TALLER CIENCIA DE DATOS

UM
UNIVERSIDAD DE MONTEVIDEO

pyxis

IDATHA

SAP

SAP next-gen

CRONOGRAMA

Bloque 1
- Conceptos básicos

Coffee break (15 minutos)

Bloque 2
- Ejemplo práctico

AMT/B/9/1
146

2

(Taken from MIND : a Quarterly Review of Psychology and Philosophy. Vol. LIX. , N.S., No. 236, October , 1950.)

COMPUTING MACHINERY AND INTELLIGENCE

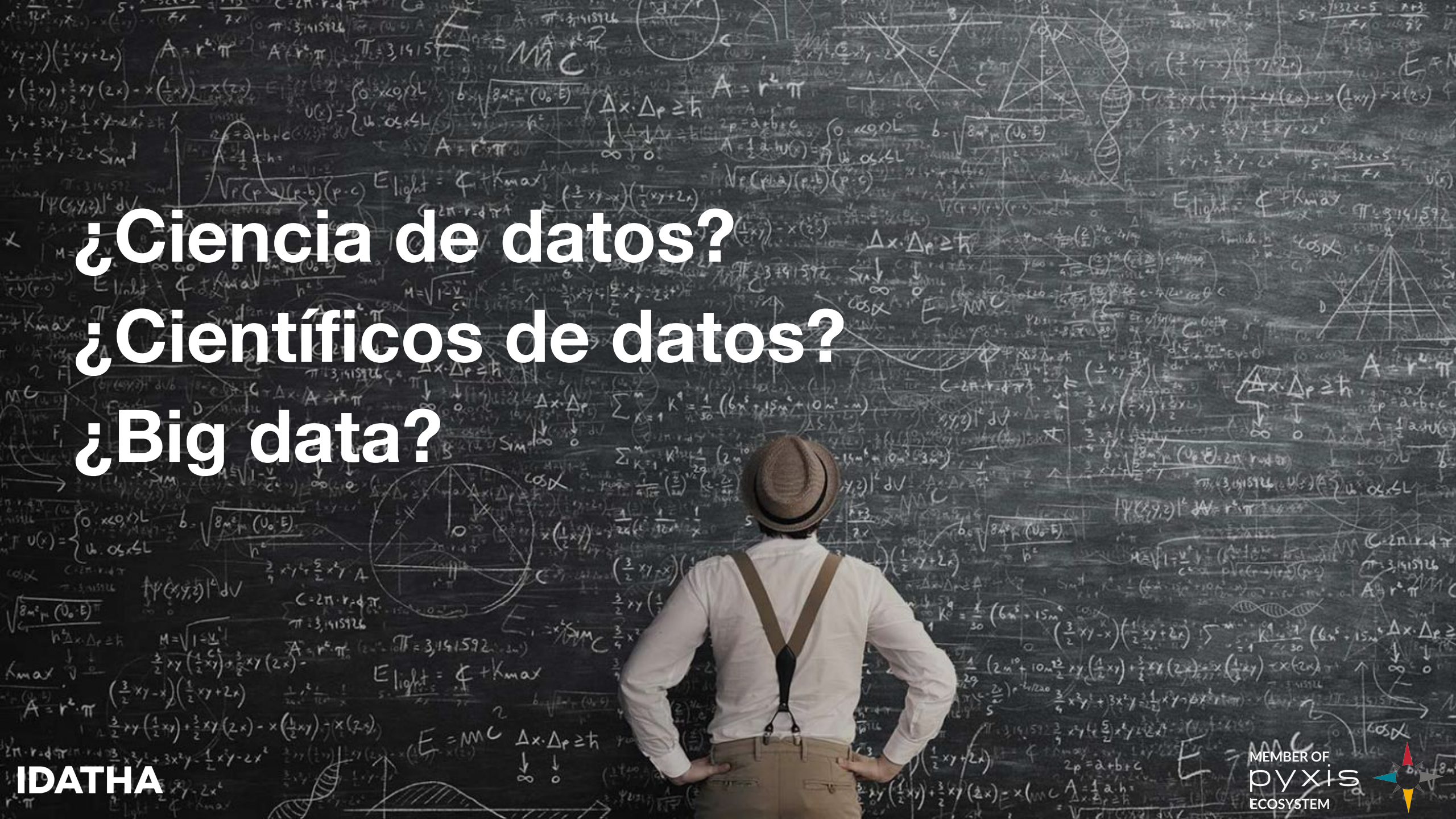
by

A. M. TURING.

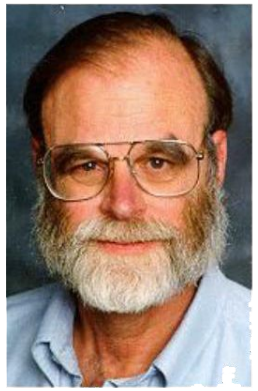
1. The Imitation Game.

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

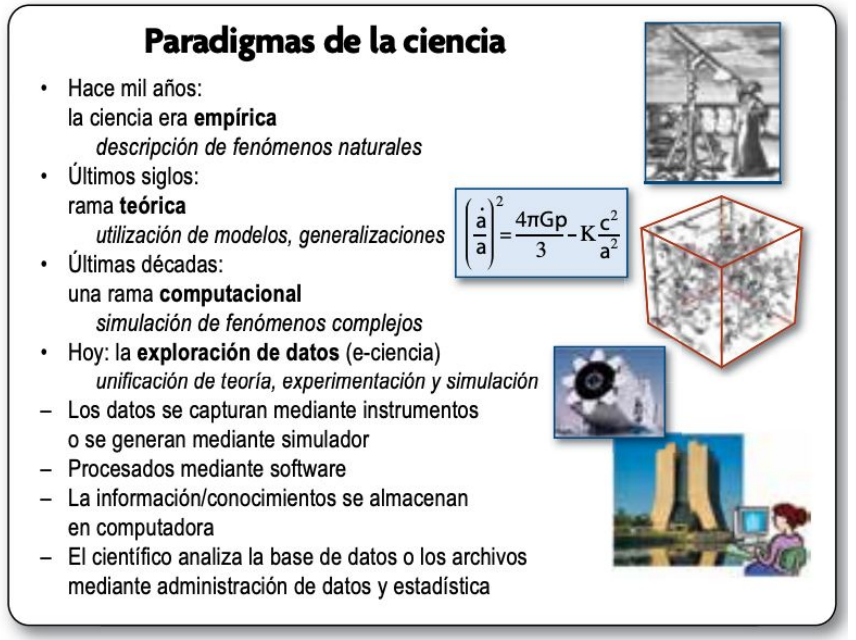

¿Ciencia de datos?
¿Científicos de datos?
¿Big data?
IDATHA
MEMBER OF
pyxis
ECOSYSTEM

Ciencia de Datos



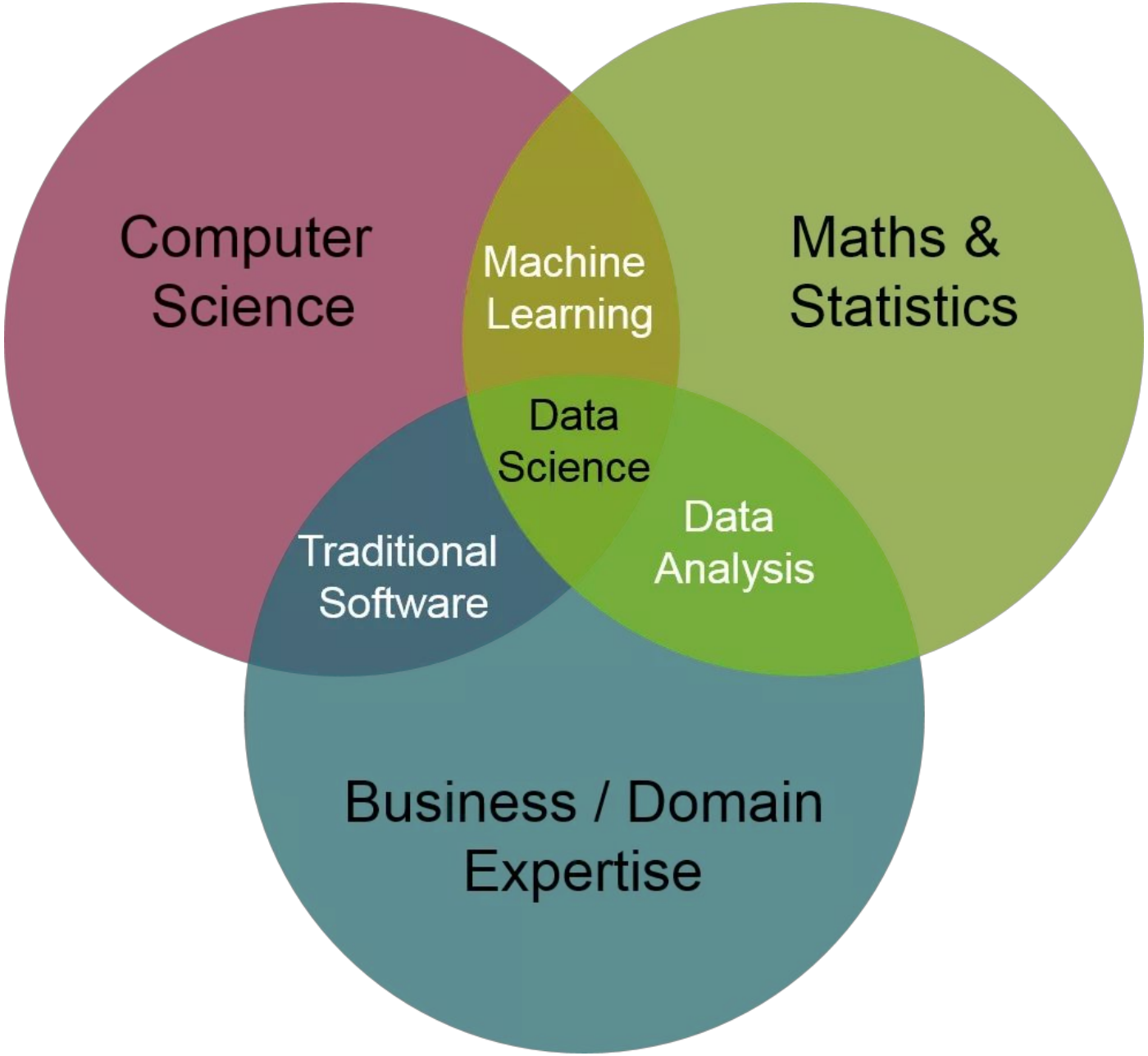
La **ciencia de datos** es un campo **interdisciplinario** que involucra **métodos científicos**, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos.

”Cuarto paradigma" de la ciencia (empírico, teórico, computacional y ahora basado en datos) y afirmó que "todo lo relacionado con la ciencia está cambiando debido al impacto de la tecnología de la información y el diluvio de datos”, Jim Gray

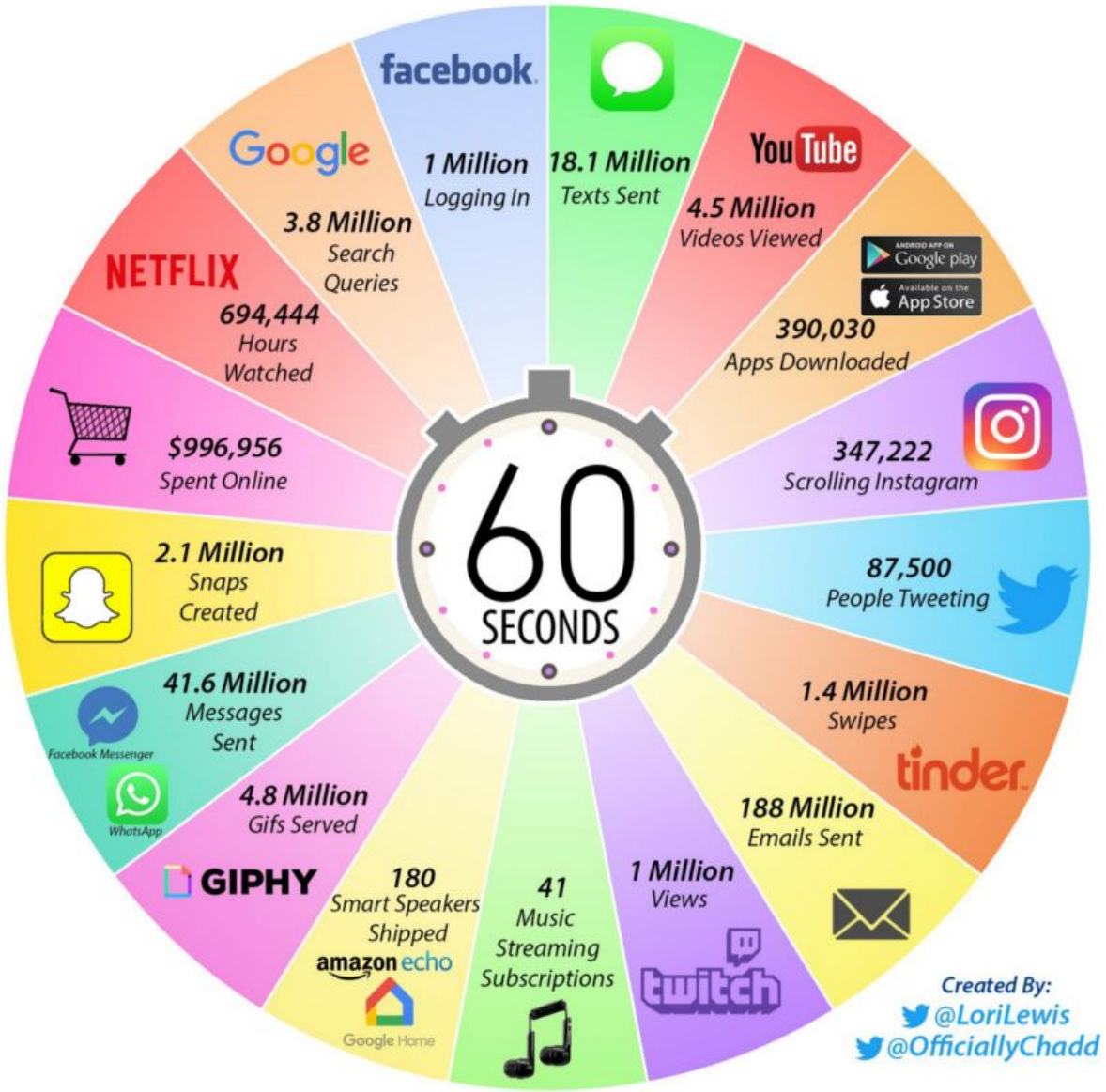


IDATHA

member of pyxis ecosystem


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

2019 This Is What Happens In An Internet Minute



IDATHA

member of pyxis ecosystem

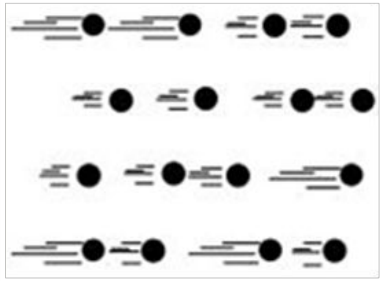
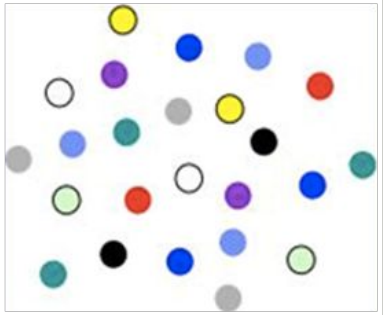
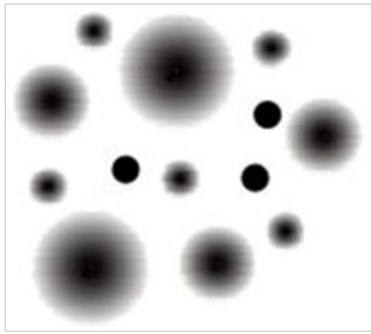
Mantenimiento Predictivo

300 sensores

200 GB de datos por día

Se transmiten y analizan en Dinamarca

BIG DATA

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

IDATHA

member of pyxis ecosystem

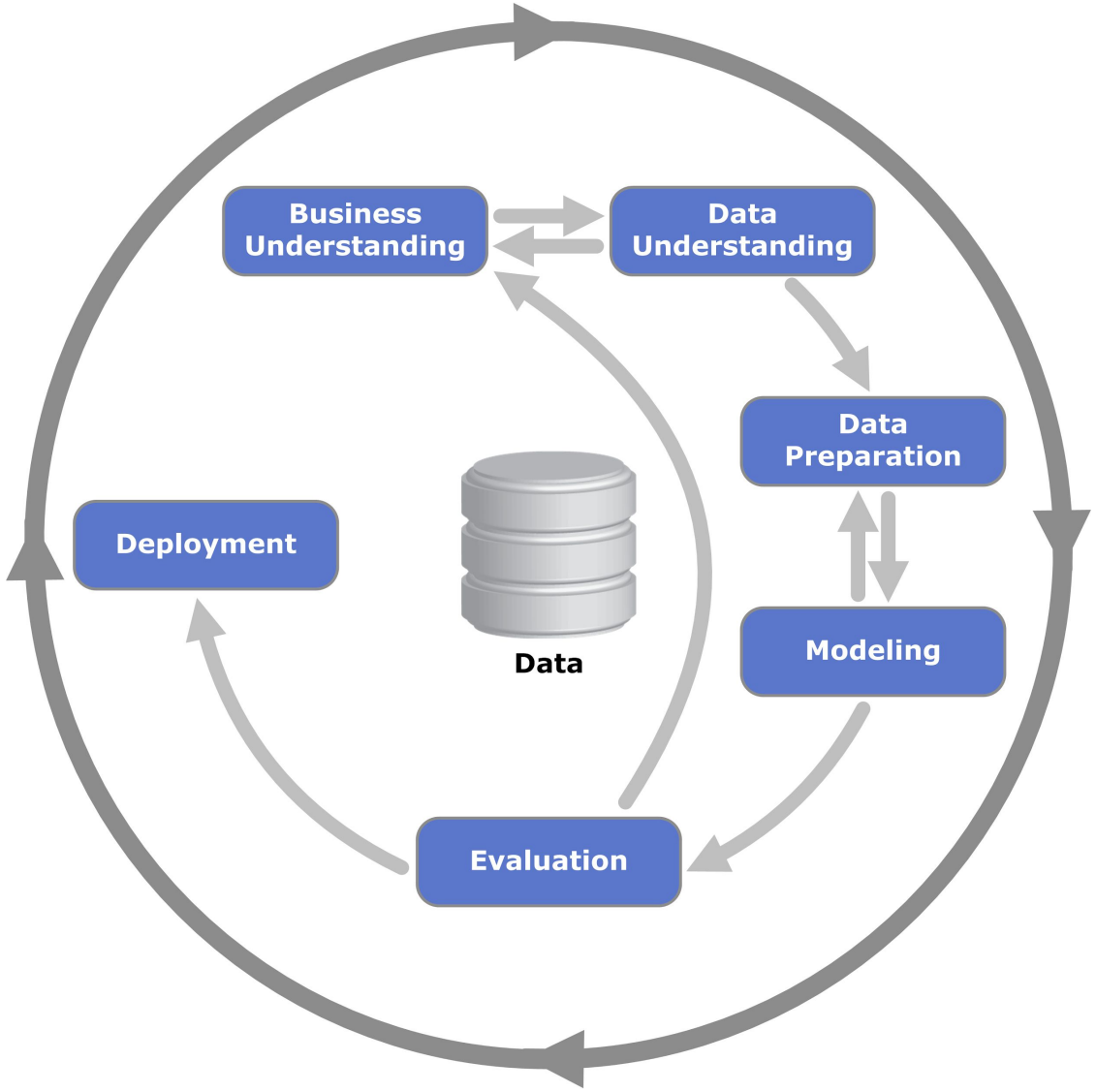
Workflow: CRISP-DM

CRISP-DM es la sigla para *CRoss Industry Standard Process for Data Mining*.

Es un modelo de proceso, propuesto inicialmente para proyectos de minería de datos y que puede ser adaptado para proyectos en ciencia de datos. El proceso es independiente del sector de la industria del cual proviene el problema que queremos resolver o de las tecnologías utilizadas.

Fue presentado por primera vez en el año 2000, a través del trabajo *CRISP-DM: Towards a standard process model for data mining*

CRISP-DM: Fases



CRISP-DM: Business Understanding

Comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.

Etapas:

- Determinar objetivos empresariales.
- Analizar la situación actual.
- Determinar objetivos de data science.
- Planificación con duración, recursos y riesgos.

CRISP-DM: Business Understanding (II)

Discutiendo con el cliente (negocio) determinar:

- Situación comercial o problema:
  - Estructura de la organización
  - Áreas involucradas/afectadas
  - Solución actual
- Objetivos comerciales específicos
  - Verificar si hay objetivos complementarios
- Criterios de rendimiento desde la perspectiva de negocio
  - Objetivos y subjetivos

CRISP-DM: Business Understanding (III)

También con el cliente relevar:

- Recursos
  - Infraestructura, datos y personal
- Requisitos, supuestos y restricciones
  - Restricciones legales o de seguridad
  - Supuestos de calidad de fuentes o despliegue de resultados
  - Acceso a datos y permisos
- Riesgos y contingencia
  - Tiempo, dinero, datos y resultados
- Costo/beneficio
  - Beneficio potencial esperado

Ejemplo: Supermercado Target

Un día, la cadena de supermercados Target decide invertir en el servicio de retención y fidelización de sus clientes, utilizando técnicas de data mining. Para ello, utilizaron datos de sus bases de datos de clientes, en las cuales almacenaban todas las compras realizadas por un cliente, junto con sus datos personales, información de crédito, etc. El objetivo era conocer mejor a sus clientes para realizar campañas de cupones de descuentos personalizados.

Luego de analizar los datos, uno de los analistas descubre ciertos patrones interesantes, como por ejemplo los productos que en la mayoría de los casos, compraban mujeres durante sus primeros meses de embarazo. Algunos hallazgos:

- Mujeres embarazadas compran grandes cantidades de lociones sin aroma, alrededor del segundo trimestre de embarazo.
- Algunas veces, en las primeras 20 semanas, las mujeres embarazadas consumen suplementos nutricionales como calcio, magnesio y zinc.

Descubrieron más de 25 productos que usualmente compran mujeres embarazadas en algún momento del embarazo y con ellos construyeron un puntaje de “predicción de embarazo”. De esta forma, comenzaron a enviar cupones de descuento en productos específicos como cremas, lociones, pañales, etc., a mujeres con un puntaje de “predicción de embarazo” alto.

Ejemplo: Supermercado Target (cont)

Tiempo después un hombre enojado llama a una tienda Target y pide para hablar con el gerente:

> “Mi hija recibió este email de ustedes! Ella todavía está en secundaria, y ustedes le están enviando cupones para ropa de bebé y pañales? Están tratando de animarla a que se embaraze?”

Tiempo después la misma persona llamó para disculparse. Su hija efectivamente estaba embarazada.

Lo interesante de este ejemplo, además del hecho de que un supermercado supo que una adolescente estaba embarazada incluso antes que ella misma, es repasar la situación y preguntarse:

¿Como cliente de Target, como me sentiría si ellos saben cosas sobre mi que incluso ni yo se?

Quizás desde el cliente (Target), faltó definir una estrategia más sensible para enviar cupones a mujeres potencialmente embarazadas, quizás esperar hasta el tercer mes de embarazo, no ser tan explícito, etc.

Fuente:
https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#64e598c86668

CRISP-DM: Data Understanding

Recolección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los mismos, identificar problemas en la calidad de los datos, descubrir primeros insights en los datos, o detectar subconjuntos de datos para formular hipótesis sobre datos ocultos. Hay un vínculo muy cercano entre las etapas de *Comprensión del negocio* y *Comprensión de los datos*

Etapas:

- Recopilar datos disponibles
- Explorar y describir los datos con tablas y gráficos.
- Verificar calidad de los datos.

CRISP-DM: Data Understanding (II)

- Obtener datos para su exploración y análisis
- Debo contar con acceso a los mismos, o tener una copia local
- Es muy importante contar con meta-data o documentación sobre los datos
- Fuentes:
  - Datos existentes: transaccionales, encuestas, registros web, CRM
  - Datos adquiridos: solicitud a secciones o áreas “dueñas” de datos
  - Datos adicionales: necesidades aún no cubiertas (proyecto en sí mismo)

CRISP-DM: Data Understanding (III)

Checklist de descubrimiento

- Cantidad de datos (general)
- Atributos relevantes o “prometedores”
- Atributos no relevantes
- Tipos de valores: tipo de dato de cada variable
- Codificación: representación de características
- Densidad de datos en atributos (suficiencia)
- Data matching y data fusion
- Valores faltantes

CRISP-DM: Data Preparation

Actividades para construir el conjunto de datos de entrenamiento. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos, y limpieza de datos para las herramientas de modelado.

Etapas:

- Selección de subconjunto de datos.
- Limpieza de datos.
- Creación de nuevos atributos (ingeniería de atributos).
- Transformaciones a atributos
- Fusión y agregado de conjuntos y registros.
- Verificación de formato de datos para el modelado.
- División en conjuntos de datos de prueba y entrenamiento.

CRISP-DM: Data Preparation (II)

Analizar los datos para determinar

- Datos relevantes a los objetivos (de negocio y de Data Science)
- Selección por filas
  - No todos los registros son relevantes (fecha, falta de información, errores)
  - Necesidad de estudiar un subconjunto específico
- Selección por columnas
  - Decisiones sobre el uso de atributos
  - Pueden existir restricciones
  - Necesidad de crear nuevos valores a partir de los actuales
- Qué datos debo limpiar
  - Recurrir al informe de calidad (si existe)

La preparación de datos insume entre el 50 y 70% del tiempo y esfuerzo del proyecto

CRISP-DM: Modeling

Se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para mejorar los resultados. Hay varias técnicas que tienen requerimientos específicos sobre la forma de los datos, por lo que puede ser necesario volver a la fase de preparación de datos.

(Mañana veremos más en profundidad esto)

CRISP-DM: Evaluation

Evaluación del modelo (o modelos) construidos, que parecen tener gran calidad desde una perspectiva del análisis de datos.

(Mañana veremos más en profundidad esto)

CRISP-DM: Despliegue

Esta fase depende de los requerimientos, pudiendo ser simple como la generación de un reporte o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización, disponibilizar un modelo predictivo como servicio, etc.

IDATHA

MEMBER OF
pyxis
ECOSYSTEM

Practise Time

IDATHA

MEMBER OF pyxis ECOSYSTEM

TALLER CIENCIA DE DATOS

Sebastián García
sgarcia@idatha.com
@dsgarcia

Emiliano Viotti
eviotti@idatha.com
@Efviodo

Manu Reynaert
mreynaert@idatha.com

Bitly http://bit.ly/idatha-ds-course

Long https://github.com/efviodo/idatha-data-science-course

IDATHA

MEMBER OF pyxis ECOSYSTEM