

Aprendizaje Automático

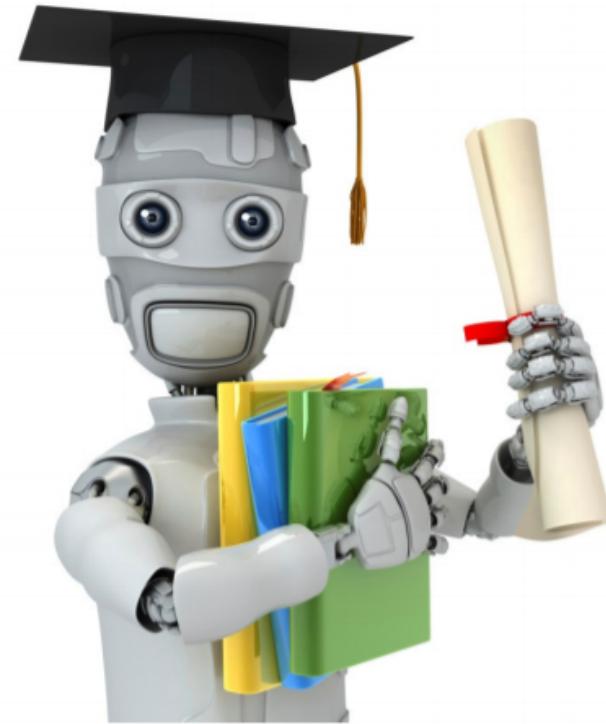
Tecnológico de Costa Rica

Programa de Ciencia de Datos

Frans van Dunné

Agenda

- **8:00 – 9:30**
 - Introducción al curso
 - Introducción a ML
 - CRISP-DM
 - Introducción al reconocimiento de patrones
 - Ejemplos y aplicaciones.
- **9:20 – 9:35**
 - Pauza
- **9:35 – 12:00**
 - Etapas de un sistema de reconocimiento de patrones.
 - Tipos de aprendizaje.
 - Discusión y cierre del dia



Introducción al curso

- **Hola yo soy ...**
 - Mi expectativa del curso es
 - Creo poder aplicar ML en ...



Frans van Dunné
Chief Data Oficer @ixpantia
Profesor @CTEC

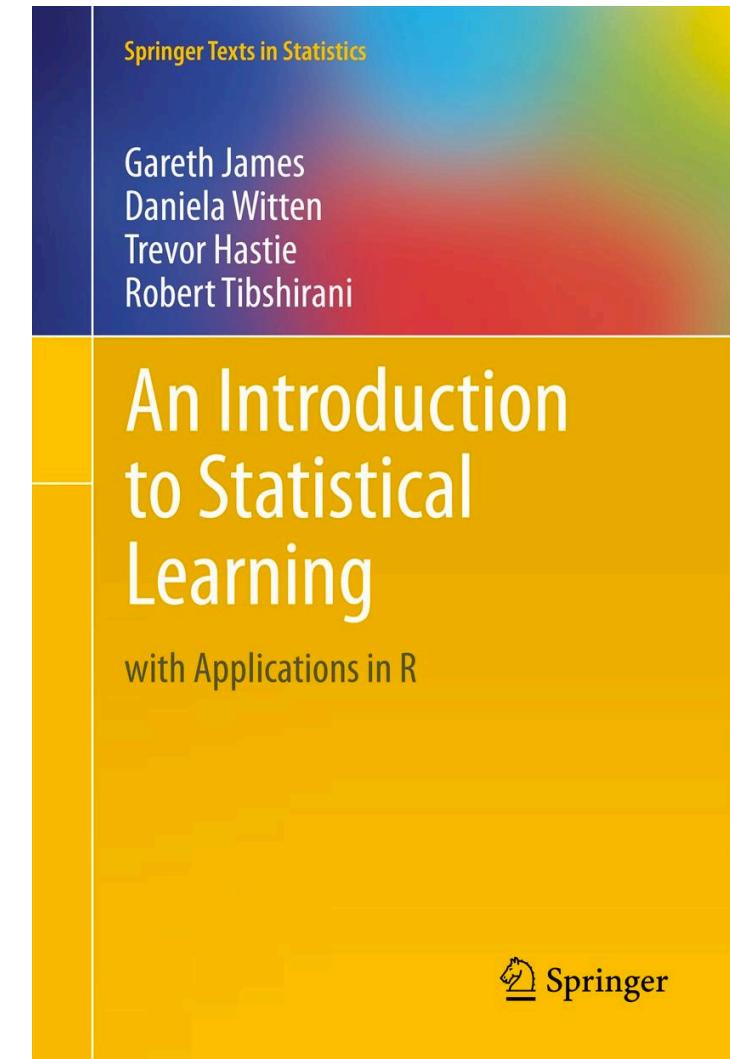
Twitter:
[@fransvandunne](https://twitter.com/fransvandunne)

Email:
frans@ixpantia.com

Introducción al curso

- **Herramientas**
 - R y Rstudio
 - git
 - Github
 - Ejemplos
 - Ejercicios y tareas
- **Fuentes información**
 - TEC Digital
 - Capita Selecta
 - Libro Introduction to Statistical Learning

bit.ly/ctec_ml_libro



Introducción

- Qué es el aprendizaje automático?

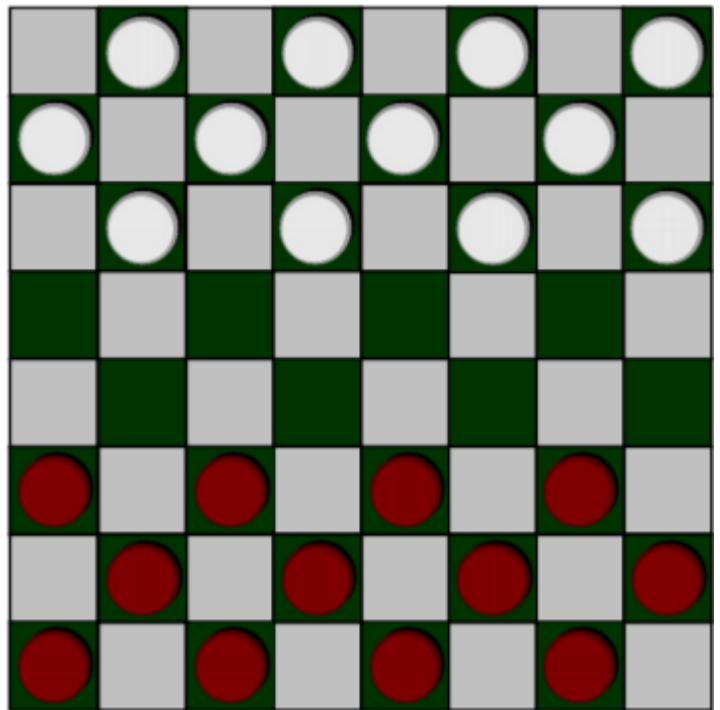
- Tom Mitchell proporciona una definición muy interesante y moderna: "Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de rendimiento P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E. "



Tom Michael Mitchell (born August 9, 1951) is an American computer scientist and E. Fredkin University Professor at the Carnegie Mellon University (CMU).

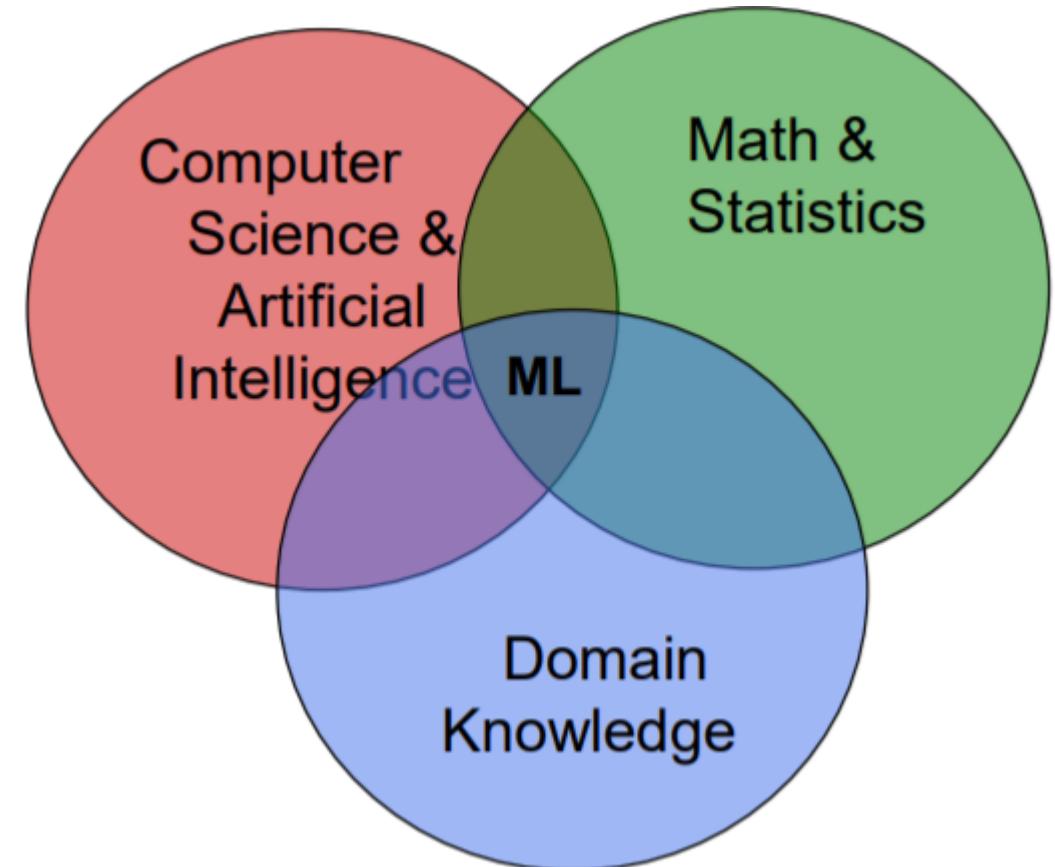
Introducción

- Ejemplo: jugar a las damas.
 - E = la experiencia de jugar muchos juegos de damas
 - T = la tarea de jugar a las damas.
 - P = la probabilidad de que el programa gane el siguiente juego.
- En general, cualquier problema de aprendizaje automático puede asignarse a una de dos clasificaciones generales:
 - **Aprendizaje supervisado y aprendizaje no supervisado.**



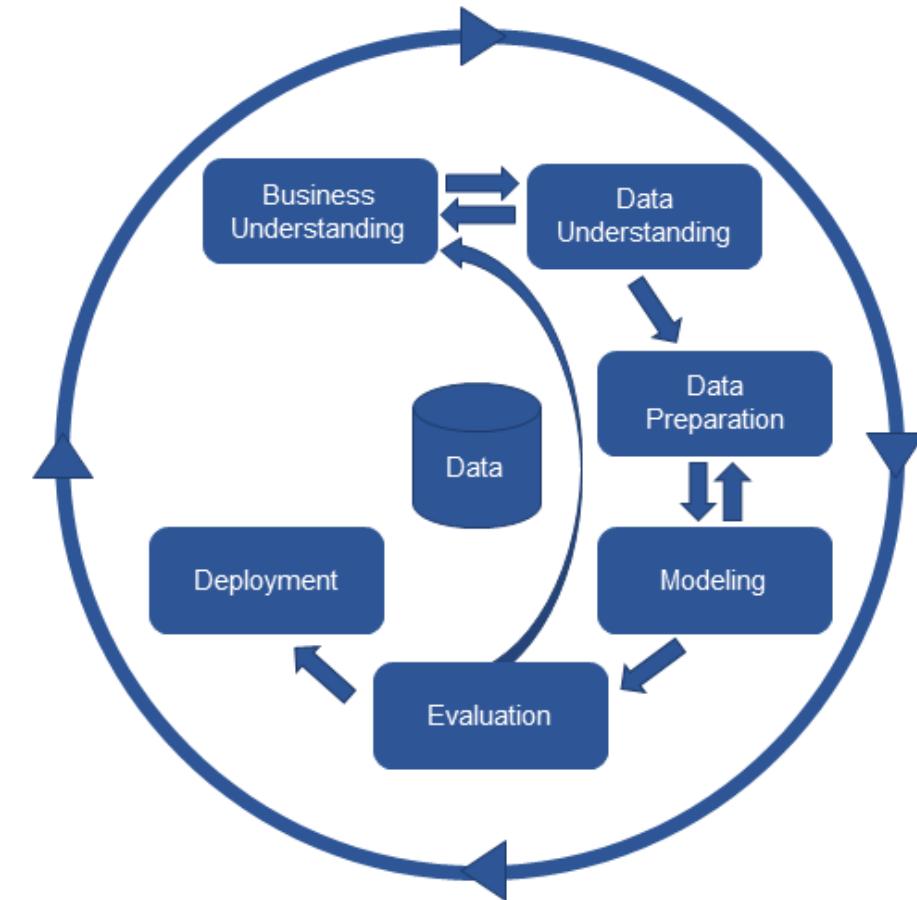
Aprendizaje automático es:

- Aprender de los datos
- No hay programación explícita
- Descubriendo patrones ocultos
- Decisiones basadas en datos



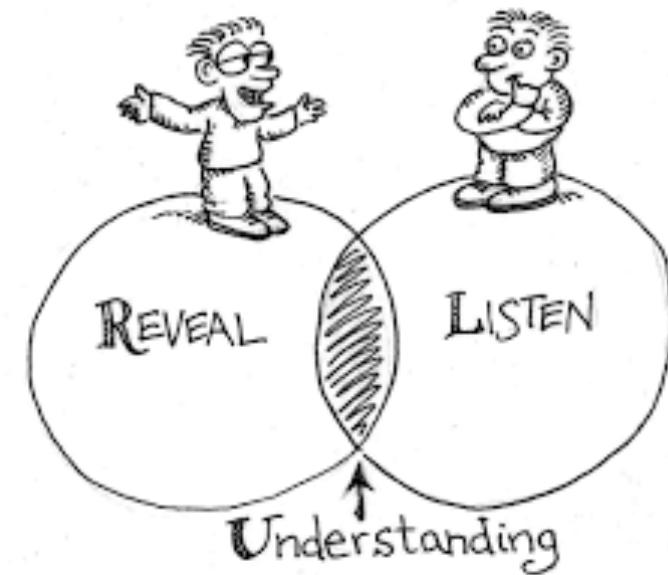
Cross Industry Standard Process for Data Mining (CRISP-DM)

- Entendimiento de negocios
- Comprensión de datos
- Preparación de datos
- Modelado
- Evaluación
- Despliegue



Fase 1 – Entendimiento del negocio

- Definir problema u oportunidad
- Evaluar la situación
- Formular objetivos

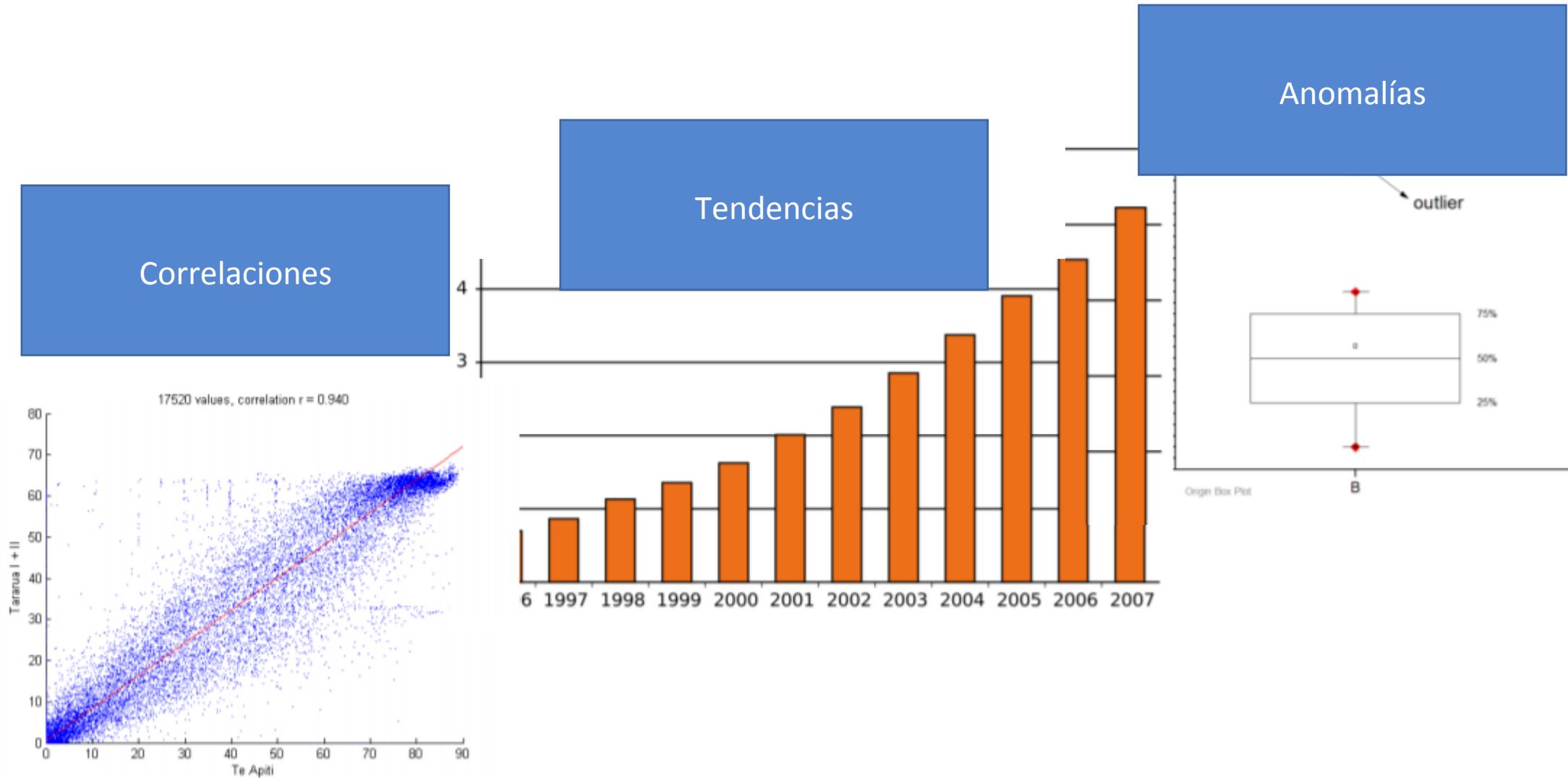


Fase 2 - Comprensión de los datos

- Adquisición de datos
- Exploración de datos
- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

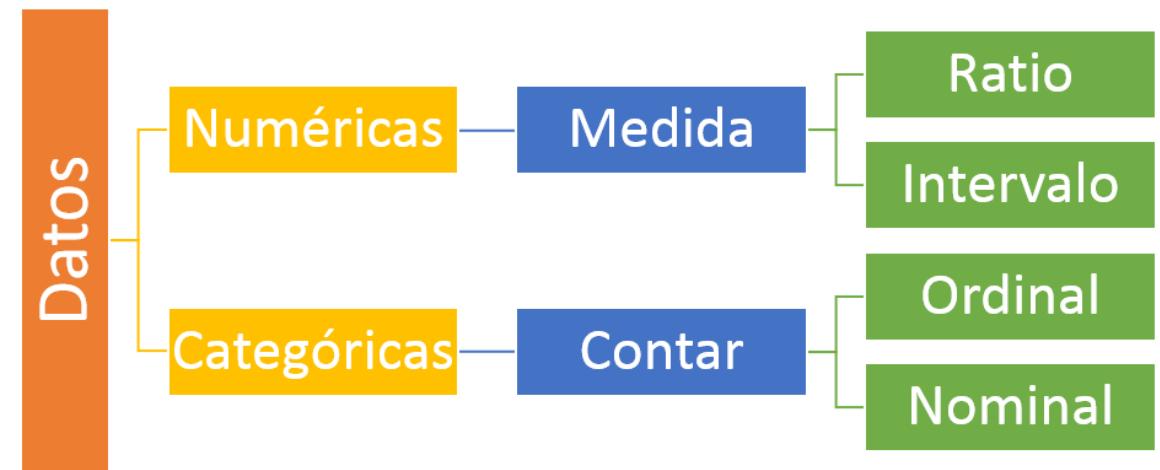


¿Por que explorar?

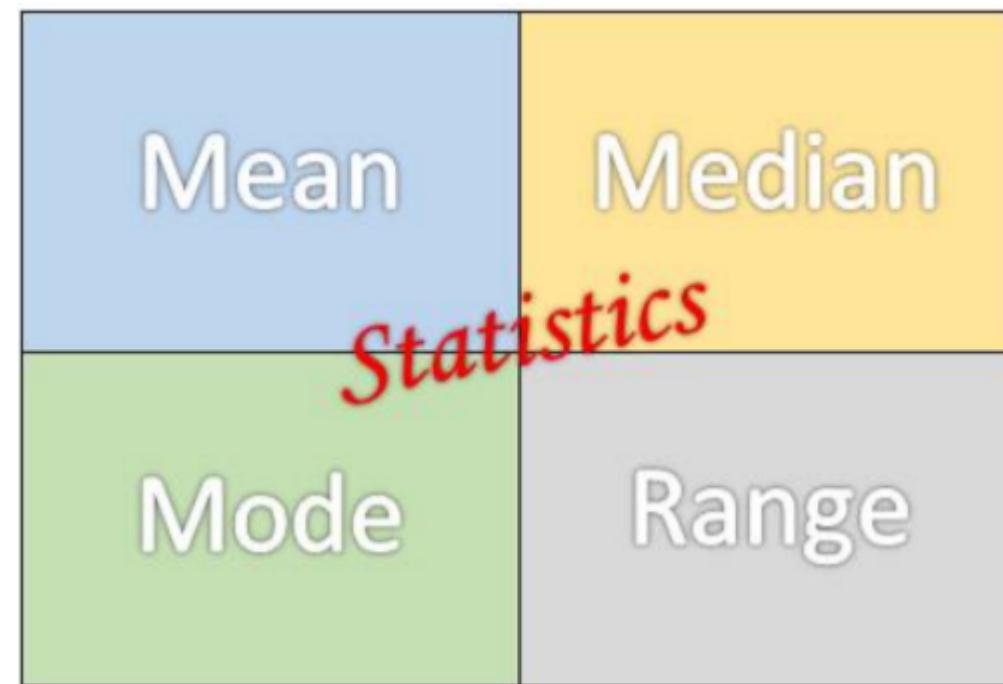
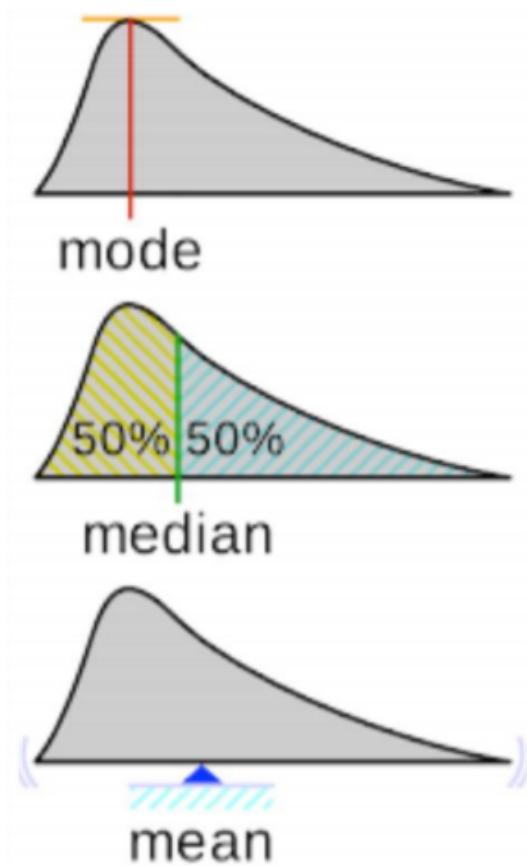


Fase 3 - Preparación de datos

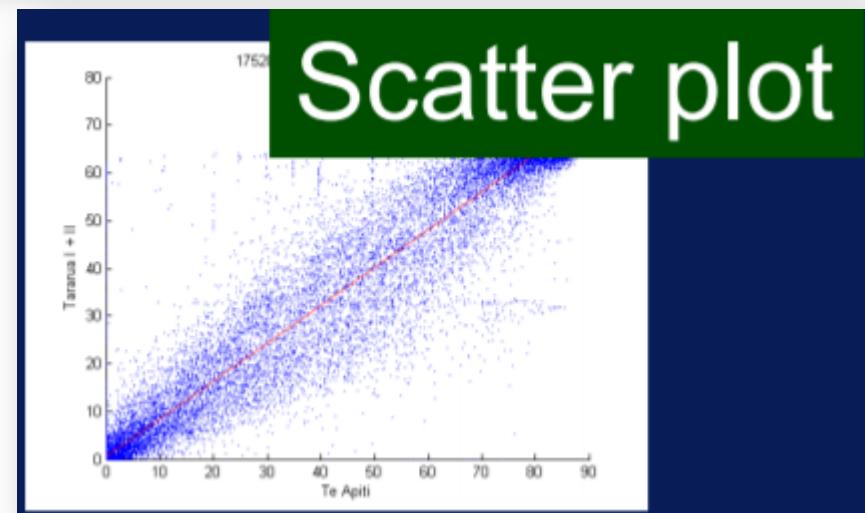
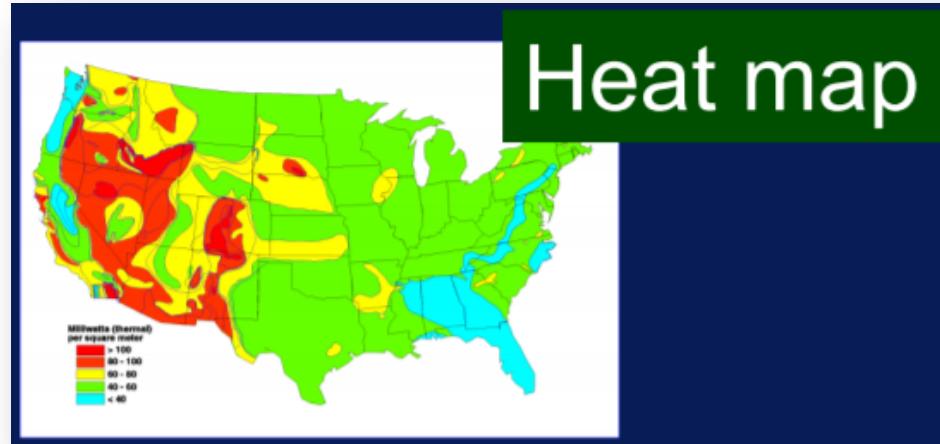
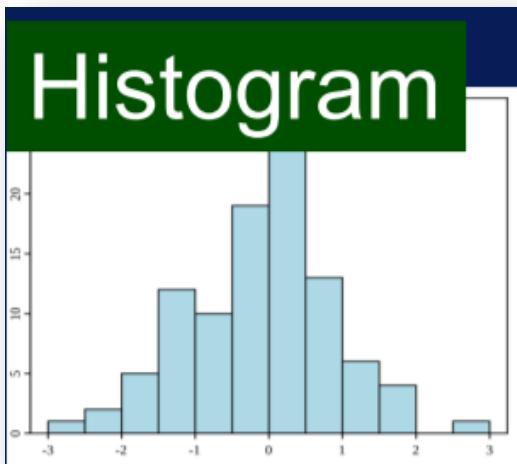
- Preparar datos para el modelado.
- Abordar problemas de calidad, seleccionar características utilizar, procesar datos para modelar



Describe tus datos



Visualiza tus datos

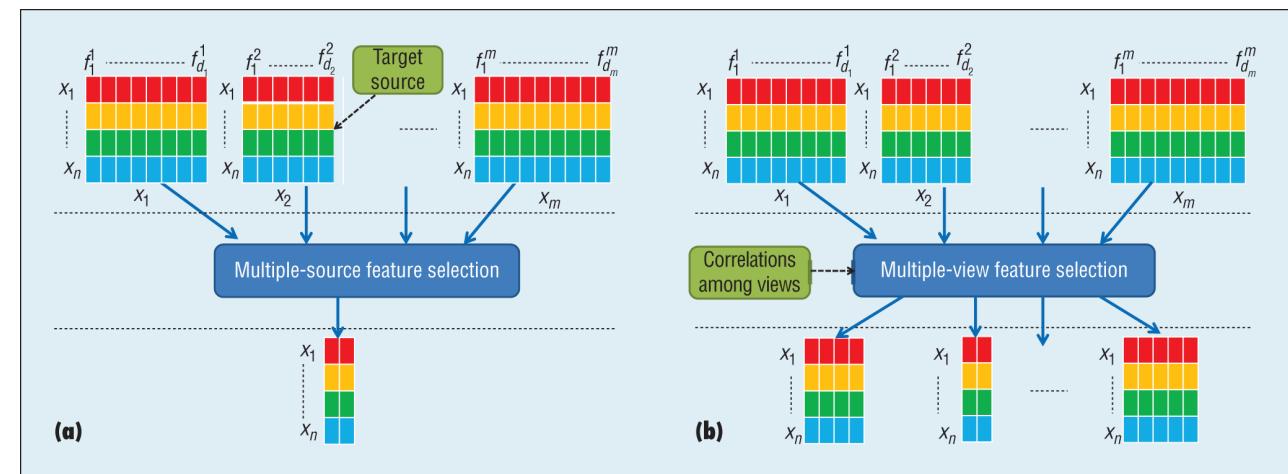


Pauza

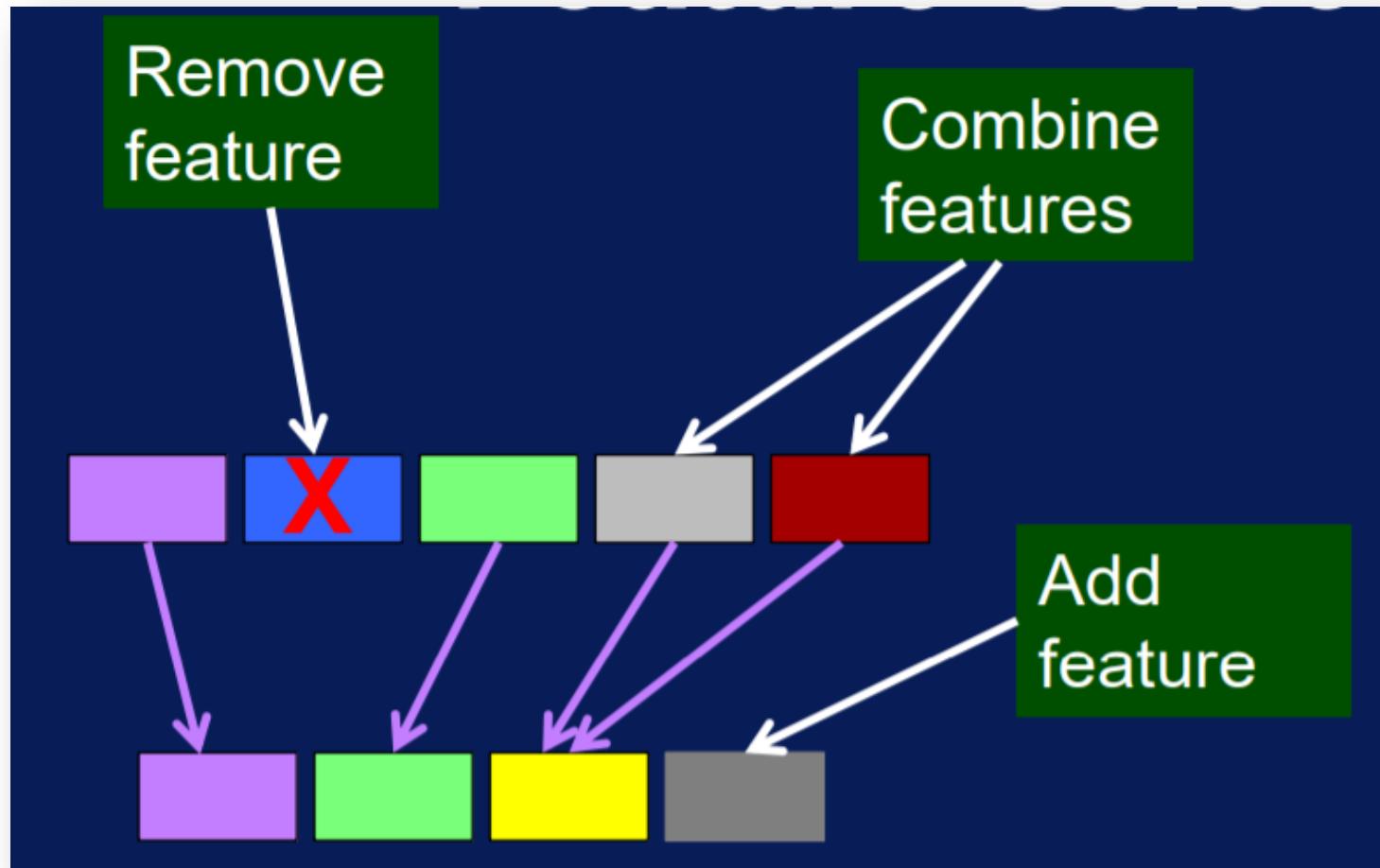
9:20 – 9:35

Limpieza de datos

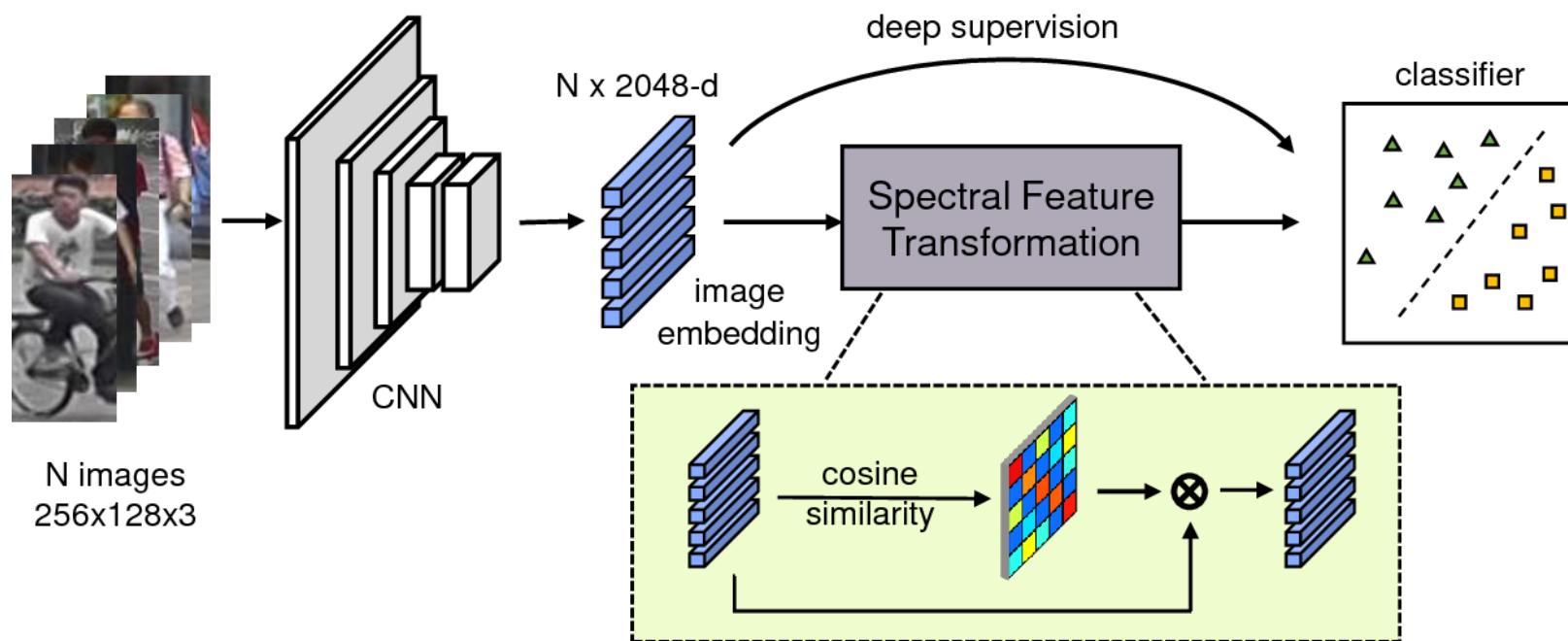
- Valores faltantes
- Datos duplicados
- Datos inconsistentes
- Ruido
- Datos anómalos
- Selección de los datos
- Construcción de datos
- Integración de datos
- Formateo de datos



Selección de características

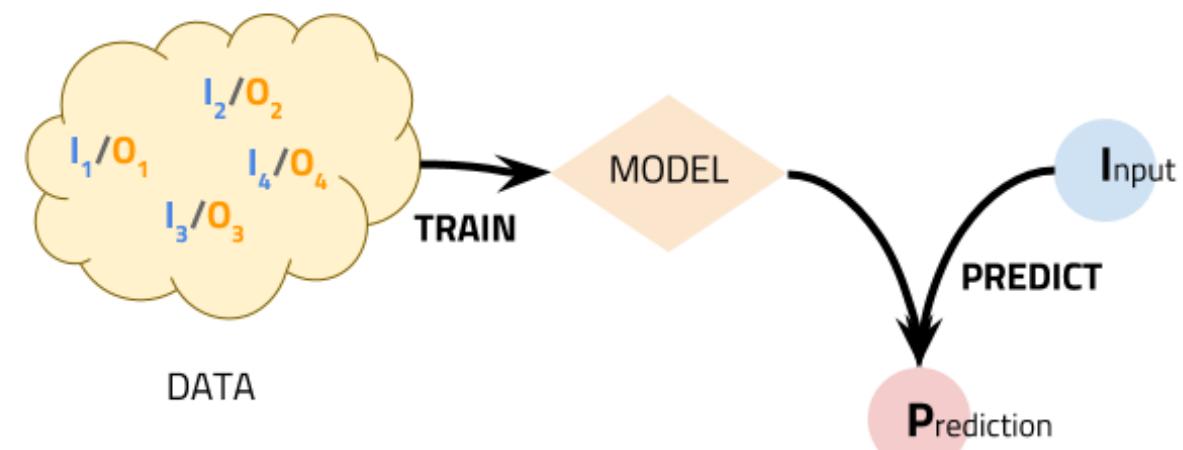


Transformación de la características



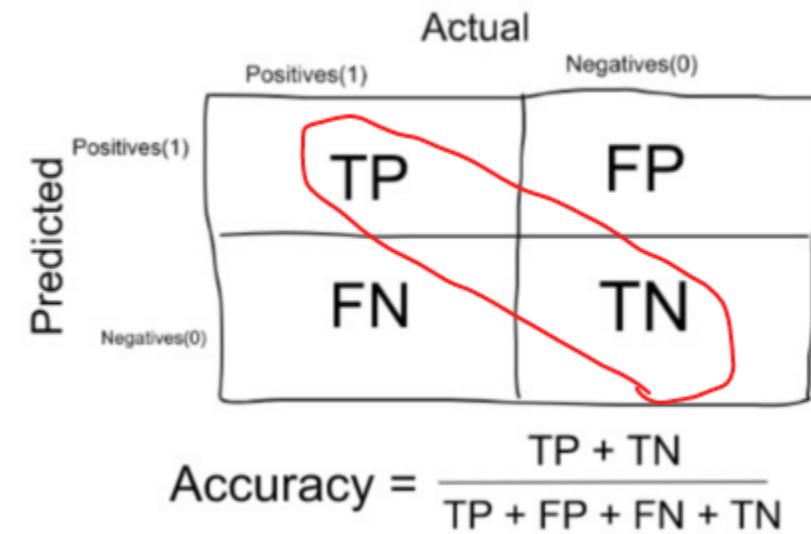
Fase 4 - Modelado

- Determinar el tipo de problema
- Seleccionar técnicas de modelado para usar
- Construcción del modelo
- Diseño de la evaluación



Fase 5 - Evaluación

- Evaluar el rendimiento del modelo
- Evaluar los resultados del modelo con respecto a los criterios de éxito
- Probabilidades en los resultados.



- Dataset:
 - 10 cats
 - 90 dogs

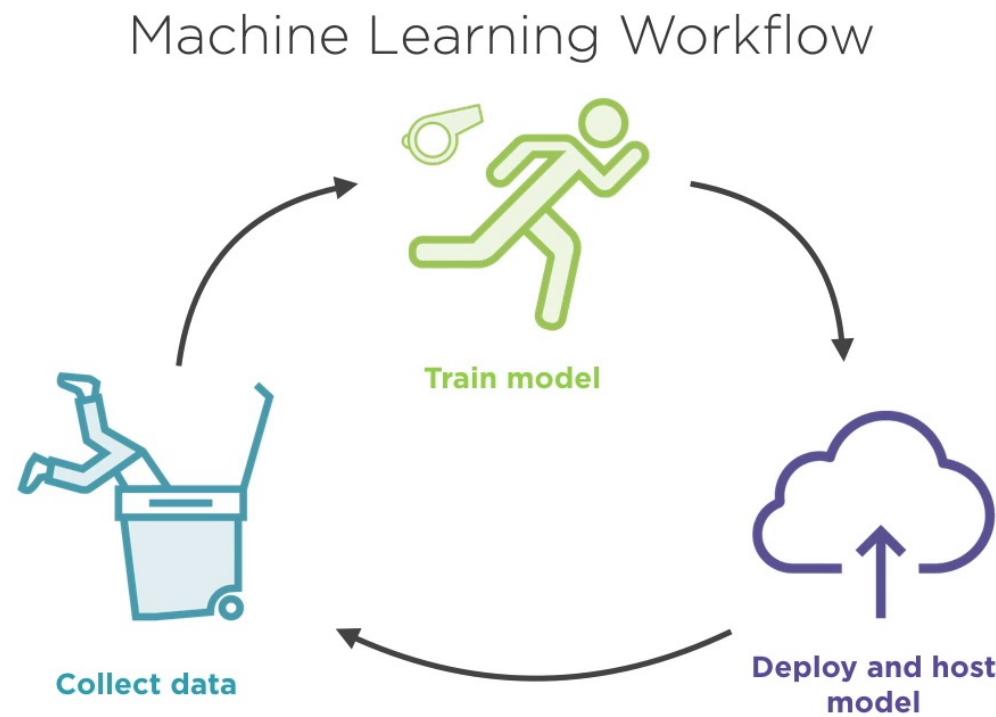
Predict always dog:
Accuracy = **0.9!**

Fase 6 - Deploy

- Producir informe final.
- Implementar modelo
- Monitorear el modelo generado



Proceso Iterativo

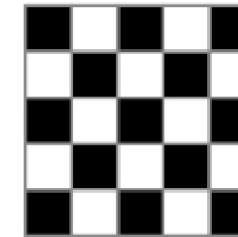


Intermezzo

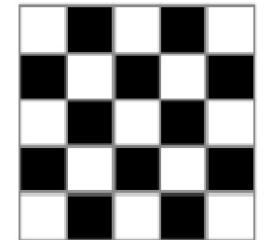
- Abrir tarea en maquina local
 - Bifurcar repositorio del curso
 - Clonar tu bifurcacion en Rstudio
 - Validar que todo esta.
- De paso
 - Tienes la ultima version de R (3.6.1)
 - Tienes la ultima version de Rstudio?

Introducción a reconocimiento de patrones

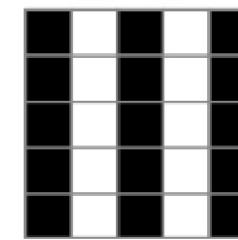
- Es una disciplina científica tiene el objetivo de clasificar, predecir, agrupar objetos en un número específico de categorías o valores.
- Dependiendo de la aplicación, estos objetos pueden ser imágenes, sonidos, olores, en general, señales producto de mediciones que deben ser clasificadas. Estos objetos se denotan con el término genérico de *patrones*.



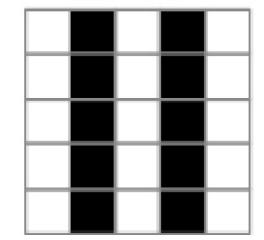
Pattern type 1, version 1



Pattern type 1, version 2



Pattern type 2, version 1



Pattern type 2, version 2

Patrones supervisado y no supervisado

- El reconocimiento supervisado asume la existencia a priori de ejemplos de clasificación que relacionan a los valores de las características con ciertas categorías.
- Sin embargo, hay situaciones en el que no se disponen de ejemplos de clasificación a priori, en te caso el reconocimiento se llama no supervisado y tienen como objeto revelar las relaciones intrínsecas de similitud entre los valores de las características creando clusters.



Problemas de reconocimiento de patrones

- **Selección de variables.**

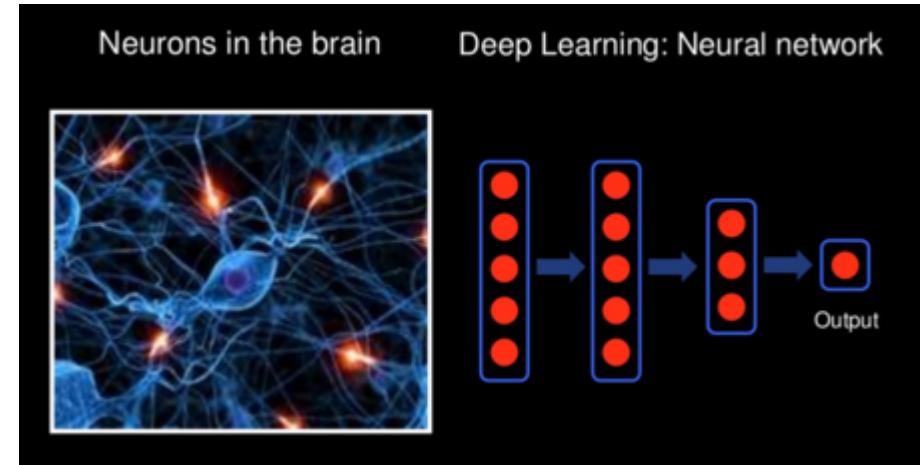
- Consiste en determinar cual es el conjunto de características más adecuado para describir a los objetos.

- **Clasificación supervisada.**

- Consiste en clasificar nuevos objetos basándose en la información de una muestra ya clasificada.

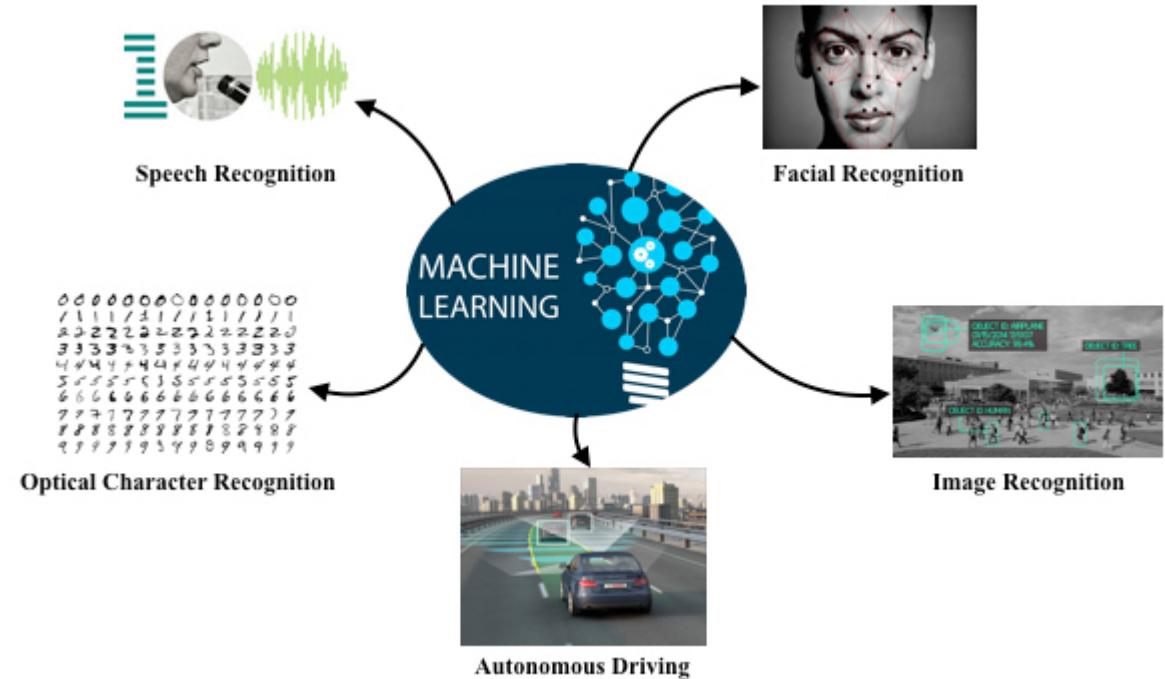
- **Clasificación no supervisada.**

- Consiste en dada una muestra no clasificada encontrar la clasificación de la misma.



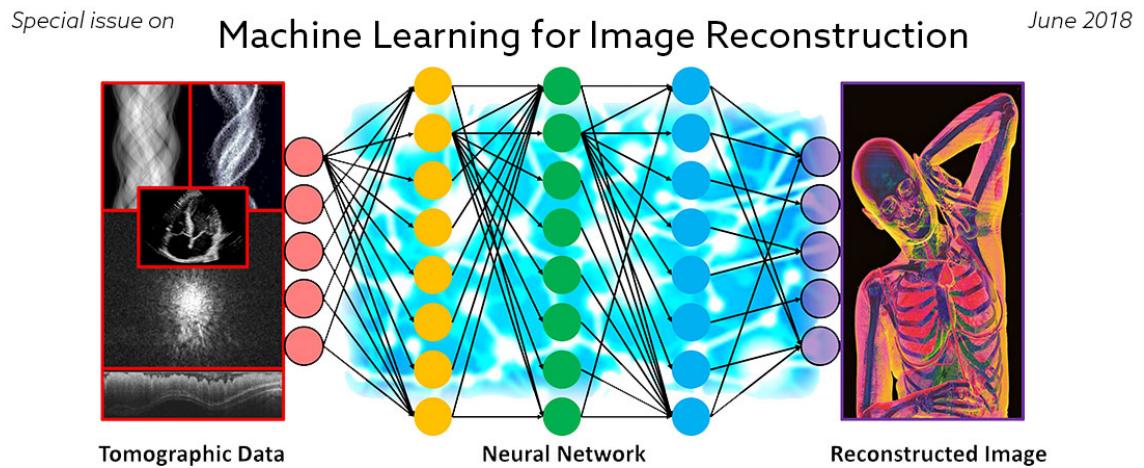
Ejemplos de reconocimiento de patrones

- Parámetros para reconocer la cara de una persona.
- Agrupación de palabras para dar sentido en una frase.
- Clasificar un fraude bancario.
- Predecir el precio de una casa.



Ejemplos de reconocimiento de patrones

- Anuncios dirigidos en aplicaciones móviles
- Análisis de los sentimientos
- Vigilancia del clima
- Detección de patrones de crimen.
- Análisis de efectividad de medicamentos.



Principales aplicaciones

- **Visión de máquina.**
 - Esta aplicaciones tienen que ver con la captura de imágenes con ayuda de cámaras digitales y la interpretación automática de lo que esta en la imagen
- **Reconocimiento de caracteres**
 - Transformación de textos impresos o manuscritos a formato digital
- **Diagnóstico por computadora**
 - Sistemas que ayudan a los doctores en el diagnóstico de enfermedades a partir de datos médicos, tales como las mamografías de Rayos X
- **Reconocimiento de voz**
 - La construcción de máquinas que puedan reconocer la información hablada



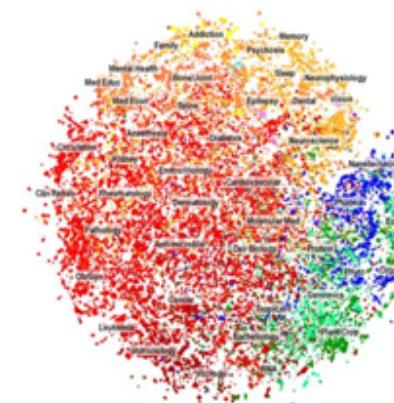
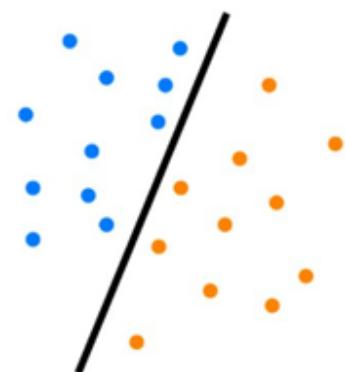
Aprendizaje supervisado y no supervisado

Supervised

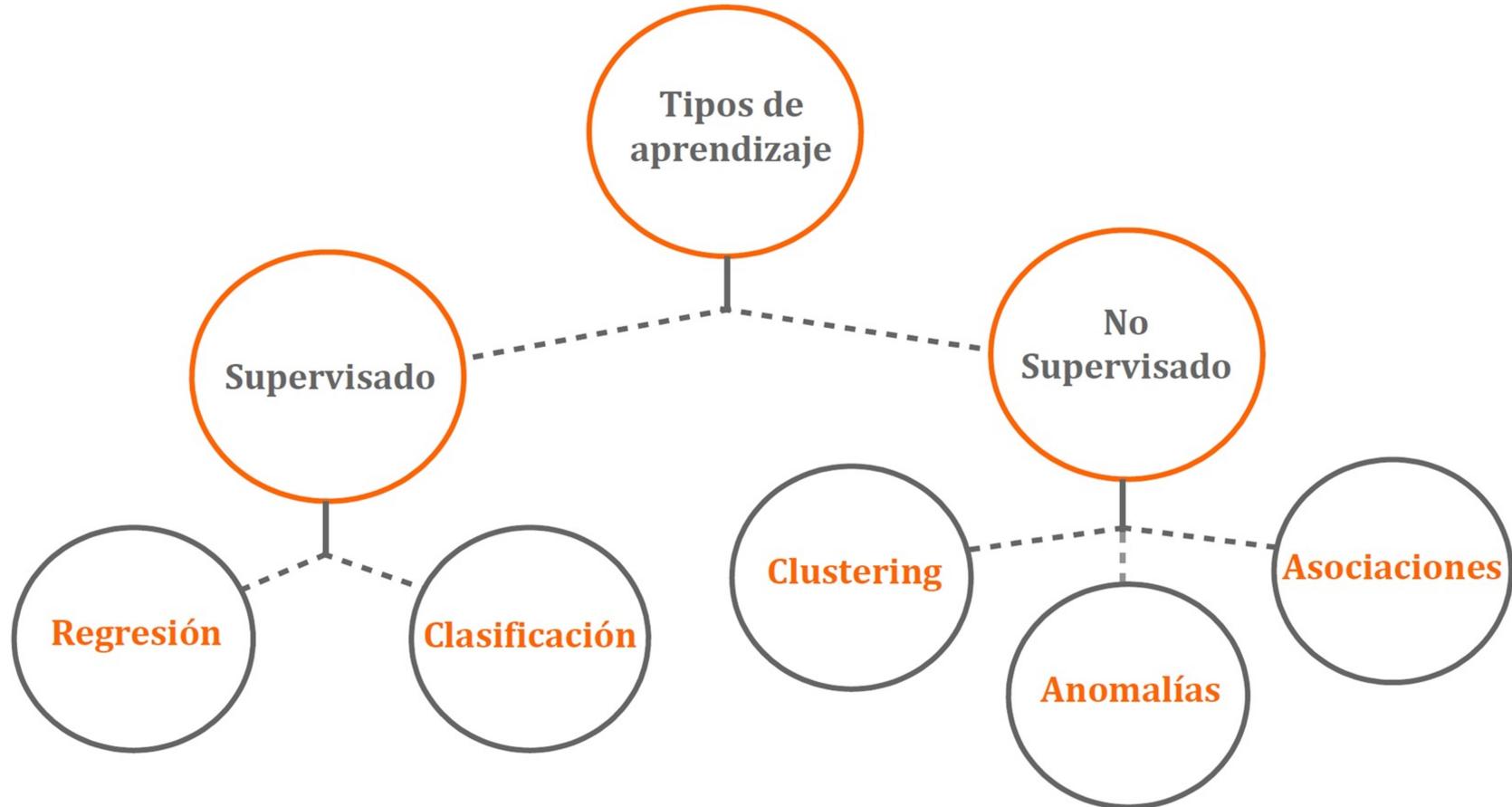
Unsupervised

Learning known patterns

Learning unknown patterns

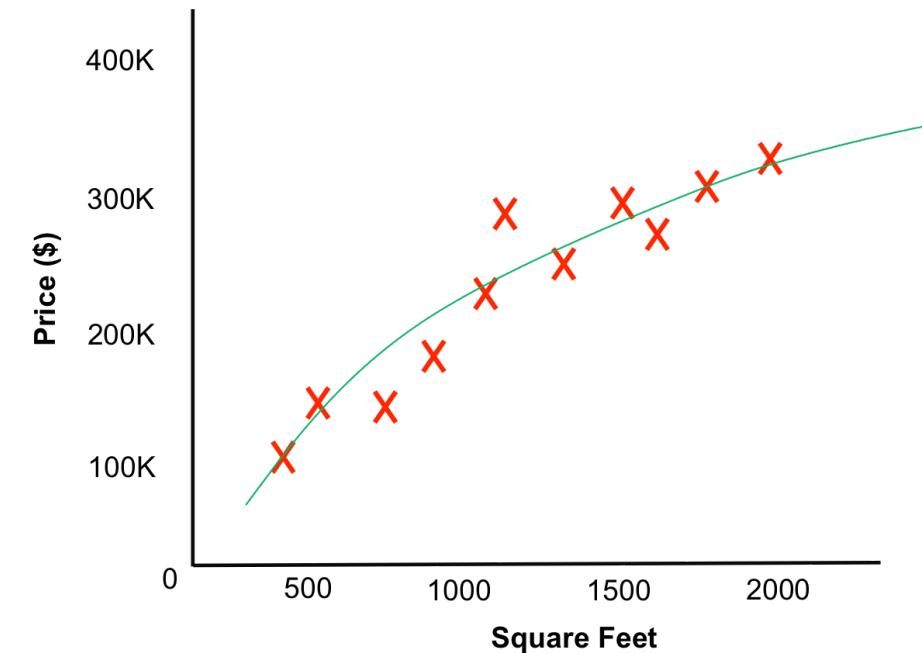


Aprendizaje supervisado y no supervisado



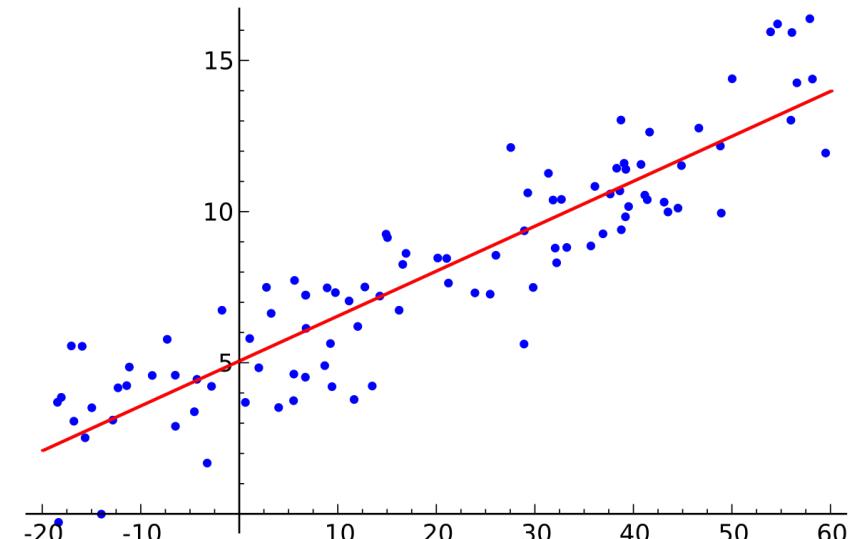
Algunos algoritmos supervisados

- Árboles de decisión
- Clasificación de Naïve Bayes
- Regresión por mínimos cuadrados
- Regresión Logística
- Support Vector Machines (SVM)
- Métodos “Ensemble” (Conjuntos de clasificadores)



Regresión

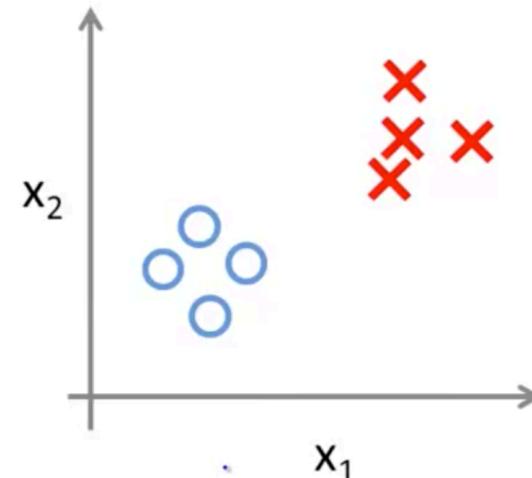
- Estimar la demanda de un producto basado en la época del año
- Predecir puntuación en un examen
- Determinar la probabilidad de la efectividad de una droga para un paciente
- Predecir la cantidad de lluvia



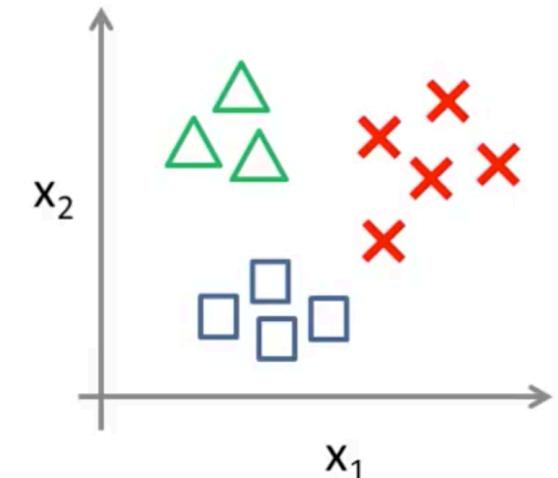
Clasificación

- Clasificar un tumor en benigno o maligno.
- Predecir si lloverá mañana
- Determine si la solicitud de préstamo es alta, riesgo medio o bajo
- Identificar el sentimiento como positivo, negativo, o neutral

Binary classification:

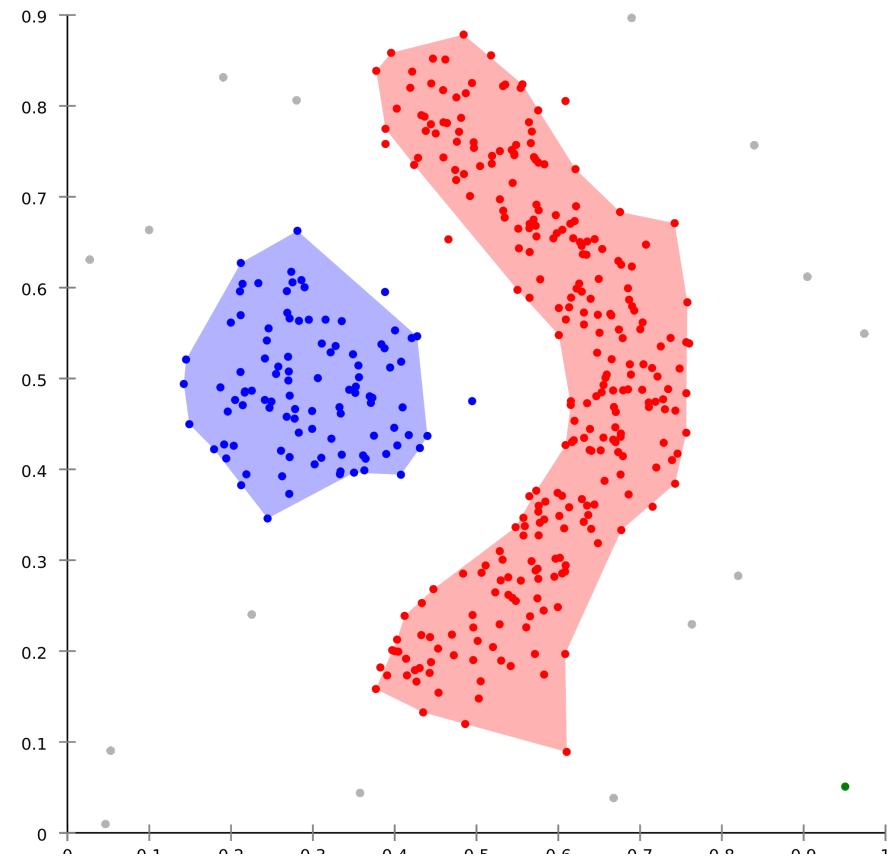


Multi-class classification:



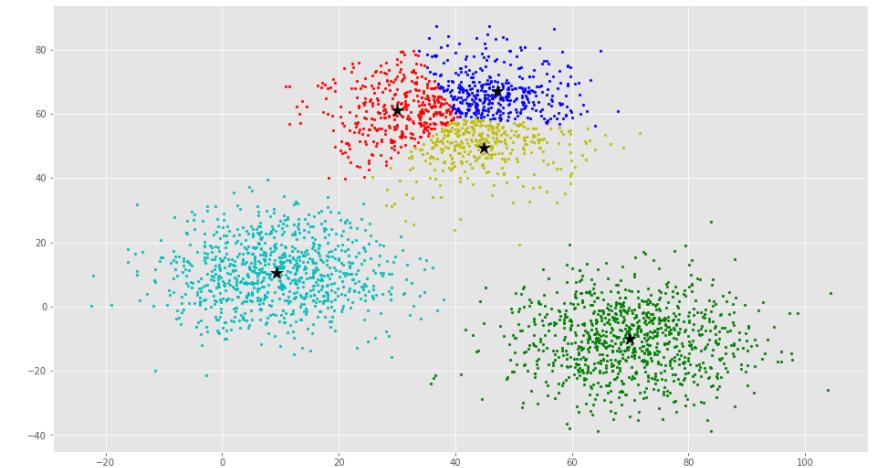
Algunos algoritmos no supervisados

- Algoritmos de clustering
- Análisis de componentes principales
- Descomposición en valores singulares
- Análisis de componentes independientes



Clustering

- Identificar áreas de similaridad.
- Topografía (desierto, hierba, etc.)
- Categorizar diferentes tipos de tejidos de imágenes médicas
- Determinar diferentes grupos de patrones meteorológicos
- Descubrir los puntos calientes de crímenes



Cierre

The screenshot shows a GitHub repository page for 'FvD / ctec_ml_2019'. The repository has 2 issues, 1 star, and 19 forks. The 'Code' tab is selected. The commit history shows several files added by 'FvD' to the 'ctec_ml_s1' branch, all made 3 hours ago. The files listed are 'img', 'Análisis de variables.Rmd', 'ctec_ml_s1.Rproj', 'kc_house_data-original.csv', 'kc_house_data.csv', and 'leeme.txt'. Each commit message is 'añade informació proyecto RStudio'.

File	Commit Message	Time Ago
img	añade informació proyecto RStudio	3 hours ago
Análisis de variables.Rmd	añade informació proyecto RStudio	3 hours ago
ctec_ml_s1.Rproj	añade informació proyecto RStudio	3 hours ago
kc_house_data-original.csv	añade informació proyecto RStudio	3 hours ago
kc_house_data.csv	añade informació proyecto RStudio	3 hours ago
leeme.txt	añade informació proyecto RStudio	3 hours ago

- Tarea para esta semana:
 - Tarea clase 1 en github

Bibliografía

- [Machine Learning Overview](#)
- [Categories of Machine Learning Techniques](#)
- [Machine Learning Process](#)
- [Goals and Activities in the Machine Learning Process](#)
- https://www.researchgate.net/publication/320771893_Big_Data_Analytics_Applications_Prospects_and_Challenges
- Clark, A.; “Machine Learning Audits in the ‘Big Data Age’,” *CIO Insights*, 19 April 2017,
www.cioinsight.com/it-management/innovation/machine-learning-audits-in-the-big-data-age.html
- Kaur, S.; “A Review of Software Development Life Cycle Models,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, iss. 11, 2015, p. 354–60,
http://ijarcsse.com/Before_August_2017/docs/papers/Volume_5/11_November2015/V5I11-0234.pdf