

# Introducción a las Ciencias de los Datos

Luis Alexander Calvo Valverde

(Material base: Saúl Calderón)

27 de febrero de 2020

# Index

- 1 Contexto**
- 2 Términos y disciplinas relacionadas con las Ciencias de los Datos**
- 3 Terminología básica**
- 4 Roles y actividades en Ciencias de Datos**
- 5 Proyectos de PARMA con Aprendizaje Automático**
- 6 Conclusiones y perspectivas**

# Tendencias del desarrollo científico-tecnológico

- Desde la aparición de las primeras computadoras a finales de la década de los 50's hasta hoy, los computadores digitales han cambiado la manera en como se producen bienes, se administran recursos, nos comunicamos e incluso entretenemos
  - Primera revolución industrial (1760 y 1830): la máquina de vapor y los sistemas mecánicos

# Tendencias del desarrollo científico-tecnológico

- Desde la aparición de las primeras computadoras a finales de la década de los 50's hasta hoy, los computadores digitales han cambiado la manera en como se producen bienes, se administran recursos, nos comunicamos e incluso entretenemos
  - Primera revolución industrial (1760 y 1830): la máquina de vapor y los sistemas mecánicos
  - Segunda revolución industrial (1850): la petroquímica, los sistemas eléctricos (radio, telégrafo)

# Tendencias del desarrollo científico-tecnológico

- Desde la aparición de las primeras computadoras a finales de la década de los 50's hasta hoy, los computadores digitales han cambiado la manera en como se producen bienes, se administran recursos, nos comunicamos e incluso entretenemos
  - Primera revolución industrial (1760 y 1830): la máquina de vapor y los sistemas mecánicos
  - Segunda revolución industrial (1850): la petroquímica, los sistemas eléctricos (radio, telégrafo)
  - Tercera revolución industrial (1950): sistemas electrónicos, informáticos, biotecnología

# Tendencias del desarrollo científico-tecnológico

- Cuarta revolución industrial (en curso?): Convergencia de distintas disciplinas en nuevas áreas transdisciplinarias como la robótica, la inteligencia artificial, nanotecnología, biotecnología, ingeniería biomédica, etc.



Figura: Tomado de <http://www.bbc.com/mundo/noticias-37631834>.

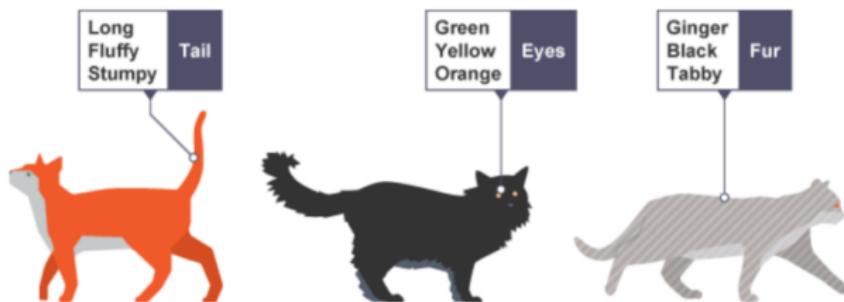
# La informática y sus perspectivas

- Automatización a gran escala con sistemas ciberfísicos, usando internet de las cosas, computación en la nube, **aprendizaje automático**, etc. para crear fábricas, procesos y servicios completamente autónomos



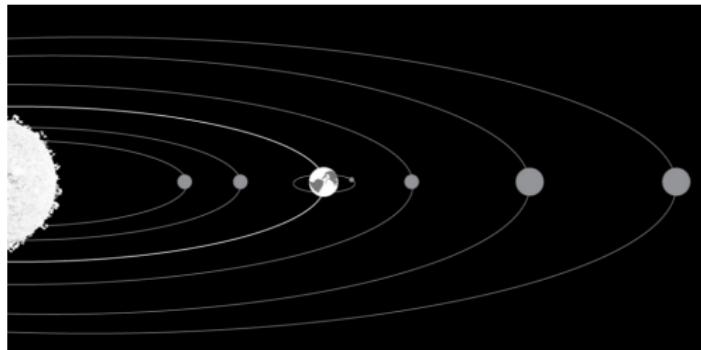
# Reconocimiento de patrones

- Diariamente realizamos muchas actividades de RP: desde buscar manzanas maduras para comer en una canasta del supermercado, hasta comprender sonidos y palabras escritas
- RP (según Richard Duda): Acto de tomar un conjunto de datos y categorizarlos o clasificarlos según un «patrón»



# Reconocimiento de patrones

- **La búsqueda de patrones ha sido una tarea fundamental en la ciencia:** P. ej, la búsqueda de repeticiones sistemáticas (patrones) en conjuntos de datos realizada por Johannes Kepler, permitió el modelado de las leyes empíricas



# Reconocimiento de patrones

- **RP como disciplina:** Se encarga de desarrollar métodos y algoritmos de descripción y clasificación de datos
- **Áreas relacionadas:**
  - **Matemática aplicada:** análisis numérico, optimización, probabilidad y estadística, matemática discreta

# Reconocimiento de patrones

- **RP como disciplina:** Se encarga de desarrollar métodos y algoritmos de descripción y clasificación de datos
- **Áreas relacionadas:**
  - **Matemática aplicada:** análisis numérico, optimización, probabilidad y estadística, matemática discreta
  - **Ingeniería eléctrica:** análisis y procesamiento de señales, procesamiento digital de señales

# Reconocimiento de patrones

- **RP como disciplina:** Se encarga de desarrollar métodos y algoritmos de descripción y clasificación de datos
- **Áreas relacionadas:**
  - **Matemática aplicada:** análisis numérico, optimización, probabilidad y estadística, matemática discreta
  - **Ingeniería eléctrica:** análisis y procesamiento de señales, procesamiento digital de señales
  - **Computación:** inteligencia artificial, aprendizaje automático, minería de datos, estructuras de datos y análisis de algoritmos, teoría de grafos, etc

# Reconocimiento de patrones

- **RP como disciplina:** Se encarga de desarrollar métodos y algoritmos de descripción y clasificación de datos
- **Áreas relacionadas:**
  - **Matemática aplicada:** análisis numérico, optimización, probabilidad y estadística, matemática discreta
  - **Ingeniería eléctrica:** análisis y procesamiento de señales, procesamiento digital de señales
  - **Computación:** inteligencia artificial, aprendizaje automático, minería de datos, estructuras de datos y análisis de algoritmos, teoría de grafos, etc
  - **Dominio específico:** Lingüística, física, biología, química, etc....

# Inteligencia Artificial y aprendizaje automático

## ■ **Inteligencia Artificial:**

- Se ocupa del diseño y construcción de sistemas capaces de percibir datos y señales para aprender, razonar y tomar decisiones de forma autónoma, y desarrolla temas como:
  - Representación del conocimiento

# Inteligencia Artificial y aprendizaje automático

## ■ **Inteligencia Artificial:**

- Se ocupa del diseño y construcción de sistemas capaces de percibir datos y señales para aprender, razonar y tomar decisiones de forma autónoma, y desarrolla temas como:
  - Representación del conocimiento
  - Búsqueda heurística

# Inteligencia Artificial y aprendizaje automático

## ■ **Inteligencia Artificial:**

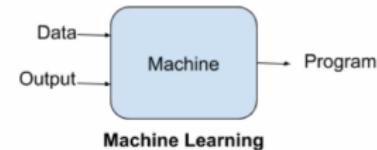
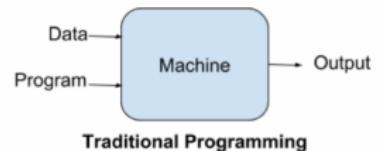
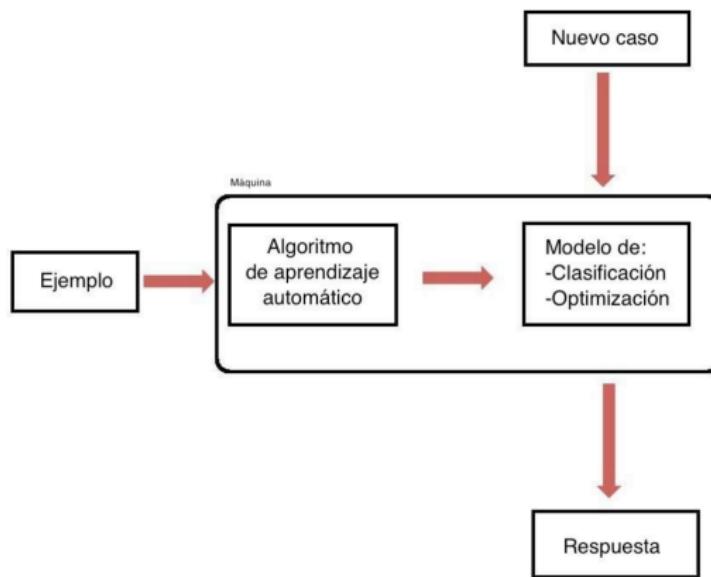
- Se ocupa del diseño y construcción de sistemas capaces de percibir datos y señales para aprender, razonar y tomar decisiones de forma autónoma, y desarrolla temas como:
  - Representación del conocimiento
  - Búsqueda heurística
  - **Aprendizaje automático y aprendizaje profundo:** Métodos y algoritmos de descripción, clasificación y regresión de **datos**

# Inteligencia Artificial y aprendizaje automático

## ■ **Inteligencia Artificial:**

- Se ocupa del diseño y construcción de sistemas capaces de percibir datos y señales para aprender, razonar y tomar decisiones de forma autónoma, y desarrolla temas como:
  - Representación del conocimiento
  - Búsqueda heurística
  - **Aprendizaje automático y aprendizaje profundo:** Métodos y algoritmos de descripción, clasificación y regresión de **datos**

# Aprendizaje automático



**Figure: ¿Porqué aprendizaje automático?** Para muchos problemas es difícil programar reglas, además de la necesidad de la adaptabilidad

# Aprendizaje profundo

- Redes neuronales con distintos patrones de conectividad y más eficientes, con muchas capas y millones de parámetros

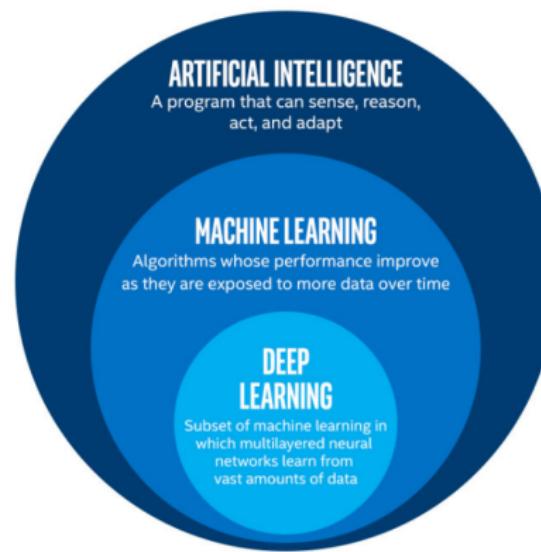


Figura: Tomado de diapositivas del Dr. Pascal Tyrrell.

# Aprendizaje profundo

- Las redes profundas escalan mejor a grandes cantidades de datos

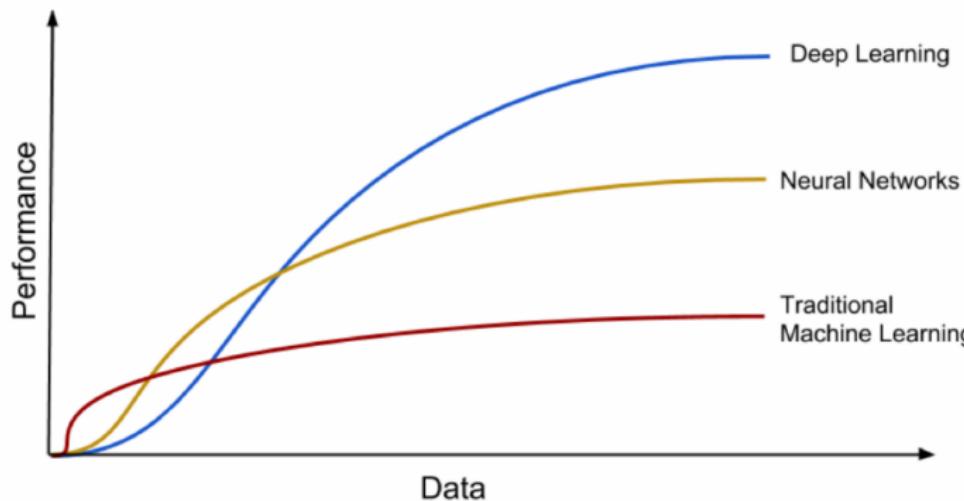


Figura: Tomado de diapositivas del Dr. Pascal Tyrrell.

# Aprendizaje profundo



Figura: Premio Turing 2019: Yann LeCun (U. New York, Facebook), Geoffrey Hinton (U. Toronto, Google) and Yoshua Bengio (U. Montreal).

# Minería de datos y Big data

- **Big data:** Técnicas computacionales para consultar grandes cantidades de datos estructurados y no estructurados, en entornos distribuidos y heterogéneos

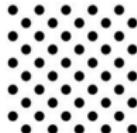
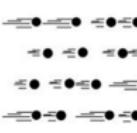
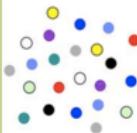
Volume	Velocity	Variety	Veracity	Value
 Data at Rest  Terabytes to Exabytes of existing data to process	 Data in Motion  Streaming data, requiring milliseconds to seconds to respond	 Data in Many Forms  Structured, unstructured, text, multimedia,...	 Data in Doubt  Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	 Data into Money  Business models can be associated to the data
<small>Adapted by a post of Michael Walker on 28 November 2012.</small>				

Figure: Datos medianos (10 GB - 1 TB), Big data más de 1 TB.

# Big data

- **Minería de datos:** Uso de algoritmos para obtener nuevos datos de valor agregado

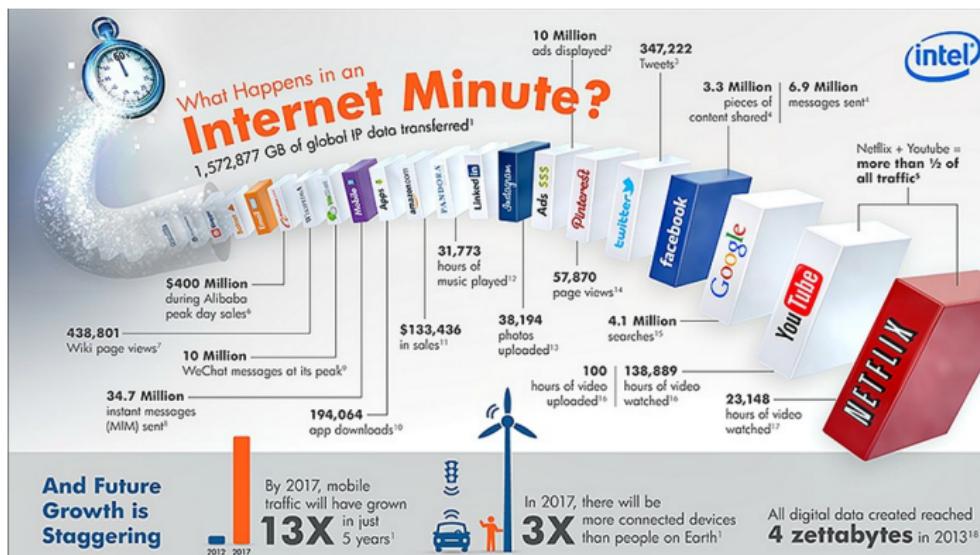
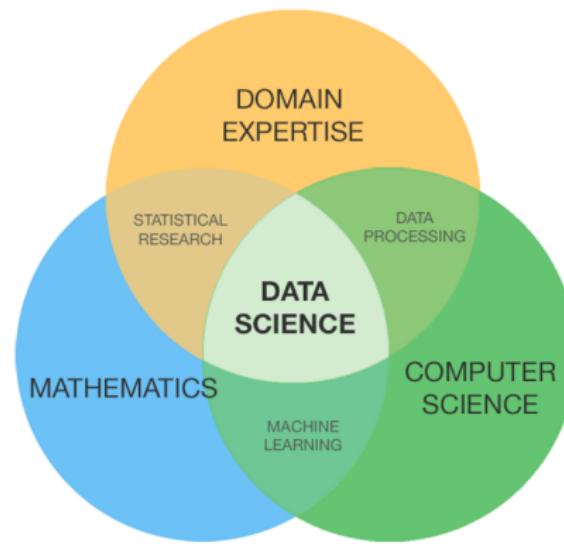


Figure: Tomado de diapositivas del Dr. Pascal Tyrrell.

# ¿Qué son las ciencias de los datos?

- Mezcla de aplicación de conocimientos en matemática, estadística, y computación, usando grandes cantidades de datos, con un **dominio específico**



Source: Palmer, Shelly. *Data Science for the C-Suite*.  
New York: Digital Living Press, 2015. Print.

# ¿Qué son las ciencias de los datos?

- **Enfasis en el dominio específico:** Medicina y agricultura de precisión, psicología, patología y radiología computacional, etc.

# ¿Qué son las ciencias de los datos?

- **Enfasis en el dominio específico:** Medicina y agricultura de precisión, psicología, patología y radiología computacional, etc.
- **Datafificación:** extraer datos de distintos aspectos de un fenómeno

## ¿Qué son las ciencias de los datos?

- **Enfasis en el dominio específico:** Medicina y agricultura de precisión, psicología, patología y radiología computacional, etc.
  - **Datafificación:** extraer datos de distintos aspectos de un fenómeno



Figura: La datafacción y el Internet de las cosas, <https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html>.

# ¿Qué son las ciencias de los datos?

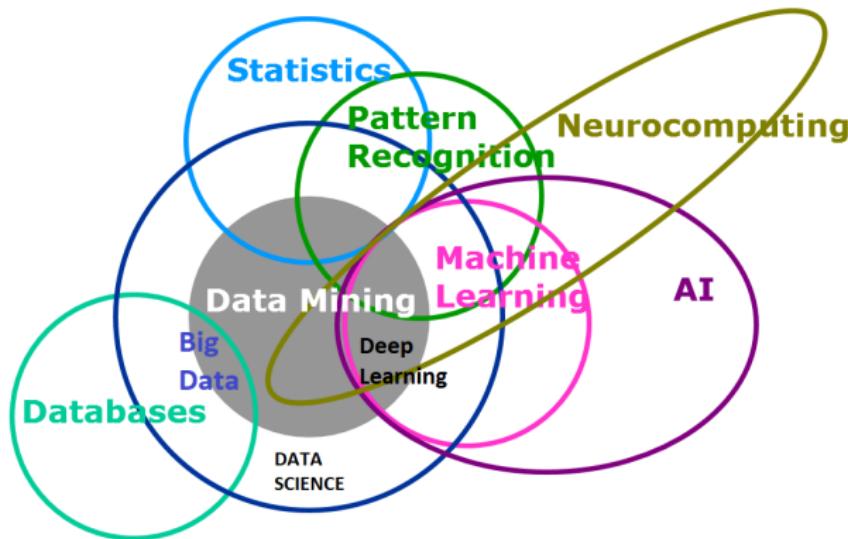


Figura: La datafificación y el Internet de las cosas, <https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html>.

# Definiciones: Característica o «Feature»

- **Problema:** clasificar imágenes de racimos en uvas Hunisa, Autumn Royal y Moscatel

# Definiciones: Característica o «Feature»

- **Problema:** clasificar imágenes de racimos en uvas Hunisa, Autumn Royal y Moscatel
- **Característica:** propiedad individual medible u observable: p. ej en el mundo real; forma, color, olor, sabor, textura, tamaño, peso, brillo



# Definiciones: Característica o «Feature»

- **Verdosidad:**  $\overrightarrow{x} \in [0 - 255]$  intensidad promedio en una imagen del color «verde» captado por una cámara digital



## Definiciones: Dimensionalidad de una muestra

- **Dimensionalidad de una muestra**  $\dim(\vec{x}_i)$ : Una muestra se representa con un arreglo de  $\dim(\vec{x}_i) = N$  valores correspondientes a múltiples características:

$$\vec{x}_i = \langle x_1, x_2, \dots, x_N \rangle$$

## Definiciones: Dimensionalidad de una muestra

- **Dimensionalidad de una muestra**  $\dim(\vec{x}_i)$ : Una muestra se representa con un arreglo de  $\dim(\vec{x}_i) = N$  valores correspondientes a múltiples características:

$$\vec{x}_i = \langle x_1, x_2, \dots, x_N \rangle$$

- **Ejemplo 1**, una muestra  $\vec{x}_a$  puede estar compuesta por los valores particulares de características de **color** y **peso**, en este caso  $N = 2$ :

$$\vec{x}_a = \langle x_1 = 253, x_2 = 40 \text{ gramos} \rangle$$

## Definiciones: Dimensionalidad de una muestra

- **Dimensionalidad de una muestra**  $\dim(\vec{x}_i)$ : Una muestra se representa con un arreglo de  $\dim(\vec{x}_i) = N$  valores correspondientes a múltiples características:

$$\vec{x}_i = \langle x_1, x_2, \dots, x_N \rangle$$

- **Ejemplo 1**, una muestra  $\vec{x}_a$  puede estar compuesta por los valores particulares de características de **color y peso**, en este caso  $N = 2$ :

$$\vec{x}_a = \langle x_1 = 253, x_2 = 40 \text{ gramos} \rangle$$

- **Ejemplo 2**, una muestra  $\vec{x}_b$  puede estar definida por los valores particulares de una imagen de  $100 \times 100$  pixeles, por lo que  $N = 10000$

# Definiciones: Patrón

- **Patrón o «plantilla»:** Regularidad discernible para una o varias características en un conjunto de objetos o **muestras**



# Definiciones: Conjunto de muestras

- **Conjunto de muestras  $X$ :** Conjunto de  $M$  muestras de un arreglo de características:

$$\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_M\}$$

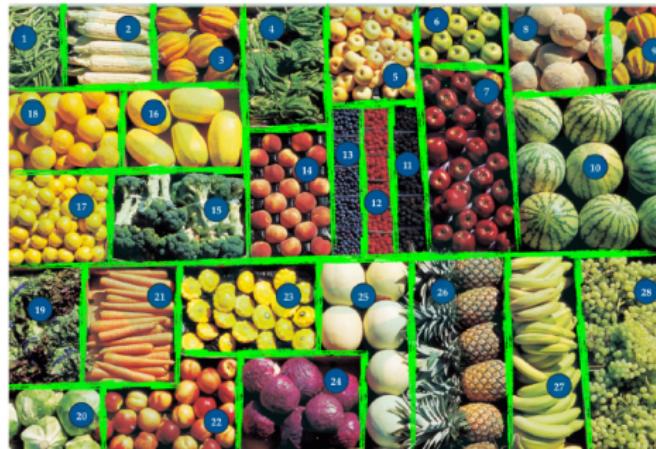
- Por ejemplo, ¿cuál es el patrón para las muestras de la verdosidad ( $\dim(\vec{x}_i) = 1 = N$ ):

$$\mathbf{X} = \{\vec{x}_1 = 253, \vec{x}_2 = 254, \vec{x}_3 = 100, \vec{x}_4 = 255\}$$



# Definiciones: Clase

- **Clase:** Abstracción de propiedades comunes o repetidas en múltiples instancias de esa clase

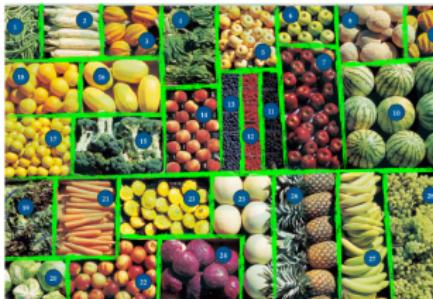


# Definiciones: Clase

- **Clase:** Para facilitar su representación una clase  $C_j$  tiene asociada una etiqueta  $j = 1, 2, \dots, J$
- **Ejemplo 1:** la clase «banano» tiene etiqueta  $j = 1$ , la clase «sandía» la etiqueta  $j = 2$ ....

# Definiciones: Clase

- **Clase:** Para facilitar su representación una clase  $C_j$  tiene asociada una etiqueta  $j = 1, 2, \dots, J$
- **Ejemplo 1:** la clase «banano» tiene etiqueta  $j = 1$ , la clase «sandía» la etiqueta  $j = 2$ ....
- **Ejemplo 2:** la clase «uva verde»  $j = 1$ , «uva morada»  $j = 2$  y «uva roja»  $j = 3$



# Definiciones: Conjunto de etiquetas

- **Conjunto de etiquetas  $T$  para un conjunto de muestras  $X$ :**  
Para un conjunto de  $M$  muestras

$$X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_M\}$$

existe un conjunto de etiquetas «correctas»:

$$T = \{t_1, t_2, \dots, t_M\}$$

## Definiciones: Conjunto de etiquetas

- **Conjunto de etiquetas  $T$  para un conjunto de muestras  $X$ :**  
Para un conjunto de  $M$  muestras

$$X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_M\}$$

existe un conjunto de etiquetas «correctas»:

$$T = \{t_1, t_2, \dots, t_M\}$$

- Para el ejemplo de las uvas:

$$X = \{\vec{x}_1 = 253, \vec{x}_2 = 254, \vec{x}_3 = 100, \vec{x}_4 = 255\}$$

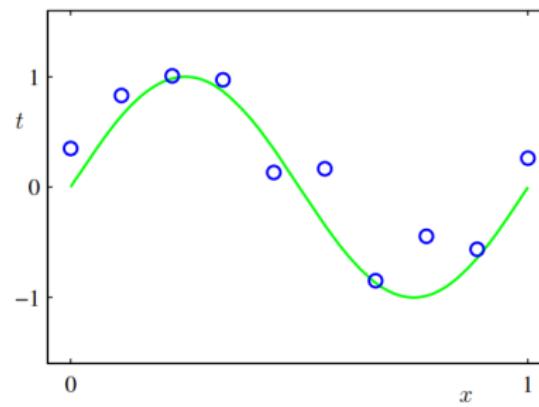
$$T = \{t_1 = 1, t_2 = 1, t_3 = 2, t_4 = 1\}$$

# Definiciones: Clasificación

- **Clasificación:** Proceso de asignar una muestra  $\vec{x}_i$  una etiqueta  $j$ , y en general para un conjunto de muestras  $X$ , obtener un conjunto de etiquetas  $T$

# Definiciones: Clasificación

- **Clasificación:** Proceso de asignar una muestra  $\vec{x}_i$  una etiqueta  $j$ , y en general para un conjunto de muestras  $X$ , obtener un conjunto de etiquetas  $T$
- Si todas las etiquetas  $j \in \mathbb{N}$  se realiza una **clasificación** y si  $j \in \mathbb{R}$  se realiza una **regresión**



# Definiciones: Un clasificador estadístico simple: Uvas verdes, moradas y rojas

- Se define el conjunto de **muestras de entrenamiento (ground truth)**  $X_e$  con sus correspondientes etiquetas  $T_e$

$$\begin{aligned}X_e &= \{\vec{x}_1 = 253, \vec{x}_2 = 254, \vec{x}_3 = 100, \vec{x}_4 = 255, \vec{x}_5 = 105\} \\T_e &= \{t_1 = 1, t_2 = 1, t_3 = 2, t_4 = 1, t_5 = 2\}\end{aligned}$$

# Definiciones: Un clasificador estadístico simple: Uvas verdes, moradas y rojas

- Se define el conjunto de **muestras de entrenamiento (ground truth)**  $X_e$  con sus correspondientes etiquetas  $T_e$

$$\begin{aligned}X_e &= \{\vec{x}_1 = 253, \vec{x}_2 = 254, \vec{x}_3 = 100, \vec{x}_4 = 255, \vec{x}_5 = 105\} \\T_e &= \{t_1 = 1, t_2 = 1, t_3 = 2, t_4 = 1, t_5 = 2\}\end{aligned}$$

- Momentos estadísticos para cada clase:

$$\begin{array}{ll}\mu_1 = 254 & \sigma_1 = 1 \\ \mu_2 = 102,5 & \sigma_2 = 3,53\end{array}$$

# Definiciones: Un clasificador estadístico simple: Uvas verdes, moradas y rojas

- Se define el conjunto de **muestras de entrenamiento (ground truth)**  $X_e$  con sus correspondientes etiquetas  $T_e$

$$X_e = \{\vec{x}_1 = 253, \vec{x}_2 = 254, \vec{x}_3 = 100, \vec{x}_4 = 255, \vec{x}_5 = 105\}$$

$$T_e = \{t_1 = 1, t_2 = 1, t_3 = 2, t_4 = 1, t_5 = 2\}$$

- Momentos estadísticos para cada clase:

$$\begin{array}{ll} \mu_1 = 254 & \sigma_1 = 1 \\ \mu_2 = 102,5 & \sigma_2 = 3,53 \end{array}$$

- Si se recibe una nueva muestra, p. ej. con valor  $\vec{x}_6 = 252$ , el clasificador hace para todas las clases  $C_j$ :

$$\begin{aligned} |\mu_1 - \vec{x}_6| &= 2 < 3\sigma_1 \\ |\mu_2 - \vec{x}_6| &= 149,5 > 3\sigma_2 \end{aligned} \Rightarrow t_6 = 1$$

# Definiciones: Clasificación

Muestra	Características			Clase
	Peso	Forma	Tamaño	
	liviano	redondo	pequeño	uva
	mediano	alargado	mediano	banano
	pesado	ovalada	grande	sandía

# Etapas básicas de un Sistema RP (SRP)

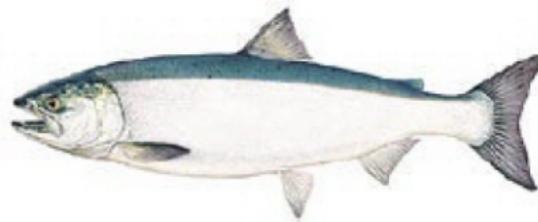


# Etapas básicas de un Sistema RP: Ejemplo

- Una empacadora de pescados necesita construir una máquina para etiquetar los róbalos (sea-bass) y salmones empacados



(a)



(b)

# Etapas básicas de un Sistema RP: Ejemplo

- Ahí van los pescados...

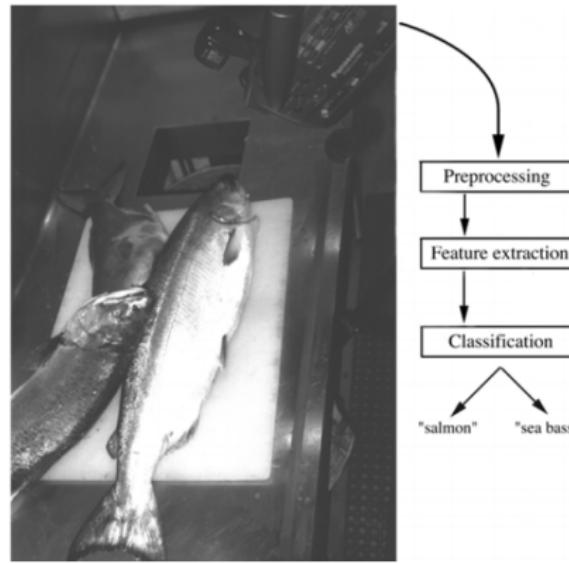


Figura: Muestras de pescados a clasificar (Tomado de Duda, Hart.)

# Etapas básicas: Preprocesamiento

- Sean los datos entrantes al sistema  $\mathcal{U} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  $\dim(\vec{u}_i) = B$

## Etapas básicas: Preprocesamiento

- Sean los datos entrantes al sistema  $\mathbf{U} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  $\dim(\vec{u}_i) = B$
- Muchas veces son preprocesados para eliminar o atenuar el «**ruido**» en los datos o rellenar **datos faltantes** o **sesgos**, aplicando una función

$$F(\vec{u}_i) = \vec{v}_i$$

que presenta como salida a  $\vec{v}_i$ , correspondiente a la muestra preprocesada, con dimensión  $\dim(\vec{u}_i) = B$

## Etapas básicas: Preprocesamiento

- Sean los datos entrantes al sistema  $\mathbf{U} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  $\dim(\vec{u}_i) = B$
- Muchas veces son preprocesados para eliminar o atenuar el «**ruido**» en los datos o rellenar **datos faltantes** o **sesgos**, aplicando una función

$$F(\vec{u}_i) = \vec{v}_i$$

que presenta como salida a  $\vec{v}_i$ , correspondiente a la muestra preprocesada, con dimensión  $\dim(\vec{u}_i) = B$

- **Función  $F$**  : Desde el procesamiento de señales: Filtros en el dominio de Fourier, Wavelets, DCT, etc

## Etapas básicas: Preprocesamiento

- Por ejemplo, si las entradas son imágenes de  $640 \times 480$  pixeles, entonces  $\dim(\vec{u}_i) = 307200$



# Etapas básicas: Extracción de características

- **Entrada:** conjunto de datos preprocesados

$V = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  
 $\dim(\vec{u}_i) = B$

## Etapas básicas: Extracción de características

- **Entrada:** conjunto de datos preprocesados

$V = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  $\dim(\vec{u}_i) = B$

- Consiste en un funcional

$$G(\vec{v}_i) = \vec{x}_i$$

con una salida  $\vec{x}_i$ , correspondiente al vector de características y dimensión  $\dim(\vec{u}_i) = N$

## Etapas básicas: Extracción de características

- **Entrada:** conjunto de datos preprocesados

$\mathbf{V} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}$ , con cada muestra de dimensión  $\dim(\vec{u}_i) = B$

- Consiste en un funcional

$$G(\vec{v}_i) = \vec{x}_i$$

con una salida  $\vec{x}_i$ , correspondiente al vector de características y dimensión  $\dim(\vec{u}_i) = N$

- **Función  $G$ :** Por cada muestra la función extrae un arreglo de características para la fácil discriminación de las clases, y **reducir la dimensionalidad**  $N \ll B$

- Procesamiento de señales, con info. específica del dominio (medicina, biología, química, etc)

## Etapas básicas: Extracción de características

- **Entrada:** conjunto de datos preprocesados

$$V = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}, \text{ con } \dim(\vec{u}_i) = 307200$$

# Etapas básicas: Extracción de características

- **Entrada:** conjunto de datos preprocesados

$$V = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_M\}, \text{ con } \dim(\vec{u}_i) = 307200$$

- **Salida:** El funcional  $G(\vec{v}_i) = \vec{x}_i = \langle x_1, x_2 \rangle$  extrae características de **ancho** del pescado, y a la «**claridad**» del **color** del pescado

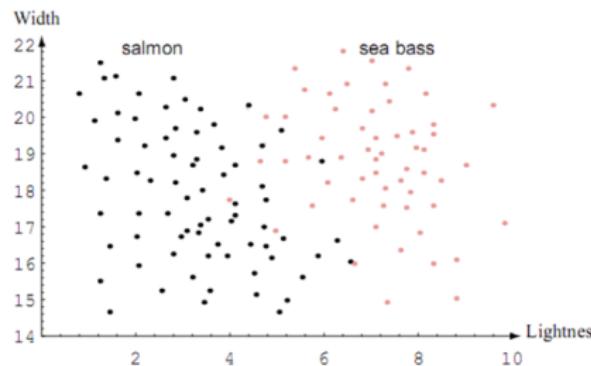


Figura: Diagrama de dispersión del espacio de muestras con dimensión  $N = 2$

## Etapas básicas: Clasificación, no paramétrica

- **K vecinos más cercanos:** Para un punto nuevo  $\vec{x}_a$ , se calculan los  $K = 6$  vecinos más cercanos, usando la distancia Euclídea,
- En este caso  $t_a = 1$  ( $C_1$  corresponde a la clase salmón)

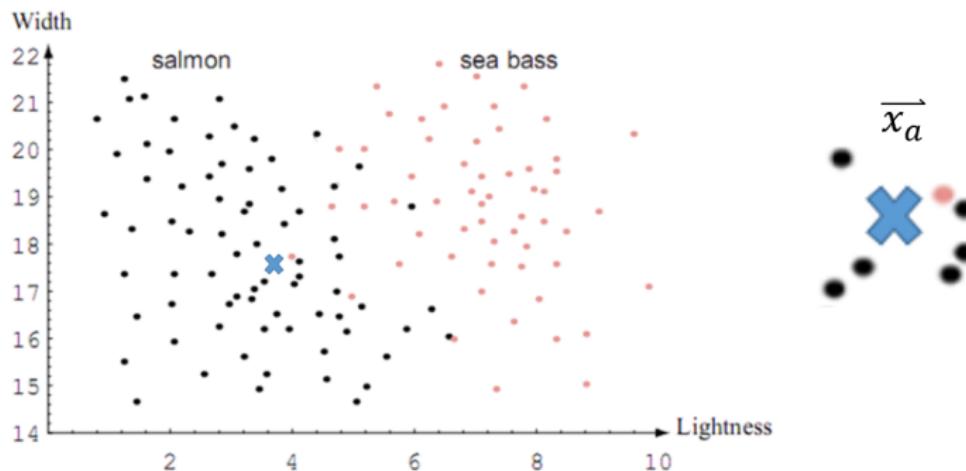


Figura: Espacio de muestras de entrenamiento  $\mathbf{X}_e$ , con  $N = 2$  ▶ ▷ ⏪ ⏪ ⏴ ⏴

# Etapas básicas: Clasificación, paramétrica lineal

- Dado un conjunto de muestras de entrenamiento  $X_e$ , **construye un hiperplano o función**  $y$  que minimice el error de clasificación
- Error cuadrático mínimo, discriminante lineal de Fisher, perceptrón, etc

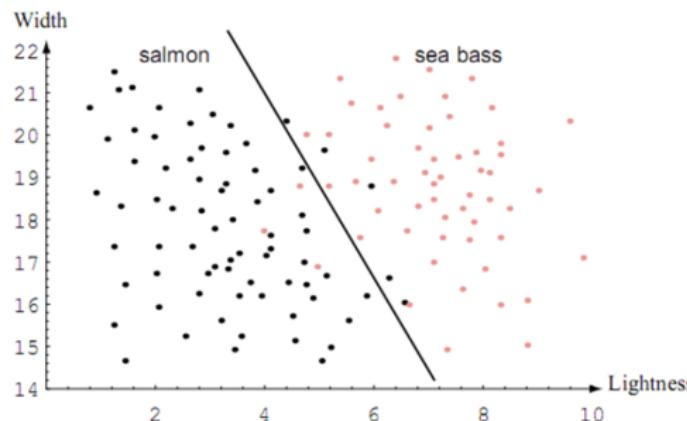
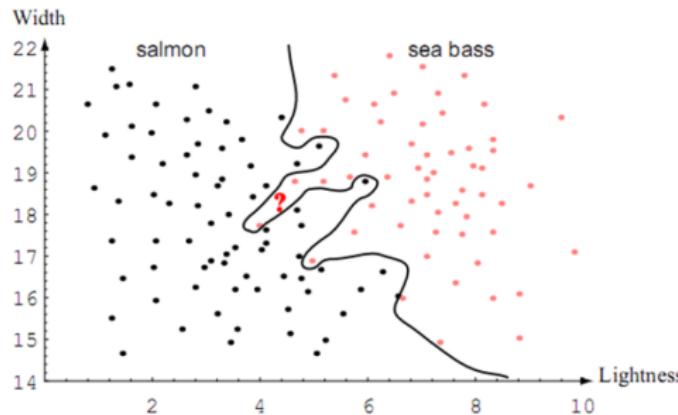


Figura: Espacio de muestras de entrenamiento  $X_e$ , con  $N = 2$

## Etapas básicas: Clasificación, superficie no lineal

- Clasificadores más sofisticados generan superficies de decisión no lineales (máquinas de soporte vectorial, redes neuronales, convolucionales, Bayesianas, etc)



**Figura: Sobre ajuste:** Una superficie que se sobreajusta, confía al 100% en  $X_e$  lo cual no es aconsejable...

## Etapas básicas: Clasificación, superficie no lineal

- Clasificadores más sofisticados generan superficies de decisión no lineales (máquinas de soporte vectorial, redes neuronales, convolucionales, Bayesianas, etc)

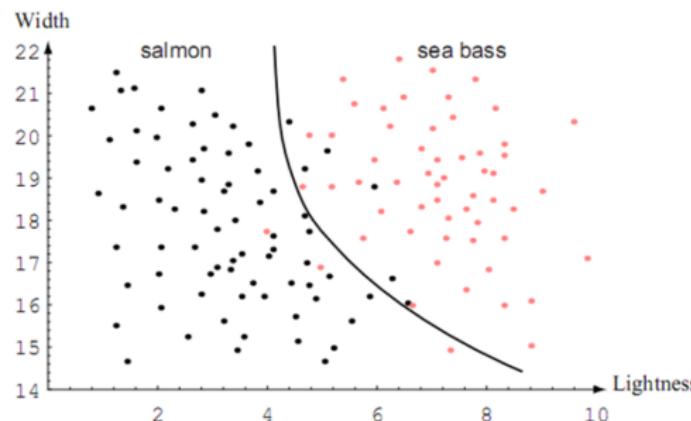


Figura: Mejores resultados con superficies regularizadas, para evitar sobreajuste

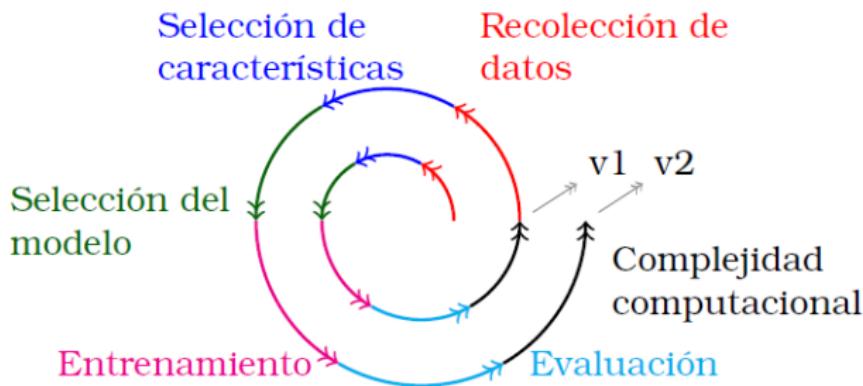
## Etapas básicas: Clasificación no supervisada

- Los clasificadores anteriores necesitan un conjunto de muestras etiquetadas  $X_e$ , por lo que se les dice **supervisados**
- **No supervisados:** «amontonan» los datos para encontrar por sí solos las muestras de cada clase  $J$

(Loading...)

Figura: Funcionamiento del algoritmo K-medias.

# Ciclo de vida de un Sistema de reconocimiento de patrones



# Proceso de un proyecto de ciencia de los datos

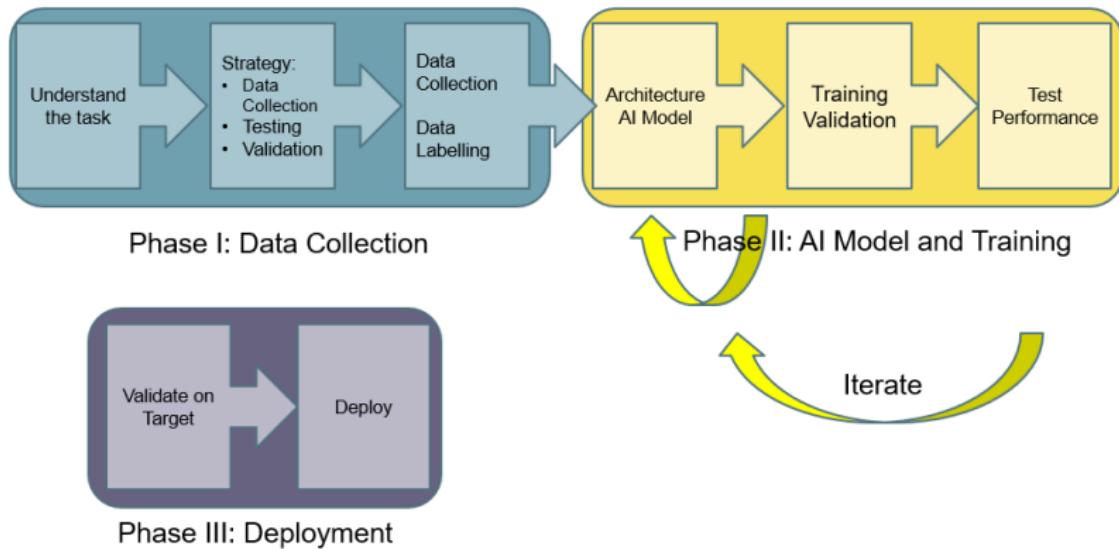


Figura: Tomado de diapositivas del Dr. Pascal Tyrrell, MiDATA.

# Roles de un científico de datos

- **Data Businessperson:** Enfocados en la generación de valor para la organización de proyectos basados en datos

# Roles de un científico de datos

- **Data Businessperson:** Enfocados en la generación de valor para la organización de proyectos basados en datos
- **Data Creatives:** Implementadores de todas las etapas, desde la definición y extracción de datos, análisis y visualización

# Roles de un científico de datos

- **Data Businessperson:** Enfocados en la generación de valor para la organización de proyectos basados en datos
- **Data Creatives:** Implementadores de todas las etapas, desde la definición y extracción de datos, análisis y visualización
- **Data Developer:** Se preocupan en cómo almacenar, obtener y aprender de los datos.

# Roles de un científico de datos

- **Data Businessperson:** Enfocados en la generación de valor para la organización de proyectos basados en datos
- **Data Creatives:** Implementadores de todas las etapas, desde la definición y extracción de datos, análisis y visualización
- **Data Developer:** Se preocupan en cómo almacenar, obtener y aprender de los datos.
- **Data Researcher:** Desarrollan técnicas del estado del arte, con entendimiento profundo de técnicas que aportan a la organización

# Roles de un científico de datos

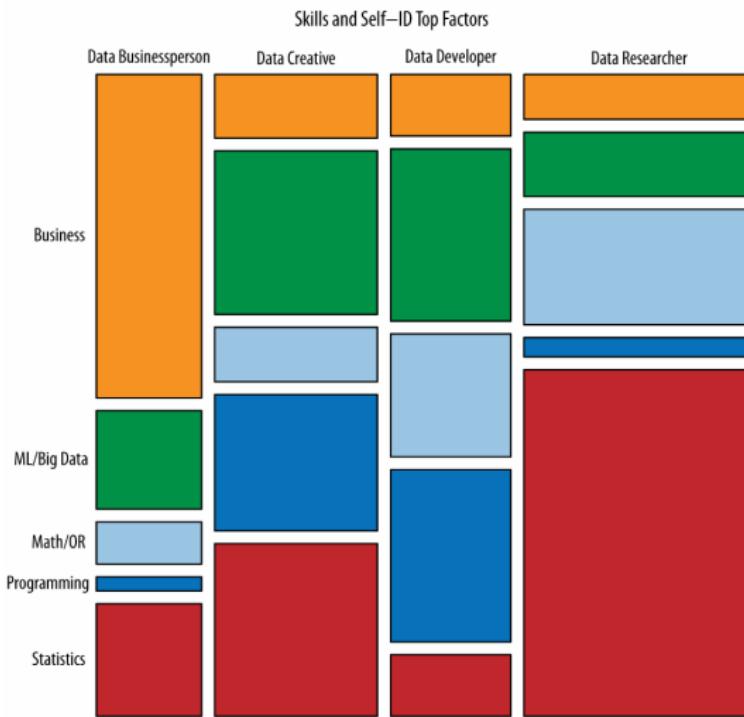


Figure 1-4. Harlan Harris's clustering and visualization of subfields of data science from *Analyzing the Analysts* (O'Reilly) by Harlan Harris, Sean Murphy, and Marek Vaizman based on a survey of several hundred data science practitioners in mid-2012

# PARMA-Group

## Objetivos

- Investigación: Estimular la investigación en el área de aprendizaje automático, proyectos con otras universidades, nacionales e internacionales

# PARMA-Group

## Objetivos

- Investigación: Estimular la investigación en el área de aprendizaje automático, proyectos con otras universidades, nacionales e internacionales
  
- Docencia: Crear y asesorar cursos relacionados para carreras de pregrado y posgrado

# Reconocimiento de patrones y aprendizaje automático en el TEC: PARMA-Group

## Objetivos

- Extensión: Realizar investigación y desarrollo de soluciones basadas en el aprendizaje automático, en conjunto con otras instituciones

# Reconocimiento de patrones y aprendizaje automático en el TEC: PARMA-Group

## Objetivos

- Extensión: Realizar investigación y desarrollo de soluciones basadas en el aprendizaje automático, en conjunto con otras instituciones
- Estimular la organización de actividades de capacitación a estudiantes y profesores, y profesionales en el área (proc. señales, aprendizaje automático, minería de datos)

# Reconocimiento de patrones y aprendizaje automático en el TEC: PARMA-Group

- Luis Alexander: Modelos de predicción para la agricultura (a partir de datos climatológicos)
- Juan Luis Crespo: Modelos de atención para máquinas, redes neuronales, dispositivo de asistencia ventricular, tratamiento de información genética, etc.

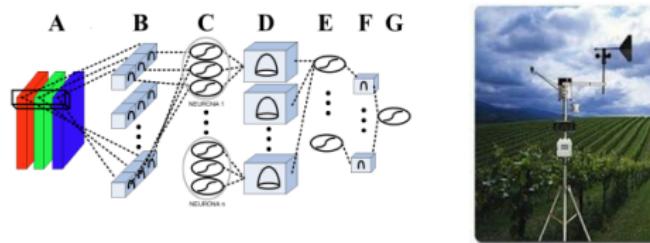


Figure: Izq. redes neuronales gaussianas, der. datos meteorológicos para la predicción de enfermedades

# Reconocimiento de patrones y aprendizaje automático en el TEC: PARMA-Group

- Geovanni Figueroa, Erick Mata y Jose Carranza: Reconocimiento de árboles a partir de imágenes digitales de corte en troncos y hojas
- Esteban Arias: Comparación de rutas metabólicas

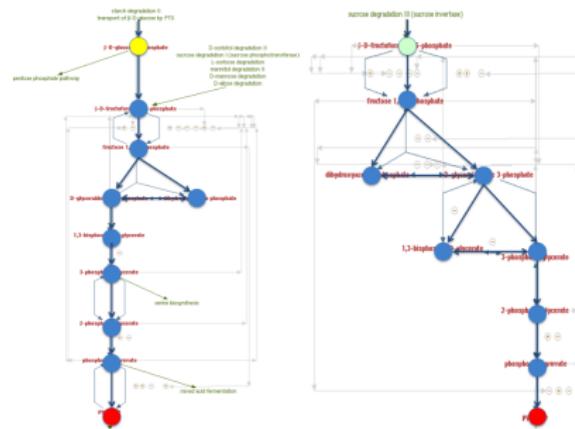
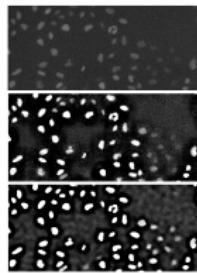


Figure: Comparación de rutas metabólicas con grafos

# Preprocesamiento, segmentación y rastreo de células

- PARMA-Group (TEC): *A first glance on the enhancement of digital cell activity videos from glioblastoma cells with nuclear staining, CONCAPAN 2016*
- DNLM-IFFT: An Implementation of the Deceived Non Local Means Filter using \\Integral Images and the Fast Fourier Transform for a Reduced Computational Cost, CIARP 2017



# Análisis automático de Histologías

- Sistema de apoyo de diagnóstico de histiologías prostaticas , en conjunto con el Dr. Jose Luis Quirós (Hospital Max Peralta) y la Dra. Tilcia López (Hospital Calderón Guardia) López

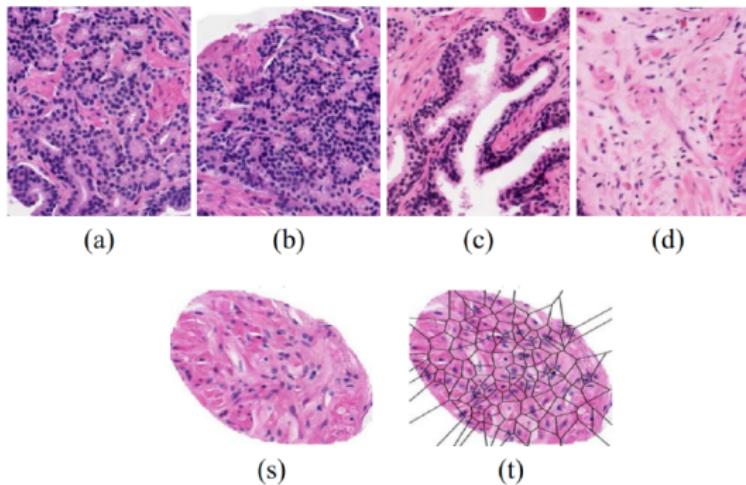


Figure: Histologías de próstata

# Análisis automático de Histologías

- **Etapa actual:** Recopilacion, etiquetado de imagenes, estudio de tecnicas de segmentacion , extraccion de caracteristicas y clasificacion

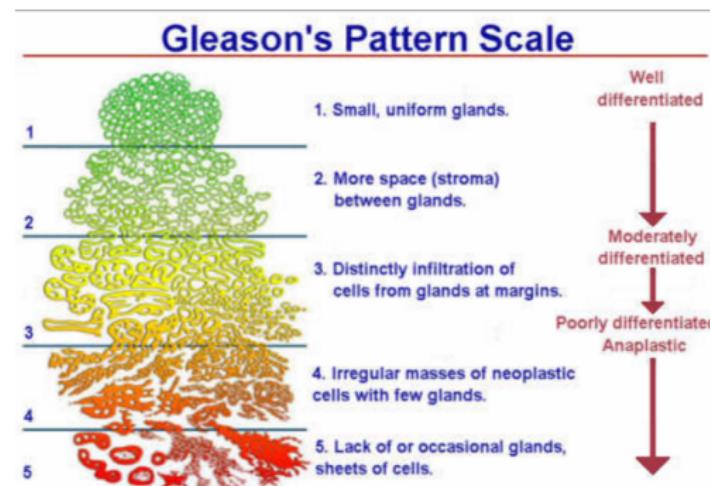


Figure: Escala de Gleason

# Estimación de productividad de cultivos a partir de imágenes multi-espectrales

- Iniciativa de proyecto conjunto con la Universidad EARTH, laboratorio de agricultura de precisión

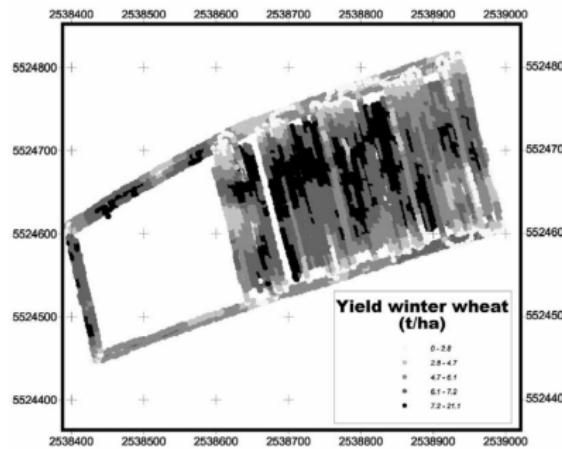


Figure: Ejemplo de estimación de productividad por hectárea

# Competencia: Estimación de edad a partir de imágenes radiológicas

- Estimar la edad ósea a partir de imágenes de rayos X, iniciativa conjunta con la U. de Toronto, y la U. de Ljubljana



Figure: Imagen del conjunto de muestras

# Identificación de Especies de Plantas usando Pliegos de Herbario

- Existen millones de fotos de pliegos de herbario hechas públicas a través de iniciativas como iDigBio.
- Tiene mucho ruido visual por el tipo de manipulación que se les da sin tener fines tecnológicos en mente.
- Hipótesis: Técnicas como Deep Learning permiten clasificar a nivel de especies, género y familia con buena exactitud.



cirad

*inria*  
INVENTEURS DU MONDE NUMÉRIQUE

Figure: Colaboraciones.

# Aprendizaje automático en el futuro

- Los algoritmos y máquinas que aprenden a partir de los datos se incorporan a las distintas esferas de la vida con mucha velocidad

# Aprendizaje automático en el futuro

- Los algoritmos y máquinas que aprenden a partir de los datos se incorporan a las distintas esferas de la vida con mucha velocidad
- **Retos:** formación en áreas relacionadas (matemáticas, ciencias de la computación, estadística) y transdisciplinar

# Aprendizaje automático en el futuro

- Los algoritmos y máquinas que aprenden a partir de los datos se incorporan a las distintas esferas de la vida con mucha velocidad
- **Retos:** formación en áreas relacionadas (matemáticas, ciencias de la computación, estadística) y transdisciplinar
- Alrededor de 7.1 millones de trabajos serán eliminados, y 2.1 millones creados (<http://money.cnn.com>)