.*

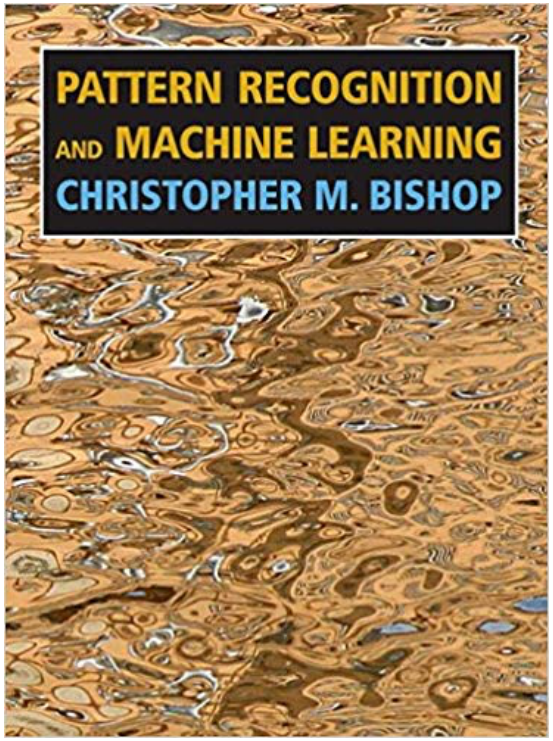


- E. Alpaydin. *Introduction to Machine Learning.*



Referencias

- C. Bishop *Pattern Recognition and Machine Learning.*



- E. Alpaydin. *Introduction to Machine Learning.*