# CLASIFICACION AUTOMATICA BASADA EN ANALISIS ESPECTRAL. CASO DE USO: PROCEDIMIENTOS CLASIFICATORIOS APLICADO A ASTEROIDES

**Tesista:** Gregorio Perichinsky[1,2]

**Director de la Tesis:** Prof. Dr. Ángel Luís Plastino[2].

(1) Facultad de Informática, UNLP. E-mail: gperichinsky@acm.org
(2) Facultad de Ciencias Exactas, Instituto de Física, Universidad Nacional de La Plata (CCT La Plata Laboratorio Protem – CONICET) E-mail: Plastino@fisica.unlp.edu.ar

**Fecha de presentación**: 28 de diciembre de 2008.
**Lugar:** Facultad de Informática de la Universidad Nacional de La Plata.

## RESUMEN

Esta tesis aborda la definición de un método numérico basado en invariantes para la clasificación automática de objetos a partir de la información de sus caracteres, focalizado en la búsqueda de las invariantes con base en una aplicación original metodológica de los principios de superposición e interferencia en el análisis de espectros, en congruencia analógica con la taxonomía numérica, por su relación lógica y con fortaleza metodológica.
Se demuestra un nuevo criterio para dar validez al método en casos no resueltos hasta ahora por la ciencia.
En esta introducción de la tesis se trata especialmente de excitar la atención, para su desarrollo, "el problema, el marco teórico y el papel asignado a las hipótesis y la realidad", dando origen a las tareas de Investigación y de Explicación Científica.
Siendo la motivación principal del enunciado verdadero, utilizando leyes y datos, es necesario conceptuar la problemática epistemológica (primera parte 1.1.) y un programa de investigación científica (PIC) como sucesión de teorías emparentadas en forma semántica y sintáctica, que se van generando en distintas disciplinas por observaciones intrigantes, que se captan históricamente y llaman la atención pues se comportan en forma desconcertante o funcionan de una manera diferente a la esperada, constituyendo familias de fenómenos intrigantes (segunda parte 1.2.).

**Palabras clave:** otu, escala, clasificación (similitud), cluster (familia), espectro, invariante.

### 1.1. De la Epistemología.

#### 1.1.1. Etapas de la investigación científica.

Se consideran las siguientes etapas básicas de la investigación científica [Gianella, A.E., 2000]: Problema, Hipótesis, Marco Teórico, Procedimientos deductivos, Consecuencias contrastables, Procedimientos de contrastación, Evaluación de los resultados y según el resultado se verifica la hipótesis o se genera una nueva hipótesis refutando la anterior. Lo que da origen a las tareas de Investigación son el problema, el marco teórico y el papel asignado a las hipótesis y la realidad y no, el mero relevamiento de datos.
De los problemas se formulan preguntas y hay que intentar responderlas o explicarlas, trascendiendo el contexto del conocimiento del estado de una disciplina, respecto a la realidad relativa.
La implicación y la generalización de los problemas y preguntas se pueden ordenar en grados: gradación [Perichinsky, G. Investigation,1995].
En las investigaciones el marco teórico, con sus componentes de una o más teorías, homogéneas o heterogéneas, está presente a través de sus hipótesis, condicionando a los interrogantes o preguntas

que se formulan, e interesarse por algo o por avanzar en determinada dirección.
Mediante una hipótesis o conjetura una vez formulado con claridad el problema, que se va a investigar, y se procederá a buscar su solución.
Las hipótesis de trabajo [Klimovsky, G. 1994] se estratifican en tres niveles: (1) Descripción de individuos u objetos (artefactos) de bajo nivel que describen, analizan, registran, enumeran y atribuyen propiedades. Los objetos toman relaciones de la Base Empírica formada por conjuntos de entidades, fenomenologías, propiedades y relaciones programadas; (2) Nivel intermedio de observaciones que generalizan, correlacionan, subsumen y clasifican; es un nivel preteórico y, (3) Hipótesis de máximo nivel y observaciones, que explican, predicen, comprenden, sistematizan, inventan soluciones y metodologías.
Esta tarea resulta tener muchas veces valor heurístico, es decir, contribuye a estimular la creatividad del científico.
Los mecanismos de producción de ideas y de resolución de problemas, "lógica del descubrimiento", recurren a procedimientos de la lógica formal, el cálculo de predicados o la teoría de conjuntos y se suelen utilizar sistemas lógico-matemáticos, premisas encadenadas, en tanto marco teórico.

**1.1.2. Consecuencias observacionales.**

Surgen consecuencias contrastables como enunciados inferidos deductivamente de las hipótesis, susceptibles de confrontación con la experiencia. El lenguaje es observacional, no teórico.
Si fueran enunciados acerca de propiedades, hechos o relaciones ya conocidos, las hipótesis los explican.
Los procedimientos de contrastación de las consecuencias observacionales, son una etapa crucial de la investigación científica.
Los momentos de aplicación son, (1) el diseño del experimento, (2) la realización del experimento y ambiente de laboratorio o de campo, y (3) el registro y evaluación de los resultados obtenidos.
En lugar de la experimentación se realizan observaciones sistemáticas o una experimentación ex post facto.
El enunciado de una ley es una hipótesis general empíricamente confirmada, inmersa en una teoría, sistema hipotético-deductivo representante de una regularidad objetiva [Bunge, M., 1969, 1983, 1999].
Los enunciados están cargados teóricamente y solo son aceptados como un acuerdo de la comunidad científica [Imre Lakatos, 1999].

**1.1.3. Desarrollo de la ciencia.**

La historia de la ciencia ha mostrado que las cosas no sucedieron de acuerdo con estos criterios tan simples; por lo cual hay que reemplazarlos por versiones refinadas de los mismos principios:
El desarrollo de la ciencia, surge de la competencia de una secuencia de teorías que comparten un núcleo duro (hard core), formado por hipótesis. Un programa de investigación científica (PIC) es una sucesión de teorías emparentadas $T_1$, $T_2$, $T_3$,…,$T_n$. Tienen en común un conjunto de hipótesis fundamentales que forman su núcleo duro, al cual se lo declara "irrefutable" por decisión de la comunidad científica.

**1.1.4. Modelo hipotético-deductivo.**

En el modelo hipotético-deductivo, dinámico y holístico [Bunge, M 1999], se rechaza todo ese desarrollo de modelos, con lo cual coincido [el autor], pues los enunciados básicos no pueden verificarse por observación o experimentación, rechaza la lógica inductiva, ya que, cualquier ley universal tendrá probabilidad cero de demostrarse, pues por inducción infinita la investigación científica no logra su verdad. Además no demarca la Ciencia de la No Ciencia, entre afirmaciones teóricas y observacionales.
Se distingue en la Historia Interna de una disciplina o teoría científica a la que incluye a las variables que pueden cambiar a la teoría, si las cuestiones metodológicas lo indican, de la Historia Externa con los elementos de Bunge.
La actividad de la epistemología como investigación de la ciencia, es una situación dialéctica con

los científicos; de aprendizaje mutuo. Para Albert Einstein, los procedimientos operacionales, no ocultan el significado de los términos teóricos, ligados a la noción de teoría no a la definición operacional, pero se aplican las técnicas operacionales para introducir conceptos.

**1.1.5. Evolución científica.**

La evolución científica, por acomodación y equilibrio, "no se puede concebir, como un acercamiento por aproximaciones sucesivas a la realidad."

"La historia de la ciencia y, en particular, la de la tecnología, es una larga y clarísima descripción de cómo los medios técnicos y los procedimientos de la ciencia para mejorarla muestran un progreso, aumento de eficacia y operatividad, pese a que a lo largo del tiempo los paradigmas se sustituyen unos a otros." [Klimovsky, G. 1994]

Es un proceso que se denomina acomodación y, a diferencia de la asimilación (semejante al de ciencia normal), es característico de las etapas de cambio en los procesos evolutivos, que finaliza cuando se alcance un nuevo estado de equilibrio, en el cual el organismo recobra las facultades de asimilación.

**1.1.6. Reduccionismo.**

Se advierte la conexión entre reducción y explicación, si existe un procedimiento para reducir una disciplina a otra y, una teoría a otra de una disciplina anterior, es el "Reduccionismo metodológico", al reducir una teoría básica a otra reducida, que implica una reducción semántica del lenguaje de una teoría básica al lenguaje de otra reducida, con el resultado que, al hacerlo, se descubra que una teoría es derivada de otra. Primero, la reducción semántica por utilizar ambas teorías vocabularios diferentes, segundo la dependencia deductiva de una con relación a la otra y tercero, porque la "máquina de deducir", no nos permitiría acceder a las locuciones de la teoría reducida a partir de las de la teoría básica. Una teoría queda explicada por aquella a la que metodológicamente se reduce.

**1.2. De la Tesis.**

Para abordar un método numérico basado en invariantes para clasificar objetos en forma automática, a partir de la información de sus caracteres, focalizado en la búsqueda de invariantes, el análisis espectral, la taxonomía computacional y la teoría de la información, del desarrollo de la ciencia y la generación de conocimiento, surge la competencia de una secuencia de teorías, de un programa de investigación científica (PIC), que tiene en común un conjunto de hipótesis fundamentales, a fin de lograr un ajuste entre teorías y resultados experimentales.

Se demuestra la validez del método en casos no resueltos hasta ahora por la ciencia.

En las investigaciones, el marco teórico con sus componentes de una o más teorías, homogéneas o heterogéneas, está presente a través de sus hipótesis, condicionando a los interrogantes o preguntas que se formulan, e interesarse por algo o por avanzar en determinada dirección.

Mediante una hipótesis o conjetura, una vez formulado con claridad el problema que se va a investigar, se procedió a buscar su solución.

La implicación y la generalización de los problemas y preguntas se pueden ordenar en grados: gradación [Perichinsky, G. Investigation,1995].

La gradación básica para la aplicación, en casos de uso, de una metodología original, de los principios de superposición e interferencia, para la generación y análisis de espectros, en congruencia analógica con la taxonomía computacional numérica y el teorema de Tchebycheff, los paradigmas de las bases de datos y herramientas emergentes de la inteligencia artificial, para verificar la fortaleza del método y además, la teoría de la información y la máxima entropía (PME).

De acuerdo a todo lo expresado surge un conjunto de explicaciones de los problemas a resolver en las distintas eras científicas.

**1.2.1. La pre ERA científica (Aristóteles)**

Se puede decir que comienza unos 2.000 años antes que Newton, Aristóteles en su Liceo y con su discípulo Teofrasto, que lo sucedió en la dirección del Liceo (peripatos), fue el fundador de la botánica. En el Liceo había una biblioteca, un zoológico y un jardín botánico. Tenía colecciones de mapas y de minerales, y varias aulas y talleres donde se estudiaba e investigaba.
Se realizaban simposios, y en ellos surgió el nombre de la Física, la Meteorología, la Economía, la Poesía, la Etica y la Política. También se enseñaba lógica, biología, medicina, astronomía, historia y sociología. Surgiendo así la competencia de una secuencia de teorías, de un programa de investigación científica (PIC), y su gradación frente a problemas.
La clasificación (taxonomía), milenios antes que Carlos de Linneo, de los vegetales, árboles, arbustos, matas y hierbas es mucho más racional que otras que se usaron antes del siglo XVII.
La organización de género, orden, división, reino, especie, familias y clases, con el principal propósito de dar esta jerarquía y plantear relaciones de evolución entre individuos.
El geocentrismo, el sistema de las esferas planetarias y los problemas de la Física como la sensación de recibir luz y calor del sol. La explicación de los griegos era que el sol y todos los cuerpos que irradian luz y calor debían arrojar corpúsculos pequeñísimos cuyo choque contra el ojo o la piel producía las sensaciones de luz y calor. Por otra parte, el arco iris atrajo su misticismo, era incapaz de explicarlo y se inclinó por una importancia sobrenatural, y aunque familiar en los casos más simples de refracción, no la conectó con el arco iris-espectro en su forma natural.

**1.2.2. La ERA científica (Newton siglo XVII).**

La clasificación (taxonomía) tomando como referencia a Michel Adanson como su iniciador, a través de sus estudios del Estado Operativo, con la cantidad de información y caracteres de los individuos, y de Carlos de Linneo, hasta nuestros tiempos (hace 300 años), con un punto débil en la ausencia de significados cuantitativos en términos clasificatorios, siguieron y explicaron más las jerarquías existentes y rangos y relaciones doctrinarias en la clasificación con la similitud o fenotípica y los orígenes o genotípica que bloquearon el avance de la similitud y de allí el planteo evolucionista de Charles Robert Darwin.
El empirismo (John Locke, siglo XVII) epistemologicamente indica que todo conocimiento depende de la experiencia y toda teoría debe verificarse experimentalmente.
Darwin explica el fenómeno de la evolución por un mecanismo de mutaciones aleatorias sucesivas. Los individuos sufren a continuación la selección natural: los mejor adaptados sobreviven y se reproducen y los otros desaparecen, siguiendo a la Filosofía zoológica (1809) e Historia de los animales invertebrados (1815-1822), de Jean-Baptiste de Lamarck.
Para Aristóteles y en casi toda la pre ERA científica los individuos eran inmutables. De todas maneras hay que llegar al siglo XX para que se reconozcan estas hipótesis.
Sir Isaac Newton representa las ideas tempranas del Color [Sawyer, R.A 1963].
La teoría corpuscular fue aceptada hasta el año (1800).
Christiaan Huygens en el siglo XVII, quien, partiendo de los fenómenos observados de la transmisión de las ondas de agua sobre la superficie de un estanque o de las ondas sonoras a través del aire, sostuvo que la luz podría ser alguna perturbación vibratoria transmitida por algún medio que llena todo el espacio interestelar, que denominó éter luminoso o transportador de la luz.
Evolucionó, por otra parte, debido a las leyes de Newton de la mecánica y de gravitación universal y de la Mecánica Celeste debido a las ecuaciones de Johannes Kepler; y al aceptar, Newton, la teoría corpuscular, la teoría del éter o teoría ondulatoria, tuvo pocos adeptos, hasta que los fenómenos de interferencia, escapaban a cualquier explicación basada en la teoría corpuscular, mientras que eran explicados por la teoría ondulatoria.
Los hallazgos experimentales fueron expresados por las ecuaciones diferenciales en derivadas parciales de James Clerk Maxwell. Las ecuaciones de Maxwell relacionan los cambios espaciales y temporales de los campos permitiendo calcularlos en cualquier momento. Al resolver las ecuaciones se "*predice*" un nuevo tipo de campo electromagnético producido por cargas eléctricas en movimiento acelerado. Este campo se propaga por el espacio con la velocidad de la luz en forma de

onda electromagnética (radiación). En 1887, Heinrich Hertz generó esas ondas por medios eléctricos, sentando las bases para la radio, el radar, la televisión y otras formas de telecomunicación: *"Concluyen en esa época con la corroboración y la respectiva explicación."*

**1.2.3. La ERA de la ciencia Moderna.**

Comienza en el siglo XX (Einstein) y en los primeros 30 años, aparecen los principales conceptos en todas las disciplinas, Teoría de la Información, Lingüística y Cibernética en Ciencias de la Computación.

La taxonomía tradicional inclusive la post Darwiniana evolucionó en conceptos y procedimientos.

Las nuevas formas sistemáticas, el desarrollo de la genética, la biogeografía y la paleontología aportan la base matemática y experimental a la teoría de Darwin constituyendo el neo darwinismo y la taxonomía (1920-1950).

Son algo más que simples generalizaciones descriptivas (se intentó hacer todas y no se hizo ninguna bien) [**el autor**].

Se deben establecer criterios para definir categorías y operaciones, para no caer en discusiones científicas sin sentido.

En los años 1950 en adelante se produce un punto de inflexión cuando H. J. LAM define al taxón y George Gaylord Simpson y Blackwelder definen a la taxonomía numérica desarrollando su teoría y metodología.

En 1962 Sneath, Sokal y Rohlf, a partir de las metodologías, logran clasificaciones precisas y publican luego los principios de la taxonomía numérica.

La evidencia taxonómica que implica la selección de objetos (organismos) de estudio, la selección y definición de caracteres taxonómicos y los criterios homológicos.

En general se trata de mantener un simbolismo uniforme para los caracteres, unidades taxonómicas operacionales (OTU) y taxones (taxones ≡ taxa: grupos de OTU´s).

La taxonomía numérica es el agrupamiento de unidades taxonómicas por métodos numéricos en TAXONES (TAXA) en base a los estados de sus caracteres.

Para las instancias de la matriz de similitud se trabaja con la distancia taxonómica, pudiendo utilizar así la distancia de Hamming y el coeficiente de coincidencia de Jaccard (Romesburg, 1984, p. 143) y finalmente el coeficiente de Cower de Sokal y Sneath, aunque siempre se termine utilizando la simple distancia Euclídea, o de la Teoría de la Información de Gluck y Corter (1985) utilizan la categoría de utilidad, la sumatoria de algún valor de instancias dividido por la cantidad de instancias, de Fried y Holyoke (1984-1988).

La gran cantidad de información e instancias llevó a la utilización de computadoras y la necesidad de generar algoritmos eficientes, como Gennari et al., 1989.

Finalmente, usando distancias Euclídeas (o con métrica de Manhattan) puedo computar la matriz de Distancia Taxonómica o de Similitud o de Semejanza o Matriz de Coeficientes de Similitud, Matriz mediante la cual deseo encontrar la estructura taxonómica de dimensiones (t x t) donde t es el número de OTU´s.

Los clusters son los conjuntos de OTU´s en el hiperespacio, fenotípicos en término de patrones.

El centro del cluster o centroide representa un objeto promedio, que es simplemente una construcción matemática, que permite la caracterización de la Densidad y la Varianza, y el Radio y Rango del taxón.

En un hiperespacio se pueden representar las posiciones de los t OTU´s en un sistema de coordenadas, si dichas posiciones son cercanas, la distancia disminuye hasta hacerse cero si coinciden, así la distancia puede ser vista como el complemento de la similitud.

A partir de los dominios normalizados se calculan la diferencia media entre caracteres, tomándose el valor absoluto de la diferencia pues esta puede ser negativa, y la distancia taxonómica donde se pueden considerar las métricas de Minkowski y de Manhattan.

El método de clustering cuyo acrónimo es SAHN resume lo encontrado anteriormente: Sequential, Agglomerative, Hierarchic and Nonoverlapping.
En las técnicas en las cuales la estrategia es la distorsión del espacio parece como si el espacio, en la inmediata vecindad de un cluster se ha contraído o dilatado. Si volvemos al criterio de admisión para un candidato que se une a un cluster existente, este espacio vecino es constante sobre todo en el método pair-group.
El tratamiento dinámico e integrado de los dominios permite una fácil normalización, **atributo - dominio - valor**, y la implementación en el modelo de Base de Datos Relacional Dinámica y su utilización en Taxonomía Numérica.
La contribución **<teórica , empírica>** es la aglomeración de objetos formando clases producidas por pasos del método obteniendo clusters y dominios con valores normalizados y la densidad y el rango en términos del radio del conjunto puede ser visualizado como una INVARIANTE CARACTERÍSTICA de los OTU´s.

**1.2.3.1. La problemática astronómica: asteroides.**
De acuerdo a los problemas astronómicos, caso de uso fundamental de esta tesis, pues examinando la distribución de los asteroides con respecto a sus elementos orbitales, en particular su movimiento principal, la inclinación y la excentricidad, se observan condensaciones en distintos lugares que parecen al azar, pero hay algunos casos en los cuales tener en cuenta solo las leyes de la probabilidad, no es tan evidente, (Hirayama, K. a partir de 1918).
Los asteroides están demasiado agrupados por tener inclinaciones cercanas o los planos orbitales tienen prácticamente el mismo polo, por ello es que se podía aventurar que existen familias de asteroides asociados.
Así para J. R. Arnold, 1969, la distribución de elementos orbitales en cinturones de asteroides no es al azar mostrando la existencia de familias se aproximan a clusters para ciertos valores especiales.
Según Arnold siguiendo la ley de Poisson el número de elementos de un conjunto debe ser menor que un cierto número esperado, con la cual no se concuerda en esta tesis pues los eventos no siguen esta ley por contradecir los grandores físicos, las características fenotípicas de caracteres o atributos de los asteroides y finalmente su genotípica u origen común.
Toda esa conclusión parece ser arbitraria pues debe prevalecer el concepto conservativo de la masa es decir la densidad y la estabilidad del entorno.
Condiciones de vecindad cercana deben ser tenidas en cuenta y las familias de alta densidad son las más estables y menos azarosas.
Investigadores han llegado a la conclusión que el problema de clasificación de los asteroides en familias está definido y prácticamente resuelto, visión simplista que esta tesis no comparte.
Los criterios de rechazo de un miembro de una familia no son claros, son arbitrarios o directamente no se exponen en los trabajos y por lógica consecuencia no son automáticos.
Se puede observar que el crecimiento en observaciones entre 1969 de Arnold, y 1990, Carusi, Masaro, 1978, Williams, 1979-1989 y Knezevic, Milani, .1990, traen discrepancias.
Las discrepancias surgen de los métodos de cómputo de los elementos propios, del criterio de rechazo de los objetos a ser clasificados, del tamaño de la muestra, de los métodos de identificación de familias y los criterios de rechazo de un miembro de una familia.
Los métodos de cómputo de los elementos propios (no efemérides) tratan la eliminación de las perturbaciones seculares de planetas sobre el verificado Cinturón principal de asteroides.
El algoritmo permite calcular un código de calidad (QC) que indica cuantas iteraciones hay que realizar para que converja.
La cantidad de asteroides que se pueden recalcular es con alrededor de 55 iteraciones y siempre que la inclinación no sea grande al igual que la excentricidad (Jet Propulsion Laboratory, California Institute of Technology).

Todo este desarrollo aparece poco claro y arbitrario, no hay un sustento formal en la relación convergencia cantidad de iteraciones y el número de asteroides.
Las familias de Hirayama se estabilizaron con valores propios de decenas de miles de asteroides nuevos descubiertos.
Se produce así un nuevo punto de inflexión, segundo, al poderse desarrollar a partir de 1990 y 1994, algoritmos de cómputo que permitían obtener clusters en forma automática sin intervención arbitraria de investigadores condicionando al algoritmo ni nivel jerárquico de fenogramas o dendrogramas, donde se sacan primero los fenogramas, el investigador decide el nivel jerárquico y luego se encuentran los clusters o familias [Crisci, López Armengol, 1983, página 69].
También se utilizaron herramientas de la inteligencia artificial como la teoría de onditas (WAVELET), Algoritmos Genéticos y redes neuronales (RRNN).
Zappala, Cellino, Farinella y Knezevic (1990-1994) y Bendjoya y Cellino con Hergenrother (1992-1996), tienen un criterio que es importante donde una clasificación de los asteroides mejorada es nombrada en las familias dinámicas, mientras analizando una base de datos de asteroides numerados cuyos elementos propios se han computado en un nuevo segundo-orden, cuarto-grado de la teoría de perturbación secular, y verificada su estabilidad en términos largos. El criterio multivariado usa la técnica de análisis de datos agrupándose en orden jerárquico. Fue aplicado para construir para cada zona del cinturón de los asteroides un "dendrograma", gráfico, en el espacio de los elementos propios, con una distancia en función relacionada a la necesaria velocidad incremental del cambio orbital después de la eyección del cuerpo del padre fraccionado.
Las familias se identifican entonces por la comparación con los dendrogramas similares, los derivados de una "casi randómica" distribución de elementos que comparan la estructura para una escala gruesa (bruta) de la distribución real.
Los parámetros de importancia asociados con cada familia, medidos como resultados de las concentraciones aleatorias, (como para transformar zonas anisótropas e in-homogéneas en zonas homogéneas e isótropas de las zonas intra-espacios (inter-gaps) en el cinturón de asteroides, modificando los atributos mecánicos como el semi-eje mayor y la inclinación) y los parámetros de robustez (estabilidad), se obtuvo repitiendo el procedimiento de la clasificación después de variar los elementos de velocidad en cantidades pequeñas al volver a computar las zonas reales de los cálculos con el cambio artificial de los coeficientes de la función de distancia.
Para tomar promedios de variación de distancias estaban armadas las designadas estalactitas, mientras tomando la anchura y la profundidad en la función de la velocidad modificada. Siendo un criterio innovador es importante analizarlo aunque no está claro la técnica del arreglo, mientras agrupándose, dentro de las zonas y los promedios de variación de velocidades, como antes de mencionó, y por otro lado se ignoran las familias de hasta cinco elementos y con solapamiento, todos son la síntesis de una instrumentación arbitraria.
Las familias más importantes y confortables son las de costumbre que juntas constituyen el 14% del cinturón principal conocido, de la población; pero 12 familias más confiables y confortables que se encontraron a lo largo del cinturón, la mayoría partió parcialmente de clasificaciones anteriores miembros son los taxonómico diferentes de los precedentes, de aquéllos con menos de cinco miembros no serán definitivamente diferentes (algo que no implica que ellos necesariamente y genéricamente serán "irreales").
Después de más de 100 años de las familias de Hirayama y los avances en Taxonomía Computacional llego a un nuevo criterio, que produce junto con intentos con Teoría de Onditas, Algoritmos Genéticos y Redes Neuronales a un tercer punto de inflexión, que comienza en 1998 pero sigue con el nuevo siglo, XXI.
Con estas motivaciones, un Criterio Espectral, en el Análisis de Clasificación, he decidido lograr el análisis espectral, las clasificaciones se extendieron a la base de datos de elementos propios de asteroides en familias. Reconozco que los trabajos de Zappala son muy importantes (clasificación automática y método jerárquico), y un punto de inflexión en los tempranos 90´s pero es diferente el

acercamiento porque trabajo en taxonomía computacional, en un hiperespacio taxonómico, y no en un criterio de composición y precedentes físicos y cosmoquímicos. Zappala y otros usan una metodología confundiendo, ambos, al tratar con sólo una variable de velocidad, un espacio transformado no claramente unívoco.
La decisión es lograr la clasificación en familias, que extienden el uso de la base de datos de elementos propios de asteroides, con un criterio de análisis espectral futuro. Incorporando así un conjunto actualizado y más grande de elementos oscilantes que se derivaron de la teoría de perturbación secular cuya exactitud (específicamente, la estabilidad en el tiempo) se ha verificado extensivamente por la integración numérica a largo plazo; en forma automática, y perjudicar la técnica de análisis de datos en los grupos del no-azar, no se usa en el espacio de elementos propios como en el criterio de Zappala y cuantitativamente la importancia estadística de estos grupos; con la robustez de las estadísticas para las familias importantes con respecto a las variaciones aleatorias pequeñas de elementos propios, todos basados en un análisis de Taxonomía Computacional.
No considero la transformación isotrópica y los conjuntos homogéneos, mientras cambiando los valores de la excentricidad y el semiejes al volver a computar los valores de las zonas de entre-espacios del cinturón de los asteroides en las velocidades en promedio, o los grupos eliminados de 5 o menos objetos y familias que se solapan, todos los cuales considero están fuera de un criterio Computacional.
Estos clusters constituyen familias, mediante el análisis estructural, basado en sus características fenotípicas, exhibiendo sus relaciones, en lo que se refiere a grados de similitud, entre dos o más OTU´s.
Entidades formadas por dominios dinámicos de atributos, cambien según los requisitos taxonómicos: la clasificación de objetos para formar familias o clusters.
Se representan aquí los objetos de Taxonomía por la aplicación de la semántica del Modelo de Base de Datos Relacional Dinámica.
Se obtienen familias de OTU´s empleando como herramienta i) las distancias Euclideas y ii) las técnicas del vecino más cercano. Así la evidencia taxonómica se reune para cuantificar la similitud para cada par de OTU´s (método pair-group) obtenido de la matriz del datos básica.
La contribución principal de la tesis presente es introducir el concepto de espectro de OTU´s, basado en los estados de sus caracteres. El concepto de espectros de familias surge, si el principio de superposición se aplica a los espectros de los OTU´s, y los grupos se delimitan a través del máximo de la relación de Bienaymé-Tchebycheff que determina Invariantes (centroide, varianza y radio).
Aplicando la técnica de dominios independientes dinámicamente integrados, para computar la Matriz de Similitud, y con el recurso de un algoritmo iterativo, se obtienen familias o clusters.
**Un nuevo criterio taxonómico se ha formulado y una aplicación astronómica ha funcionado.**

### 1.2.3.2. Robustez del Método Espectral utilizando Data Mining y la Entropía Máxima.

Un nuevo acercamiento a la Taxonomía Computacional mediante Data Mining.
Machine Learning es el campo dedicado al desarrollo de métodos de cálculo donde el aprendizaje subyacente procesa y se aplica, a los sistemas de aprendizaje basados en computadora, en problemas prácticos de Sistemas Complejos y Dinámicos. Data Mining intenta resolver esos problemas relacionados a la búsqueda de modelos interesantes y las regularidades importantes en las grandes bases de datos. Utiliza métodos y estrategias de otras áreas, incluso de Machine Learning. Al aplicar esas técnicas para resolver un problema de Data Mining, se dice que ésta es Inteligente.
En la tesis se analiza los TDIDT (Top Down Induction Trees), la familia de inducción y en particular al algoritmo C4.5. El intento es determinar el grado de eficacia logrado por el algoritmo de C4.5 cuando es aplicado en datos para generar a modelos válidos de datos en problemas de clasificación con la Ganancia de Entropía (PME).

El algoritmo de C4.5 genera los árboles de decisión y la decisión gobierna a los datos pre-clasificados. "Divida y gobierne" es el método que se usa para construir árboles de decisión. Este método divide los datos de entrada en subconjuntos según algunos criterios preestablecidos. Entonces funciona en cada uno de estos subconjuntos, que los dividen de nuevo, hasta que todos los casos presentes en un subconjunto pertenezcan a la misma clase.
Antes de hacer cada partición, el sistema analiza todas las posibles pruebas que pueden dividir el conjunto de datos y pueden seleccionar la prueba con la ganancia de información superior o la proporción de ganancia de entropía superior (ME).
El principio de máxima entropía (PME) se aplica ampliamente, no sólo en la física, en general en los procesos de cualquier naturaleza, dónde se quiere obtener información que empieza con un conjunto de datos incompleto, o usando la cantidad más pequeña en las suposiciones anteriores. El PME proporciona una formulación alternativa de la Mecánica Estadística, elegante y compacta, formulada por Jaynes, que presenta al PME como un método canónico para construir la densidad principal, relativa a variables cuyos valores medios se conocen "a priori". Los Sistemas Dinámicos utilizan en los últimos años esta metodología.
Para Albert Einstein, los procedimientos operacionales, ligados a la noción de teoría, se aplican a las técnicas operacionales para introducir conceptos.
Los modelos funcionales, sistémicos, holísticos y homeostáticos hipotético-deductivos, mantienen una acomodación en estado de equilibrio donde unos paradigmas suceden a los otros, de acuerdo a la historia interna y externa de las disciplinas y sus teorías.
El reduccionismo metodológico permite solucionar problemas dentro del instrumentalismo y el realismo mediante construcciones de experimentos de laboratorio y de campo, integrando disciplinas en forma semántica y sintáctica.
Albert Einstein, superó las dificultades del éter fibroso, en 1905, lo expuso en forma revolucionaria, vinculando los resultados de Max Planck (1900). Einstein supuso que la energía emitida por cualquier radiador no sólo se conserva en *cuantos*, al trasladarse por el espacio, como se había supuesto, sino también que una fuente dada podía emitir y absorber energía radiante únicamente en esas unidades.
Einstein encontró una ecuación que predecía correctamente todos los hechos observados y contrastada en el *Ryerson Laboratory* (1904-1915) como ecuación exacta y de validez general, constituyó el progreso más conspicuo de la física experimental y del futuro de la ciencia, por la dualidad corpúsculo ondulatorio. Se reconcilió la teoría ondulatoria con la corpuscular frente a esos fenómenos nuevos y perturbadores, imaginando toda la energía electromagnética, trasladándose a lo largo de líneas de fuerza, concebidas como verdaderas cuerdas, que se extenderían a través del espacio.

### 1.2.4. Estructura del Programa de Investigación Científica.

La Descripción de los Problemas fundamentales no resueltos son la clasificación de objetos en familias o clusters mediante un método automático de formación de regiones, el cual es un proceso de agrupamiento de objetos en clases teniendo en cuenta sus relaciones y atributos comunes, a los efectos de realizar estudios de sus características y las propiedades estructurales y de su comportamiento relativo, en la tesis se da solución a los mismos y la utilización de Data Mining para corroborar su fortaleza. La utilización del Teorema de Tchebycheff, para el manejo de invariantes, y de la Espectroscopia para llegar a los espectros de objetos y de familias, sin tener que utilizar algunas maniobras arbitrarias o no claramente explicadas, en otros criterios.
Así se llega a plantear la operatoria para la solución del Problema(i) se introduce la matriz de datos, se define el proceso de su construcción y se describe el proceso de normalización asociado. (ii) Se presenta la matriz de similitud y (iii) se definen los espectros de OTU´s y familias. (iv) Se analiza la caracterización, la dispersión y la normalización del rango de la dispersión. (v) Con el refinamiento de invariantes utilizando el teorema de Tchebycheff, la distribución de masa y la aplicación de

equiprobabilidad para la Entropía Máxima (PME). (vi) Se despliega la algoritmia iterativa asociada que incluye el análisis de conservación del espacio y la distorsión ínter cluster por vecindad. (vii) La Representación Taxonómica en Base de Datos, Dominios Estandarizados de OTU´s en Bases de Datos Relacionales Dinámicas y Data Mining para testear la fortaleza de la solución propuesta y la Algoritmia.

Con la Fenomenología Física se plantean los principios de interferencia y superposición y las analogías asociadas, se vincula el análisis espectral y la dinámica de fasores con el hiperespacio taxonómico y su dinámica vectorial de distancias normalizadas.

Los Sistemas Complejos y Dinámicos explican y describen sus herramientas de Mecánica Estadística y los métodos de Ciencias de la Computación desde la Mecánica Clásica y Cuántica y se plantea la Teoría de la Información, la Entropía, el Principio de máxima Entropía, la Distancia de Hamming y la Ganancia Entrópica con su aplicación en Data Mining.

Como un caso de uso se realiza una aplicación con una experimentación de la clasificación automática de cuerpos celestes, en particular de familias de asteroides. El Principio de Máxima Entropía, la Metodología de Ganancia de la Información relativa a la Entropía y Data Mining y el Testeo para mostrar la fortaleza del método. Se realiza un enfoque de la Ingeniería de Requerimientos, desde de la Ingeniería de Software, para el caso de experimentación se exponen los planteos y conjeturas o hipótesis de Hirayama y desde Arnold hasta Zappala, y su problemática de identificación de cuerpos celestes y de elementos propios no perturbados. Se plantean los matices de la implementación para la atención del caso de estudio en la matriz de datos, la matriz de similitud y la estructuración. Finalmente se presentan los resultados desde el "***nuevo criterio"*** de Análisis Espectral de Taxonomía Computacional, para la familia María con sus instancias que contrastan con las Familias de Hirayama.

## CONCLUSIONES APORTES ORIGINALES

- **Definición de un método numérico basado en invariantes para la identificación automática de pertenencia de OTU´s a una familia.**
- **Aplicación de la metodología a asteroides según las familias de Hirayama.**
- **Identificación, testeo y puesta a punto de las invariantes: centroide, varianza y radio del método definido.**
- **Comprobación del método para familias de asteroides en forma no arbitraria de identificación de elementos.**
- **Definición de medidas que permiten determinar la estabilidad de un familia por sensibilidad de atributos.**
- **Contrastación de las familias propuestas por Hirayama, por el método numérico descripto en la tesis (objetivo).**
- **Utilización original del método de superposición e interferencia de espectros para confirmar que los resultados obtenidos por el método numérico propuesto son congruentes (consistentes).**
- **Contrastación con Cúmulos, Nebulosas y Galaxias, comprobándose la separación de Galaxias gemelas a las cuales el método las separó.**
- **Continuar investigando el poder separador del Método y con más Cúmulos.**
- **Desarrollar con Ingeniería de Software y de Requerimientos un tipo de Sistema Experto.**