MPGI UC

MAGISTER EN PROCESAMIENTO Y GESTIÓN DE LA INFORMACIÓN

Internet y sociedad: comunicación y cultura en la era digital

César Antonio Aguilar

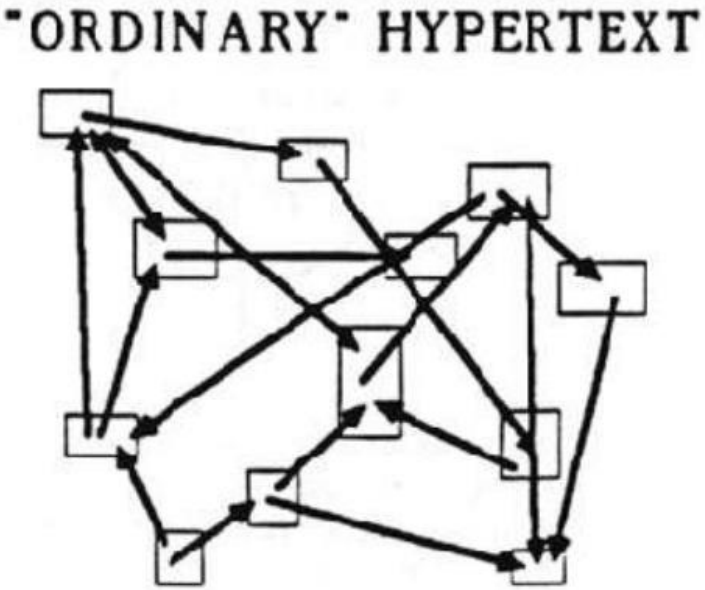
Facultad de Lenguas y Letras

17/09/2020

Cesar.Aguilar72@gmail.com

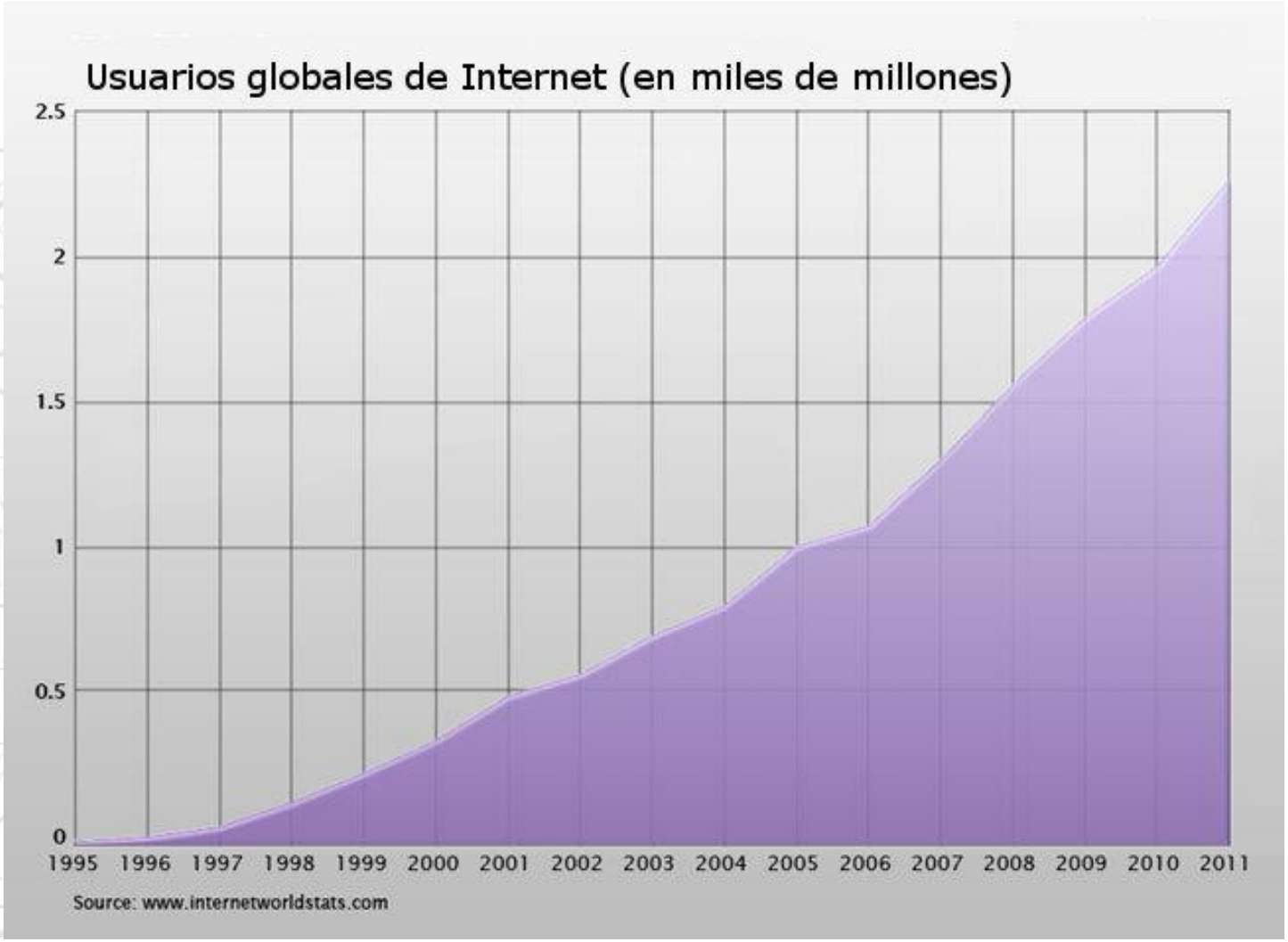
Síntesis de la clase anterior (1)

En la clase anterior revisamos brevemente cómo es que el desarrollo de Internet los hipertextos ha tenido impacto en las Humanidades Digitales.



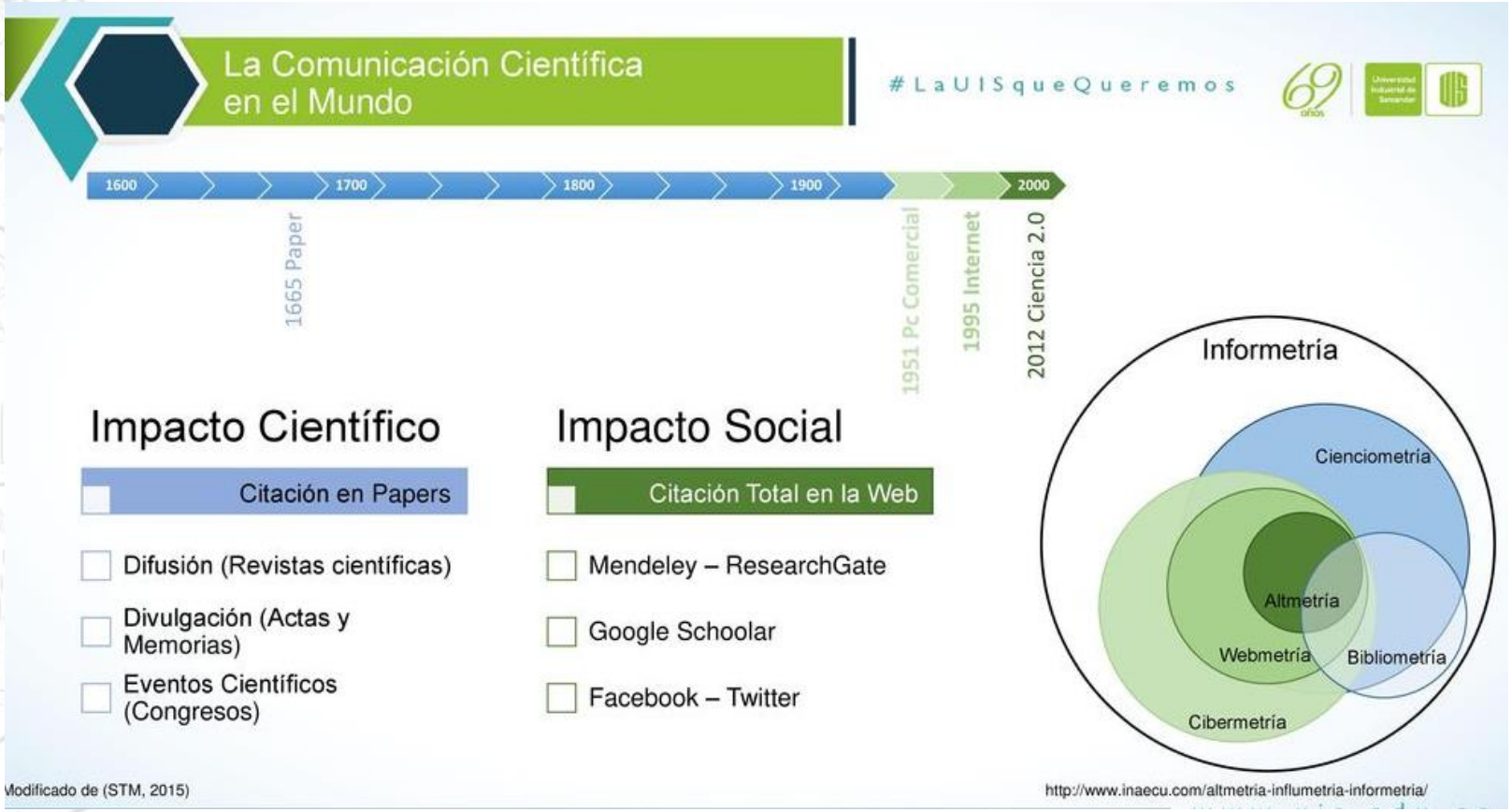
Síntesis de la clase anterior (2)

Dados estos elementos, el periodo que va del año 1999 al 2000 puede considerarse como el inicio de la masificación de la WEB:



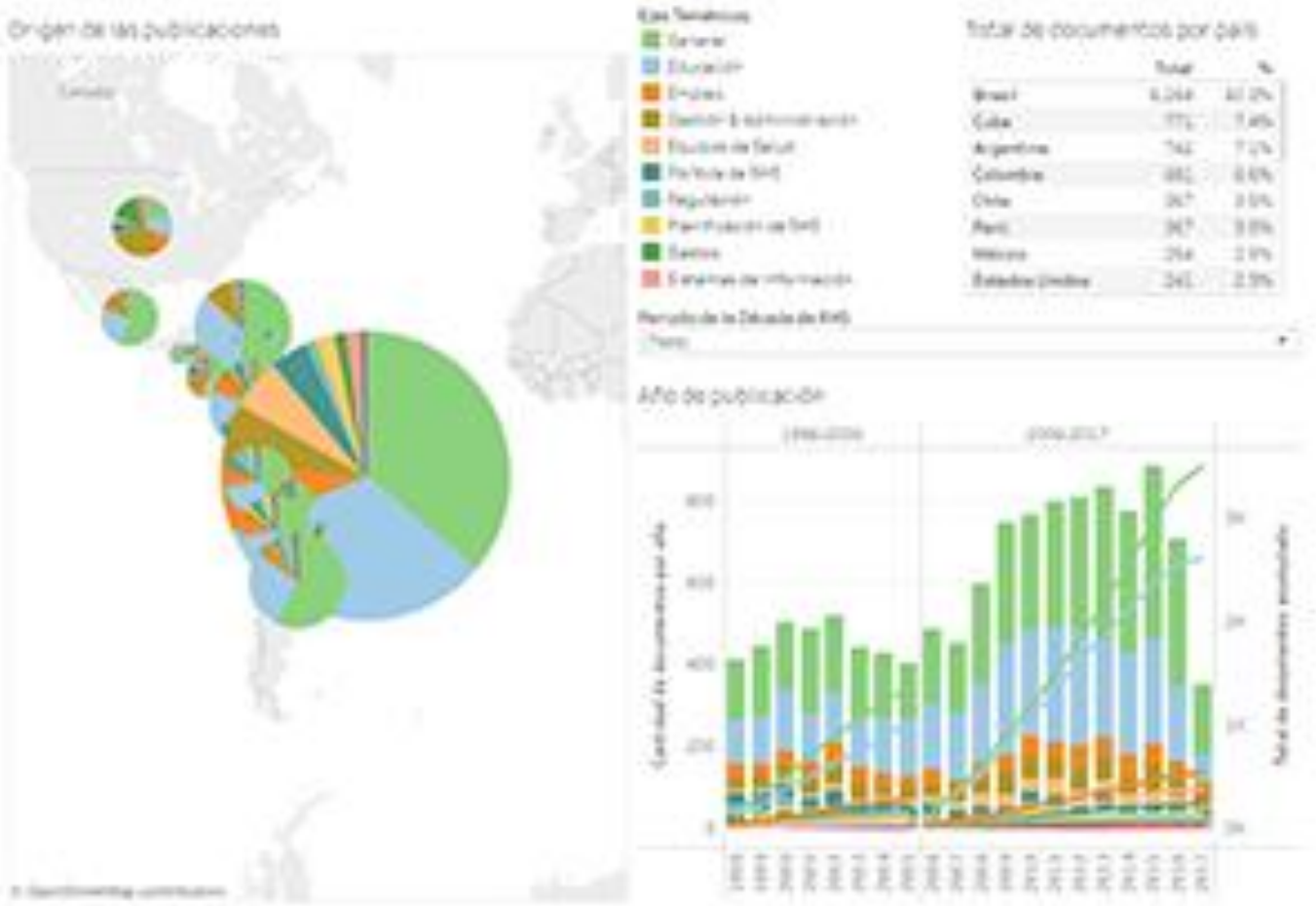
Síntesis de la clase anterior (3)

Esta masificación trajo consigo nuevas posibilidades y problemas a resolver. Un área de investigación que logró consolidarse en estos años fue la **infometría**, enfocada en la medición del impacto que tiene la producción científica en la sociedad actual:



Síntesis de la clase anterior (4)

La infometría ha puesto en relieve el uso de indicadores numéricos para analizar el comportamiento de la producción académica entre individuos, instituciones, áreas de conocimiento y zonas geográficas:



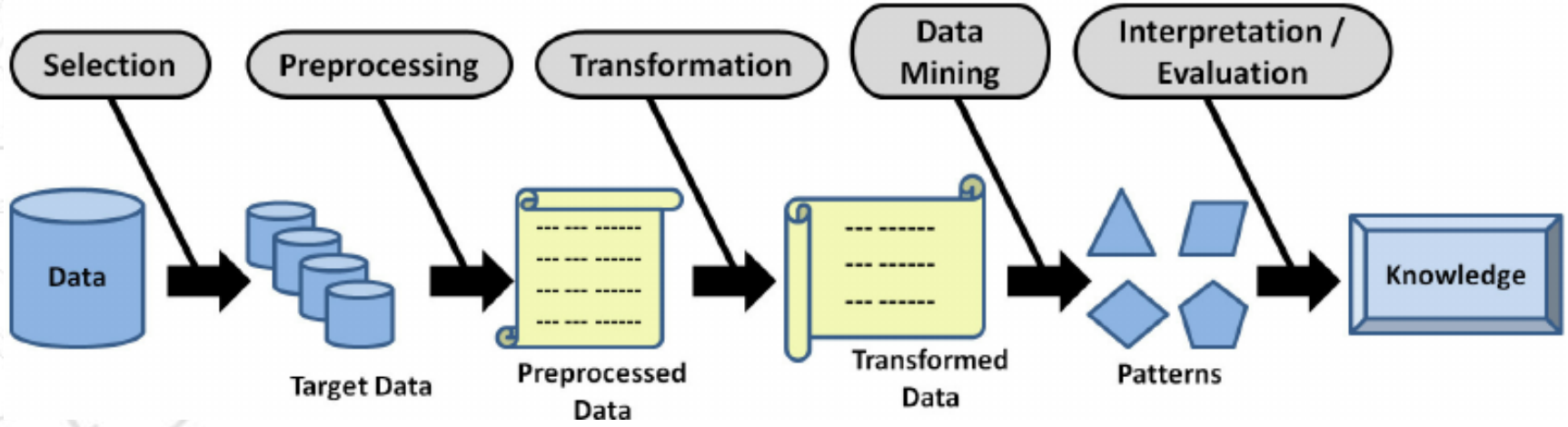
Descubriendo nuevo conocimiento (1)

Conforme se ha generado más información en la WEB, se han generado nuevas líneas de investigación, además de todas aquellas que conforman la infometría. Una de las más relevantes y conocidas es el **descubrimiento de conocimiento**, también conocida como **minería de datos**.

Descubriendo nuevo conocimiento (2)

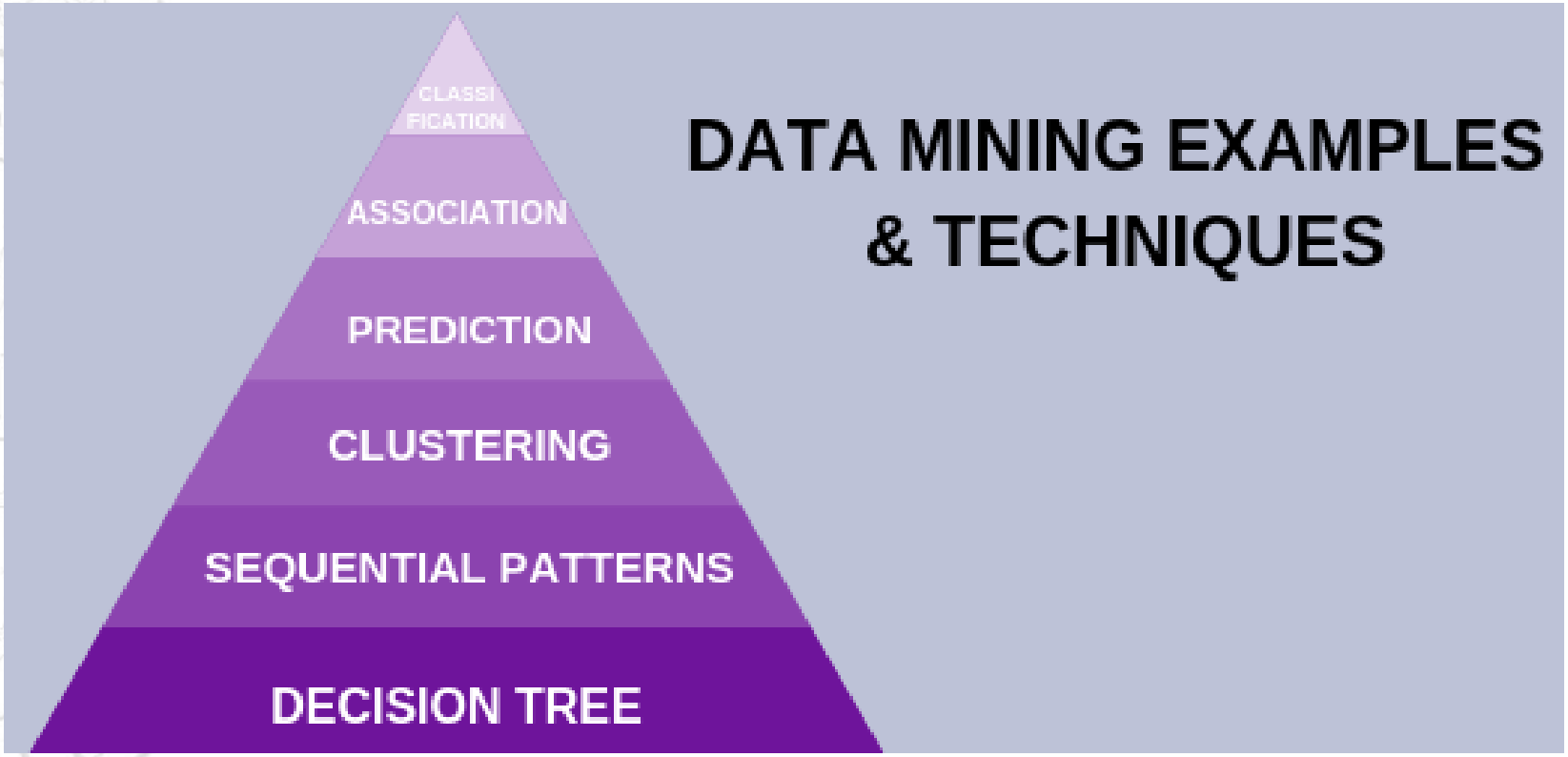
A grandes rasgos, hay que ambos términos se refieren prácticamente a la misma de investigación, aunque existen algunas diferencias:

En el caso del **descubrimiento de conocimientos** (ing.: *Knowledge Discovery in Database*, **KDD**), se focaliza el resultado final de un proceso de inferencia, el cual ayuda detectar información relvante en grandes colecciones de datos, es decir:



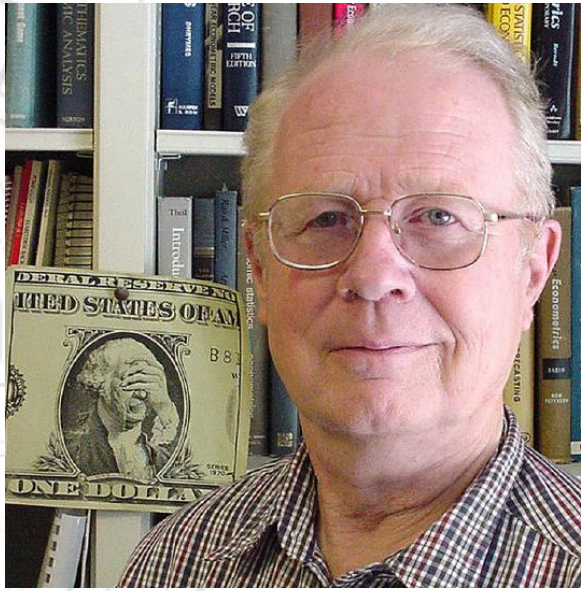
Descubriendo nuevo conocimiento (3)

Por su parte, la **minería de datos**, en estricto sentido, originalmente se refería a una fase del KDD, en concreto el reconocimiento de patrones recurrentes en los datos extraídos. Para esto, era necesario aplicar distintos métodos para concretar tal reconocimiento:



¿Cómo nace la minería de datos? (1)

Conforme a lo que sería una *arqueología* del KDD y la minería de textos, los primeros trabajos en esta línea de investigación se llevaron a cabo en la década de los 60 en el área de economía, bajo la iniciativa de **Michael C. Lowell**, profersor norteamericano afiliado a la **Wesleyan University.**

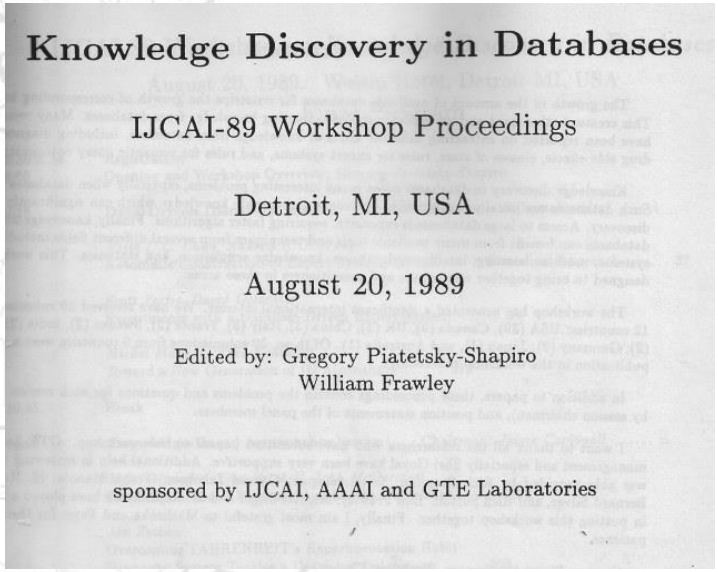


Michael C. Lowell (1930)

Para saber más:
http://mlovell.web.wesleyan.edu/

¿Cómo nace la minería de datos? (2)

Tras el trabajo pionero de Lowell, para 1990 ya se estaban acuñando los términos *descubrimiento de conocimiento* y *minería de datos* en varios medios académicos. De hecho, esta línea de trabajo se inauguró "oficialmente" en 1989, con un *workshop* organizado por **Gregory Piatetsky-Shapiro**.


Knowledge Discovery in Databases

IJCAI-89 Workshop Proceedings

Detroit, MI, USA

August 20, 1989

Edited by: Gregory Piatetsky-Shapiro
William Frawley

sponsored by IJCAI, AAAI and GTE Laboratories

Gregory Piatetsky-Shapiro (1958)

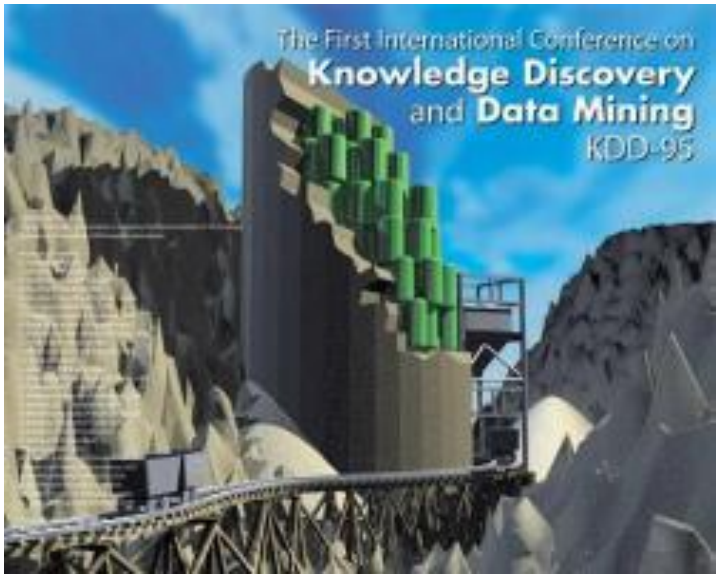
Para saber más:
www.kdnuggets.com/meetings-past/kdd89/index.html

¿Cómo nace la minería de datos? (3)

En 1995, dos computólogos norteamericanos, **Usama Fayyad** y **Ramasamy Uthurusamy** organizaron la primera conferencia internacional sobre KDD y minería de datos, la cual tuvo una excelente acogida, además de que consolidó los vínculos que había con el campo de la inteligencia artificial.

Usama Fayyad
(1963)

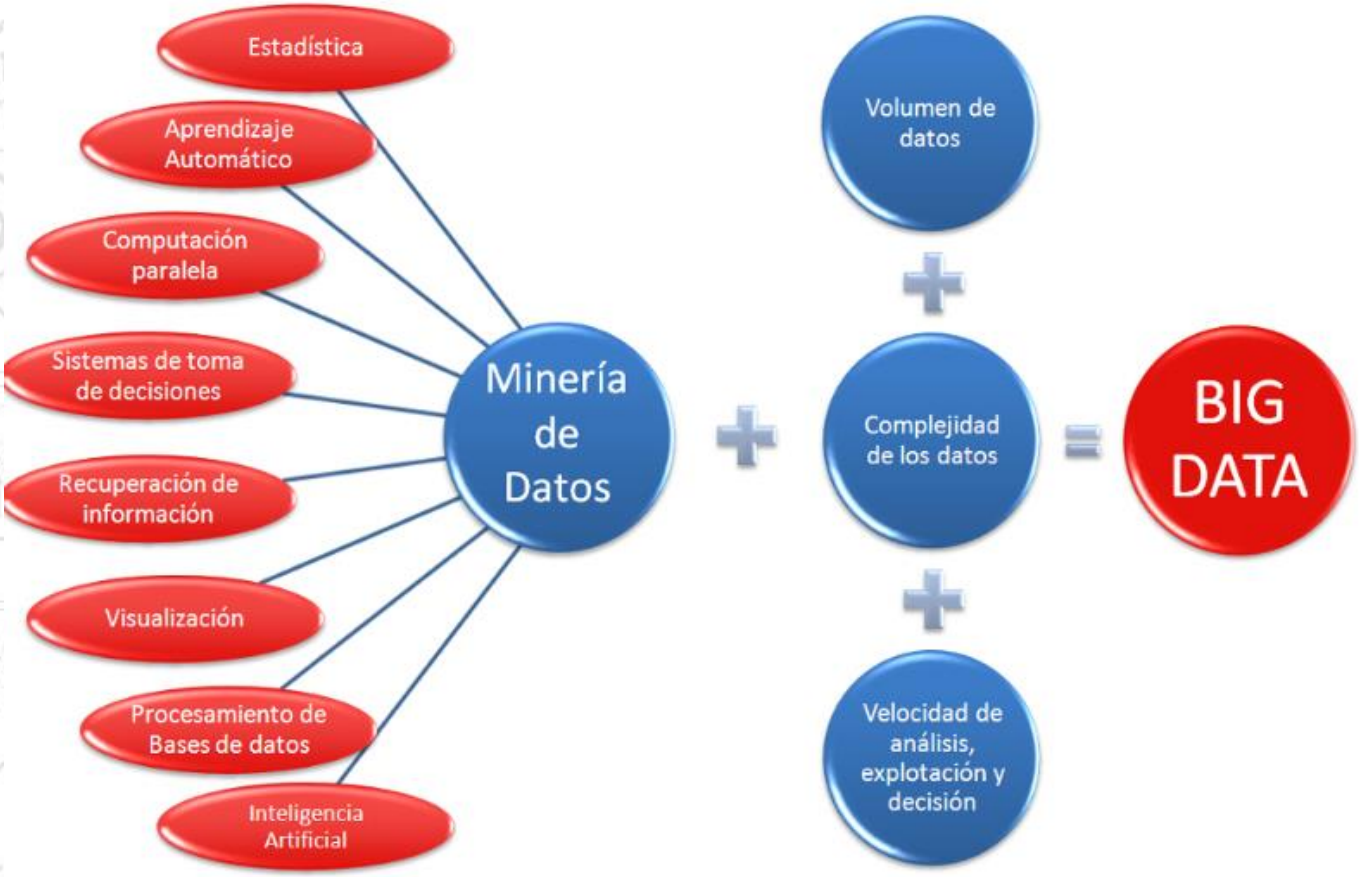
Ramasamy Uthurusamy



Para saber más:
https://dl.acm.org/citation.cfm?id=3001335&picked=prox

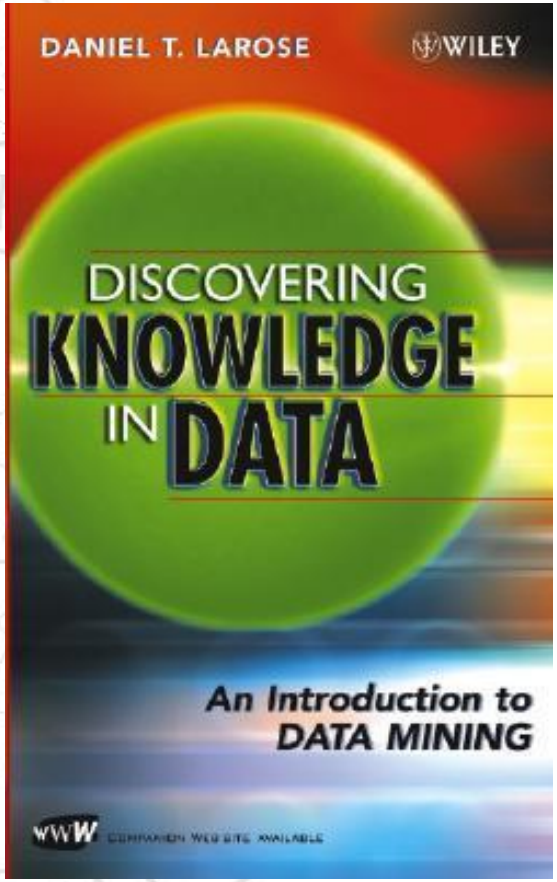
¿Cómo nace la minería de datos? (4)

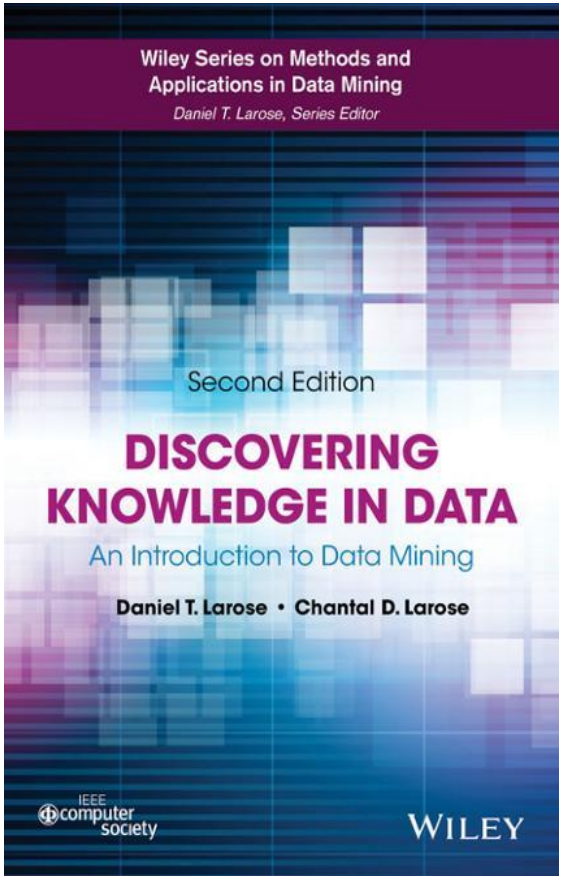
Desde 1995 a la fecha, gracias al impulso que le dieron Fayyad y Uthurusamy, la minería de datos se ha constituido como una línea de investigación que considera los siguientes modelos y procesos:

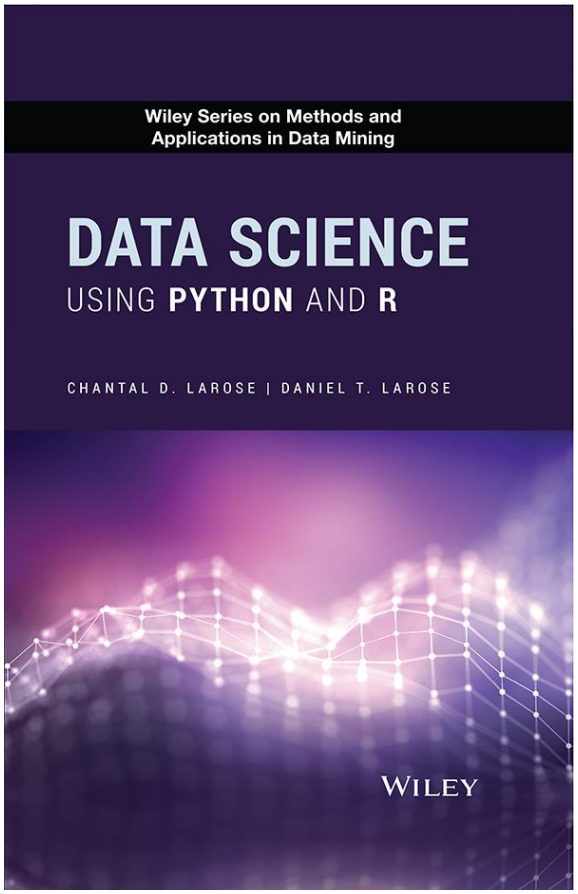


¿De dónde viene la idea? (1)

Muchos textos sobre minería de datos, en aras de ir directamente al grano sobre los modelos y aplicaciones que la constituyen, omiten buena parte de los antecedentes históricos que subyacen en su marco teórico, p. e.:







¿De dónde viene la idea? (2)

Sin embargo, resulta interesante inferir cuáles fueron los caminos que se siguieron para concebir una propuesta como la minería de datos, la cual plantea una serie de cuestionamientos interesantes a lo que se conoce . como la **teoría clásica de la probabilidad**. En este sentido, podemos considerar que la minería de datos es una extensión de la **inferencia bayesiana.**

Blaise Pascal
(1623-1662)

Versus

Thomas Bayes
(1701-1761)

Todo comenzó con un juego... (1)

El interés por el estudio del azar se puede rastrear desde tiempos antiguos, por lo menos desde la cultura sumeria (3500 A.C.), y ha estado estrechamente relacionado a los juegos de azar.

La cuestión a tratar aquí es la siguiente: ¿existe alguna manera de predecir quién puede ganar o perder un juego de azar, considerando todos los elementos que participan en dicho juego: piezas del juego, reglas, combinaciones, números de jugadas, etc.?

Todo comenzó con un juego... (2)

Si bien estos descubrimientos han sido importantes, la mayoría de los historiadores de las matemáticas prefieren ubicar el nacimiento de esta teoría en el s. XVII, a partir de una serie de experimentos que realizaron en colaboración **Blaise Pascal** y **Pierre de Fermat**, para resolver una cuestión relacionada con los juegos de azar.

Blaise Pascal
(1623-1662).

Pierra de Fermat
(1601-1665).

Todo comenzó con un juego... (3)

La anécdota detrás de estas preguntas es la siguiente: en una ocasión, un hombre llamado Antoine Gombaud de Méré (1607-1684), quien era aficionado a los juegos de azar, logró contactar a Pascal y Fermat, con el fin de plantearles un reto matemático, formulado los siguientes términos:

Supóngase que dos jugadores acuerdan jugar un determinado número de partidas, digamos, jugar al mejor de siete, y se les interrumpe antes de que puedan terminar la tanda. ¿Cómo debería repartirse el dinero apostado si, por ejemplo, uno ha ganado cinco partidas y el otro cuatro?

Todo comenzó con un juego... (4)

Las deducciones que desarrollaron Fermat y Pascal para resolver esta duda los llevaron a relacionar varios fenómenos:

Procesos combinatorios: el hecho de que alguno de los jugadores gane o pierda se relaciona con series de combinaciones entre las juagadas que ya ganaron, y las posibilidades que al menos uno de los dos tiene de ganar (esto es, el que más se acerca al triunfo).

Relación condicional: la pregunta de la Méré no intenta adivinar quién será el ganador (o en su defecto, qué hace falta para ganar el juego), sino más bien establece una **condición:** la repartición de la apuesta entre dos jugadores está determinada por el número de jugadas ganadas que llevan, dentro de una serie de 7 juegos.

Todo comenzó con un juego... (5)

Considerando las 2 cuestiones anteriores, una definición que se puede dar sobre lo que es la probabilidad es la siguiente:

La probabilidad es el conjunto de todas las posibilidades que tenemos de obtener un resultado, en función del número total de posibles maneras que tenemos para obtener tal resultado.

Esta definición se la debemos a **Gerolamo Cardamo**, quien vislumbró el problema antes que Pascal y Fermat.

Gerolamo Cardamo (1501-1576).

Todo comenzó con un juego... (6)

Siguiendo el razonamiento planteado por Cardano, Fermat propone la siguiente solución:

Si el Jugador A lleva ganadas 5 partidas de 7, en tanto que el Jugador B lleva 4 de 7, A necesita 2 jugadas más para llegar a 7; en tanto que B requiere 3 para llegar al mismo número…, y de hacerlo, necesitará una más para desempatar con A.

Por ello, pongamos como margen un máximo de 4 jugadas para que cualquiera de los dos gane el juego.

Todo comenzó con un juego... (7)

Ahora bien, dado ese margen de 4 jugadas, ¿cuáles son las combinaciones que requieren A y B para ganar el juego? De acuerdo con Fermat, el conjunto total de combinaciones son:

$$(a,a,a,a),(a,a,a,b),(a,a,b,a),(a,b,a,a),(b,a,a,a),(a,a,b,b),(a,b,a,b),(a,b,b,a),$$

$$(b,a,a,b),(b,a,b,a),(b,b,a,a),(a,b,b,b),(b,a,b,b),(b,b,a,b),(b,b,b,a),(b,b,b,b)$$

Por convención, establezcamos que las jugadas que representa *a* son aquellas que gana el jugador A, en tanto que las que representa *b* son las que gana el jugador B.

Todo comenzó con un juego... (8)

El total de combinaciones dentro de estas 4 jugadas es de 16. Ahora, dentro de estas 16 jugadas, el Jugador A puede optar por 11 que le permiten ganar la partida (recordemos, está a dos juegos de triunfar). Las combinaciones ganadoras para A son:

1. {a, a, a, a} 2. {a, a, a, b} 3. {a, a, b, a} 4. {a, b, a, a}
5. {b, a, a, a} 6. {a, a, b, b} 7. {a, b, b, a} 8. {b, b, a, a}
9. {b, a, a, b} 10. {b, a, b, a} 11 {a, b, a, b}

Por su parte, B sólo puede optar por 5 combinaciones:

1. {b, b, b, b} 2. {b, b, b, a} 3. {b, b, a, b} 4. {b, a, b, b}
5. {a, b, b, b}

Todo comenzó con un juego... (9)

Finalmente, para establecer el margen de probabilidad que tienen ambos jugadores para ganar la partida completa en 4 juegos, considerando que el jugador A lleva ventaja sobre el jugador B, hacemos el siguiente cálculo:

Jugador A: 11/16 = 0,69

Jugador B: 5/16 = 0,31

Así, la repartición de la apuesta podría ser del siguiente modo: si ambos jugadores ponen 50 monedas de oro cada uno, de modo que la suma total da 100 monedas, el jugador A tiene derecho a 69 monedas, porque es el que estuvo más cerca de ganar; en contraste, el jugador B tiene derecho a 31 monedas, dado que está más alejado del triunfo en comparación con A.

Causas y efectos... (1)

Por su parte, el Padre Bayes se enfocó en tratar de entender cómo es que un fenómeno puede ser condicionado por otro, de tal suerte que un efecto observado puede se provocado por una causa no visible pero coincidente.

Un ejemplo sencillo es el siguiente: si nos duele la cabeza (A), y sabemos que estamos resfriados (B), ¿existe la probabilidad de que tal dolor sea provocado por el resfrío, o hay que considerar otra causa?

Causas y efectos... (2)

Si bien el ejemplo anterior parece sencillo, nos permite plantear lo siguiente: si partimos de una hipótesis que nos ayuda a explicar un fenómeno dado, es necesario contrastarla consu negación. Así, para resolver este constraste, resulta necesario colectar evidencia que sustente empíricamente alguna de las dos hipótesis.

A medida que la evidencia se acumula, el grado de creencia en una hipótesis se va modificando. De este modo, lashipótesis con un grado de creencia muy alto se pueden tomar como verdaderas, mientras que las que tienen un bajo grado de creencia muy bajo son vistas como falsas.

Causas y efectos... (3)

¿Qué pasa si añadimos más variables relacionadas con el fenómeno que estamos analizando? Pues las cosas se ponen más interesantes. Veamos: ¿qué factores influyen para que una persona enferme de cáncer en los pulmones?

(a) Fumar

(b) Respirar aire contaminado

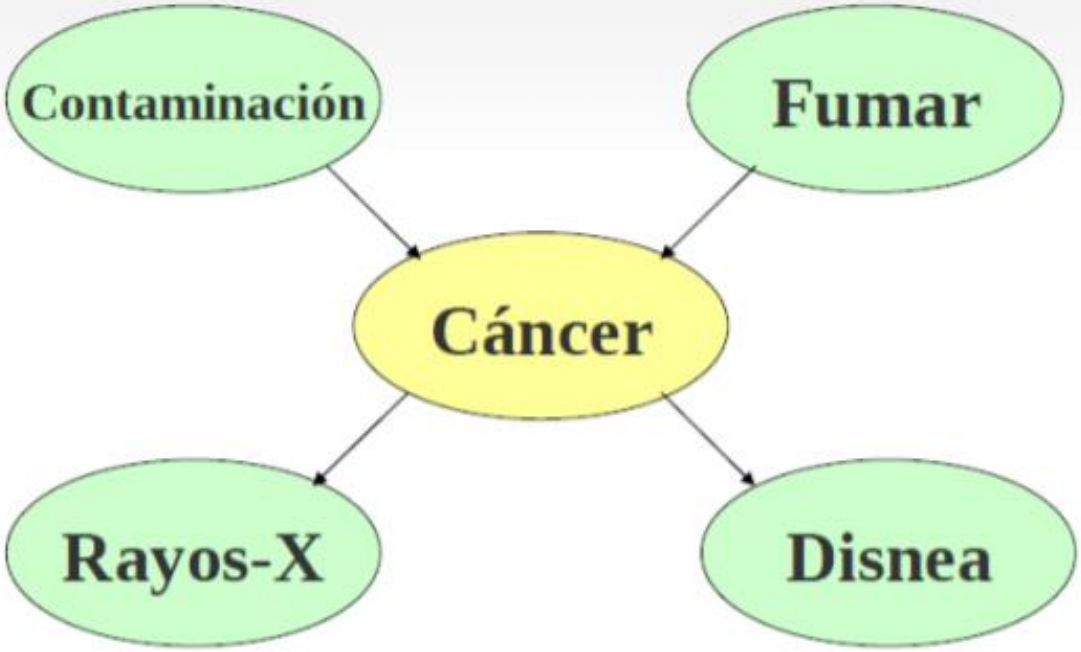
Una vez resuelta esta duda, hagamos otra pregunta: ¿cuál es la mejor forma de detectarlo?

(i) A través de una radiografía

(ii) Detectando síntomas de disnea (esto es, dificultades para respirar)

Causas y efectos... (4)

Tenemos entonces dos posibles causas (contaminación *vs.* fumar), un efecto (contraer cáncer), y dos maneras de detectarlo (vía rayos-X *vs.* Padecer de disnea).



Preguntas maliciosas:

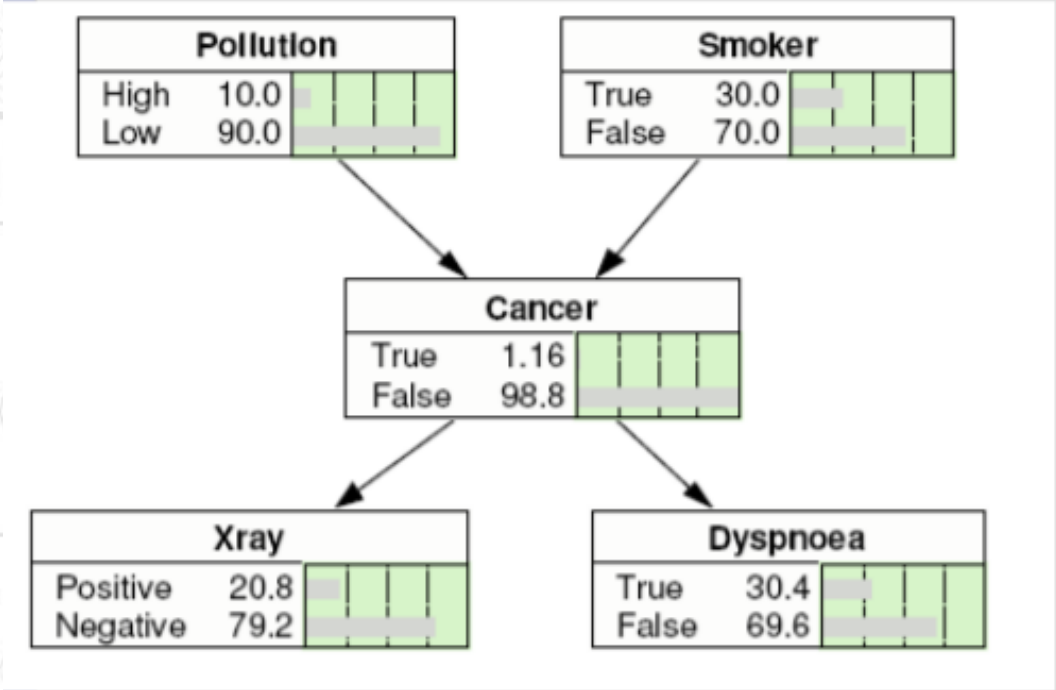
1. ¿Cuál sería el peor paciente posible?

2. Si fumo, pero no sufro disnea, y los rayos X no me detectan nada en el pulmón, ¿puedo llegar a sufrir cáncer?

3. ¿Todos los habitantes del DF, por el hecho de estar expuestos a la contaminación, pueden sufrir cáncer?

Causas y efectos... (5)

Siguiendo este razonamiento, parece que en un 30% de los casos de fumadores, ésta puede contraer cáncer, y podemos detectarlo en un 30.4% de los casos considerando si tienen problemas para respirar.



Imaginemos que la combinación de estos dos factores nos da lo siguiente:

1. Fumar CAUSA cáncer.

2. La disnea ES-UN EFECTO del cáncer.

Si evito fumar, puede ayudar a que no padezca disnea, y la combinación de ambos hace que tenga pocas probabilidades de padecer cáncer.

Causas y efectos... (6)

Nuevamente, partimos de lo que hemos visto sobre probabilidades condicionales: lo que queremos establecer es si en una cadena de eventos dada, ¿podemos reconocer cuáles son dependientes entre sí, y cuáles son marginales? Esto podemos determinarlo con la siguiente fórmula:

Esto quiere decir: "sea {A1, A2, A3...An} un conjunto de sucesos mutuamente excluyentes y exhaustivos, tales que la probabilidad de cada uno de ellos es distinta de cero".

A esto le añadirmos: "sea B un suceso cualquiera del que se conocen las probabilidades condicionales P(B | Ai)".

Al final, siguiendo la fórmula, deducimos:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)}$$

- $P(A_i)$ son las probabilidades a priori.
- $P(B \mid A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i \mid B)$ son las probabilidades a posteriori.

Esto se cumple $\forall i = 1 \ldots n$

Causas y efectos... (7)

Para mayores detalles, este tipo de inferencias se basan en el **teorema de Bayes**, el cual nos dice que:

Sea $\{A_1, A_2, A_3 \ldots A_i \ldots A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B \mid A_1)$. Entonces, la probabilidad $P(A_1 \mid B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i|B)$ son las probabilidades a posteriori.

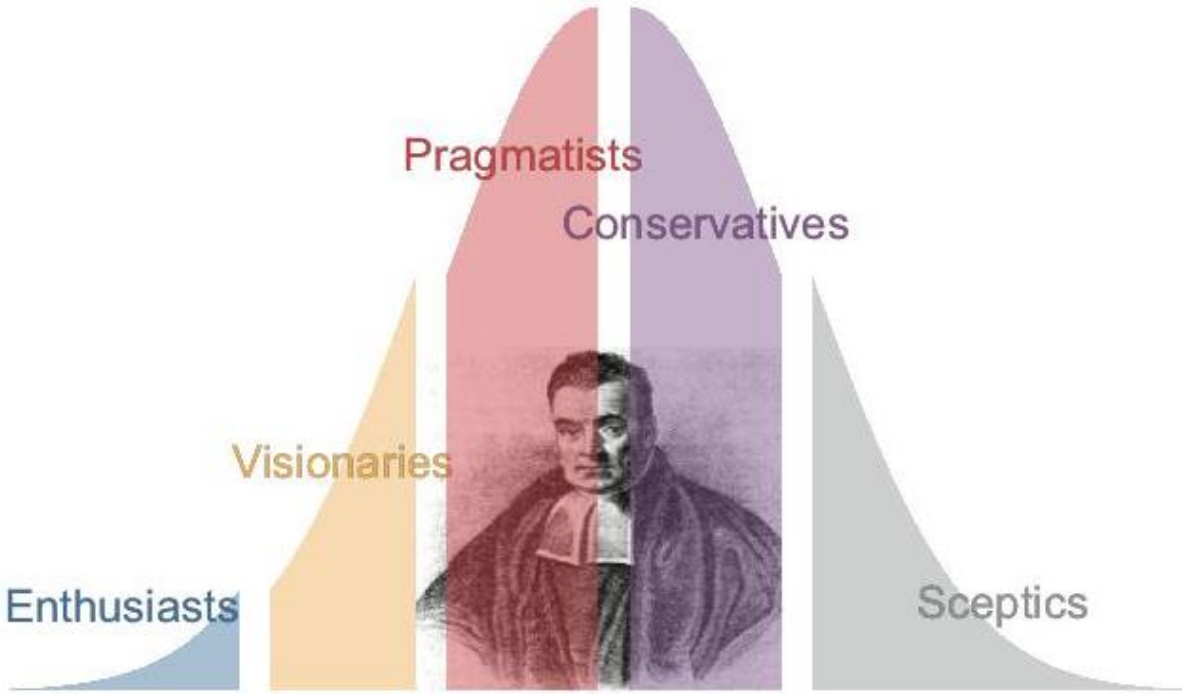
Dejemos que hablen los datos... (1)

Una de las cualidades que muchos investigadores valoran del teorema de Bayes es su *ingenuidad* para hacer inferencias que permitan asociar causas y efectos más probables.

Dejemos que hablen los datos... (1)

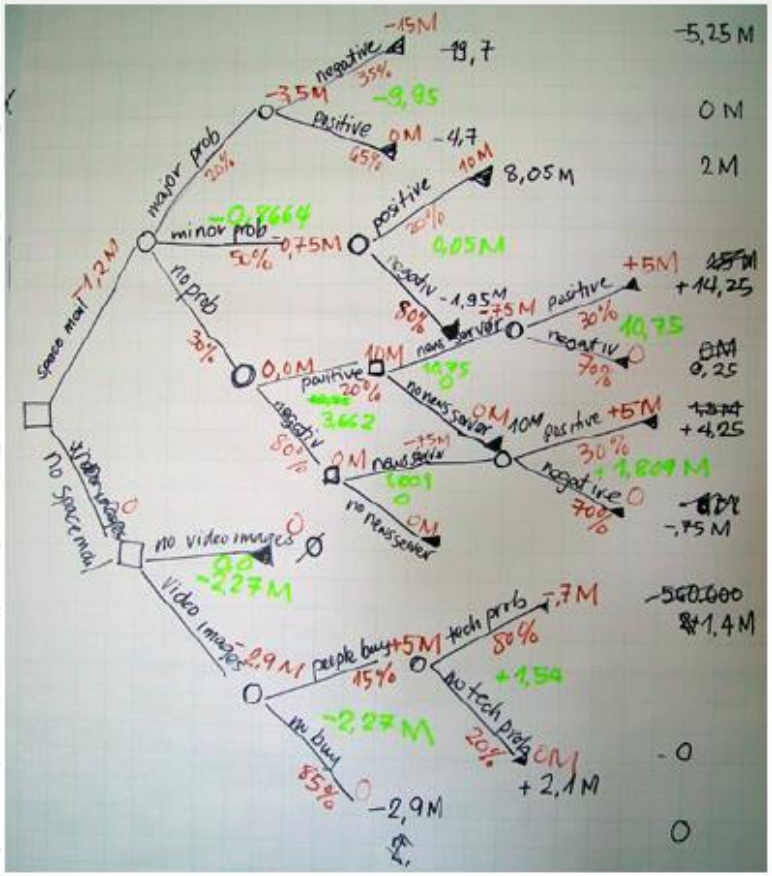
Dadas estas relaciones de probabilidad entre causas y efectos, el teorema bayesiano ha sido empleado con éxito para resolver problemas de **agrupamiento** (*clustering*) **y clasificación de datos.**

Viendo el gráfico, nos damos cuenta que podemos matizary generar sub-conjuntos de rasgos que definan a aquellas personas que se encuentran entre dos extremos: ser entusiastas, o ser escépticos.



Dejemos que hablen los datos... (2)

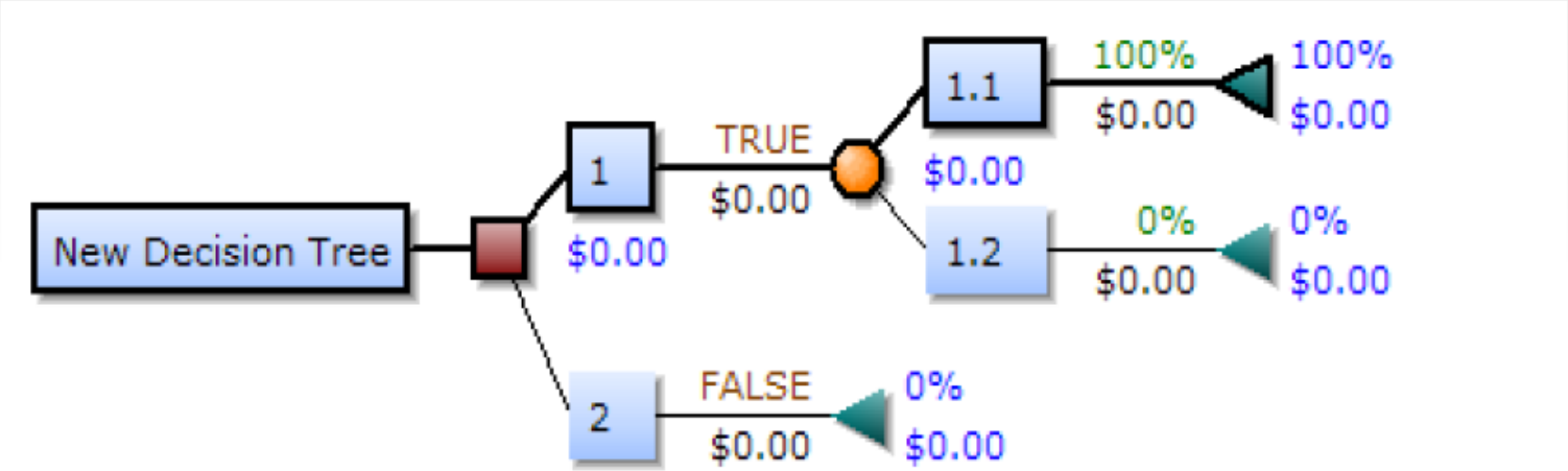
Para abordar estos problemas de clasificación, se han creado varios modelos computacionales basados en este teorema. Uno de los más conocidos son los **árboles de decisión**.



Un árbol de decisión es un modelo de predicción que opera siguiendo un principio muy similar al de las redes bayesianas. La idea aquí es: supongamos que tenemos un conjunto de hipótesis qué explican una secuencia de fenómenos. Cada uno de estos fenómenos puede, o no puede ocurrir. Como en la lógica clásica, si ocurre tal fenómeno, aparte de validar nuestras hipótesis, tal fenónomeno podrá dar lugar a otros que estén relacionados. En caso de que no tenga lugar algún fenómeno, entonces mi hipótesis es falsa, y por tanto no me ayuda a establecer una secuencia de relaciones.

Dejemos que hablen los datos... (3)

Visto en abstracto, parece que los árboles de decisiones son mecanismos demasiado complejos. Empero, si los vemos en un terreno práctico, en realidad operan de forma más simple.



Supongamos que trabajamos en un supermercado, y de repente detectamos que durante los meses de marzo a mayo hay una tendencia, casi inexplicable, a comprar pañales para bebés y cervezas.
Aprovechando esto, decidimos poner en oferta estos dos productos.
¿De dónde nos llegó está extravagante idea?

Dejemos que hablen los datos... (4)

Preguntas maliciosas:

1. ¿Quiénes hacen estas compras: hombres y/o mujeres?

2. ¿Hay algún rango de edad en estos clientes: 20, 30, 40, 50 años, p. e.?

3. ¿Estas compras se hace todos los días, los fines de semana, o incluso en fechas específicas?

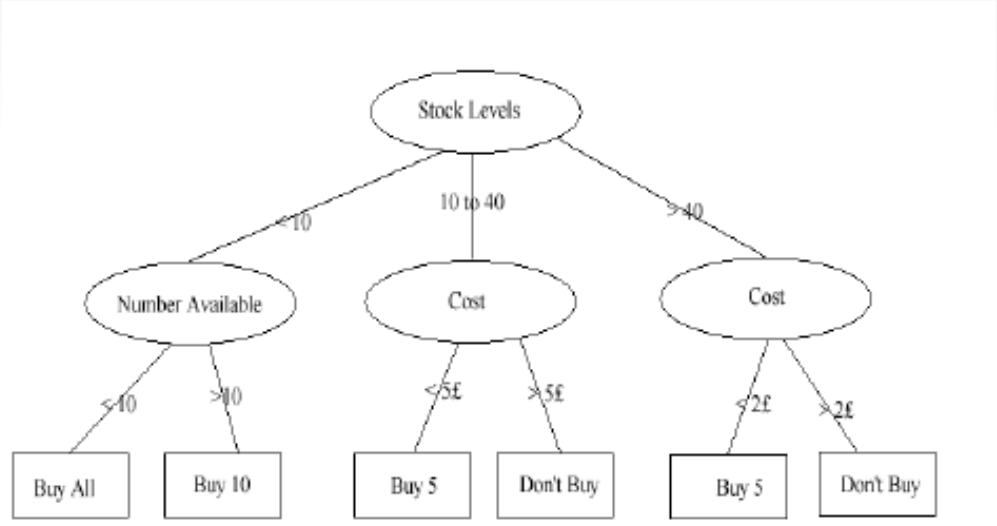
4. ¿Desde cuándo hemos observado esta tendencia: hace un mes, un año, dos años, diez años...?

5. ¿Hay alguna preferencia por cierta marca de pañales y/o cerveza?

6. La suma de estos factores, ¿nos haría pensar que conviene poner los dos anaqueles cerca, o los dejamos alejados uno del otro?

Dejemos que hablen los datos... (5)

El ejemplo del supermercado es uno de los más típicos para explicar procesos de minería de datos para la toma de decisiones. Lo que queremos hacer es determinar, p. e., comportamientos de compra en clientes, dada regularidad observada en la asociación de una serie de factores.



Ahora, si quisiéramos evitarnos ese proceso de analizar y tomar decisiones, y se lo asignáramos a un sistema de cómputo, entonces lo que tendríamos que hacer es ayudarlo a razonar cómo se relacionan estos factores, de tal modo que si ocurre uno, éste obligará a hacer una elección concreta. Así, dada la probabilidad de que ocurre el fenómeno *X*, ¿debo hacer la elección A, o la elección B?

Minería de textos (1)

Tras el impulso que tomó la minería de datos en los años 90 —lo que a su vez ha ayudado a revalorar los métodos probabilísticos basados en el teorema de Bayes—, durante la primera década del 2000 fue emergiendo una línea de investigación paralela, la **minería de textos**.

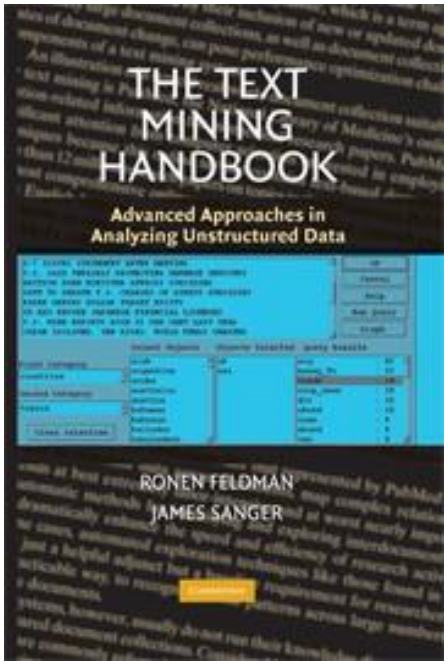
La minería de textos se puede concebir como el proceso enfocado en la detección y de patrones en grandes repositorios de documentos. Estos patrones representan información relevante contenida en tales documentos.

Minería de textos (2)

Casi de manera informal, se pueden reconocer a **Ronen Feldman** y **James Sanger** como los primeros que comenzaron a acuñar tal término en un manual que publicaron en 2006 sobre el tema:

Ronen Feldman

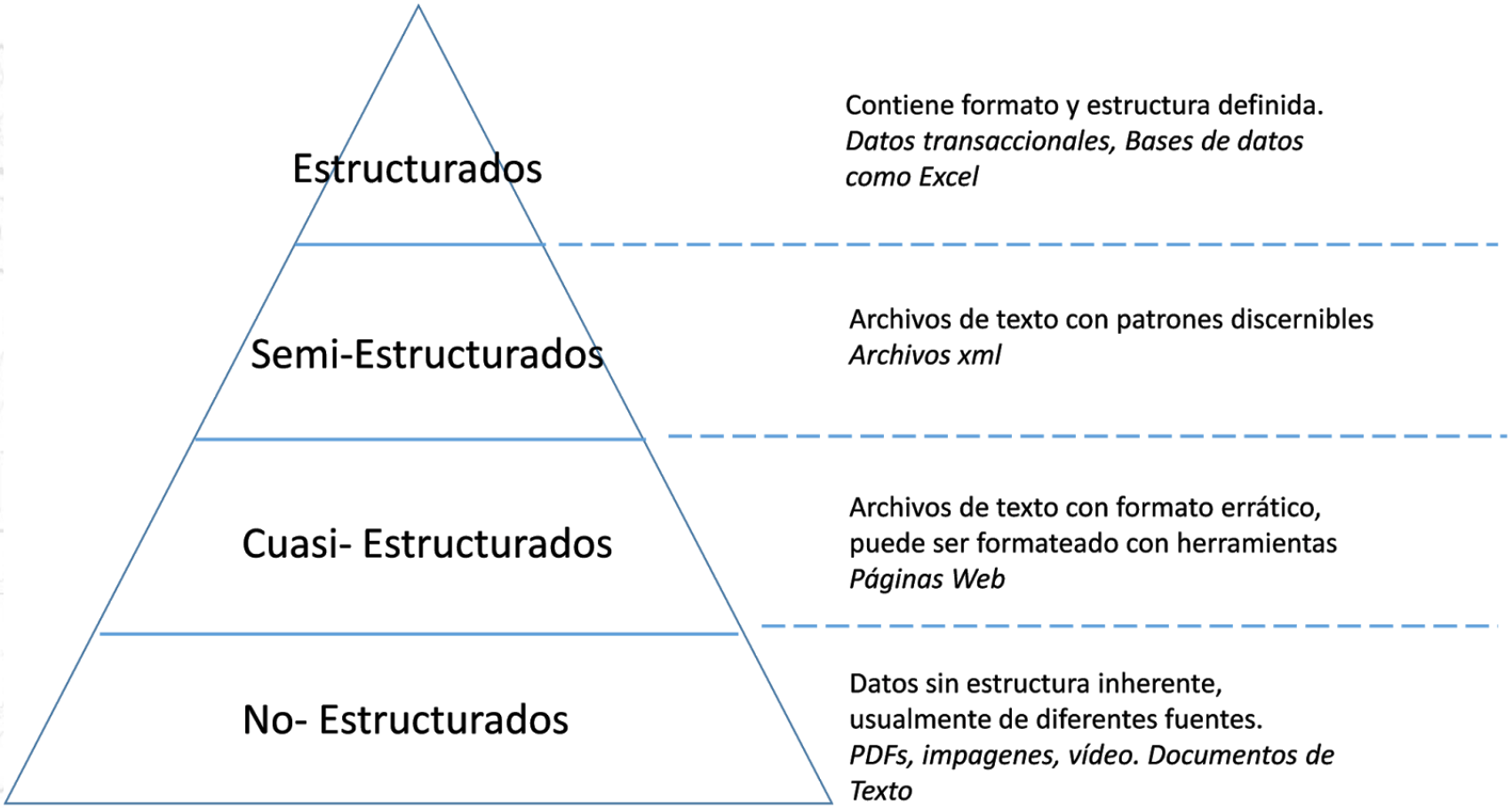
James Sanger



Para saber más:
www.cambridge.org/cl/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/text-mining-handbook-advanced-approaches-analyzing-unstructured-data?format=HB

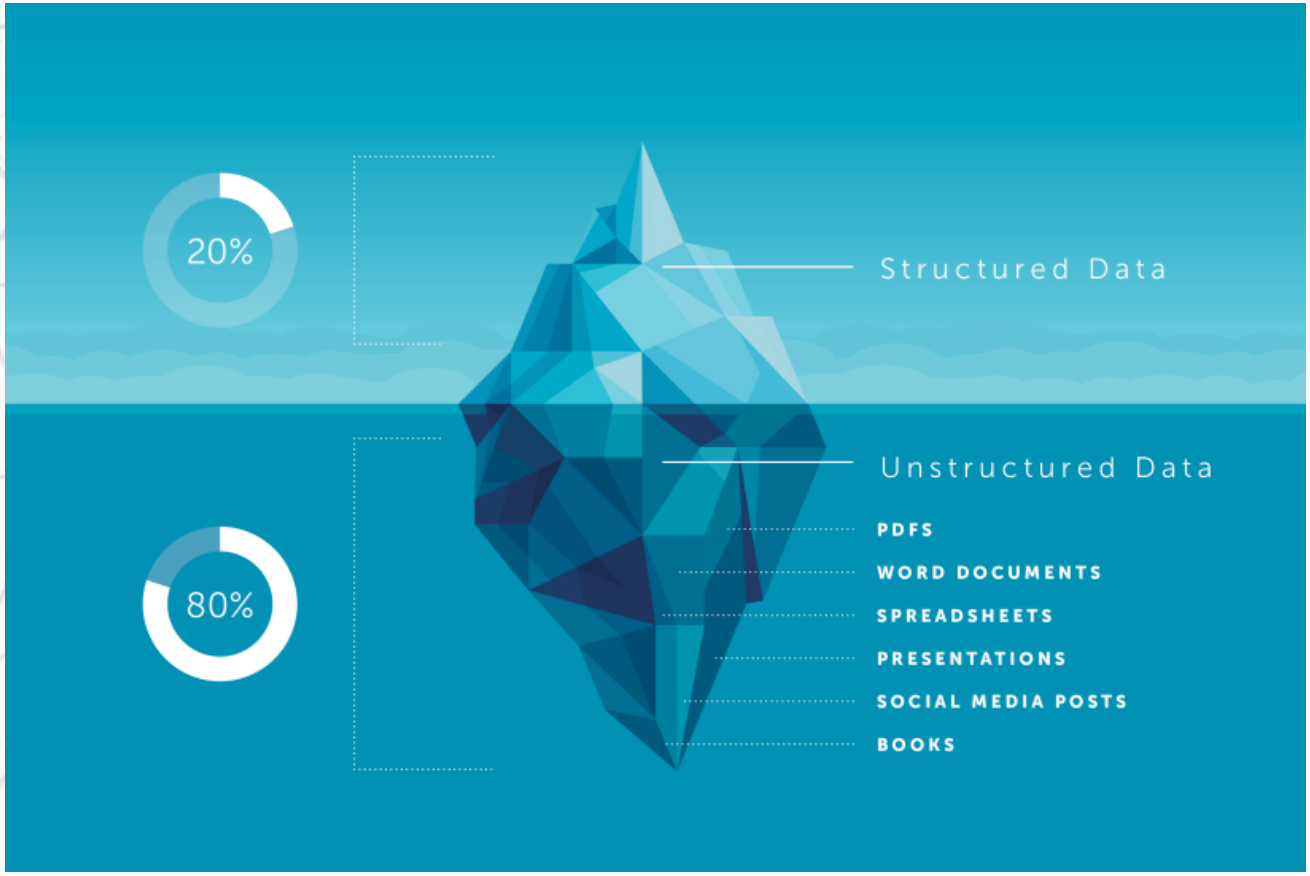
Minería de textos (3)

La génesis de la minería de textos se debe a cuestiones esencialmente prácticas, que tienen que ver con la necesidad de procesar datos no-estructurados, esto es:



Minería de textos (4)

Esta diferencia entre datos estructurados y no-estructurados no tendría que representar un problema..., salvo que hoy en día tan solo la WEB (no contemos todavía repositorios de documentos que van desde bibliotecas digitales..., hasta nuestros computadores) está constituida principalmente por datos no-estructurados:



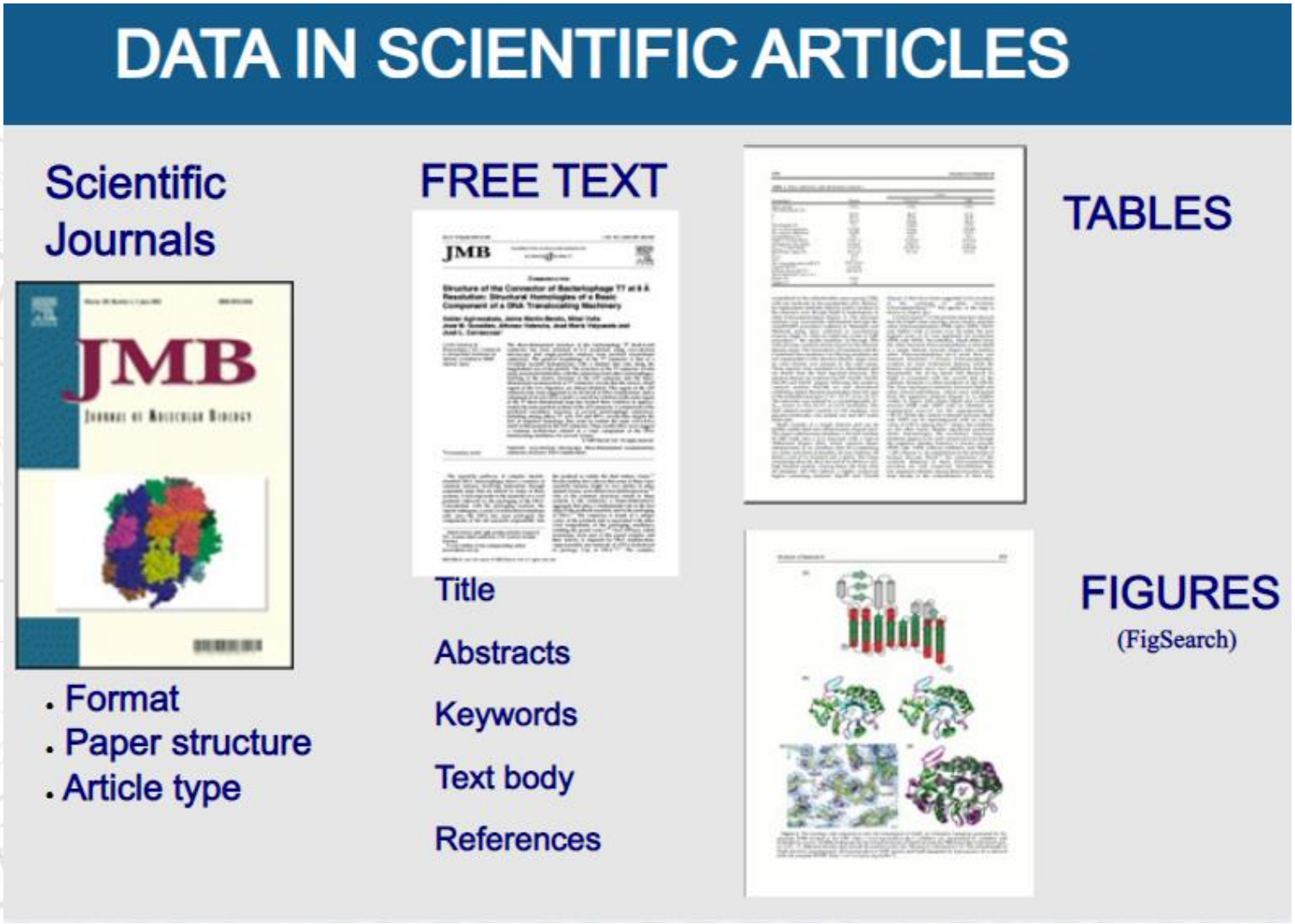
Minería de textos y bioinformática (1)

Hay muchas áreas de conocimiento en donde la minería de textos se ha aplicado para resolver el procesamiento de datos no-estructurados. Una de ellas es la **bioinformática**.

A grandes rasgos, y retomando un poco de lo que hablamos sobre infometría, a la bioinformática le interesa procesar datos reportados en la litertarua que se publica constantemente:

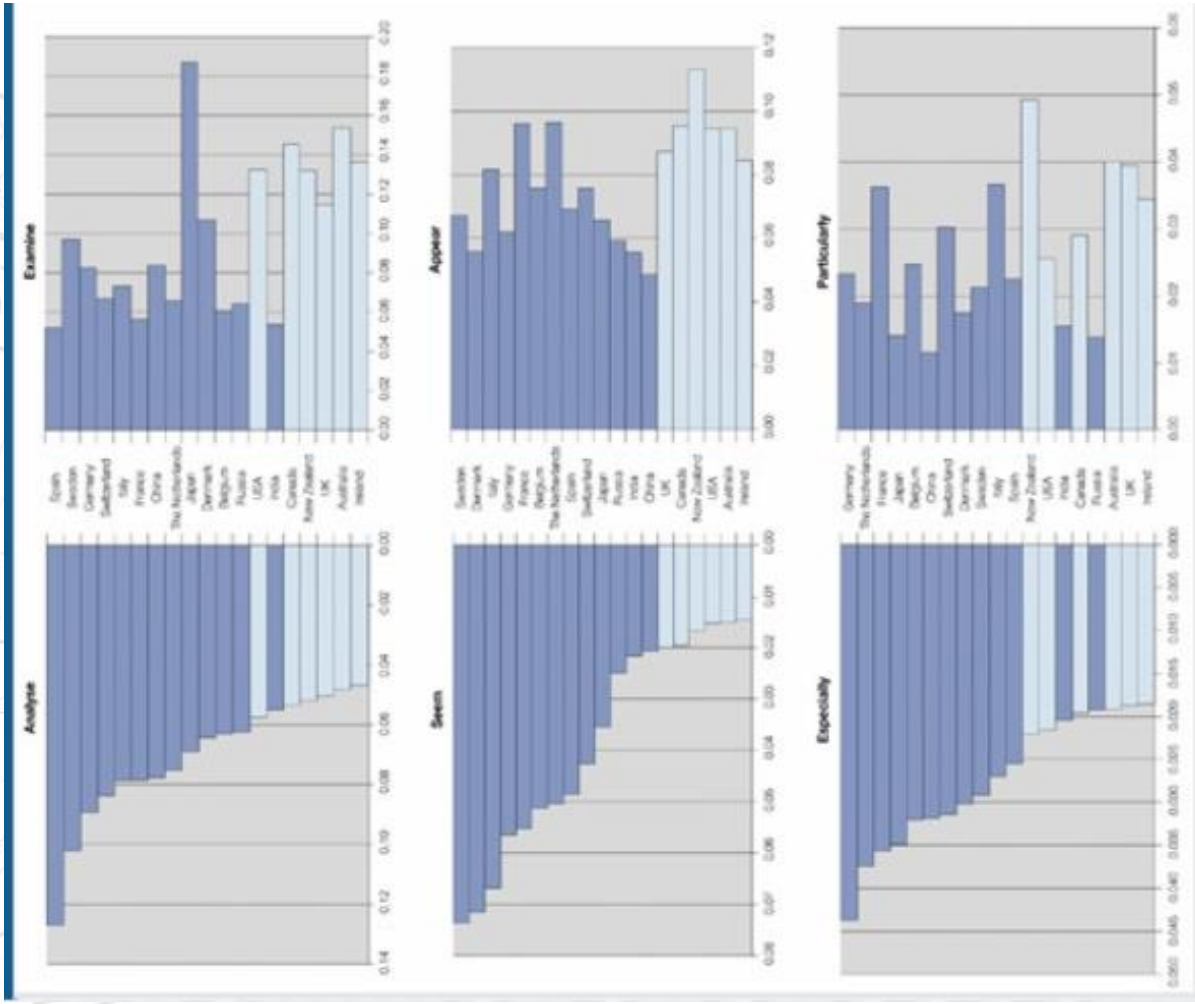
Minería de textos y bioinformática (2)

Si nos concentramos en uno de sus productos, los artículos científicos, lo que interesa analizar es lo siguiente:



Minería de textos y bioinformática (3)

Si consideramos a profundizar en el contenido de los textos, comenzamos a tener algunos problemas, p. e., la lengua: ¿analizamos solo textos en inglés, o debemos atender otros idiomas?:

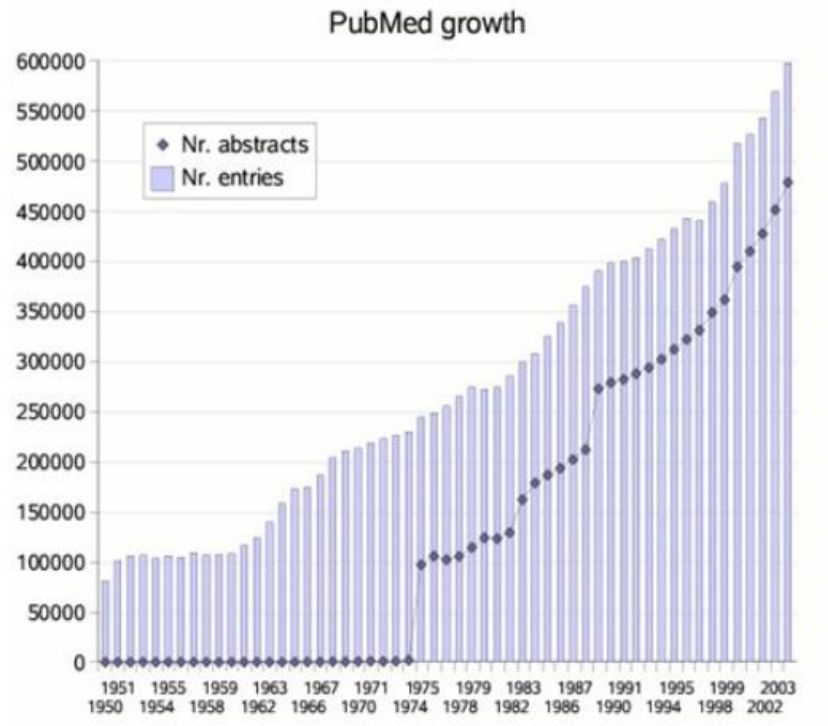


- Most in English.
- Different native languages.
- Different word usage (preferences)

Netzel R, Perez-Iratxeta C, Bork P, Andrade MA
The way we write. EMBO Rep. 2003 May;4(5):446-51

Minería de textos y bioinformática (4)

Por otro lado, tenemos el problema de la cantidad de documentos: hasta el día de hoy, de acuerdo con la información que brinda **PubMed**, cuenta con alrededor de 30 millones de referencias a artículos sobre literatura biomédica. Por supuesto, esta cantidad crece de manera constante:



PubMed

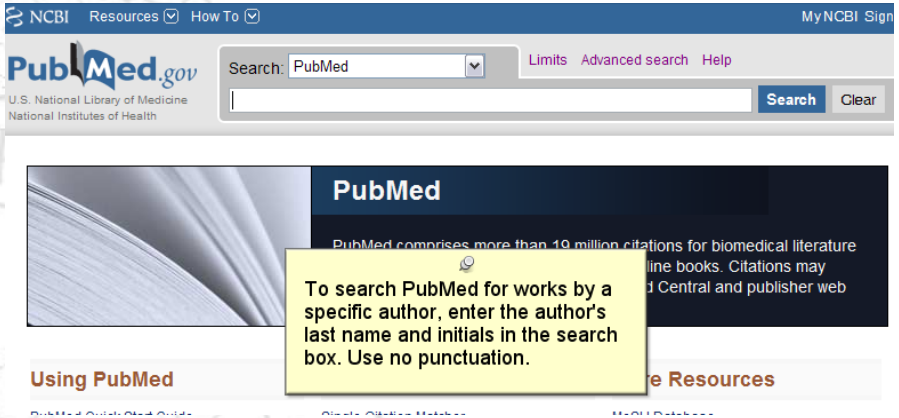
~ 450,000 new abstracts/a

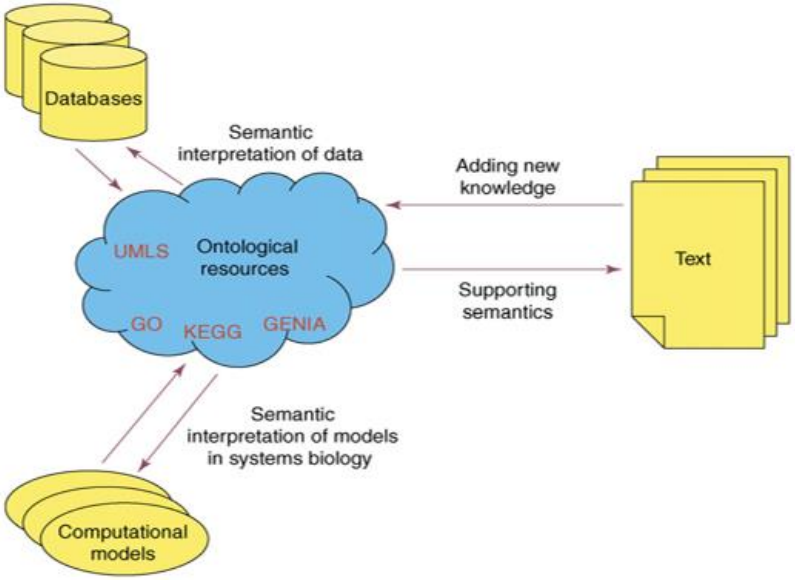
> 4,800 biomedical journals

Para saber más:
www.ncbi.nlm.nih.gov/pubmed/

Minería de textos y bioinformática (4)

Tomando en cuenta lo anterior, muchos investigadores han desarrollado varios proyectos de minería de textos que toman como objeto de análisis los repositorios textuales de PubMed, p. e.: el **National Center of Text Mining** (NaCTeM) en Manchester, Inglaterra.

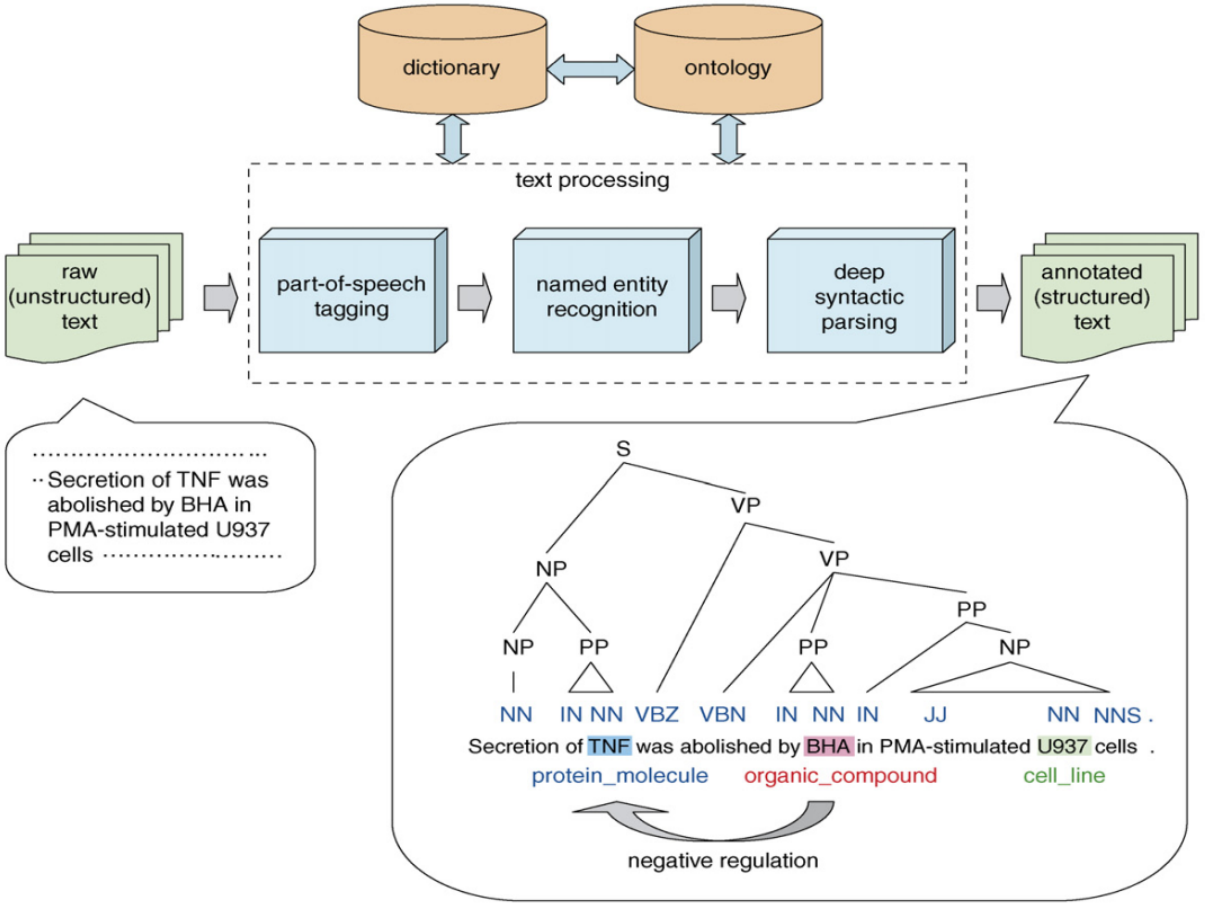




www.nactem.ac.uk

Minería de textos y bioinformática (5)

Dos investigadores de este centro, **Sophia Ananiadou** y **John McNaught**, han desarrollado un método estándar para hacer minería de textos en biomedicina, que es el siguiente:



Para saber más:
https://us.artechhouse.com/Text-Mining-for-Biology-and-Biomedicine-P932.aspx

Gracias por su atención

Blog del curso:
https://cesaraguilar.weebly.com/internet-y-sociedad.html