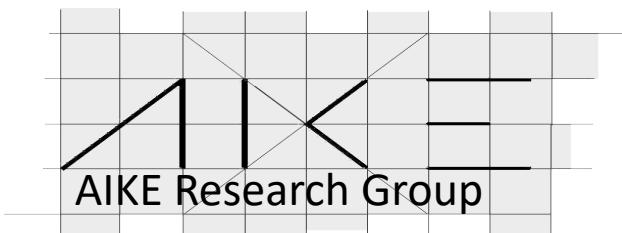


Saturdays.AI  
Murcia

# Desarrollo de modelos predictivos con machine learning

José Tomás Palma Méndez

Inteligencia Artificial e Ingeniería del Conocimiento -INTICO  
Universidad de Murcia



# LA IA no es una tecnología nueva

- Los primeros trabajos datan de la década de los 50 del s. XX.
  - Programas que juegan a las damas y el ajedrez
- 1956 Dartmouth Conference
  - Se acuña el término Inteligencia artificial
- Edad de oro: del 1956 hasta mediados de los 70
  - General Problem solver, otros demostradores de teoremas, STRIPS
  - Razonamiento automático
  - Lenguaje natural



# ¿Qué estamos usando de la IA?

- La IA es mucho más amplia que la imagen que se nos está transmitiendo
  - Planificación
  - Reconocimiento de patrones
  - Visión artificial
  - Procesamiento del lenguaje natural
  - Sistemas basados en conocimiento, ...
- El boom de la IA se centra en un área
  - Aprendizaje computacional y más concretamente en la Minería de datos



# ¿A qué es debido este boom?

- Ordenadores cada vez más potentes:
  - Superordenadores, cloud computing, multinúcleos, gpus....
- Datos, datos, datos ....
  - Sociedad datificada
  - Redes sociales, IoT, bases de datos, imágenes, videos, textos, ....
  - Tecnología para tratar con dichos datos → Big Data



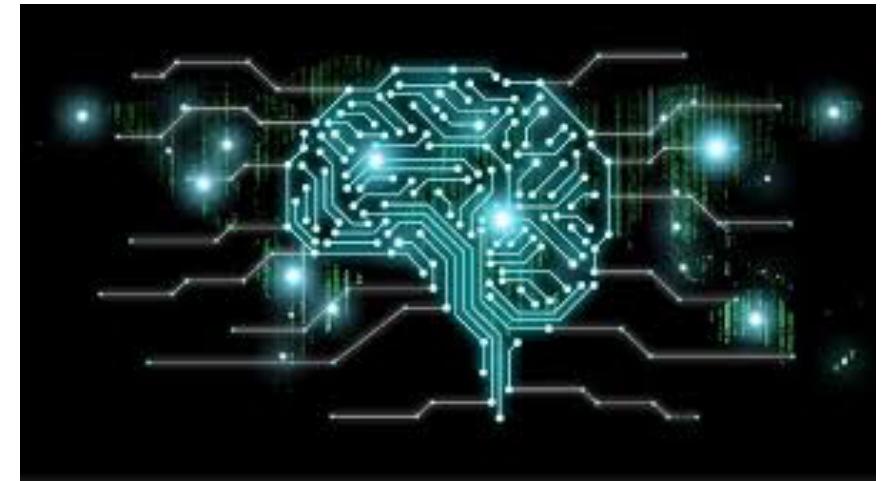
# Algunos Hitos mediáticos

- 1997 IBM Deep Blue bate a Kasparov
- 2011 IBM Watson bate a dos antiguos campeones en Jeopardy
- 2012 Google Brain crea un sistema para reconocer gatos en YouTube
- 2016 AlfaGo bate al campeón mundial de Go
- 2017 AlfaZero bate a Stockfish al ajedrez



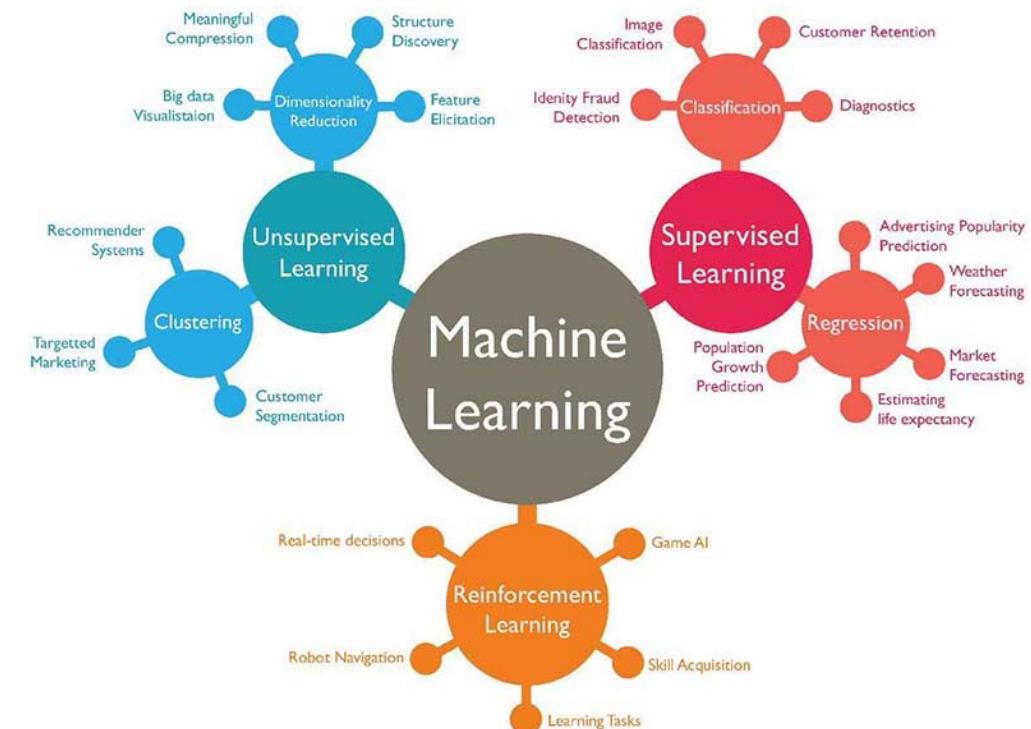
# ¿Qué es el Machine Learning (ML)?

- Disciplina de la IA centrada en la extracción de conocimiento y patrones a partir de una serie de observaciones (datos).
- Extracción de conocimiento sin necesidad de programar los sistemas para que lo hagan
  - Permite obtener valor de los datos sin realizar grandes esfuerzos de programación
- Requieren cantidades importantes de datos.



# Técnicas de Machine Learning

- Supervisado
  - Datos etiquetados
  - Modelos predictivos de clasificación y regresión
- No supervisado
  - Datos no esquematizados
  - Técnicas para extraer estructuras y patrones de los datos.
- Aprendizaje por refuerzo
  - Cuando el sistema es retroalimentado sobre lo buena o mala que son algunas decisiones.



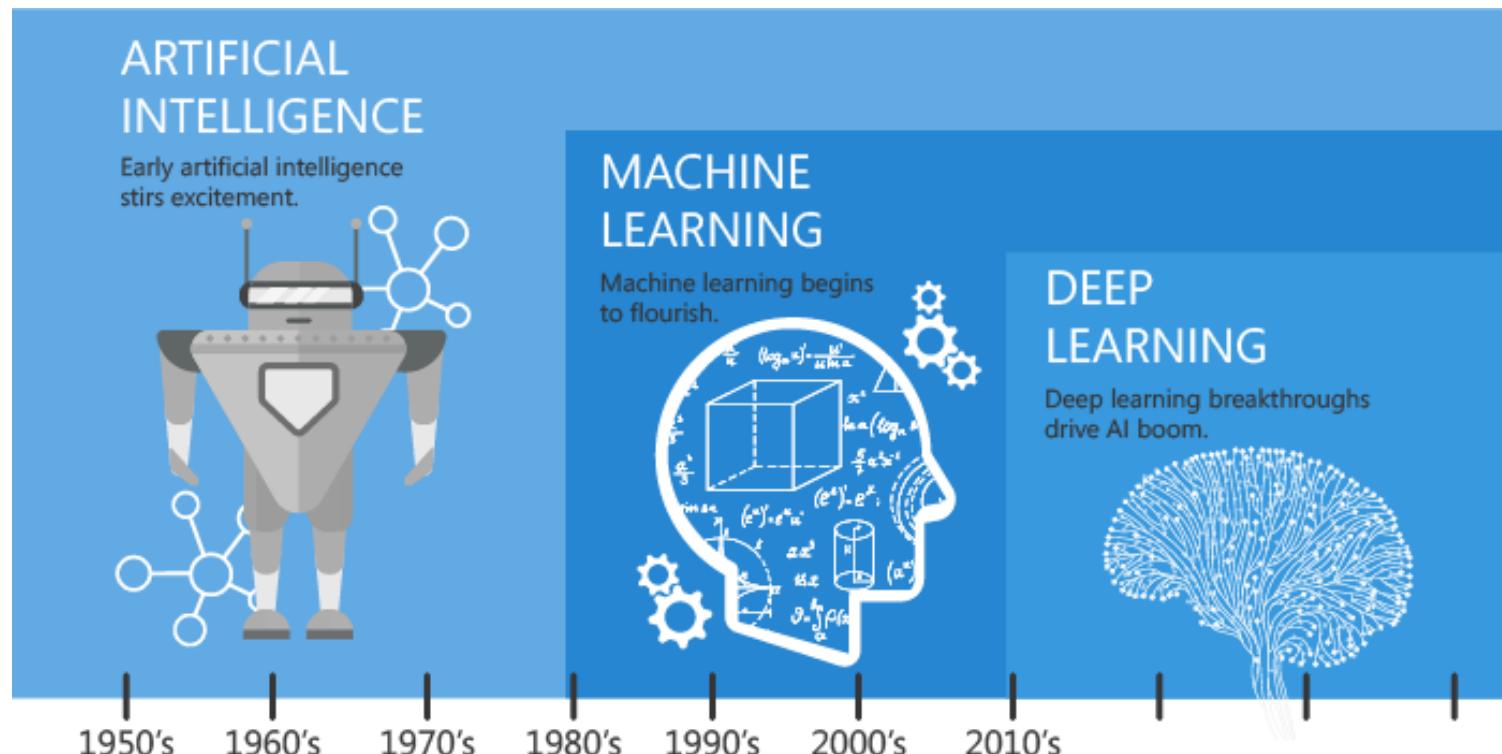
<https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/>

# ¿Qué podemos hacer?

- Gestión de clientes:
  - Predecir perdidas de clientes (customer churn)
  - Customer Scoring
  - Predicción de demanda
  - Predecir decisiones de los clientes
- Industria
  - Mantenimiento predictivo
  - Predicción de consumo energético
  - Predicción de roturas de stock



# ¿Es el machine learning una tecnología nueva?



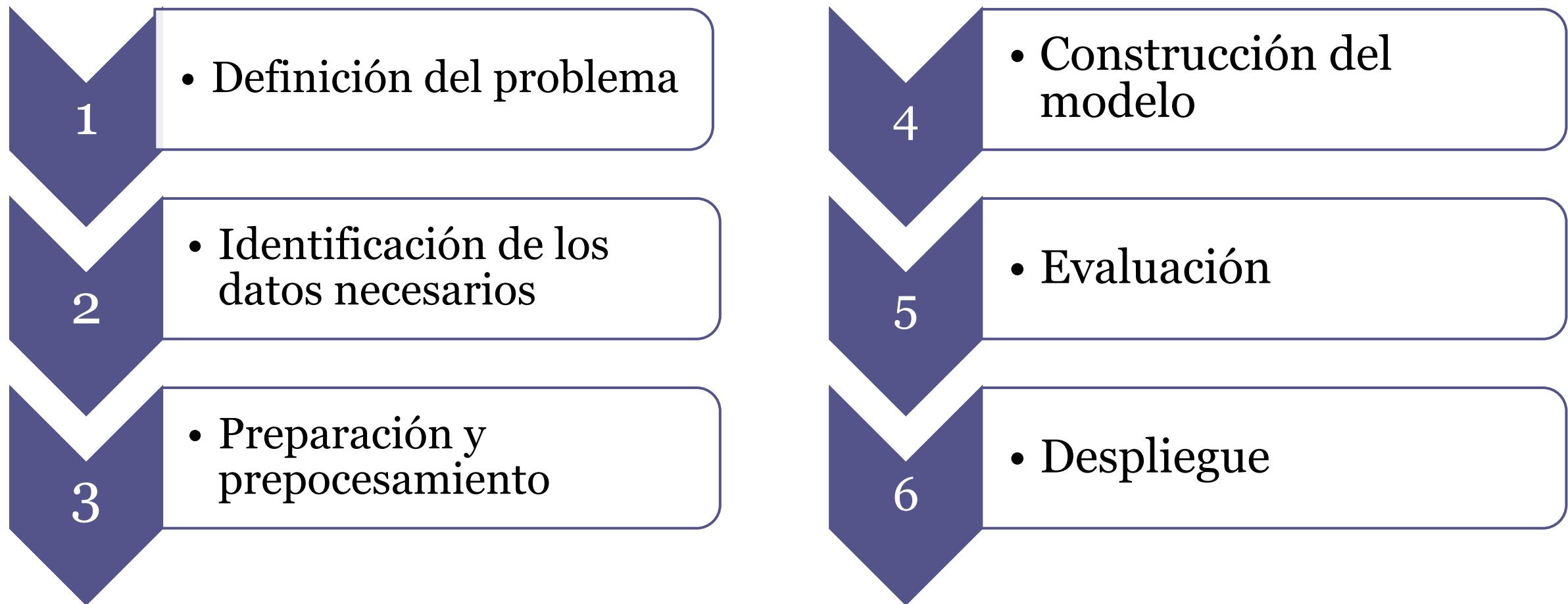
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

# ¿Es la minería de datos una tecnología nueva?

- Las técnicas más usadas no son tan modernas:
  - 50's primeras redes neuronales
  - 1974 Máquinas de soporte de vectores → 1995 consolidación
  - 1980 Fundamentos del aprendizaje profundo (Deep learning)
  - 1986 ID3 inducción de árboles de decisión y reglas
  - 1989 aprendizaje por refuerzo



# Fases de un proyecto de Minería de Datos



# 1. Definición del problema

- Identificación los objetivos organizacionales:
  - ¿Qué problema estamos intentando resolver?
    - Adquisición de nuevos clientes
    - Retención de clientes
    - Reducción de los costes de mantenimiento y operacionales, ...
  - ¿Qué medidas vamos a utilizar para determinar si se han cumplido los objetivos
- Identificación los objetivos de modelo predictivo
  - De objetivos organizacionales a objetivos de minería de datos
    - Primera aproximación a qué datos son necesarios
    - ¿Cuáles van a ser las entradas y las salidas del problema?
    - Definir la metodología a utilizar
    - Definir objetivos de desarrollo alcanzables

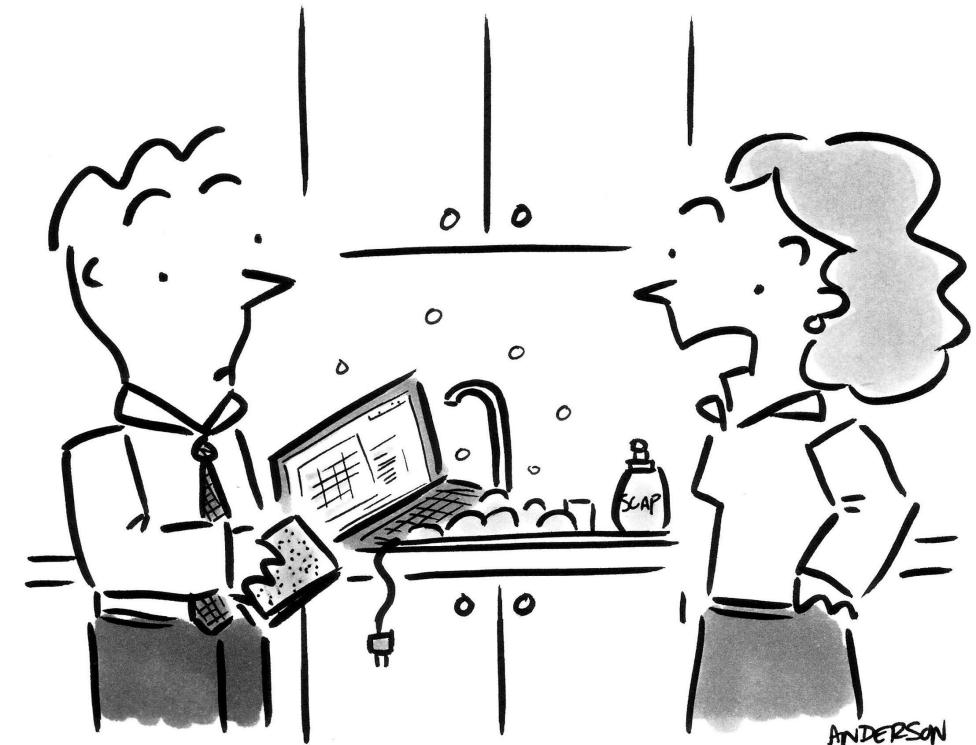
## 2. Identificación de los datos necesarios

- Recoger, describir y explorar los datos
  - Identificar las fuentes de datos
  - Analizar y comprender los metadatos
  - Analizar la calidad de los datos
    - Bad data vs. Good data
  - Herramientas de análisis datos, estadística, técnicas de visualización
  - Interacción entre los ingenieros de datos y los expertos de la organización



### 3. Preparación y preprocesamiento de datos

- Construcción del modelo de datos para la fase de desarrollo de los modelos predictivos
  - Integración de datos y limpieza de los datos
  - Transformación del formato de los datos
  - Creación de atributos derivados
  - ETL
- Importancia de los datos:
  - Es el elemento principal del proceso
    - Podemos buenos resultados sin las mejores técnicas, pero no sin “buenos” datos
  - El modelo reflejará los patrones que hay en los datos.
  - La preparación de datos es la parte más costosa del todo el proceso



"This is not what I meant when I said 'we need better data cleansing!'"

### 3. Preparación y preprocesamiento de datos

- Hay que tener cuidado con los datos sesgados
  - **Sesgo en la muestra:** los datos no representan el entorno de operación:
- Solución
  - Cubrir todos los casos a los que puede ser expuesto el sistema
  - Examinar el dominio de los atributos y construir una muestra balanceada y con una distribución equilibrada

### 3. Preparación y preprocesamiento de datos

- Hay que tener cuidado con los datos sesgados
  - **Sesgo en la muestra:** los datos no representan el entorno de operación:
  - **Sesgo por exclusión:** Eliminar atributos que creemos que son irrelevantes
- Solución
  - Analizar y estudiar todos los atributos
  - Buscar una segunda opinión
  - Estudiar la relación de los atributos con el atributo objetivo
  - Utilizar herramientas para hacer este análisis
    - Determinar la importancia de los atributos

### 3. Preparación y preprocesamiento de datos

- Hay que tener cuidado con los datos sesgados
  - **Sesgo en la muestra:** los datos no representan el entorno de operación:
  - **Sesgo por exclusión:** Eliminar atributos que creemos que son irrelevantes
  - **Sesgo del observador:** muchas veces tendemos a ver lo que esperamos ver o lo que queremos ver
- Solución
  - Analizar bien los procesos de toma de datos.
  - Asegurarse de que las personas que hagan la experimentación estén bien formadas.
  - Asegurarse que el protocolo de experimentación es correcto y está documentado.

### 3. Preparación y preprocesamiento de datos

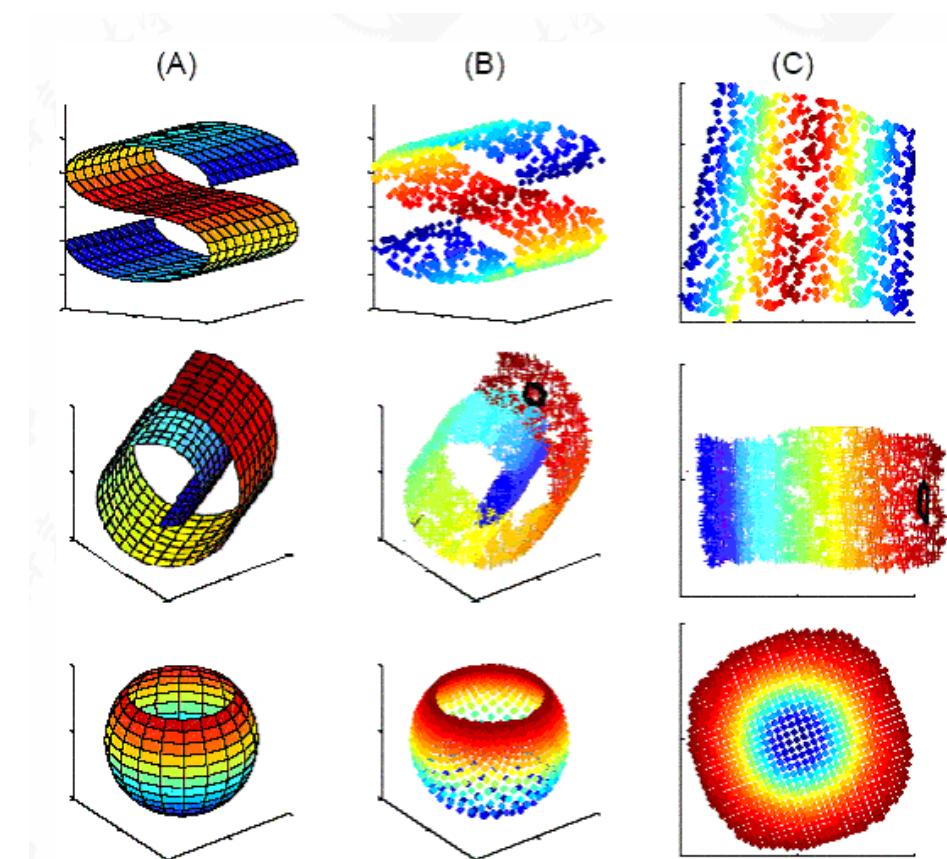
- Hay que tener cuidado con los datos sesgados
  - **Sesgo en la muestra:** los datos no representan el entorno de operación:
  - **Sesgo por exclusión:** Eliminar atributos que creemos que son irrelevantes
  - **Sesgo del observador:** muchas veces tendemos a ver lo que esperamos ver o lo que queremos ver
  - **Sesgos por prejuicios** resultado de influencias culturales o estereotipos
- Solución
  - Evitar distribuciones influenciadas por cuestiones de género o culturales
  - Buscar distribuciones más paritarias.

### 3. Preparación y preprocesamiento de datos

- Hay que tener cuidado con los datos sesgados
  - **Sesgo en la muestra:** los datos no representan el entorno de operación:
  - **Sesgo por exclusión:** Eliminar atributos que creemos que son irrelevantes
  - **Sesgo del observador:** muchas veces tendemos a ver lo que esperamos ver o lo que queremos ver
  - **Sesgos por prejuicios** resultado de influencias culturales o estereotipos
  - **Sesgo por medición:** fallos sistemáticos en los equipos de medida
- Solución
  - Tener varios equipos de medidas.
  - Adecuada calibración.
  - Comprobación continua de los equipos de medida

# 4. Proceso de construcción del modelo

- 4.1 Preprocesado de datos
  - Análisis del estado del arte
  - Feature Engineering/Extraction
  - Adaptar los datos a las técnicas utilizadas
    - Variables dummy.
    - Discretización.
    - Eliminación de ruido.
    - Cambios de escala....
    - Valores ausentes
- 4.2 Selección de variables.
  - Filtros
  - Wrappers
  - Reducción de la dimensionalidad



<http://jntsai.blogspot.com/2015/04/ammai-nonlinear-dimensionality.html>

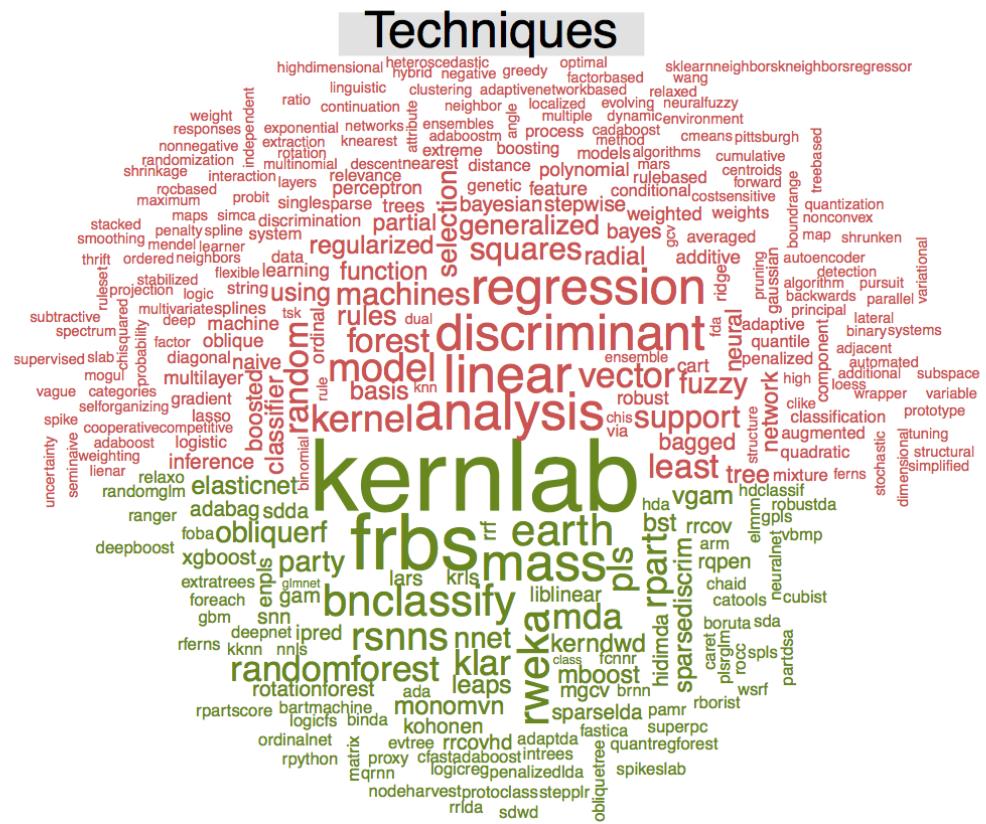
# 4. Proceso de construcción del modelo

- 4.3 Creación de los conjuntos de training y test:
  - Necesario para evaluar la capacidad de generalización del modelo
  - Nos permitirá detectar el overfitting.
- 4.4 Selección de la técnica de muestreo:
  - Hold-out
  - Cross-validation
  - Bootstrap
  - A tener en cuenta
    - Estratificación
    - Repetición



## 4. Proceso de construcción del modelo

- 4.5 Selección de las diferentes técnicas y aplicarlas
    - Análisis del estado del arte
    - ¿Necesitamos modelos interpretables?
    - Analizar y comprender los parámetros de los modelos
    - ¿Qué características de los modelos afectan a los datos?
  - 4.7 Búsqueda de hiperparámetros
    - ¿Qué parámetros hay que determinar?
    - Búsqueda de hiperparámetros
      - Empezar por un espacio pequeño



<https://alastairrughworth.github.io/Doing-machine-learning-in-R/>

# 5. Evaluación

- Evaluar si los modelos generados cumplen la expectativas.
  - **Primera fase:** sólo expertos en minería de datos
  - Seleccionar los criterios de evaluación
    - Medidas de rendimiento
  - Puede implicar volver a la anterior fase
  - Seleccionar los mejores/mejor modelo
  - Detectar el overfitting evaluando en el conjunto de test.

## Performance metrics

For each class (or for two class problems):

Precision / PPV	$tp / (tp + fp)$	$\text{green} / (\text{green} + \text{red})$
Recall / Sensitivity	$tp / (tp + fn)$	$\text{green} / (\text{green} + \text{orange})$
Specificity	$tn / (tn + fp)$	$\text{blue} / (\text{blue} + \text{red})$
Accuracy	$(tp+tn) / (tp + fp + fn + tn)$	$(\text{green} + \text{blue}) / (\text{green} + \text{orange} + \text{red} + \text{blue})$
F1-score	$2 * \text{prec} * \text{sens} / (\text{prec} + \text{sens})$	

Basic elements for each class:

- true positives
- false positives
- false negatives
- true negatives



## 5. Evaluación

- **Segunda fase:** analizar el modelo en el contexto organizacional
  - ¿El modelo satisface los objetivos organizacionales?
  - ¿Hemos tenido en cuenta todos los aspectos organizacionales?



# 6. Despliegue

- Integrar el modelo en la infraestructura tecnológica
  - Debe estar planificado con antelación
  - Facilitado por las herramientas utilizadas
  - Análisis de rendimiento
- Mantenimiento
  - ¿Cuándo se debe reentrenar el modelo?
  - Utilizar los indicadores establecidos en la primera fase
  - Considerar la aportación de nuevos datos



# 6. Herramientas

Lenguajes	Plataformas	Big Data	Open Source tools
<ul style="list-style-type: none"><li>• Python<ul style="list-style-type: none"><li>• Scikit-learn</li><li>• Tensorflow</li><li>• Keras,</li><li>• Theano,</li><li>• PyTorch</li></ul></li><li>• R<ul style="list-style-type: none"><li>• Caret</li><li>• Keras</li></ul></li><li>• C++,C#,Java, ..</li></ul>	<ul style="list-style-type: none"><li>• Google Cloud Machine Learning Platform</li><li>• Microsoft Machine Learning Studio</li><li>• AWS machine learning</li><li>• Watson Machine Learning</li><li>• Oracle Machine Learning</li></ul>	<ul style="list-style-type: none"><li>• Apache Spark MLlib</li><li>• Hadoop</li><li>• Node.js</li></ul>	<ul style="list-style-type: none"><li>• Accord.NET</li><li>• KNIME</li><li>• Weka</li><li>• Orange</li><li>• Shogun</li></ul>

# 6. Herramientas

Lenguajes	Plataformas	Big Data	Open Source tools
<ul style="list-style-type: none"><li>• Python<ul style="list-style-type: none"><li>• Scikit-learn</li><li>• Tensorflow</li><li>• Keras,</li><li>• Theano,</li><li>• PyTorch</li></ul></li><li>• R<ul style="list-style-type: none"><li>• Caret</li><li>• Keras</li></ul></li><li>• C++,C#,Java, ..</li></ul>	<ul style="list-style-type: none"><li>• Google Cloud Machine Learning Platform</li><li>• Microsoft Machine Learning Studio</li><li>• AWS machine learning</li><li>• Watson Machine Learning</li><li>• Oracle Machine Learning</li></ul>	<ul style="list-style-type: none"><li>• Apache Spark MLlib</li><li>• Hadoop</li><li>• Node.js</li></ul>	<ul style="list-style-type: none"><li>• Accord.NET</li><li>• KNIME</li><li>• Weka</li><li>• Orange</li><li>• Shogun</li></ul>

## 8. Otros aspectos a considerar

- Aspectos éticos
  - Algunos modelos pueden afectar directamente a las personas:
    - Hay que ser mucho cuidado con los sesgos en los datos.
      - Compas: doblaba la tasa de falsos positivos para los afroamericanos
      - Allegheny Family Screening Tool: sistema para decidir sobre la custodia de niños bajo abusos
    - Garantizar el acceso a la IA.
  - Aspectos legales
    - GDPR:
      - Derecho a la explicación
      - Datos personales
      - Origen de los datos.
    - Lawful
    - Ethical
    - Robust



# Conclusiones

- Necesidad de equipos multidisciplinares
- Es importante la formación
  - No se pueden utilizar las herramientas de forma automática
  - Hay que saber bajo qué condiciones se pueden aplicar cada modelo
  - Hay que saber interpretar los resultados.
- Gracias a las herramientas la aplicación del machine learning se ha convertido en un proceso de ingeniería.
  - Importancia de la selección de la herramienta y lenguaje
- A la hora de desplegar infraestructuras tecnológicas tener en cuenta las posibles y futuras aplicaciones.

# Gracias por su atención

José Tomás Palma Méndez ([jtpalma@um.es](mailto:jtpalma@um.es))

Grupos de Investigación en Inteligencia e Ingeniería del Conocimiento.

Facultad de Informática  
Universidad de Murcia



Saturdays.AI  
Murcia