UNIVERSIDAD DE JAÉN

Investidura del

Excmo. Sr. D. Antonio Pascual Acosta

como Doctor *Honoris Causa*

LAUDATIO

a cargo del
Prof. Dr. D. Emilio Damián Lozano Aguilera
Profesor Titular del Área de Estadística
e Investigación Operativa

DISCURSO DE INVESTIDURA

del
Excmo. Sr. D. Antonio Pascual Acosta

Jaén, 27 de junio de 2018

UNIVERSIDAD DE JAÉN

Investidura del

Excmo. Sr. D. Antonio Pascual Acosta

como Doctor *Honoris Causa*

LAUDATIO

a cargo del
Prof. Dr. D. Emilio Damián Lozano Aguilera
Profesor Titular del Área de Estadística
e Investigación Operativa

DISCURSO DE INVESTIDURA

del
Excmo. Sr. D. Antonio Pascual Acosta

Jaén, 27 de junio de 2018

DISCURSO DE INVESTIDURA

No son los tiempos tan malos ni el terruño tan estéril como afirman los de afuera y lo que es peor, algunos de los de casa. Quizás no demos todo el fruto conveniente; pero flores ya hay, y viéndolas y admirándolas, aunque el fruto no responda totalmente a nuestras esperanzas, obligados nos sentimos todos a conservar y cuidar el árbol.

Benito Pérez Galdós
Prólogo Tercera Edición La Regenta de Clarín

SEÑOR RECTOR MAGNÍFICO DE LA UNIVERSIDAD DE JAÉN.
SEÑORAS Y SEÑORES VICERRECTORAS, VICERRECTORES,
DECANAS, DECANOS Y MIEMBROS
DEL CLAUSTRO UNIVERSITARIO.
AUTORIDADES
PERSONAL DOCENTE, ESTUDIANTES,
Y PERSONAL DE ADMINISTRACIÓN Y SERVICIOS.
SEÑORAS Y SEÑORES.

Estoy emocionado y feliz por encontrarme en Jaén, en mi tierra, con mi gente, entre mis amigos, en una mañana inolvidable para mí.

A lo largo de mi vida, Jaén ha sido una referencia esencial y determinante de mi biografía más personal e íntima. Aquí nací y me crié, en este escenario, en la calle Salido, en la Calle Goya en el Barrio de Peñamefécit, en la Calle Castilla en el Barrio del Arrabalejo, trascurrieron años inolvidables y, cuando tuve la necesidad de salir para continuar mis estudios en la Universidad, dejé atrás tantos familiares, amigos, experiencias y recuerdos, tal cúmulo de raíces, que hicieron ya imposible el desamor o el olvido.

Durante muchos años, por razones de trabajo y familiares, he vuelto con bastante frecuencia a mi Jaén. Hoy regreso de nuevo, aunque este viaje, a diferencia de otros, está envuelto en un exceso inmerecido de generosidad, la que han tenido los que han decidido concederme el máxi-

mo galardón que una Universidad puede otorgar, como es la distinción con el grado de Doctor Honoris Causa.

Nunca se me hubiera ocurrido pensar que viviría o protagonizaría un hecho semejante, que recibiría un honor de este calibre, a todas luces excesivo, respecto a mis méritos. Hoy que me veo honrado con esta distinción tan importante para mí, cambiaré el signo de mi habitual inclinación y pensaré que, si la generosidad de tan docto colectivo así lo ha estimado oportuno, no es correcto debatir su pertenencia, ni siquiera conmigo mismo y me limitaré como no podría ser de otra manera a expresar mi sincera gratitud a los miembros de esta Universidad. Pues me preocupa sobremanera caer en ese pecado, uno de los más severos de cuantos se puedan cometer, según Cervantes, por medio de Don Quijote, al escribir que *"uno de los pecados que a Dios más ofende es la ingratitud"*.

Mi gratitud, en primera instancia, al Rector D. Juan Gómez Ortega, que tanto y tan bien está haciendo por la consolidación y proyección de esta Universidad, al Departamento de Estadística, al Consejo de Gobierno, al Claustro Universitario y a todos los que habéis participado en esta decisión tan grata para mí.

A Emilio Lozano Aguilera quiero mostrarle públicamente mi agradecimiento por aceptar la propuesta para que realizara la *laudatio,* y agradecerle también sus afectuosas palabras, fruto sin duda del cariño que nos profesamos y de los muchos años que llevamos trabajando juntos.

Y a todos ustedes, señoras y señores, y especialmente a los que os habéis desplazado a Jaén para acompañarme en este acto, os agradezco vuestra asistencia y vuestra amable atención.

He de deciros que tengo sentimientos contrapuestos. Por un lado, la enorme alegría de ver entre vosotros a mis hijos, a mis hermanos y a mis sobrinos. Y, por otro lado, una profunda tristeza por no poder estar conmigo mi padre, Manuel Pascual, pues hoy sería la persona más orgullosa de la tierra. Aunque se dedicó al mundo de la empresa, su vocación era la Universidad, y de hecho fue uno de los primeros profesores del Colegio Universitario de Jaén. Él me inculcó el amor por la Universidad, la cultura del esfuerzo y el respeto a los demás, y a otras ideas diferentes a las tuyas.

La puesta en marcha de esta Universidad, en 1993, no hubiese sido posible sin la calidad docente e investigadora, el empuje y el tesón de tantos profesores del Colegio Universitario de Jaén. Sin el apoyo y colaboración de tantas personas que, ya sea desde la Junta de Andalucía, desde la Diputación Provincial y el Ayuntamiento de Jaén, desde los partidos políticos, desde las organizaciones empresariales y sindicales, o desde los medios de comunicación, ayudaron a hacer realidad un instrumento que ha transformado social, cultural y económicamente, a la ciudad de Jaén y a toda su provincia, y que ha ayudado a abrir en Jaén una ventana al mundo.

Pienso que la grandeza de la actividad pública, reside en la posibilidad de poder llevar a cabo proyectos generales de transformación y mejora de la sociedad en la que se vive desde la perspectiva de las distintas organizaciones sociales y políticas de que se dispone en un sistema democrático.

Es más, cuanto más amplia es la responsabilidad que se ha desempeñado, más difícil resulta en ocasiones visualizar en hechos concretos, las consecuencias que esas actuaciones han tenido en la vida de los pueblos.

Ustedes me están proporcionando hoy, no obstante, una ocasión inmejorable para comprobar de una sola ojeada una de las consecuencias que más gratamente pueden obtenerse en el desempeño de una tarea pública. Estamos viviendo la experiencia de una comunicación entre personas que incluso desde perspectivas contrapuestas, hemos desempeñado diversos papeles en la organización de esta sociedad, pero hemos coincidido en algo que nos une a todos los que estamos en este acto: El amor por nuestra tierra, el reconocimiento de sus múltiples encantos y al mismo tiempo una pasión, que me recuerda a lo que el filósofo alemán Spengler denominó el "alma fáustica de Occidente"[1], una pasión por conseguir que Jaén supere sus deficiencias y esas frustraciones históricas que aún perviven en la conciencia de algunos de nosotros, de Jaén como tierra de frontera en la que solo se piensa como lugar de paso y sepamos mirar con seguridad, ilusión y optimismo su futuro.

Quisiera resaltar, al mismo tiempo, con ocasión del acto que estamos celebrando, lo que significa también de reconocimiento, por encima de las personas, a la tarea pública. Creo que es el momento de reflexionar sobre la necesidad y la grandeza del trabajo que muchos hombres y mujeres están llevando a cabo en los distintos partidos políticos y organizaciones sociales por lo que supone de dedicación y compromiso con intereses y objetivos que trascienden lo particular y que se refieren a la dinámica y a la vida del conjunto de cada pueblo.

Pienso que aún siguen vigentes, al menos para mí lo están, aquellas características que el filósofo e historiador alemán, considerado por muchos como uno de los padres de la sociología moderna, Max Weber, señalaba como esenciales para el ejercicio de una tarea pública en una sociedad democrática: pasión, sentido de la responsabilidad y mesura.

Pasión, porque aunque es necesario mantener la cabeza siempre fría para decidir con la máxima objetividad y equilibrio posibles, entre los diversos intereses en juego, sin embargo la política no puede convertirse en un mero cálculo estadístico o de probabilidades, sino que es preciso identificarse con las situaciones difíciles que viven los hombres y mujeres y creerse vivamente las vías que se elijan para resolverlas.

Responsabilidad, porque la toma de decisiones que afectan a intereses colectivos siempre debe estar contrape-

sada con la capacidad de responder personalmente ante los errores cometidos, y

Mesura, porque, aunque no se puede realizar lo posible si no se busca y se sueña en lo imposible, es necesario ser conscientes de las limitaciones y condiciones que rodean a cada momento histórico para alcanzar lo que sea posible conseguir a través de la negociación y el acuerdo entre las partes implicadas.

No tengo reparo alguno, lo confieso, en calificar mi trabajo universitario como una verdadera vocación: desde siempre, influenciado como decía antes por mi padre, o por mi maestro, el profesor D. Rafael Infante Macías, del que tantas cosas buenas he aprendido, o por profesores de la Universidad de Granada como D. Ramón Gutiérrez Jáimez, D. Alfonso Guiraúm Martín o D. Luis Esteban Carrasco, desgraciadamente alguno de ellos ya no están entre nosotros, que desde las aulas universitarias granadinas inculcaron en mí, con su ejemplo y dedicación, el entusiasmo por la docencia y la investigación, en definitiva, por la pertenencia a una comunidad que, desde hace siglos, está firme y fundamentalmente empeñada en ampliar las fronteras del conocimiento humano y, de esta forma, promover el progreso de la sociedad.

Deseo pues, en esta intervención, hacer un reconocimiento profundo y sincero a la que considero la mayor y más noble aventura de la historia. La humanidad no sería la que es, ni habríamos alcanzado el nivel de bienestar y el

grado de autonomía del que hoy gozamos, sin las sucesivas aportaciones de estudiosos, científicos e investigadores, de miles de hombres y mujeres que, a contracorriente muchas veces y en frecuente conflicto con dogmas, directrices o inercias, probaron intuiciones, fundamentaron conocimientos y aplicaron técnicas innovadoras que, poco a poco, han ido abriendo y despejando el camino del futuro.

Después de más de cuarenta años dedicado a la Educación, ya sea como profesor o como responsable de la Educación Superior, he llegado a la conclusión de que lo esencial en materia de enseñanza, el elemento clave de todo proyecto educativo de carácter personalista, reside en preguntarse por los fines de la educación o dicho más coloquialmente, toda pretensión de educar exige plantearse, ¿qué tipo de ciudadanos o ciudadanas queremos formar?, ¿cuál es el perfil de la persona que queremos se haga cargo en el futuro de los destinos de nuestra sociedad en cualquiera de sus ámbitos, cultural, social, económico o político?

Tras abrir estos dos interrogantes debemos afrontar, a continuación, el análisis de la realidad social y los cambios que están condicionando toda previsión de futuro.

Soy consciente de que solo si el sistema educativo es capaz de asumir y hacerse cargo del contexto real en el que nos movemos y en el que han de moverse quienes ahora están en periodo de formación, será posible ofrecer los resor-

tes y los códigos de desciframiento adecuados y a la altura de las exigencias de cada época. Solo así los educandos de hoy estarán en condiciones de enfrentarse a sus responsabilidades, podrán participar creativamente en la construcción de un mundo nuevo y transformar la realidad conforme a los criterios y principios axiológicos que favorecen una auténtica convivencia democrática.

Vivimos en sociedades cada vez más complejas, en las que es preciso prestar mucha atención a las nuevas tendencias que están surgiendo en el ámbito de la economía y en su relación con la sostenibilidad; a los modos en que se están reestructurando los ámbitos sociales, culturales o empresariales, con particular acento en el fenómeno de la inmigración; y, en fin, a la evolución del área de las comunicaciones y al espectacular desarrollo de las tecnologías de la información.

Manuel Castells, hace ya algunos años, en su obra *La era de la información. Economía, Sociedad y Cultura*[2], sostenía que *"teniendo en cuenta el peso de las nuevas tecnologías, la hegemonía de los planteamientos económicos y la importancia de la ciencia y de la innovación en todas las esferas de actividad, se puede ya hablar de una transición del industrialismo al informacionalismo, como característica definitoria del momento histórico que estamos viviendo."*

Se trataría, en definitiva, del paso de una sociedad configurada conforme al modelo que inauguró la revolución

industrial clásica a otra que gira en torno al conocimiento como fuente de energía y materia prima esencial.

De esta transformación se ha dicho que constituye una nueva era de dimensiones insospechadas, en la que los cambios que se están produciendo a escala planetaria, arrastran otros efectos y provocan otras dinámicas en el plano regional y local.

Hoy día, vivimos experiencias de pérdida de homogeneidad y asistimos a importantes movimientos de población con millones de personas que emigran o buscan refugio, huyendo de la guerra, del subdesarrollo y la explotación. Las relaciones comerciales y las modernas comunicaciones promueven e introducen costumbres e influencias no fácilmente controlables. La sociedad tradicional se tambalea al menos en algunas de sus dimensiones y sugiere distintos modelos de familia y nuevos papeles o funciones para mujeres y hombres.

En el ámbito de las relaciones internacionales, las certezas hasta ahora existentes dan paso a un grado de incertidumbre considerable ante problemas o conflictos como el terrorismo y el fundamentalismo, el cambio climático, la pobreza y la desesperación que reinan en grandes zonas del mundo.

Dar una respuesta positiva a estos fenómenos es especialmente complicado. Sobre todo si, al mismo tiempo, hay que enfrentarse, como ocurre en estos momentos, a la salida de una crisis económica y financiera de proporciones

globales, que tanto ha influido de manera negativa en la vida de casi todos.

En buena parte de la población, tal cúmulo de mutaciones, tantas y tan rápidas, tan desconocidas y sorprendentes, dan lugar a una sensación de pérdida, de inseguridad, de miedo a lo desconocido, de amenaza incluso.

Esta situación, sin duda bastante generalizada, es la que lleva al sociólogo alemán Ulrich Beck, a definir el mundo contemporáneo como "la sociedad del riesgo"[3], en la que la mayoría instalada ve, cómo aquello que a sus ojos es valioso y constituye el fundamento de su posición y del sentido de su vida, empieza a ser cuestionado, discutido y hasta atacado.

Ese miedo que, en ocasiones es desasosiego moral, encuentra una vía de salida y un asidero en las justificaciones teórico-ideológicas que pretenden arropar y dar carta de naturaleza al statu quo. Nos encontramos, así, con el discurso del "fin de la historia", que intenta avalar al capitalismo como el único modelo posible de sociedad y a otras variantes de lo que se ha dado en llamar "pensamiento único", cuyo principal objetivo es convencernos de la imposibilidad de alternativas a la hegemonía y a las inercias del mercado.

Conjurar el miedo y la incertidumbre consiste, para muchas personas, en adoptar como modo de vida aquello que se predica como inevitable. De ahí que nos encontremos con un nuevo tipo de conservadurismo que, en muchas ocasiones, no es identificable con las formas clásicas y con-

sagradas de la mentalidad o del pensamiento conservadores. Se trata más bien de una actitud acrítica, instrumental, pragmática, que invade por contagio multitud de esferas de las vidas política, económica o social y que, por supuesto, afecta también a la educación.

Son las voces que conciben los centros educativos como si fueran "empresas", que se deben programar, dirigir y gestionar con criterios de eficacia y rentabilidad; que piensan en los alumnos y en sus familias como "clientes" a los que ofrecer productos terminados, competitivos y diversificados. Son, en definitiva, los que sostienen que educar, es fundamentalmente, proveer de "mano de obra" al tejido productivo y, en consecuencia, el talante que se exige al alumnado es el del agresivo competidor, orientado a triunfar a costa de lo que sea, sin discutir jamás los supuestos del sistema. Esta actitud es, además, compatible e inseparable de la condición de sumiso consumidor y de legitimador del poder establecido.

Ni que decir tiene, que el proyecto educativo en el que creo y que siempre he defendido no es, ni quiere ser, un servicio sometido a las demandas perentorias del mercado, sino una apuesta con perspectiva de futuro, que considera a cada alumno y alumna como persona, aún en fase de formación, y a los que, por tanto, hay que ofrecerles medios humanos y materiales, conocimientos científicos y culturales, así como sólidos principios éticos y valores cívicos que

les permitan, una vez interiorizado todo este bagaje, crecer en madurez, en autonomía y criterio propio, para juzgar o actuar y, de esta forma, avanzar en el dominio de los procedimientos y destrezas que han de servirles a lo largo de su vida para integrarse en la sociedad como miembros activos, responsables, útiles y creativos.

Esto significa que educar es, ante todo, abrir horizontes, impulsar el acceso al mayor número de oportunidades, promover la capacidad de elección, facilitar la resolución de los problemas y de las necesidades que se planteen ahora o en el porvenir y es también, lógicamente, motivar e ilusionar a los jóvenes, ayudarles a ser ellos mismos y a estar preparados para analizar el entorno en el que cada cual se mueve, con el fin de intervenir en él como un factor de cambio y transformación.

En mi campo, la Estadística, he vivido el cambio que en la misma se ha producido desde los años 70 del siglo pasado hasta la actualidad. Pienso que sigue siendo válida la definición de Hampel que afirma que la estadística es la ciencia y el arte de extraer información útil y relevante de un conjunto de datos. Afecta pues al trabajo del estadístico la recolección de los datos, el procedimiento de la misma, así como el análisis y conclusiones que se pueden extraer en función del objeto del estudio. Si entonces hablábamos de grandes muestras cuando el tamaño era mayor que 30, hoy

día, es fácil acceder a grandes volúmenes de datos debido a las capacidades de almacenamiento de los ordenadores, que además pueden ser heterogéneos y sobre los que no existen modelos a priori, pero que pueden ser analizados gracias a la capacidad de cálculo de esos mismos ordenadores. Así, tenemos disponibles para aprendizaje y uso por parte del alumnado datos, no de 30 o más casos como hace 40 años, sino de tamaño que hoy llamamos medio como por ejemplo los datos sobre viajes realizados por taxis en N.Y. (datos NYC TLC) que ocupan aproximadamente 150GB y contienen 1.200 millones de líneas.

Ahora bien, los datos por sí solos prácticamente carecen de valor. El valor fundamentalmente se obtiene cuando se dispone de herramientas capaces de transformar dichos datos en conocimiento, como son las proporcionadas por la Minería de Datos, que suele definirse como un conjunto de herramientas estadísticas basadas en procedimientos multivariantes y ejecutables mediante algoritmos computacionales a partir de los cuales se estructuran procedimientos tanto descriptivos como predictivos, que permiten extraer el valor real de los datos. La Minería de Datos ha de hacer frente al enorme volumen de datos del que se dispone actualmente, requiere de una potente estructura de almacenamiento y la capacidad de obtención de resultados de un modo rápido y eficaz, prácticamente en tiempo real ya que, al no disponer de un modelo subyacente, la incorporación

de nuevos datos produce a menudo modificaciones en los resultados. Es en esta situación cuando surge el concepto de Big Data, refiriéndose al conjunto de infraestructuras y procedimientos de almacenamiento, reservándose el concepto de Ciencia de los Datos al conjunto de técnicas de análisis de los mismos.

Para definir qué es el Big Data, recurriré a una aplicación que lo ilustre, utilizada por muchos autores con esta finalidad. Se trata de la aplicación recogida en Ginsberg y otros (2009)[5], referida a la gripe, enfermedad que según la OMS en su estimación de diciembre de 2017 produce unas 650.000 muertes anuales en el mundo. En el año 2009 esta situación genérica se complicó al aparecer una nueva variante del virus, causante de lo que se denominaría "gripe aviar", que se expandió de una forma muy rápida, haciendo temer que se produjera una pandemia de aún peores consecuencias.

Para prevenir, controlar y abordar esta posible pandemia, en USA, los Centros para el Control y Prevención de Enfermedades (CDC), tenían como misión recoger toda la información que se generaba en torno a la enfermedad y consolidarla, publicando la información resultante en los ámbitos regional y nacional, lo que se producía con un retardo de una a dos semanas.

Ante este hecho en Google Inc. idearon un método para la vigilancia de la enfermedad que era capaz de generar informes sobre la misma con un retardo de un día y con una

precisión a priori muy buena. Para ello utilizaron el motor de búsqueda de Google, donde con "todas las reservas de la privacidad individual", analizaron las consultas que los ciudadanos hacían en la red relativas a la enfermedad citada, valoraron 50 millones de preguntas candidatas a considerar en el proyecto, con información procedente de los CDC en años anteriores, y utilizaron también 450 millones de modelos diferentes para validar cada una de las preguntas, utilizando para todo ello sistemas informáticos distribuidos con cientos de ordenadores. A partir de aquí validaron y ajustaron el modelo final.

Si se analiza lo descrito, queda claro que se disponía de un volumen importante de información, que además no estaba centralizada, con sistemas informáticos también distribuidos y también en volumen importante.

A partir de aquí se podría dar una definición de Big Data, pero aunque existen muchas definiciones, no hay una aceptada de forma generalizada. Muchas de las que se han dado, han sido proporcionadas por multinacionales del mundo informático como IBM, Intel, Oracle, etc, tal como lo recogen Ward y Barker (2013)[6], quienes tras analizarlas concluyen que en ellas concurren los siguientes factores: el volumen y la variabilidad de los tipos de información a utilizar, la componente tecnológica que va desde el almacenamiento de la información hasta el software necesario y las técnicas a utilizar para el tratamiento de la información.

Laney (2001)[7] en su trabajo sobre comercio electrónico resalta que la explotación de la información debe someterse a las denominadas tres uves: Volumen, Velocidad y Variedad. El Volumen identifica la cantidad y la dimensión de los datos a utilizar, la Velocidad tanto para la información que se incorpora como para la que se genera, y la Variedad referida a la posibilidad de utilizar simultáneamente diferentes tipos de datos y de fuentes que la generan. A estas tres uves iniciales, diversos autores han ido añadiendo otras para caracterizar mejor el proceso. Así, por ejemplo, Microsoft e IBM han añadido una cuarta V que corresponde a la Veracidad de la información.

El Big Data está cada vez más presente en nuestra sociedad. En él confluyen diversos actores y también diferentes especialidades científicas y técnicas. Así el demandante de Big Data, que es quien plantea el problema para satisfacer demandas de usuarios o clientes, tendrá que apoyarse en las ingenierías de hardware y de software y, por supuesto, en Ciencias como las Matemáticas y la Estadística.

Yin y Kaynak (2015)[8] afirman que:

"El Big Data se refiere incluso a conjuntos de datos cuyo tamaño está más allá de las posibilidades que las bases de datos y las herramientas de software permiten hoy en día, capturar, almacenar, gestionar y analizar."

El Big Data va un paso más allá de lo que, hasta hace poco, hemos denominado información a analizar. La necesidad de disponer incluso en tiempo real, de información

precisa, junto a la capacidad y el desarrollo de sistemas para medir el flujo y el tipo de información y por otra parte, la necesidad de encontrar procedimientos y herramientas de medición y análisis de este tipo de datos, han dado lugar al establecimiento de nuevas líneas de investigación interdisciplinares.

La primera es la de Almacenamiento, absolutamente necesaria para el tratamiento del Big Data, por el volumen de datos a tratar, la tipología de estos datos, las distintas ubicaciones físicas y soportes tecnológicos y el desarrollo de *software* que permita gestionar el correcto almacenamiento de la información.

La segunda es la de la Arquitectura de Almacenamiento, donde hay que considerar, cómo se gestiona el flujo de información y la interacción de los sistemas que se están utilizando.

La tercera es el Cálculo Distribuido, ya que como se tendrán sistemas informáticos distribuidos, será necesario disponer de software capaz de abordar el tratamiento de dichas estructuras de conexiones y hacer los sistemas más transparentes y homogéneos entre sí.

Y por último el Análisis y tratamiento de la información almacenada, fundamentalmente mediante el Cálculo de Probabilidades y la Estadística Matemática, ya que lo que generalmente se pretenderá será realizar planificaciones adecuadas, diagnósticos certeros, predicciones fiables, y propuestas personales.

En torno al Big Data existe una corriente de opinión que podría resumirse diciendo que disponer de la V relativa a Volúmenes importantes de información, garantiza un éxito seguro de las conclusiones que se obtengan del análisis de la citada información. Así Anderson (2008)[9], afirma que el método científico de hipótesis, modelo y contrastes, está llegando a ser obsoleto frente a la utilización masiva de los datos. O como se recoge en el trabajo de Elliot y Valliant (2017)[10], "En un mundo de Big Data, se dispone de gran cantidad de datos, que son más rápidos y fáciles de obtener que disponer de una muestra probabilística", de lo que parece desprenderse que podría considerarse obsoleto el tradicional muestreo aleatorio.

Sin embargo esto no es así. Basta recordar el método desarrollado por Google Inc. para determinar la demanda de vacunas para la gripe al que antes me he referido. Años más tarde, en 2013, la revista Nature publicó un artículo en el que reflejaba que el método que se diseñó en Google Inc. sobreestimaba los niveles de gripe. Y en 2014, en la revista Science se publicó un trabajo titulado *The parable of Google flu: Traps in Big Data analysis,* en el que se recoge que el método predecía en más del doble el número de visitas a médicos para la vacuna de la gripe que las cifras que proporcionaron los CDC. Se señalaban posibles fallos del método, afirmando que los errores que se generaban no eran aleatorios, y se ponía de manifiesto que la falta de transparencia del método diseñado conducía a la imposibilidad de replicarlo, afirmándose, que incluso teniendo acceso a los datos de que dispuso Google, sería imposible replicar el análisis del artículo original.

Esto muestra que el hecho de disponer de volúmenes importantes de datos puede conducir a ignorar cuestiones fundamentales, básicas del método científico, como la obtención de la información, la fiabilidad de los datos, su dependencia, o la replicabilidad. Y esto no es nuevo, baste con recordar algunos de los ejemplos que tradicionalmente se esgrimen en el campo de la Estadística.

En el año 1916 se incorpora a la realización de predicciones sobre campañas electorales en USA la revista "Literary Digest", obteniendo predicciones más o menos correctas en las elecciones de 1920 a 1932. El método que utilizaban para obtener la opinión de los ciudadanos consistía en realizar encuestas por correo y las personas a las que se le enviaba el cuestionario eran seleccionadas de marcos predefinidos, según se recoge en Moon (1999)[11].

Pero en las elecciones presidenciales de 1936, Gallup utilizó un nuevo método que fue recogido en el "Washington Post" en octubre de 1935 bajo el titular "Los sondeos son unas elecciones nacionales a pequeña escala". Bradburn y Sudman (1988)[12], en págs. 20 y 21, reproducen las páginas del periódico, donde se dice:

"El sondeo recogido hoy y todos los sondeos realizados por el American Institute of Public Opinion son una elección nacional en una escala pequeña. Las papeletas

proceden de todos los estados de la Unión, de grandes y pequeñas ciudades y distritos rurales, de todos los niveles de rentas. A todos los grupos se les ha asignado la misma proporción de votos que tienen en la elección nacional. El número de votos de cada estado está en proporción directa a su población y voto electoral..."

Con este método y con solo 2.500 entrevistas, el "American Institute of Public Opinion" predijo de forma correcta la victoria de Roosevelt en esas elecciones, en tanto que "Literary Digest", con una muestra de 2.400.000 respuestas a los 10.000.000 de cuestionarios remitidos a hogares que disponían de teléfono o tenían registrado a su nombre al menos un coche, erró en la predicción adjudicando la victoria al candidato republicano Alfred Landon.

Este hecho muestra que disponer de una gran cantidad de datos, 2.400.000 frente a 2.500, no supone garantía de éxito en las conclusiones que se obtengan.

En 1948 las encuestas electorales en EE.UU. sufrieron un nuevo revés. Dicho año los candidatos republicano y demócrata, eran Dewey y Truman. Las principales empresas dedicadas en ese momento a la realización de sondeos electorales, Crossley, Gallup y Roper predijeron la victoria de Dewey por al menos un 5% de diferencia; sin embargo el resultado fue el contrario, ganó Truman por una diferencia del 4,5%, lo que hizo que el "Social Science Research Council" creara una comisión, dirigida por Mosteller, para analizar las razones que pudieron dar lugar al citado error. Esta comisión

llegó a una serie de conclusiones de las que quizá la más significativa sea la debilidad del muestreo por cuotas para obtener la muestra representativa, ya que los entrevistadores tenían tendencia a entrevistar a las personas con mayor nivel educativo. Por consiguiente una de las recomendaciones de la comisión, fue que se utilizara el muestreo probabilístico en las encuestas electorales.

En Neyman (1934)[13] se abordan ideas estadísticas novedosas para el momento, justificando económica y temporalmente las necesidades de realizar muestreos, afirmando que la "muestra representativa general" debería ser en realidad un "método representativo de muestreo y método consistente de estimación".

También en Kruskal y Mosteller (1980)[14], se analizan multitud de situaciones prácticas que aparecen la literatura anterior justificando la necesidad de la utilización del muestreo probabilístico.

Por tanto, el hecho de disponer de grandes volúmenes de datos no garantiza por sí solo la obtención de conclusiones acertadas, por lo que no nos podemos olvidar de los fundamentos existentes para la obtención y tratamiento de la información de que se disponga, y que muchas de las técnicas que se han desarrollado para el análisis de datos, siguen siendo válidas en la era del Big Data con las adaptaciones oportunas.

Tradicionalmente en el campo de la Estadística Matemática, ha sido y es un deseo de cualquier investigador o persona que tome decisiones, disponer de grandes volúmenes de datos, llegándose incluso a pensar que disponer de ellos supone tener muy avanzada la resolución del problema que se estudia. Sin embargo la exigencia del rigor científico conduce de forma irremediable a tener que disponer de técnicas que sean capaces de sintetizar y generar conclusiones que sean absolutamente válidas dentro de unos márgenes de error adecuados y aceptables.

Todo esto introduce en el Big Data, cuestiones relativas a qué debemos analizar, qué va a suponer disponer de grandes volúmenes de información y qué novedades hay que introducir en el análisis de datos.

Disponer de grandes masas de información, conduce a tener que tratar con muchas características, rasgos o variables de los elementos u objetos de que se disponga para el estudio o análisis correspondiente, y por tanto podemos afirmar que en la mayoría de las ocasiones se dispondrá de datos hiperdimensionales, es decir datos donde el número de características, rasgos o variables a analizar es muy elevado.

Conviene recordar al respecto lo que Bellman ya puso de relieve en la década de los 60 del siglo pasado, y que describió con la expresión *la maldición de los datos hiperdimensionales* y aunque él lo hacía por la dificultad de optimizar en espacios productos por una enumeración exhaustiva,

dicha afirmación vuelve a ser vigente por razones o problemas como los que se describen a continuación.

Comenzaré planteando cuestiones generales sobre el hecho de disponer de datos hiperdimensionales. Ante esta situación, hay que admitir por lo general que los datos han sido generados por distintas fuentes y por diversos métodos, por lo que resultará conveniente estudiar las fuentes que los han generado y los correspondientes métodos utilizados para obtenerlos; para poder concluir sobre la fiabilidad y la homogeneidad de los datos.

Por otro lado, puede decirse que los datos de que se dispondrá vendrán afectados de errores en la mayoría de las ocasiones, más aún hoy en día donde se pueden estar utilizando simultáneamente datos generados por un dispositivo de precisión y otros obtenidos desde las redes sociales, de las cuales se sabe, que muchas de las informaciones que se obtienen, no son lo transparentes y reales que debieran ser. Por consiguiente, al manejar datos hiperdimensionales los errores mencionados pueden acumularse, por lo que habrá que comprobar si dicha acumulación genera un error global importante.

Además, manejar datos hiperdimensionales, generalmente lleva consigo que en la mayoría de las dimensiones no se disponga de datos de algunos elementos u objetos, lo que se conoce en el campo de la Estadística como el problema de las celdas vacías, que afecta al método a utilizar y

a la fiabilidad de las conclusiones que puedan derivarse de los datos.

Hay que valorar también cuestiones tan generales como la caducidad de la información, ya que puede ocurrir que en la información de que se dispone, no todos los datos tengan el mismo ciclo de vida, es decir la misma validez temporal. Y por supuesto hay que valorar también el problema de la seguridad y privacidad de la información, ya que al disponer de muchas informaciones que por sí mismas de modo aislado garantizan la privacidad, al componerlas puede llegarse a identificaciones de tipo personal, lo que puede venir agravado por el tipo de arquitectura distribuida de las bases de datos. En Mantelero (2017)[15], se trata el problema de la protección de la información de las personas y el tratamiento de los datos personales en el Big Data, reflexionando sobre las divergencias en su tratamiento del Consejo de Europa y de la Unión Europea.

A estas consideraciones de carácter general han de unirse otras como las que se describen a continuación, con las que pretendo poner de manifiesto peculiaridades que hacen dudar de la posible aplicación sin adaptación de diversos métodos utilizados para el análisis de datos y de las interpretaciones que puedan derivarse de ellos.

Trabajar con datos en dimensiones reducidas, sean del tipo que sean, siempre ha supuesto considerar su estructura geométrica y el comportamiento geométrico de los mismos, que han coincidido con lo que intuitivamente se

esperaba de ellos. Ahora bien, trabajar con datos hiperdimensionales en el marco del Big Data, plantea situaciones anómalas o al menos paradójicas.

Comencemos por algo tan simple como es el volumen de un cuerpo geométrico, como puede ser el cubo en tres dimensiones, o el hipercubo en más dimensiones.

Sea $C^n(s)$ el hipercubo centrado en el origen del espacio n-dimensional R^n y con longitud de lado $2s$, su volumen n-dimensional viene dado por $Vol.\ (C^n(s)) = 2s^n$. Y de esta expresión se concluye que si la dimensión aumenta, $n \uparrow \infty$, y el lado es inferior a uno, $s < \frac{1}{2}$, entonces el volumen se reduce tendiendo a cero, $Vol.\ (C^n(s)) \rightarrow 0$, en tanto que si el lado es superior a uno el volumen aumenta tendiendo a infinito, $s > \frac{1}{2}\ Vol.\ (C^n(s)) \rightarrow \infty$. Por último si el lado vale 1 el volumen permanece constante, $s = \frac{1}{2}\ Vol.\ (C^n(\frac{1}{2})) \rightarrow 1$, y además la diagonal del cubo $(C^n(\frac{1}{2}))$ es $\sqrt{n}$ es, que converge a infinito. Esta diversidad de casos indica que se pueden presentar situaciones extremas y paradójicas, que la intuición respecto de las mismas se pierde, y que en consecuencia es esperable que se presenten problemas para la interpretación de la información de la que se disponga.

Uno de los problemas clásicos que se aborda en el análisis de datos es el del "vecino más próximo", el cual consiste en que dada una colección de puntos (datos) y un punto adicional fijado, todos dentro de un espacio métrico n-dimensional, se debe encontrar el punto de la colección más

próximo al punto fijado. El interés de este problema radica en su gran aplicabilidad, pero no solo por lo que puede significar en los análisis de las bases de datos, sino también porque es fundamento en la construcción de técnicas para la identificación de observaciones *outliers,* o para el análisis de conglomerados, o bien en aplicaciones tan reales como la detección del fraude, la comparación de imágenes, o la información geográfica, por ejemplo.

El problema del vecino más próximo en el caso hiperdimensional ha sido tratado entre otros por Beyer et al. (1999)[16], que demuestra bajo condiciones muy generales, que cuando la dimensión del espacio se incrementa, todos los puntos convergen a la misma distancia del punto fijado, es decir la distancia del vecino más próximo se asimila a la distancia del vecino más lejano, por consiguiente el vecino más próximo en estructuras hiperdimensionales, o bien no tiene significado alguno o bien no discrimina situaciones diversas. En ese mismo trabajo y con el fin de dar una visión práctica de lo que ello significa, presenta resultados empíricos en los que se observa, que una dimensión mayor que 15, ya plantea los problemas descritos.

Dentro del ámbito de la Estadística la distribución normal es relevante, tanto por sus aplicaciones en diversos campos, como desde el punto de vista estadístico y matemático, ya que muchos de los modelos estudiados en la literatura convergen a ella.

En la distribución normal n-dimensional de vector media cero y matriz de varianzas y covarianzas la identidad, la función de densidad tiene su única moda en el origen y el valor de dicha moda es $\frac{1}{(2\pi)^{\frac{n}{2}}}$, por tanto, cuando aumenta la dimensión, este valor tiende a cero, y esta distribución que en bajas dimensiones tiende a concentrar la probabilidad en el entorno del origen, al incrementar la dimensión de la distribución se va desvaneciendo, es decir la probabilidad tiende a desplazarse hacia las colas de la misma. Este hecho queda patente en la siguiente tabla dada por Wang (2012) [17], donde calcula la probabilidad de que la norma de un punto sea mayor que dos, $P[\|Z\| \geq 2]$, para dimensiones 1, 2, 5, 10, 20 y 100 y se observa que la probabilidad tiende de forma rápida hacia los extremos de la distribución cuando la dimensión se incrementa.

n	1	2	5	10	20	100
Probabilidad	0,04550	0,13534	0,54942	0,94734	0,99995	1,00000

Entre otras peculiaridades en el tratamiento de datos hiperdimensionales, puede citarse el hecho de que cuando se trabaja en el campo de la Estadística, el tamaño de la muestra a utilizar siempre es mayor que la dimensión del espacio en el que se está trabajando, sin embargo en el caso que nos ocupa ello no suele ocurrir; es más, las leyes de los grandes números, se abordan desde el hecho de que el tamaño muestral tiende a infinito y la dimensión del espacio permanece fija, y cuando se abordan situaciones en espa-

cios hiperdimensionales, el tamaño de la muestra tiende a ser finito y la dimensión del espacio infinita, por lo que hay que cuestionarse la aplicación de dichas leyes.

En el mismo sentido cabe citar el trabajo de Puts et al. (2015)[18], quienes desde el Instituto de Estadística de Holanda (Statistics Netherlands) han abordado el problema clásico de depuración de datos en una estructura de Big Data para datos generados por sensores de carretera. Trabajando con varios millones de datos, y aplicando los modelos y métodos que tenían por su experiencia al tratar volúmenes relativamente grandes de datos administrativos, observaron que muchas de esas técnicas no podían ser aplicadas no solo por la variedad de los datos sino también por su volumen, teniendo que generar nuevos métodos y técnicas para tratar la situación que generaba el Big Data.

He pretendido con estos ejemplos poner de manifiesto la precaución que hay que tener al aplicar sin modificaciones a situaciones de datos hiperdimensionales métodos y técnicas construidos para situaciones donde la dimensión del espacio es pequeña.

En mi vida profesional, como decía al principio, debo indicar que he estado viviendo la "Revolución de los Datos", y en esta revolución existen dos grandes épocas, la que se denominó la época del Análisis de Datos y que explotó en torno a los años 60 del siglo pasado y la que se conoce como Big Data y que podemos situar a finales de ese siglo. Las diferencias entre estas dos situaciones están muy bien

diferenciadas por Donoho (2000)[19], quien precisaba que el análisis de datos estaba pensado para disponer de datos u observaciones de fenómenos particulares, esos datos correspondían generalmente a valores de un vector de variables observadas y donde se suponía que las variables eran bien elegidas y localizadas en su ubicación, y sin embargo en el Big Data, por lo descrito, se dispone de datos de muy distinta naturaleza, donde se desconoce inicialmente el interés de los mismos, y además están deslocalizados.

El hecho de citar estos dos momentos precisos en esa revolución, es debido al hecho de que actualmente en el Big Data, están ocurriendo, salvando las diferencias entre ambos momentos, situaciones muy similares a las ya vividas en el Análisis de datos.

Situaría el inicio del "Análisis de Datos" en el trabajo de Tukey (1962)[20], cuya importancia tal como indica Jones (1992)[21] se pone de manifiesto por el hecho de que fue publicado en los Annals of Mathematical Statistics, sin contener en dicho trabajo teorema alguno ni condiciones de optimalidad. En dicho trabajo Tukey ponía de manifiesto que el análisis de datos era un campo más amplio que el de la Estadística Matemática, y transmitía el hecho de huir de la matematización, para introducirse abiertamente en el desarrollo de métodos y técnicas construidas ad hoc para los análisis que se desee realizar, es más, se desprendía de su exposición que el analista de datos no tenía por qué tener una formación rigurosa como la que se adquiere desde el

campo de las matemáticas. De hecho se ha visto a lo largo de estos años, cómo biólogos, médicos, psicólogos, pedagogos, etc., se han implicado sobremanera en el trabajo que se articula en torno al análisis de datos.

Sin embargo también se ha visto en ese tiempo que muchas de las técnicas desarrolladas, han debido abordarse desde el fundamento matemático, para salvar muchos de los errores y planteamientos en los que se estaba incurriendo. Basta considerar la teoría de la robustez, que apareció para salvar hipótesis iniciales que se asumían y que sin embargo no se cumplían en los datos, y todos los desarrollos realizados sobre modelos especiales, etc., ya que conviene recordar que en el momento en que se producen las reflexiones de Tukey, podría decirse que toda la Estadística Matemática o el tratamiento de los datos se desarrollaba en torno a la hipótesis de estructuras normales y con unos medios informáticos muy elementales tanto de *software* como de *hardware.*

Ahora está ocurriendo algo similar con el Big Data: se está abordando el tratamiento de la información que se genera desde un esquema que podía denominarse tecnológico-informático, generando resultados de los que a medio o largo plazo podríamos desdecirnos, como ya he puesto de manifiesto en ejemplos previos; y lo que se debería hacer es que el esfuerzo tecnológico-informático vaya unido a planteamientos rigurosos que aborden cuestiones como

las indicadas al plantear las estructuras hiperdimensionales.

En definitiva, aunque es correcto admitir que en el Big Data pueden encontrarse estudios o situaciones que puedan abordarse sin el rigor y la precisión necesaria, motivado bien porque no están desarrollados lo fundamentos que se necesitan para dichos estudios o bien porque resulta imposible desarrollarlos por la perentoriedad del tema, es imperativo que se busque una colaboración estrecha entre Ciencia y Tecnología, en mi opinión, se pone de manifiesto más que nunca la necesidad de equipos pluridisciplinares, formados por al menos, el proveedor de los datos, el gestor de los datos desde el punto de vista tecnológico, y el analista de los datos, quién debería tener una formación científica rigurosa en Estadística Matemática.

Y por ello me gustaría finalizar después de lo expuesto, precisando que el tratamiento de los datos no empieza en este siglo, que debemos mirar y aprender del pasado porque ello nos ayudará en el presente a encarar el futuro.

Señoras y señores, como dice Barry Mazur (Profesor de la Universidad de Harvard), las matemáticas son una larga conversación entre aquellos que resuelven los problemas y aquellos que construyen las teorías, entre las zorras y los erizos que pueblan todas las ciencias, por utilizar una expresión del poeta griego Aquiloco el cual escribía: *"muchas cosas sabe la zorra pero el erizo sabe una sola y grande"*.

Yo, tengo la suerte de pertenecer a un gran equipo que ha tratado de resolver problemas reales y en ocasiones

también ha tratado de diseñar sus propios modelos teóricos. Nuestra aportación, grande o pequeña, al desarrollo de la ciencia ha sido factible gracias al espíritu de colaboración y a la conjunción del esfuerzo con el compañerismo. Son muchos los nombres y apellidos, los hombres y mujeres que han participado, que han colaborado para hacer posible la obtención de los resultados. Muchas de estas personas se encuentran esta mañana aquí, y saben de sobra de mi reconocimiento a su trabajo, a su esfuerzo y su buen hacer. Mi gratitud y mi cariño especialmente a tres personas que han estado siempre a mi lado, en los momentos buenos y en las situaciones difíciles, que de todo ha habido, Joaquín Muñoz, Andrés González y Luis Parras.

No quisiera retener por más tiempo vuestra atención pero sí quiero agradecerles de corazón a todos ustedes su presencia en este acto, lo que sin duda entiendo como una manifestación de cariño y amistad.

Termino reiterando que me siento muy honrado por esta distinción, que significa para mí un estímulo al considerarme unido estrechamente a un conjunto de personas comprometidas con lo que Unamuno calificó como uno de los mayores servicio que se pueden prestar a una nación, entregarse de una manera callada y persistente a cultivar lo que él denominó *"el heroísmo del trabajo y el culto a la verdad"*.

Muchas gracias.

Bibliografía

[1] Spengler, O. (1923): La decadencia de Occidente. Bosquejo de una morfología de la Historia Universal. Calpe, Madrid.

[2] Castells, M.(1998): La era de la información. Economía, Sociedad y Cultura. Vol. 1. La sociedad red. Madrid. Alianza, 2.ª. reimpr.

[3] Beck, U. (1998).- La sociedad del riesgo. Hacia una nueva modernidad. Paidos. Barcelona.

[4] Fisher, R. A., and F. Yates: Statistical Tables for Biological, Agricultural and Medical Research
https://es.wikipedia.org/wiki/Sociedad_de_la_informaci%C3%B3n

[5] Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature. Vol. 457, pp. 1012 – 1014.

[6] Ward, J.S. and Barker, A. (2013). Undefined by data: A survey of big data definitions. arXiv: 1309.5821v1 [cs.DB] .

[7] Laney, D. (2001). 3–D data management: Controlling data volume, velocity and variety. Application Delivery Strategies. META Group Research Note, February 6. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[8] Yin, S. and Kaynak, O. (2015). Big Data for modern industry. Challenges and trends. Proceedings of the IEEE. Vol. 103, pp. 143 – 146.

[9] Anderson, Ch. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired (science). https://www.wired.com/2008/06/pb-theory/
[10] Elliot, M.R. and Valliant, R. (2017). Inference for nonprobabilistic samples. Statistical Science. Vol. 32, pp. 249 – 264.
[11] Moon, N. (1999). Opinión Polls. History, theory and practice. Ed. Manchester University Press.
[12] Bradburn, N.M. and Sudman, S. (1988). Polls and Surveys. Understanding what they tell us. Ed. Jossey – Bass Publishers.
[13] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method purposive selection. Journal of the Royal Statistical Society. Vol. 97, pp. 558 – 625.
[14] Kruskal, W. and Mosteller, F. (1980). Representative sampling. IV: the history of the concept in Statistics, 1895 – 1939. International Statistical Review. Vol. 48, pp. 169 – 195.
[15] Mantelero, A. (2017). Regulating big data. The guidelines of the Council of Europe in the context of the European data protection framework. Computer Law & Security Review. Vol. 33, pp. 584 – 602.
[16] Beyer, K.; Goldstein, J.; Ramakrishnan, R. and Shaft, U. (1999).When is "nearest neighbors" meaningful? Proceedings International Conference Database Theorie, pp. 217 – 235.
[17] Wang, J. (2012). Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Ed. Springer.

[18] Puts, M.; Daas, P. and de Waal, T. (2015). Finding errors in big data. Significance, Vol. 12, pp- 26 – 29.
[19] Donoho, D.L. (2000). High – dimensional data analysis: The curses and blessing of dimensionality. Lecture Delivered at the "Mathematical Challenges of the 21st Century" Conference of the American Math. Society, Los Angeles.
[20] Tukey, J.W. (1962). The future of data analysis. The Annals of Mathematical Statistics. Vol. 33, pp. 1-67.
[21] Jones, L.V. (1992). Introduction to Tukey (1962), The future of data analysis. In Breakthroughs in Statistics. Volume II. Methodology and Distribution. Editors Kotz, S and Johnson, N.L. Ed. Springer Verlag.