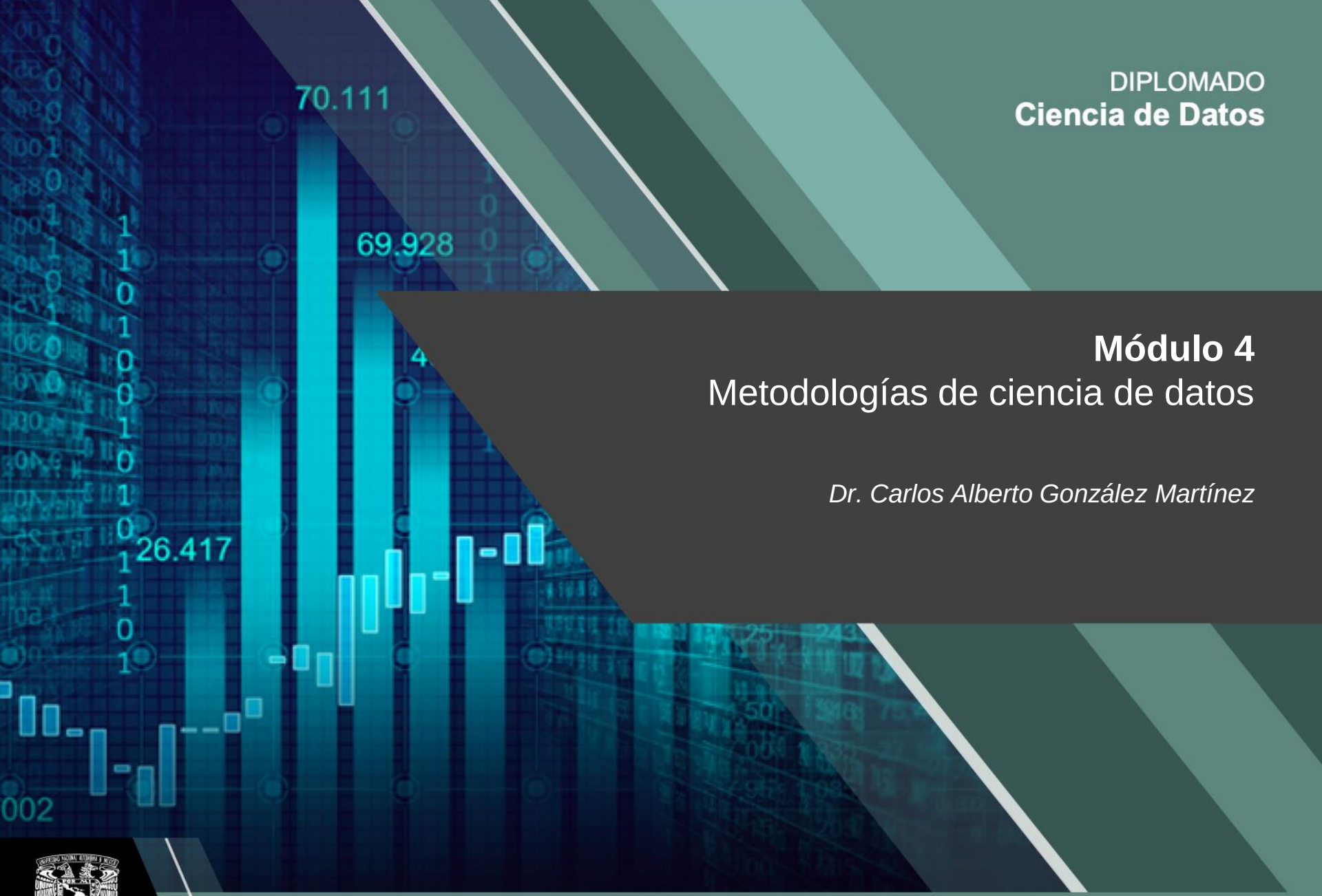


DIPLOMADO
Ciencia de Datos

Módulo 4
Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

AGILE DATA SCIENCE

gmc

Objetivo

El participante identificará la forma en que los desarrolladores pueden ejecutar un proyecto de ciencia de datos de forma sistemática.

AGILE DATA SCIENCE

Contenido

1. **La ciencia de datos ágil y el marco ágil para crear una práctica de ciencia de datos impulsada por el ROI**

(con base en el trabajo de Favio Vázquez titulado Marco ágil para crear una práctica de ciencia de datos impulsada por el ROI, 2018)

2. **Características de los proyectos de ciencia de datos**
3. **Manifiesto para la ciencia de datos ágiles**

AGILE DATA SCIENCE

1. La ciencia de datos ágil y el marco ágil para crear una práctica de ciencia de datos impulsada por el ROI

Agile Data Science (ADS) es una nueva metodología para el desarrollo de productos analíticos.

El desarrollo de Agile de proyectos de ciencia de datos es la forma en que los desarrolladores pueden ejecutar un proyecto de ciencia de datos de forma sistemática, con control de versiones y en colaboración dentro de un equipo de proyecto.

El llamado **Business Science Problem Framework (BSPF)** es una nueva iteración de un marco ágil, para implementar la ciencia de datos de una manera que permita que la toma de decisiones siga un proceso sistemático, que conecta los modelos que crea con el ***retorno de la inversión (ROI)*** y muestre el valor que sus mejoras aportan al negocio.

AGILE DATA SCIENCE

1. **La ciencia de datos ágil y el marco ágil para crear una práctica de ciencia de datos impulsada por el ROI**

Aspectos relacionados con la implementación de la metodología ágil, para una práctica de ciencia de datos impulsada por el ROI .

- Definición del problema de ciencia de datos.
- Contar con un marco de problemas de ciencia empresarial BSPF, un marco de proyecto de ciencia de datos que ha reducido a la mitad el tiempo de entrega de proyectos de ciencia de datos que generan ROI.
- Agilidad en la ciencia de datos: investigaremos qué son los métodos ágiles y cómo ayudan en la ciencia de datos para el proceso empresarial (Vazquez, F., 2018, p. 3).

AGILE DATA SCIENCE

1. **La ciencia de datos ágil y el marco ágil para crear una práctica de ciencia de datos impulsada por el ROI**

El marco ágil para la ciencia de datos, es un proceso iterativo aplicado para la solución de problemas de ciencia de datos. Agilidad es una palabra para hacer eso en el mundo de la TI, y es por eso que se considera al BSPF un marco ágil.

El proceso de un ***flujo de trabajo de ciencia de datos ágil****,* propuesto por Russell Jurney, es una forma asombrosa de comprender cómo y por qué la ciencia de datos, junto con la agilidad, nos ayudan a ir más allá, a ver más y a resolver problemas de una manera creativa.

AGILE DATA SCIENCE

2. Características de los proyectos de ciencia de datos

Reproducible: Necesario para facilitar la prueba del trabajo y el análisis de otros.

Falible: la ciencia de datos y la ciencia no buscan la verdad, sino el conocimiento, por lo que cada proyecto puede ser sustituido o mejorado en el futuro. Ninguna solución es la solución definitiva.

Colaborativo: el científico de datos no existe solo. Necesita un equipo (incluso un equipo virtual, como colaborar en un proyecto de código abierto). Este equipo hará las cosas posibles para crear inteligencia y soluciones. La colaboración es una gran parte de la ciencia y la ciencia de datos no debería ser una excepción.

AGILE DATA SCIENCE

2. Características de los proyectos de ciencia de datos

Creativo: la mayor parte de lo que hacen los científicos de datos son nuevas investigaciones, nuevos enfoques o la adopción de diferentes soluciones, por lo que su entorno debe ser muy creativo y fácil de trabajar. La creatividad es crucial en la ciencia. Es la única forma en que podemos encontrar soluciones a problemas difíciles y complejos.

Cumple con las regulaciones: en este momento hay muchas regulaciones sobre ciencia, no tanto sobre ciencia de datos, pero habrá más en el futuro. Es importante que los proyectos que estamos construyendo puedan conocer estos diferentes tipos de regulaciones, para que podamos crear una solución limpia y aceptable a los problemas.

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

El **Manifiesto para la ciencia de datos ágiles** evita sacar conclusiones muy rápido. Antes se llevan a cabo los análisis requeridos. Es un proceso iterativo.

La siguiente imagen muestra los aspectos del manifiesto para la ciencia de datos ágiles:

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

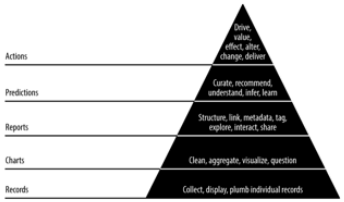
The Agile data science manifesto

Iterate, iterate, iterate

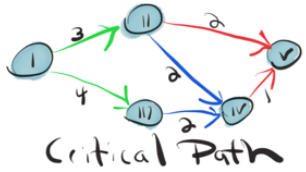
Ship intermediate output

Perform experiments, not tasks

Listen to the data



Respect the data-value pyramid



Find the critical path

Get meta

gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

La iteración es el acto esencial en la elaboración de aplicaciones de análisis, lo que significa que a menudo nos quedamos al final de un *sprint* con cosas que no están completas.

Si no enviamos la salida incompleta o intermedia al final de un sprint, a menudo terminaríamos enviando nada en absoluto. Y eso no es *agile*. Yo lo llamo el "ciclo de la muerte", donde se puede perder un tiempo infinito perfeccionando cosas que nadie quiere.

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Los buenos sistemas se auto documentan y, en la ciencia de datos ágil, documentamos y compartimos los activos incompletos que creamos mientras trabajamos.

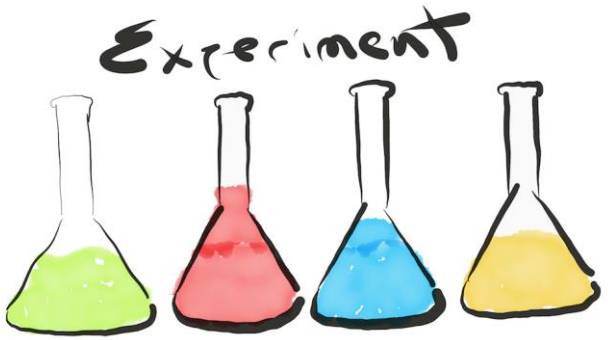
Comprometemos todo el trabajo con el control de la fuente. Compartimos este trabajo con los compañeros de equipo y, lo antes posible, con los usuarios finales. Este principio no es obvio para todos. Muchos científicos de datos provienen de antecedentes académicos, donde años de intenso esfuerzo de investigación se destinaron a un solo artículo grande llamado tesis que resultó en un título avanzado.

gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Realiza experimentos, no tareas
En ingeniería de software, un gerente de producto asigna un gráfico a un desarrollador para que lo implemente durante un sprint.



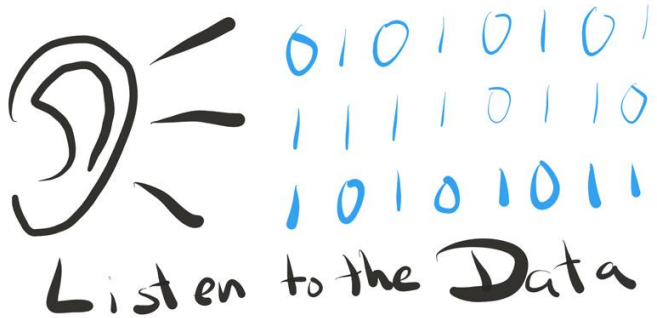
El desarrollador traduce la asignación en un GROUP BY de SQL y crea una página web para ello. ¿Misión cumplida? Equivocado. Es poco probable que los gráficos que se especifican de esta manera tengan valor.

La ciencia de datos se diferencia de la ingeniería de software, en que es en parte ciencia y en parte ingeniería.

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Escuche los datos
Lo que es posible es tan importante como lo que se pretende. Es tan importante saber qué es fácil y qué es difícil como lo que se desea.



En el desarrollo de aplicaciones de software, hay tres perspectivas a considerar: **las de los clientes, los desarrolladores y la empresa**. En el desarrollo de aplicaciones analíticas, existe otra perspectiva: **la de los datos**. Sin comprender lo que los datos "tienen que decir" sobre cualquier característica, el propietario del producto no puede hacer un buen trabajo.

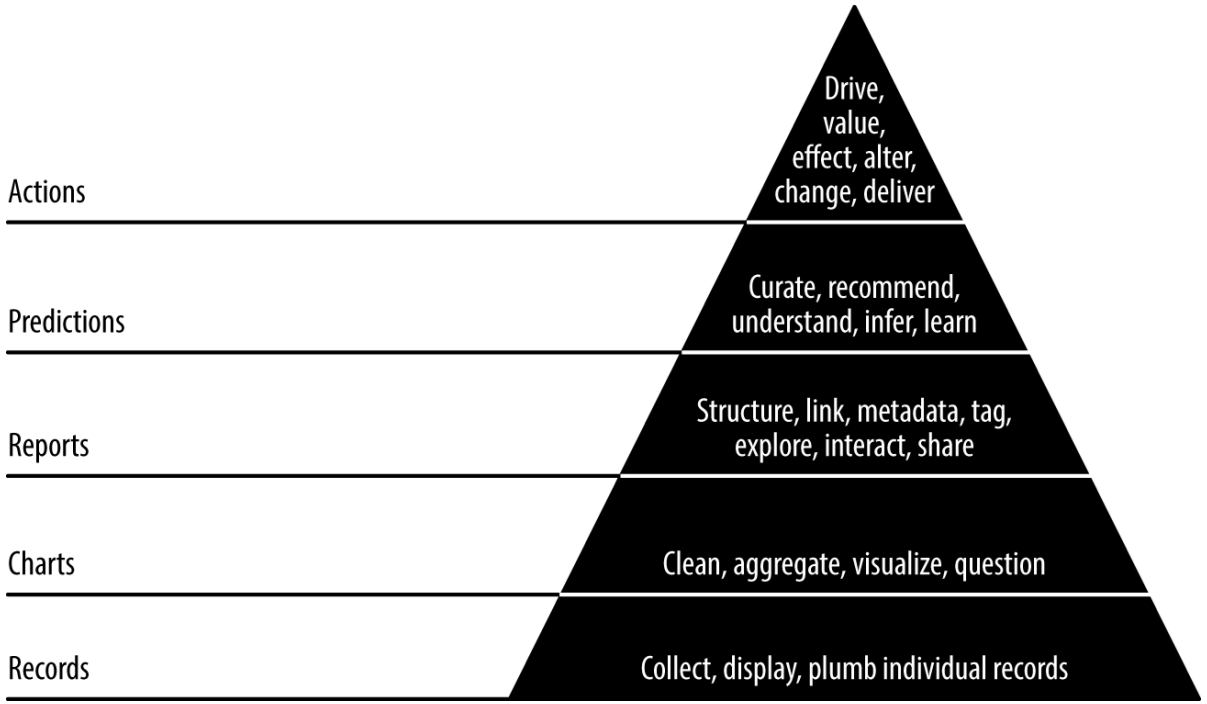
La opinión de los datos siempre debe incluirse en las discusiones del producto, lo que significa que deben basarse en la visualización a través del análisis exploratorio de datos en la aplicación interna que se convierte en el foco de nuestros esfuerzos.

gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Pirámide de valor de datos



gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Pirámide de valor de datos

Respete la pirámide del valor de los datos
La pirámide de valor de datos es una pirámide de cinco niveles, modelada según la jerarquía de necesidades de Maslow.

Expresa la creciente cantidad de valor creado al refinar los datos sin procesar en tablas y gráficos, seguidos de informes y luego predicciones, todo lo cual está destinado a permitir nuevas acciones o mejorar las existentes

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Pirámide de valor de datos

Los niveles de la pirámide de valor de los datos se leen de abajo hacia arriba.

El primer nivel de la pirámide de valores de datos (**registros**) se trata de plomería: hacer que un conjunto de datos fluya, desde donde se recopila hasta donde aparece en una aplicación.

La capa de **gráficos y tablas** es el nivel donde comienza el refinamiento y el análisis.

La capa de **informes** permite una exploración inmersiva de datos, donde realmente podemos razonar sobre ellos y conocerlos.

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Pirámide de valor de datos

La capa de **predicciones** es donde se crea más valor, pero crear buenas predicciones significa ingeniería de características, que los niveles inferiores abarcan y facilitan.

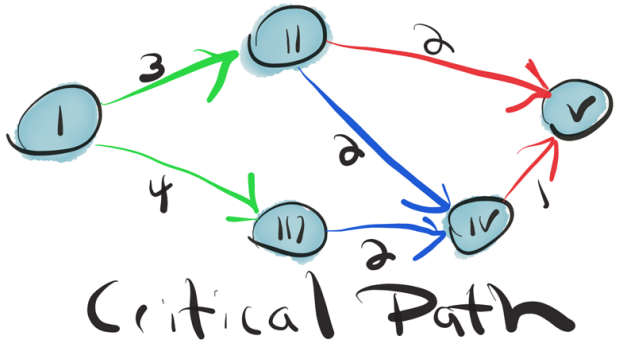
El nivel final, **acciones**, es donde está teniendo lugar la locura de la inteligencia artificial (IA). Si su conocimiento no permite una nueva acción o mejora una existente, no es muy valioso.

La pirámide del valor de los datos estructura nuestro trabajo. La pirámide es algo a tener en cuenta, no una regla a seguir. A veces te saltas pasos, a veces trabajas hacia atrás.

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Encuentra el camino crítico
Para maximizar nuestras probabilidades de éxito, debemos concentrar la mayor parte de nuestro tiempo en ese aspecto de nuestra aplicación, que es más esencial para su éxito.



¿Pero qué aspecto es ese? Esto debe descubrirse mediante la experimentación. El desarrollo de productos analíticos es la búsqueda y la consecución de un objetivo en movimiento. Una vez que se determina un objetivo, por ejemplo, se debe hacer una predicción, debemos encontrar el ***camino crítico*** para su implementación y, si resulta valioso, para su mejora. Los datos se refinan paso a paso a medida que fluyen de una tarea a otra.

gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Obtener meta
Si no podemos enviar fácilmente buenos activos de productos en un cronograma comparable al desarrollo de una aplicación normal, ¿qué enviaremos?

Si no enviamos, no somos ágiles. Para resolver este problema, en la ciencia de datos ágiles "obtenemos meta". La atención se centra en documentar el proceso de análisis, en lugar del estado final o el producto que buscamos. Esto nos permite ser ágiles y enviar contenido intermedio, a medida que escalamos iterativamente la pirámide de valor de datos para seguir el camino crítico hacia un producto excelente.

gmc

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

Estos siete principios funcionan juntos para impulsar la metodología de ciencia de datos ágil. Sirven para estructurar y documentar el proceso de análisis exploratorio de datos y transformarlo en aplicaciones analíticas.

Debo enfatizar que este es un marco ágil, no que la ciencia de datos sea ágil. Esto sigue las palabras de Dave Thomas, uno de los creadores del Manifiesto para el desarrollo de software ágil.

- No es un programador ágil, es un programador que programa con agilidad
- No trabajas en un equipo ágil, tu equipo exhibe agilidad
- No usa herramientas ágiles, usa herramientas que mejoran su agilidad

AGILE DATA SCIENCE

3. Manifiesto para la ciencia de datos ágiles

- No es un científico de datos ágil, es un científico de datos que sigue un marco con agilidad

Referencias bibliográficas

Jurney, R., (2017). *Un manifiesto para la ciencia de datos ágiles.* Recuperado de https://www.oreilly.com/radar/a-manifesto-for-agile-data-science/

Vazquez, F., (2018). *Marco ágil para crear una práctica de ciencia de datos impulsada por el ROI.* Recuperado de https://www.business-science.io/business/2018/08/21/agile-business-science-problem-framework.html

Contacto

Carlos Alberto González Martínez

Jefe de departamento de correlaciones, cruces y alertas (C5i)

gmcmmiv@hotmail.com